UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Maestría y Doctorado en Ciencias Bioquímicas

IDENTIFICACIÓN DE INTERVENCIONES GENÉTICAS PARA LA REDIRECCIÓN DE RECURSOS A FUNCIONES SINTÉTICAS.

TESIS

QUE PARA OPTAR POR EL GRADO DE:
Maestra en Ciencias

PRESENTA:
ELISA MÁRQUEZ ZAVALA

TUTOR PRINCIPAL
Dr. JOSÉ UTRILLA CARRERI
Centro de Ciencias Genómicas, UNAM

MIEMBROS DEL COMITÉ TUTOR
DR. ALVARO RAÚL LARA RODRÍGUEZ
División de Ciencias Naturales e Ingeniería, UAM
DRA. ROSA MARÍA GUTIÉRREZ RÍOS
Instituto de Biotecnología, UNAM

Ciudad de México. mayo, 2024

PROTESTA UNIVERSITARIA DE INTEGRIDAD Y HONESTIDAD ACADÉMICA Y PROFESIONAL
(Graduación con trabajo escrito)

De conformidad con lo dispuesto en los artículos 87, fracción V, del Estatuto General, 68, primer párrafo, del Reglamento General de Estudios Universitarios y 26, fracción I, y 35 del Reglamento General de Exámenes, me comprometo en todo tiempo a honrar a la Institución y a cumplir con los principios establecidos en el Código de Ética de la Universidad Nacional Autónoma de México, especialmente con los de integridad y honestidad académica.

De acuerdo con lo anterior, manifiesto que el trabajo escrito titulado:

IDENTIFICACIÓN DE INTERVENCIONES GENÉTICAS PARA LA REDIRECCIÓN DE RECURSOS A FUNCIONES SINTÉTICAS.

que presenté para obtener el grado de ----Maestría--------es original, de mi autoría y lo realicé con el rigor metodológico exigido por mi programa de posgrado, citando las fuentes de ideas, textos, imágenes, gráficos u otro tipo de obras empleadas para su desarrollo.

En consecuencia, acepto que la falta de cumplimiento de las disposiciones reglamentarias y normativas de la Universidad, en particular las ya referidas en el Código de Ética, llevará a la nulidad de los actos de carácter académico administrativo del proceso de graduación.

Atentamente

Elisa Márquez Zavala
312265470

El presente trabajo fue realizado en el Programa de Biología de Sistemas y Biología Sintética del Centro de Ciencias Genómicas de la Universidad Nacional Autónoma de México bajo la dirección del Dr. José Utrilla Carreri.

Se contó con apoyo económico de los proyectos UNAM–DGAPA-PAPIIT IN213420, Newton Advanced Fellowship Project NA160328 y de una Beca Nacional CONACYT de maestría con clave: 2019-000037-02NACF-22530.

Agradecimientos

Primero, me gustaría expresar mi más sincero agradecimiento a los miembros de mi comité, el Dr. José Utrilla Carreri, el Dr. Álvaro Raúl Lara Rodríguez y la Dra. Rosa María Gutiérrez Ríos, quienes siempre estuvieron dispuestos a brindarme su apoyo tanto académico como personal. En especial, al Dr. José Utrilla Carreri, con quien siempre me sentí en confianza para compartir cualquier preocupación. Su orientación me ayudó a descubrir y sentirme más segura sobre el camino que deseo seguir en la ciencia y en la vida. Su dedicación y apoyo constante han sido invaluables para mi desarrollo académico y personal. Es una persona excepcional, cuyo compromiso con sus estudiantes va más allá de lo esperado, creando un ambiente de trabajo motivante y agradable en su grupo, donde he vivido momentos muy gratificantes.

También quiero agradecer a los miembros del jurado, el Dr. Juan Enrique Morett Sánchez, el Dr. Mauricio Alberto Trujillo Roldán, la Dra. Marcela Ayala Aceves, el Dr. Leonardo Collado Torres y el Dr. Alfredo Martínez Jiménez, por su valiosa retroalimentación y tiempo dedicado a mejorar este trabajo. Su gran calidad humana y comprensión hicieron de este proceso una experiencia enriquecedora y alentadora.

Quiero también agradecer a mi familia, especialmente a mis padres, Víctor y María Guadalupe, por creer en mí y apoyarme a lo largo de mi carrera. Gracias a ellos, he tenido la oportunidad de conocer a personas y lugares maravillosos que han ampliado mi perspectiva de la vida. Si hoy tengo la oportunidad de graduarme, se lo debo en gran parte a ellos y quiero que sepan que este logro no es solo mío. A mi hermana Nayeli, le agradezco por todo su apoyo y por siempre escucharme.

A quienes se han convertido en mi familia, Citlali, Valeria, Amaranta, Analí, Carmina y Dilan, gracias por las muchas experiencias compartidas durante la carrera. Aprendí mucho de ustedes y siempre supe que podía contar con un espacio seguro para hablar sin sentirme juzgada. A pