Bioinformatic Characterization of Mutations Associated with Enhanced Acetate Metabolism in *Escherichia coli*

by
Madison Yahn

A thesis presented to the Honors College of Middle Tennessee State University in partial fulfillment of the requirements for graduation from the University Honors College

Fall 2025

Thesis Committee:

Elliot Altman, Thesis Director

Mary Farone, Second Reader

Rebecca Seipelt-Thiemann, Thesis Committee Chair

APPROVED:

_____

Elliot Altman, Thesis Director
[Professor, Department of Biology]

_____

Mary Farone, Second Reader
[Professor, Department of Biology]

_____

Rebecca Siepelt-Thiemann, Thesis Committee Chair
[Professor, Department of Biology]

ABSTRACT

The conversion of non-food lignocellulosic biomass, such as from trees or grasses, into sugars that can be converted into ethanol by fermentation is key to the future U.S. energy infrastructure. The accumulation of acetate during lignocellulosic biomass pretreatment inhibits microbial fermentation and limits bioethanol yields. This study aimed to characterize *Escherichia coli* strains engineered to efficiently utilize acetate as a sole carbon source to support acetate detoxification prior to yeast fermentation. Eight mutant strains were generated using either spontaneous or ethyl methanesulfonate (EMS) mutagenesis, then characterized using whole-genome sequencing and comparative bioinformatic analysis. Variant detection was performed using two pipelines, Bowtie2 + bcftools and Bowtie2 + DeepVariant, to evaluate differences in sensitivity and precision. The mutants exhibited enhanced growth on acetate despite lacking mutations in acetate metabolism genes (*acs, pta, ackA*). Instead, changes occurred in genes associated with replication and potentially global regulation of gene expression (*dnaA*), acid resistance (*gadX*), glycogen metabolism (*glgX*), membrane stability (*ytcA*), and proton export (ECOLC_RS20895). These mutations collectively improved stress tolerance, pH homeostasis, and metabolic efficiency. The results suggest that global physiological remodeling, rather than direct enzymatic modification, underlies enhanced acetate assimilation in *E. coli*, providing a genetic foundation for designing robust microbial strains for bioethanol production.

PREFACE

The utilization of *E. coli* to detoxify the lignocellulosic hydrolysates used in the production of bioethanol provides a viable solution to offset our reliance on non-renewable resources. I chose this topic while working with Dr. Altman and Nicole Gammons due to my interest in genetics. This research journey has led to not only the development of new technological skills and the ability to perform genomic analysis, but also has helped me prepare myself for post-graduate work.

I want to thank Dr. Altman and my committee for taking me through this experience and always being there for the hard questions and help. I also want to thank Nicole Gammons for all her endless help on not only my thesis, but during my time in college as well. Nicole has been one of the best people I have met during my time at MTSU and has kept me going despite the stress and weight of everything going on all at once.

I would also like to thank the honors college faculty. I appreciate the belief in me to succeed not only as an honors student, but as a Buchanan Fellow the past three years. It has made my college experience so much less stressful and expanded my opportunities.

Lastly, thank you to my friends and everyone around me who has had faith in me, it means the world.

This thesis is organized in the steps taken during the process, from introduction to results and conclusion to discuss what exactly was found. My honors thesis has been an irreplaceable experience during college. Thank you all.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACKA	Acetate kinase
ACS	Acetyl-CoA synthetase
AD	Allele depth
AF	Allele frequency
ATP	Adenosine triphosphate
BCFtools	Bayesian variant-calling tools suite
bp	Base pairs
COA	Coenzyme A
DNA	Deoxyribonucleic acid
DNAA	Chromosomal replication initiator protein
DP	Read depth at variant sites
EMS	Ethyl methanesulfonate
GADX	Acid resistance transcriptional activator
GLGX	Glycogen debranching enzyme
GT	Genotype
IGV	Integrated Genome Viewer
INDEL	Insertion or deletion variant
NADH	Nicotinamide adenine dinucleotide
PBS	Phosphate-buffered saline

PCR	Polymerase chain reaction
PE150	Paired-end 150-base sequencing reads
PTA	Phosphotransacetylase
QUAL	Variant quality score
RNA	Ribonucleic acid
SNP	Single nucleotide polymorphism
SnpEff	SNP Effect Predictor (variant annotation tool)
TCA	Tricarboxylic acid cycle
WGS	Whole-genome sequencing
YtcA	YtcA family lipoprotein

LIST OF TERMS

Acetate assimilation	The way *E. coli* turns acetate into acetyl-CoA so it can be used for energy and growth.
Acetate toxicity	The harmful effect of too much acetate (or acetic acid), which lowers cell pH and slows down growth.
Adaptive evolution	A process where bacteria slowly develop useful mutations that help them survive better in tough conditions.
Bioethanol	Ethanol fuel made by microbes from plant materials like wood, grass, or crop waste.
Bioinformatics pipeline	A series of computer programs used to process DNA sequencing data and find genetic changes.
Bowtie2	A computer tool that matches DNA reads from sequencing to a known reference genome.
DeepVariant	A software program that uses artificial intelligence to find mutations in DNA sequences very accurately.
Ethyl methanesulfonate	A chemical that causes DNA changes (mutations) to help study or improve bacteria.
Genome annotation	Adding labels or descriptions to genes in a genome to explain what they do.
Genome stability	The ability of a cell to keep its DNA organized and unchanged over time.
Glyoxylate shunt	A backup energy pathway that helps *E. coli* survive when growing on acetate instead of sugar.
Lignocellulosic hydrolysate	The liquid mix of sugars and byproducts made when plant biomass is broken down for biofuel production.
Mutagenesis	Any method used to intentionally create mutations in an organism's DNA.
Proton export	The cell's way of pumping out hydrogen ions to keep its internal pH from getting too acidic.
Single Nucleotide Polymorphism	A single base change in the DNA sequence that can affect how a gene works.
Stress response	How a cell protects itself when conditions are harsh, like high acid or lack of nutrients.
Variant calling	The computational identification of genetic variants by comparing sequencing reads to a reference genome.

Whole-genome sequencing	The process of determining the complete DNA sequence of an organism's genome

INTRODUCTION

Several U.S. legislative mandates have called for the increased production of bioethanol from non-food lignocellulosic biomass, such as trees and grasses, to offset our reliance on gasoline which is produced from the non-renewable resource petroleum. Over the past two decades, federal energy policies have steadily promoted the transition toward cleaner, renewable fuels to enhance energy security and environmental sustainability. These policies have attracted public and private investment in biofuel technologies by establishing clear production targets and long-term market incentives.

On the public side, government funding, grants, and loan guarantees have supported biofuel research and the construction of biorefineries. Examples include the U.S. Department of Agriculture's Biorefinery Assistance Program established through Farm Bill acts, the Department of Energy's Bioenergy Technologies Office, and federal research initiatives created under the Biomass Research and Development Act of 2000 (Pub. L. No. 106-224, 2000). These efforts have been complemented by private-sector investment, as production targets and incentives introduced through the Renewable Fuel Standards of 2005 and 2007 (Pub. L. No. 109-58, 2005; Pub. L. No. 110-140, 2007) created a stable, long-term market for renewable fuels. In response, major energy and biotechnology companies such as POET, DuPont, and Shell have invested heavily in building commercial scale biorefineries and developing new enzyme, feedstock, and

fermentation technologies. Together, these public and private initiatives have accelerated progress toward the large-scale production of lignocellulosic ethanol.

Since 2000, a series of federal laws have established and strengthened the United States' commitment to renewable fuel production. Early acts such as the Biomass Research and Development Act of 2000 (Pub. L. No. 106-224, 2000) (**Table 1**) and the Farm Security and Rural Investment Act of 2002 (Pub. L. No. 107-171, 2002) (**Table 1**) funded research and infrastructure for producing fuels from plant materials such as trees or grasses. The Energy Policy Act of 2005 (Pub. L. No. 109-58, 2005) introduced the first Renewable Fuel Standard, later expanded by the Energy Independence and Security Act of 2007 (Pub. L. No. 110-140, 2007) (**Table 1**), which set ambitious production targets for cellulosic and other advanced biofuels. Subsequent legislation, including the Food, Conservation, and Energy Act of 2008 (Pub. L. No. 110-246, 2008), the Agricultural Act of 2014 (Pub. L. No. 113-79, 2014), the Energy Act of 2020 (Pub. L. No. 116-260, 2020), and the Inflation Reduction Act of 2022 (Pub. L. No. 117-169, 2022) (**Table 1**), continued to support bioenergy innovation through research funding, biorefinery programs, and tax incentives.

Table 1: U.S. Policies Supporting Bioethanol Development (2000–2022)

Together, these policies have created a strong base for developing lignocellulosic ethanol and other advanced biofuels in the United States. This national move toward renewable fuels helps lower greenhouse gas emissions, reduces dependence on imported oil, and supports the use of sustainable energy sources made from agricultural and forestry waste instead of food crops. Unlike first-generation biofuels made from corn or sugarcane, lignocellulosic ethanol production makes use of non-edible plant matter, offering a more sustainable and carbon-neutral alternative.

Lignocellulosic biomass is composed primarily of cellulose, hemicellulose, and lignin (Jayme, 1944) (Wise & Ratliff, 1947). To obtain fermentable sugars, the biomass must undergo physical, chemical, or enzymatic pretreatment, which breaks down complex polymers into a sugar mixture known as a lignocellulosic hydrolysate (Sun & Cheng, 2002). This hydrolysate typically contains sugars such as glucose, xylose, arabinose, mannose, and galactose (Mosier et al., 2005) (Sun & Cheng, 2002). However, during pretreatment, the acetyl groups attached to hemicellulose are hydrolyzed, releasing acetic acid as a byproduct (Mosier et al., 2005) (Sun & Cheng, 2002). This is a serious roadblock to the utilization of lignocellulosic biomass to produce ethanol. When lignocellulosic biomass is broken down to release the sugars which can be fermented into ethanol by yeasts, the microbial inhibitor acetic acid is also released (Palmqvist & Hahn-Hägerdal, 2000).

In the production of lignocellulosic hydrolysates acetic acid is generated from the deacetylation of hemicellulose and other acetylated polysaccharides during acid or steam

pretreatment (Mosier et al., 2005). Once released, acetic acid can readily diffuse into yeast cells in its undissociated form, where it dissociates inside the cytoplasm and lowers intracellular pH (Pampulha & Loureiro-Dias, 1989). This disrupts metabolic balance, drains ATP through proton pumping, and slows growth (Pampulha & Loureiro-Dias, 1989). The inhibitory concentration (IC_{50}) of acetic acid for *Saccharomyces cerevisiae*, the most widely used yeast in ethanol fermentation, typically ranges between 50 and 150 mM, depending on strain and fermentation conditions (pH < 5.0) (Pampulha & Loureiro-Dias, 1989). Simply salting out the acetic acid as acetate does not solve the problem as acetate is just as toxic to microorganisms as acetic acid. Therefore, detoxification of acetate before fermentation is essential to achieving efficient ethanol yields.

Previous research has demonstrated the usefulness of the substrate-selective approach to create effective *Escherichia coli* strains that can detoxify the acetate from sugars derived from lignocellulosic biomass so they can subsequently be fermented by yeasts into ethanol (Eitemen *et al.* 2008). This strategy relies on engineering *E. coli* strains that can utilize acetate as a carbon source but cannot consume the fermentable sugars present in the hydrolysate, thereby protecting the sugar substrate for yeast fermentation. We aimed to create more effective *E. coli* strains that can metabolize acetate so they could be used to detoxify the acetate from lignocellulosic hydrolysates. *E. coli* can use acetate as a sole carbon source if acetate is converted into acetyl-CoA, which can be further metabolized via the tricarboxylic acid cycle (Wolfe, 2005) (**Figure 1**). Acetate can be converted into acetyl-CoA in two steps by acetate kinase (*ackA*) and

phosphotransacetylase (*pta)* or in a single step by acetyl-CoA synthetase (*acs*) (Brown *et al.* 1977) (**Figure 2**). These reactions are detailed below.

Acetate kinase (*ackA*) reaction
Acetate + ATP → acetyl phosphate + ADP

Phosphotransacetylase (*pta*) reaction
acetyl phosphate + CoA → acetyl-CoA + Pi

Acetyl-CoA synthetase *(acs*) reaction
Acetate + CoA + ATP → acetyl-CoA + AMP + 2Pi

]
‘

Figure 1: Central Metabolism via Glycolysis and the TCA Cycle. Bacteria such as *E. coli* can utilize acetate, a common byproduct of metabolism, as a sole carbon source if it is converted into acetyl-CoA, which is then metabolized through the tricarboxylic acid (TCA) cycle. Excessive acetate accumulation disrupts the tricarboxylic acid cycle and oxidative phosphorylation, ultimately inhibiting yeast growth and ethanol production, particularly in the context of lignocellulosic hydrolysates, making acetate detoxification crucial. (Source Gammons)

Mutations that enhance acetate utilization are frequently located in genes directly involved in acetate uptake and metabolism, such as *ackA*, *pta*, and *acs*, or in global regulators that influence carbon flow through central metabolic pathways, including *crp*, *iclR*, and *arcA* (Wolfe, 2005). Together, these genes coordinate how *E. coli* converts acetate into acetyl-CoA, integrates it into the tricarboxylic acid (TCA) cycle, and

balances energy production with stress resistance. Mutations that increase the expression or activity of these enzymes can improve the rate at which acetate is metabolized, allowing cells to grow faster when acetate is the sole carbon source.

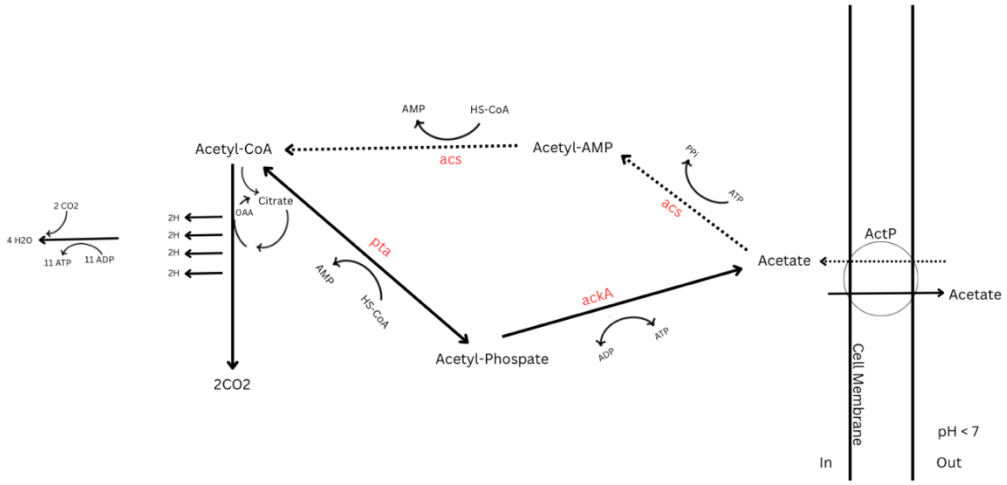

Figure 2: Pathways for Acetate Utilization and Acetyl-CoA Production. *E. coli* can convert acetate into acetyl-CoA via two different pathways. One pathway is a two-step process involving a single enzyme acetyl-CoA synthetase (*acs*), while the other is a two-step pathway involving two enzymes acetate kinase (*ackA*) and phosphotransacetylase (*pta*).

E. coli mutants that grow more robustly on acetate can at least in theory be easily generated if one or more of the enzymes mentioned in the previous paragraph are overproduced or enzymatically altered by spontaneous or chemically induced mutagenesis. *E. coli*, a versatile model organism and industrial host, possesses the metabolic capacity to utilize acetate as a carbon source, but its efficiency under stress conditions is often limited by regulatory and metabolic constraints. To address this challenge, adaptive evolution, spontaneous mutagenesis, and chemical mutagenesis were employed to generate variants with improved tolerance and metabolic response (**Figure 3**).

The industrial wild-type parent ALS1229 (ATCC8739) strain was chosen for this research since in a study by Rajaraman et al., 2016 that examined the ability of promising *E. coli* strains to utilize acetate as a sole carbon source, the ALS1229 strain was clearly the most superior at doing so. The industrial wild-type parent ALS1229 strain from which the mutants were isolated could only grow at maximal concentration of 1% acetate. Through the adaptive evolution strategy, we were able to induce mutant *E. coli* strains to grow on 3% acetate in M9 media by increasing the acetate concentration by 0.25% every 4-6 days. Using the spontaneous growth strategy, we increased the tolerance of ALS1229 to 5% sodium acetate from an overnight culture grown in LB media. With the EMS mutagenesis strategy, we were able to obtain mutants that could grow on up to 4.5% sodium acetate M9 plates.

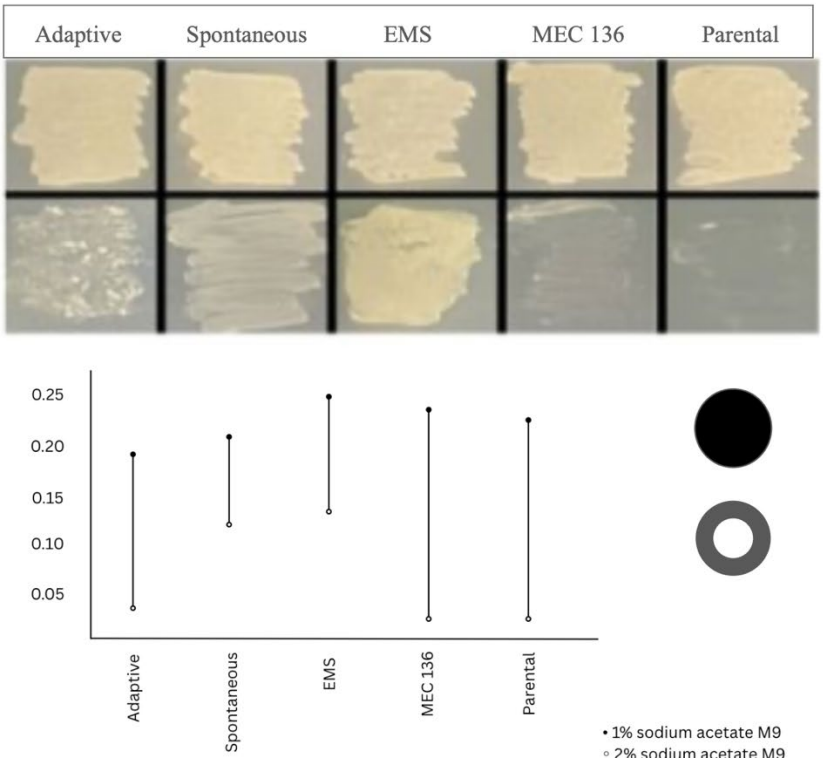

Figure 3: Comparison of mutant, parental, and previously best strain. Comparing the new acetate mutants to the *E. coli* parent and the best previously isolated acetate mutant by Rajaraman at al. 2016. Growth of each mutant type on minimal 1% or 2% acetate plates at 48 hours. Growth rates of each mutant type on 1% and 2% minimal acetate media. This study clearly showed that the new spontaneous of EMS mutants were superior to the *E. coli* parent or the best previously isolated acetate mutant by Rajaraman et al. 2016. (Source Gammons)

In *E. coli* ATCC 8739, beneficial mutations that improve acetate utilization generally fall into several recurring functional categories. First, many adaptive changes activate acetate assimilation genes such as *acs*, *ackA*, *pta*, *aceA*, and *aceB*, which encode key enzymes that convert acetate into acetyl-CoA and channel it into the tricarboxylic acid (TCA) and glyoxylate cycles (Brown et al., 1977; Wolfe, 2005). Brown et al. (1977) first described how the *ackA–pta* pathway interconverts acetate and acetyl-CoA, forming the biochemical foundation for acetate metabolism. Wolfe (2005) later synthesized decades of data into the concept of the acetate switch, explaining how *E. coli* transitions between acetate excretion and assimilation depending on environmental carbon availability. When these genes are upregulated or more active, cells oxidize acetate more efficiently, generating more energy and biosynthetic precursors when acetate is the only available carbon source (Castaño-Cerezo et al., 2009; Enjalbert et al., 2017). Enjalbert et al., 2017, further demonstrated that flux through the *pta–ackA* pathway is thermodynamically controlled and that increased activity in these enzymes improves growth efficiency on acetate.

Second, some mutations disable or reduce the activity of transcriptional repressors such as *iclR*, *arcA*, and *cra* (Wolfe, 2005). Wolfe (2005) showed that these regulators normally repress key enzymes in acetate metabolism under glucose-rich conditions. Rajaraman et al. (2016) explored how *E. coli* adapts to using acetate as its only carbon source by combining growth experiments, long-term evolution, and gene expression analysis. Among eighteen tested strains, ATCC 8739 showed the fastest growth on acetate, with a rate of about 0.41 h^{-1} in media containing 85 mM acetate. This strain was

then continuously cultured under selective conditions in a chemostat, which produced an evolved strain, MEC136, that grew faster than the original. Genome sequencing of MEC136 identified a single mutation in the *rpoA* gene, which encodes a component of RNA polymerase, implying that improved growth resulted from changes in the regulation of gene expression rather than direct alterations to metabolic enzymes. Transcriptome profiling supported this idea, showing only a few genes with altered expression between the evolved and parent strains. Overall, the study showed that *E. coli* can become more efficient at growing on acetate mainly through subtle adjustments in how genes are regulated, rather than through extensive genetic or metabolic changes.

Third, beneficial mutations often adjust the balance between stress response and growth by modifying regulators such as *rpoS*, *gadX*, and *hns* (Wolfe, 2005; De Mets et al., 2019). Wolfe (2005) described how acetate accumulation lowers intracellular pH and activates the acid stress regulon, which includes *gadX* and *hns*. De Mets et al. (2019) later revealed that small regulatory RNAs fine-tune acetate metabolism and coordinate it with the TCA cycle, linking stress adaptation directly to metabolic control. Under acetate stress, excessive activation of these systems protects cells but can slow growth; thus, the reduced or modulated expression of rpoS, gadX, or hns.

Finally, subtle changes in replication-related genes such as *dnaA* can stabilize chromosome replication and overall cell physiology during acetate stress (Seong et al., 2020). Seong et al. (2020) showed that adaptive laboratory evolution under acetate selection can produce compensatory mutations that restore energy balance by optimizing ATP use, which may also synchronize DNA replication with metabolic state. Rajaraman

et al. (2016) observed similar coordination between replication control and carbon metabolism, supporting the interpretation that *dnaA* variants help maintain genomic stability under acetate stress. Collectively, these findings show that mutations across metabolic, regulatory, and stress-response pathways reprogram *E. coli* for more efficient acetate assimilation (Wolfe, 2005; Pinhal et al., 2019). These coordinated genetic changes enable *E. coli* to oxidize acetate more rapidly, generate energy more efficiently, and maintain homeostasis under acid or energy-limited conditions (Rajaraman et al., 2016; Pinhal et al., 2019).

Whole-genome sequencing (WGS) has become a standard method for examining genetic variation and adaptative mutations in microorganisms. Early work by Chain et al. (2009) emphasized how advances in sequencing technologies established consistent standards for bacterial genome projects, allowing accurate comparison across studies. Land et al. (2015) reviewed two decades of bacterial genome sequencing and highlighted how WGS has transformed our understanding of microbial physiology, evolution, and metabolic diversity by enabling genome-wide comparisons of mutations and gene content. Li (2011) introduced a statistical framework for single nucleotide polymorphism (SNP) identification and mutation discovery that remains a foundation for many current bioinformatics pipelines. Similarly, Nielsen et al. (2011) discussed challenges in identifying accurate genotypes from next generation sequencing data, underscoring the importance of quality control and algorithmic precision in variant detection. Barrick and Lenski (2013) applied these approaches to experimental evolution studies, showing how genome sequencing can track mutations over time and reveal the genetic basis of

adaptive traits in bacterial populations. Collectively, these studies demonstrate how WGS and reliable variant-calling methods have made it possible to connect genetic changes with functional and adaptive outcomes in microbial systems.

Accurate variant detection is vital in SNP calling to prevent misinterpretation of biological data and the unnecessary expenditure of lab time and resources. Commonly used pipelines typically involve mapping sequencing reads to a reference genome followed by the detection of genetic variants within the aligned sequences. Selecting the appropriate sequence aligner and variant caller can be difficult because there are numerous tools (with numerous versions) available.

Most benchmarking studies of these bioinformatics tools have been conducted on human datasets (Barbitoff et al., 2022; Lin et al., 2022; Bush et al., 2020; Schilbert et al., 2020; Zhao et al., 2020). Recent research comparing different short-read aligners and variant callers has shown that the variant caller has a greater impact on detection accuracy than the aligner (Barbitoff et al., 2022).

The accuracy of variant discovery in bacterial genomes depends on the computational pipeline used, particularly during sequence alignment and variant calling. Bacterial genomes are compact, haploid, and often contain highly repetitive or homologous regions, plasmids, and horizontally transferred elements. These characteristics complicate read alignment and variant detection, since even minor mapping inaccuracies can produce false SNP calls or cause true variants to be missed.

Therefore, the choice of alignment and variant-calling tools directly determines the accuracy and biological interpretability of detected mutations in prokaryotic systems.

For bacterial short-read alignment, commonly used tools include Bowtie2 (Langmead & Salzberg, 2012) and BWA (Li & Durbin, 2009). Bowtie2 and BWA are preferred aligners for microbial genomes because they efficiently handle short Illumina reads and provide high alignment accuracy in the absence of introns. Bowtie2 is well suited for *E. coli* and similar prokaryotic species due to its ability to manage high read depth and small genome size without sacrificing speed or precision. Minimap2, though primarily designed for long-read data from platforms like Oxford Nanopore or PacBio, is also used for hybrid assemblies and has shown strong performance in detecting large indels and structural variants in bacterial isolates (Li, 2018).

After alignment, variant calling determines which genomic positions differ from the reference. Tools such as BCFtools (Li, 2011), DeepVariant (Poplin et al., 2018), FreeBayes (Garrison & Marth, 2012), GATK (McKenna et al., 2010), and VarScan2 (Koboldt et al., 2012) are frequently used for this step. BCFtools remains one of the most widely adopted programs for microbial variant analysis because it is lightweight, customizable, and integrated with SAMtools for depth-based filtering (Li, 2011). However, traditional statistical callers like BCFtools and VarScan2 rely heavily on base quality and read depth metrics, which can limit precision when distinguishing true variants from sequencing noise, particularly in repetitive or low-coverage regions common in bacterial genomes.

Recent research indicates that deep learning–based tools, such as DeepVariant, have surpassed conventional variant callers in accuracy when applied to bacterial genome data. While both traditional pipelines and modern neural network-based workflows are used for bacterial variant analysis, their performance characteristics differ. Barbitoff et al. (2022) demonstrated that differences in variant-calling accuracy across pipelines are driven more by the variant caller than the aligner, with machine learning models achieving higher precision in SNP and INDEL detection. Hall et al. (2024) further confirmed this trend, showing that on *E. coli* and Klebsiella pneumoniae nanopore datasets, DeepVariant produced variant calls with accuracy equal to or better than Illumina short-read pipelines, even at lower sequencing depths (10X). These results indicate that DeepVariant's neural network can adapt effectively to prokaryotic genomes, accurately recognizing and interpreting sequencing noise unique to bacterial data.

THESIS STATEMENT

The purpose of this study was to identify mutations present in *E. coli* strains that allow the strains to be capable of detoxifying acetate in lignocellulosic hydrolysates, thus enabling efficient ethanol production. By enhancing acetate metabolism through mutagenesis strategies, we have created *E. coli* mutants that can grow robustly on acetate as a sole carbon source, which, with mutations identified, will be crucial in overcoming the inhibitory effects of acetate which would negatively impact the fermentation process.

This identification step is crucial in the generation of methods to enable the fermentative production of ethanol from lignocellulosic sugars on a global scale. Additionally, I compared two variant calling pipelines in genome analysis to determine the differences in quality as well as accuracy in identifying the mutants that resulted in *E. coli* strains that were superior in utilizing acetate as a sole carbon source. This research is important for advancing bioethanol production, reducing reliance on non-renewable resources, and enabling more sustainable fuel alternatives.

METHODS

Two widely used variant-calling pipelines for bacterial WGS data (**Table 2**) were compared in order to identify the mutations that occurred in the best acetate-utilizing *E. coli* mutants, the spontaneous induced mutants X7, X8, X9, X10, and the EMS induced mutants X3, X4, X5, X6. Focusing on the collection of statistics that are directly relevant to comparing SNP detection between Pipeline 1 and Pipeline 2, to evaluate differences in variant detection sensitivity and precision. Pipeline 1 utilized Bowtie2 for alignment, followed by bcftools mpileup and bcftools call for variant detection. The indexed BAM files, genome sequence, and annotation files were used as inputs to identify single nucleotide polymorphisms (SNPs) and small insertions and deletions (INDELs). Variants were subsequently annotated with SnpEff to predict their potential functional effects, and variant positions were validated using RStudio and visualized in the Integrated Genome Viewer (IGV) to confirm read depth (greater than 100×) and genotype confidence. For pipeline 2, DeepVariant was supplied with Bowtie2-generated BAM files as input which ensured consistent mapping across both pipelines. However, because DeepVariant does

not output detailed alignment metrics (such as overall alignment rate, concordant/discordant pairs, or mapping quality distributions), these statistics were obtained exclusively from Bowtie2. The two pipelines were compared across all *E. coli* strains to evaluate performance in variant detection and annotation.

Table 2: Comparison of Variant-Calling Pipeline Characteristics. Pipeline 1 and Pipeline 2 differ in the middle detection step.

The core metrics used to compare variant calling performance included the total number of SNPs and INDELs, read depth (DP) at variant sites, variant quality scores (QUAL), predicted genotypes (GT), allele frequencies (AF), and allele depths (AD), as well as the SnpEff-derived total annotated variants and SNP/INDEL ratios. These statistics are outlined below:

Alignment:

1. Read Depth at variant positions (DP).

Variant Calling:

2. The total number of SNPs detected by each pipeline.
3. The depth of coverage (DP) at SNP sites to assess the reliability of the variant calls.
4. The variant quality (QUAL) score, which reflects the confidence in SNP calls.
5. The predicted genotype (GT) for each SNP, indicating how confident the tool is in the genotype assignment.
6. The variant allele frequency (AF), which measures the proportion of reads supporting the alternative allele at each SNP position.
7. The read counts for the reference and alternative alleles (AD) to understand how many reads support each allele.

Annotation:

8. The SNP/INDEL ratio, which compares the number of SNPs to insertions and deletions detected.

The core metrics are summarized and used to evaluate variant-calling performance in *E.coli* genome analyses below in **Table 3**. Each metric reflects a distinct aspect of variant detection accuracy, coverage, and confidence. Together, these values help compare pipeline sensitivity (bcftools) versus precision (DeepVariant) and provide context for interpreting SNP and INDEL calls across all analyzed strains.

Table 3. Key Variant-Calling Metrics and Their Analytical Roles.

Read depth (DP) is used in analysis to measure how reliable the sequencing data is at each genomic position. Higher DP values mean stronger evidence for a variant, while low values suggest the site may not have been read enough times to trust the result. Analysts use DP to filter out low-confidence regions and confirm that important variants are well supported by the data.

SNP and INDEL counts are used to compare the number and types of variants detected between samples or pipelines. This helps identify whether one dataset has more genetic variation or if a particular tool is over or under calling certain mutation types. The QUAL score is used to assess confidence in variant calls. Analysts often set a QUAL threshold (for example, ≥30) to keep only variants with a low probability of being false positives. Genotype information (GT) is used to determine which alleles are present in each sample. It helps confirm whether a mutation is homozygous, heterozygous, or absent, which is important for interpreting genetic traits or comparing strains. Allele frequency (AF) is used to estimate how common a variant is within a sample or population. It helps distinguish true mutations (high AF) from sequencing noise (low AF). Allele depth (AD) provides read counts for each allele, allowing analysts to verify whether the allele frequency and genotype are consistent with the data. It's often used to check for uneven coverage or contamination. The SNP/INDEL ratio is used as a quality metric to

summarize mutation patterns across the genome. Abnormally low or high ratios can indicate issues with variant calling or sequencing quality. Annotated variants are used in downstream functional analysis to connect mutations to genes and predict their biological effects. This step helps determine which variants are most likely to affect protein function or gene regulation.

This dataset integrates three key sources of information that together describe the alignment, variant calling, and annotation performance for each *E. coli* strain. First, Bowtie2 mapping statistics include total reads, mapped reads, overall alignment rate, and the number of properly paired alignments. These values provide a direct measure of sequencing quality and alignment accuracy, indicating how efficiently each sample's reads were mapped to the *E. coli* ATCC 8739 reference genome. The mapping summary outputs were generated immediately after alignment and before variant calling to ensure that only high-quality alignments were used in subsequent analyses. Second, bcftools mpileup and call outputs provided variant-level information such as the total number of SNPs and INDELs, read depth at variant positions (DP), variant quality scores (QUAL), predicted genotypes (GT), allele frequencies (AF), and allele depths (AD). These metrics represent the variant detection data and confidence parameters used for downstream filtering, comparison, and annotation. They serve as the basis for evaluating how many high-confidence variants were detected and how strongly supported each was by sequencing reads. Finally, SnpEff annotations classified each identified variant by its genomic context, such as coding, intergenic, or regulatory regions, and predicted its functional consequence (e.g., synonymous, missense, nonsense, or frameshift). SnpEff

also produced per-strain totals for annotated variants and calculated the SNP/INDEL ratio, which was used to assess the overall variant composition and to identify any potential biases toward over or under calling specific variant types. These three data sources provide variant detection and annotation outcomes for each pipeline.

Genomic DNA was extracted from eight *E. coli* mutant strains of *E. coli* ATCC 8739. Four mutants (X3–X6) were generated through ethyl methanesulfonate (EMS) mutagenesis, while the remaining four (X7–X10) were obtained via spontaneous mutagenesis. High-throughput sequencing was performed by Novogene on the Illumina NovaSeq 6000 PE150 platform, producing paired-end reads with an average coverage depth of approximately 100X. Raw reads were assessed for quality using FastQC to identify potential issues such as low-quality bases or adapter contamination, after which trimmed, reads were aligned to the *E. coli* ATCC8739 reference genome (NCBI accession GCF_000019385.1) using Bowtie2, and alignments were sorted by reference coordinates.

RESULTS

The goal of this project was to compare the sensitivity and precision of two bioinformatics pipelines used for variant discovery. As the first step in each pipeline is alignment, we first investigated alignment metrics for the common pipeline aligner, Bowtie2. **Figure 4** summarizes the Bowtie2 alignment performance across eight *E. coli* mutant strains (X3–X10) against the *E. coli* ATCC 8739 reference genome.

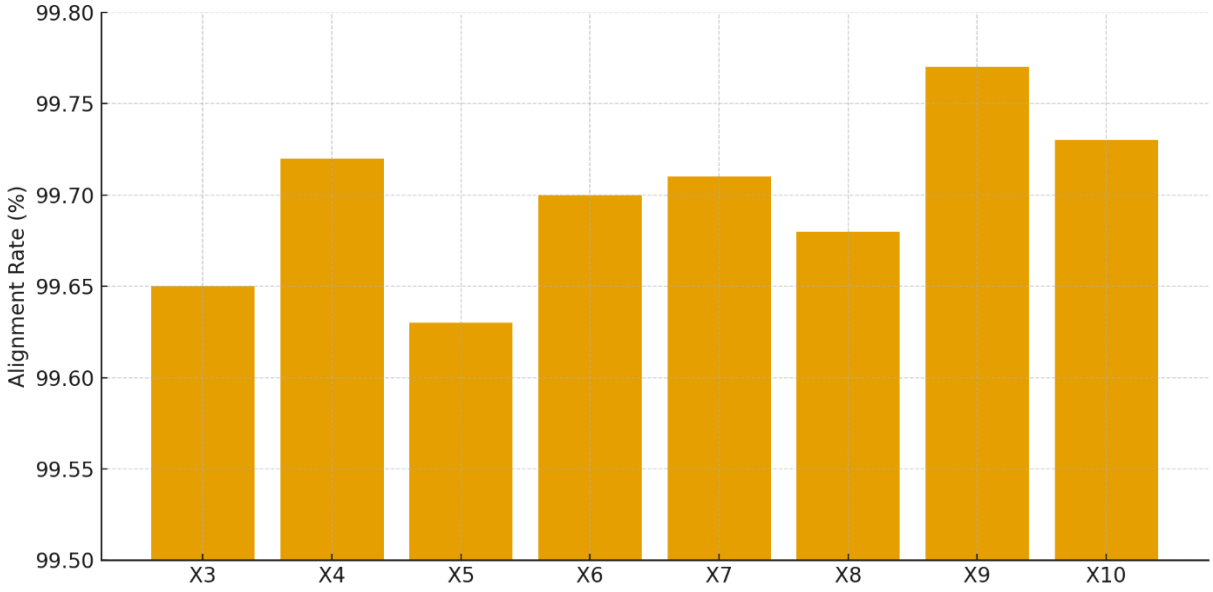

Figure 4: **Bowtie2 Alignment Summary** for eight *E. coli* mutant strains (X3–X10) aligned to the *E. coli* ATCC 8739 reference genome. Each strain shows total read count, alignment distribution, and overall mapping rate. All samples achieved >99.6% overall alignment, indicating high sequencing quality and strong compatibility between the Illumina read data and the reference genome, ensuring reliability for downstream variant calling and annotation analyses.

Next, as each pipeline uses a different variant detection tool, we compared the metrics for variant calling by mpileup and bcftools (pipeline 1) and DeppVariant (pipeline 2). **Table 4** presents a comparative summary of variant profiles obtained using two Bowtie2-based analysis pipelines, highlighting consistent SNP detection trends but differing sensitivity between bcftools and DeepVariant-based workflows.

Table 4. Comparison of variant-calling metrics. Between two *E. coli* analysis pipelines, comparison of results is detailed below.

ANALYSIS

In comparison, DeepVariant generated slightly fewer total variants but with higher variant quality (QUAL) scores and tighter depth distributions, reflecting improved precision and confidence modeling. Average QUAL scores, which reflect confidence in variant calls strains exceed 180 mean QUAL in most strains, indicating high-quality calls (**Figure 5**). Slightly lower scores in X8 and X9 suggest modest reductions in read support or alignment clarity for those samples.

Figure 5. Average variant quality (QUAL) per strain generated using two SNP-calling pipelines.

Next, we were interested in the comparison of call type, so we compared the mutation type reported for mpileup and DeepVariant (**Figure 6**). Bcftools reported marginally higher SNP and INDEL counts along with allele frequency (AF) and allele depth (AD) metrics, indicating greater sensitivity and providing transparency for manual variant validation. Overall, both pipelines are consistent for *E. coli* SNP and INDEL detection, with DeepVariant favoring accuracy and bcftools favoring interpretability.

Across all *E. coli* strains analyzed, comparison of Pipeline 1 (Bowtie2 + bcftools + SnpEff) and Pipeline 2 (Bowtie2 + DeepVariant + SnpEff) revealed differences in both the quantity and quality of detected variants. Pipeline 1 consistently identified a greater

total number of variants, particularly single-nucleotide polymorphisms (SNPs), resulting in a higher SNP/INDEL ratio.

Figure 6. **SNP/INDEL Ratio per Yeast Strain.** This measures a variant-type distribution. Bcftools shows a consistent SNP bias (ratios >2 across strains), aligning with its known sensitivity toward single-base substitutions relative to indels.

One feature that could impact variant calling is depth of coverage. To examine this feature, we compared the DP metric for both pipelines. This pattern indicates that bcftools is more sensitive but also more prone to over-calling variants, especially in regions of low coverage, repetitive sequences, or ambiguous read alignments. The depth of coverage (DP) remained comparable between pipelines, as both analyzed the same

sequencing data; however, DeepVariant exhibited a narrower DP distribution, suggesting more uniform read support across variant sites.

Figure 7. Average Read Depth (DP) at Variant Sites. This summarizes the mean sequencing depth at variant positions. All samples maintain >130× coverage, demonstrating uniform data quality across libraries and providing strong support for variant detection confidence.

Next, we considered differences in allele metrics. The QUAL values reported by bcftools were more variable and included many moderate-confidence sites, whereas DeepVariant's QUAL scores were tightly clustered at higher values due to its neural network–based confidence modeling. Differences in allele-level statistics further illustrated this contrast: DeepVariant's variants typically corresponded to allele frequencies (AF) near 1.0, consistent with true fixed haploid polymorphisms, while bcftools occasionally reported intermediate AF values (0.4–0.8), indicating uncertain or

partially supported sites. Likewise, bcftools' allele depth (AD) distributions were broader, reflecting the inclusion of lower confidence reads.

Bcftools exhibited higher sensitivity in detecting variants, capturing a broader range of potential polymorphisms, while DeepVariant achieved superior precision and confidence in each call. These complementary strengths indicate that bcftools is best suited for exploratory or high-throughput analyses, where maximizing detection is the priority, whereas DeepVariant is better optimized for confirmatory or publication-grade studies, where accuracy and reliability are essential in *E. coli* genomics. Annotation results derived from SnpEff confirmed that bcftools favored SNP detection and yielded a higher SNP/INDEL ratio, whereas DeepVariant produced a more balanced mix of SNPs and INDELs, particularly in coding and regulatory regions. These differences can affect how genetic changes are interpreted. Pipelines that detect too many SNPs may include extra or uncertain calls, which can make it seem like there are more mutations in important genes than there really are. This can lead to false positives changes that appear to affect protein function but don't. Conversely, more cautious pipelines, such as those using DeepVariant, focus on high-confidence variants that are strongly supported by the sequencing data. Although they may find fewer total variants, the ones they do identify are more likely to be real and biologically meaningful, especially in genes or regulatory regions where accuracy matters most. Choosing between the two approaches depends on the goal: finding all possible changes versus finding the ones most likely to truly affect function. Taken together, these findings demonstrate that while bcftools offers higher sensitivity and broader variant detection, DeepVariant provides greater accuracy and

confidence in variant calls, making it more suitable for functional genomics and comparative analyses requiring precise variant characterization.

We have isolated eight *E. coli* mutants (4 EMS and 4 spontaneous) with an enhanced ability to utilize acetate, which can be applied in the acetate detoxification process our laboratory has developed for lignocellulosic hydrolysates. The genetic changes identified in these *E. coli* strains are not concentrated in canonical acetate metabolism genes such as *acs*, *pta*, or *ackA*. Instead, they primarily affect genes involved in stress response, global regulation, membrane integrity, and transport (**Table 5**). Mutations in regulators like *gadX* and *ytcA*, and in membrane-associated or transposase genes (e.g., ECOLC_RS04560, ECOLC_RS20895), suggest adaptive modifications that help the cell maintain pH balance, membrane stability, and metabolic flexibility when acetate is the sole carbon source. One mutation, ECOLC_RS18465, encodes a DUF905 domain-containing protein that may modulate acetate and acetyl-CoA flux, providing the most direct link to acetate metabolism. However, the broader pattern of mutations points to indirect adaptation, cells may tolerate the acidic by-products of acetate utilization more effectively and maintain redox balance through improved proton export and altered global regulatory control. These mutations likely enhance acetate utilization not by altering core acetate pathway enzymes, but by optimizing cellular physiology and stress resilience to support growth in an acetate-only environment.

Table 5 Mutations identified in eight EMS or spontaneous *E. coli* acetate mutants.

Sequencing and variant analysis of parental and mutant *E. coli* strains were conducted using NovaSeq 6000 (PE150). Quality checks were performed with FastQC, followed by read trimming and alignment using Bowtie2. Reads were sorted by reference coordinates, and variant calling was done with bcftools, which generated indexed .bam and .vcf files. Variants were annotated and their effects analyzed using SNPeff. Variant positions were confirmed in RStudio, and unique variants in evolved strains were identified. Coverage depth and variant verification were conducted using IGV.

Across both EMS-induced and spontaneous mutants, many mutations occur in genes associated with stress response, metabolism, and genome stability. **Table 6** shows the distribution of the eight mutations that occurred in the eight mutated strains X3 – X10. In the EMS mutants, frequent changes in *dnaA* and *gadX* indicate altered control of DNA replication and improved acid resistance, likely supporting growth under acetate stress. Additional mutations in *glgX*, *ytcA*, and *ECOLC_RS14775* may influence glycogen metabolism, membrane transport, and phage repression, enhancing stress tolerance and cellular stability. In contrast, the spontaneous mutants often carried mutations in dnaA, *ECOLC_RS04560*, and *ECOLC_RS18465*, genes associated with genome maintenance and acetate or acetyl-CoA regulation, reflecting distinct routes of metabolic adaptation. Overall, these results suggest that the EMS and spontaneous

mutants evolved through different but complementary mechanisms to improve acetate utilization, stress tolerance, and metabolic flexibility. These strains represent promising candidates for further engineering to enhance detoxification efficiency.

Table 6: Mutations present in each strain.

	Mutants							
Mutations	**X3**	**X4**	**X5**	**X6**	**X7**	**X8**	**X9**	**X10**
dnaA	x			x			x	X
gadX		x	x	x				
glgX					x			
ytcA							x	
ECOLC_RS04560						x	x	x
ECOLC_RS14775		x	x	x				
ECOLC_RS18465							x	
ECOLC_RS20895		x	x	x				

CONCLUSION

This study demonstrates that *Escherichia coli* can acquire improved capacity for acetate utilization through diverse genetic routes that primarily target regulatory, stress-response, and structural stability pathways rather than direct enzymatic components of the acetate assimilation network. Whole-genome sequencing of eight mutants, four derived from EMS mutagenesis and four from spontaneous mutagenesis, revealed consistent alterations in genes involved in acid resistance (*gadX*), replication control (*dnaA*), membrane maintenance (*ytcA)*, glycogen turnover (*glgX*), and mobile-element regulation (*ECOLC_RS04560*), along with mutations in poorly characterized loci such as

ECOLC_RS18465 and *ECOLC_RS20895* that likely modulate proton balance and acetyl-CoA flux.

Collectively, these mutations represent a coordinated physiological adaptation that allows *E. coli* to maintain intracellular pH, stabilize membrane integrity, and optimize ATP usage during growth on acetate as the sole carbon source. Rather than increasing flux through the *acs–pta–ackA* pathway directly, the mutants appear to improve the energetic context in which acetate metabolism occurs, reducing proton leakage, minimizing ATP drain, and preventing genome instability under prolonged acid stress. This systems-level reprogramming aligns with previous adaptive evolution studies, supporting the notion that robust acetate assimilation in bacteria is achieved through global regulatory tuning rather than single-enzyme modification.

The comparative analysis of variant-calling pipelines further established that while the traditional Bowtie2 + bcftools workflow provides broader variant detection and interpretability, the Bowtie2 + DeepVariant pipeline offers greater precision and confidence in bacterial SNP and INDEL identification. Integrating both approaches allowed comprehensive and reliable discovery of functionally relevant mutations across all strains. From a biotechnological perspective, the findings highlight a valuable route toward developing acetate-detoxifying biocatalysts for lignocellulosic ethanol production. Strains combining enhanced acid tolerance, controlled replication, and metabolic stability can efficiently remove inhibitory acetate from hydrolysates before fermentation by yeast, thereby increasing ethanol yield and process reliability.

In summary, the evolved *E. coli* mutants generated in this study exemplify how regulatory and structural adaptations, not just metabolic rewiring, can yield substantial physiological improvements. These insights expand our understanding of bacterial acetate metabolism and provide a genetic foundation for engineering industrial strains capable of thriving in harsh, acetate-rich environments. Future work should focus on reconstructing individual mutations to verify their specific contributions, performing transcriptomic and metabolomic analyses to map downstream regulatory effects, and evaluating performance in mixed-culture fermentations with yeast. Such integrative studies will bridge molecular genetics and process engineering, advancing the sustainable production of bioethanol from lignocellulosic feedstocks.

REFERENCES

Agricultural Act of 2014, Pub. L. No. 113-79, 128 Stat. 649 (2014).

Andrews, S. (n.d.). FastQC: A quality control tool for high throughput sequence data (Galaxy Version 0.74+galaxy1). Babraham Bioinformatics. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Barbitoff, Y. A., Abasov, R., Tvorogova, V. E., & others. (2022). Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery. *BMC Genomics, 23*, 155. https://doi.org/10.1186/s12864-022-08365-3

Biomass Research and Development Act of 2000, Pub. L. No. 106-224, 114 Stat. 428 (2000).

Brown, T. D. K., Jones-Mortimer, M. C., & Kornberg, H. L. (1977). The enzymic interconversion of acetate and acetyl-coenzyme A in *Escherichia coli*. *Journal of General Microbiology, 102*, 327–336. 10.1099/00221287-102-2-327

Bush, S. J., Foster, D., Eyre, D. W., Clark, E. L., De Maio, N., Shaw, L. P., ... & Walker, A. S. (2020). Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism–calling pipelines. *Gigascience, 9*(2), giaa007. 10.1093/gigascience/giaa007

Castaño-Cerezo, S., Pastor, J. M., Renilla, S., Bernal, V., Iborra, J. L., & Cánovas, M. (2009). An insight into the role of phosphotransacetylase (pta) and the acetate/acetyl-CoA node in Escherichia coli. *Microbial cell factories*, *8*(1), 54.

Castaño-Cerezo, S., Pastor, J. M., Renilla, S., Bernal, V., Iborra, J. L., & Cánovas, M. (2009). An insight into the role of phosphotransacetylase (pta) and the acetate/acetyl-CoA node in Escherichia coli. *Microbial cell factories*, *8*(1), 54.

Chain, P. S. G., et al. (2009). *Genome project standards in a new era of sequencing. Science*, 326(5950), 236–237.

Chong, H., Yeow, J., Wang, I., Song, H., & Jiang, R. (2013). Improving acetate tolerance of Escherichia coli by rewiring its global regulator cAMP receptor protein (CRP). *PloS one*, *8*(10), e77422.

Cingolani, P., Platts, A., Wang, leL., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). 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. https://doi.org/10.4161/fly.19695

Dale, B. E., Henk, L. L., & Shiang, M. I. N. G. (1985). Fermentation of lignocellulosic materials treated by ammonia freeze-explosion.

De Mets, F., Van Melderen, L., & Gottesman, S. (2019). Regulation of acetate metabolism and coordination with the TCA cycle via a processed small RNA. *Proceedings of the National Academy of Sciences*, *116*(3), 1043-1052.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ... & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Eitman, M. A., & Lee, S. A., & Altman, E. (2008). A co-fermentation strategy to consume sugar mixtures effectively. *Journal of Biological Engineering, 2*, 3.
Energy Act of 2020, Pub. L. No. 116-260, 134 Stat. 1182 (2020).
Energy Independence and Security Act of 2007, Pub. L. No. 110-140, 121 Stat. 1492 (2007).
Energy Policy Act of 2005, Pub. L. No. 109-58, 119 Stat. 594 (2005).
Enjalbert, B., Millard, P., Dinclaux, M., Portais, J. C., & Létisse, F. (2017). Acetate fluxes in Escherichia coli are determined by the thermodynamic control of the Pta-AckA pathway. *Scientific reports*, *7*(1), 42135.
Farm Security and Rural Investment Act of 2002, Pub. L. No. 107-171, 116 Stat. 134 (2002).
Food, Conservation, and Energy Act of 2008, Pub. L. No. 110-246, 122 Stat. 1651 (2008).
Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907.
Hall, M. B., Wick, R. R., Judd, L. M., Nguyen, A. N., Steinig, E. J., Xie, O., ... & Coin, L. (2024). Benchmarking reveals superiority of deep learning variant callers on bacterial nanopore sequence data. *Elife*, *13*, RP98300.
Inflation Reduction Act of 2022, Pub. L. No. 117-169, 136 Stat. 1818 (2022).
Jayme, G. (1944). *Untersuchungen über Lignocellulose. Cellulose-Chemie, 22*, 89–96.
Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods, 12*(4), 357–360. https://doi.org/10.1038/nmeth.3317
Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., ... & Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research, 22*(3), 568–576. https://doi.org/10.1101/gr.129684.111
Land, M. L., et al. (2015). *Insights from 20 years of bacterial genome sequencing. Functional & Integrative Genomics*, 15(2), 141–161.
Landmead, B., & Salzberg, S. L. (2023). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9, 357–359. https://botwtie-bio.sourceforge.net//bowtie2/index.shtml
Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2 (Version 2.5.3). *Nature Methods, 9*(4), 357–359. https://doi.org/10.1038/nmeth.1923
Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*(4), 357–359. https://doi.org/10.1038/nmeth.1923
Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics, 27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509
Li, H. (2011). *A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics*, 27(21), 2987–2993.
Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics, 34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics, 25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools (Version 1.19.2). *Bioinformatics, 25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., & others. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lin, Y. L., Chang, P. C., & Hsu, C., & others. (2022). Comparison of GATK and DeepVariant by trio sequencing. *Scientific Reports, 12*, 1809. https://doi.org/10.1038/s41598-022-05833-4

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research, 20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Mosier, N., Wyman, C., Dale, B., Elander, R., Lee, Y. Y., Holtzapple, M., & Ladisch, M. (2005). Features of promising technologies for pretreatment of lignocellulosic biomass. *Bioresource technology*, *96*(6), 673-686.

National Center for Biotechnology Information (NCBI). (2008, March 11). *Escherichia coli strain ATCC 8739, complete genome (GCF_000019385.1)*. NCBI Datasets. https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000019385.1/

National Center for Biotechnology Information (NCBI). (2022). NCBI Datasets.

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature reviews. Genetics*, *12*(6), 443–451.

Palmqvist, E., & Hahn-Hägerdal, B. (2000). Fermentation of lignocellulosic hydrolysates. I: inhibition and detoxification. *Bioresource technology*, *74*(1), 17-24.

Pampulha, M. E., & Loureiro-Dias, M. C. (1989). Combined effect of acetic acid, pH and ethanol on intracellular pH of fermenting yeast. *Applied microbiology and biotechnology*, *31*(5), 547-550.

Pinhal, S., Ropers, D., Geiselmann, J., & De Jong, H. (2019). Acetate metabolism and the inhibition of bacterial growth by acetate. *Journal of bacteriology*, *201*(13), 10-1128.

Poplin, R., Chang, P. C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., ... & DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology, 36*(10), 983–987. https://doi.org/10.1038/nbt.4235

Rajaraman, E., Agarwal, A., Crigler, J., Seipelt-Thiemann, R., Altman, E., & Eiteman, M. A. (2016). Transcriptional analysis and adaptive evolution of *Escherichia coli* strains growing on acetate. *Applied Microbiology and Biotechnology, 100*, 7777–7785.

Schilbert, H. M., Rempel, A., & Pucker, B. (2020). Comparison of read mapping and variant calling tools for the analysis of plant NGS data. *Plants (Basel, Switzerland), 9*(4), 439. https://doi.org/10.3390/plants9040439

Seong, W., Han, G. H., Lim, H. S., Baek, J. I., Kim, S. J., Kim, D., Kim, S. K., Lee, H., Kim, H., Lee, S. G., & Lee, D. H. (2020). Adaptive laboratory evolution of Escherichia coli lacking cellular byproduct formation for enhanced acetate utilization through compensatory ATP consumption. *Metabolic engineering*, *62*, 249–259.

Shafin, K., Pesout, T., Chang, P. C., Nattestad, M., Kolesnikov, A., Goel, S., ... & Paten, B. (2021). Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nature Methods, 18*(11), 1322–1332.

Sun, Y., & Cheng, J. (2002). Hydrolysis of lignocellulosic materials for ethanol production: a review. *Bioresource technology*, *83*(1), 1-11.

Wise, L.E. and Ratliff, E.K. (1947) Quantitative Isolation of Hemicelluloses and Summative Analysis of Wood. Analytical Chemistry, 19, 459-462.

Wolfe, A. J. (2005). The acetate switch. *Microbiology and molecular biology reviews*, *69*(1), 12-50.

Zhao, S., Agafonov, O., Azab, A., Stokowy, T., & Hovig, E. (2020). Accuracy and efficiency of germline variant calling pipelines for human genome data. *Scientific Reports, 10*(1), 20222.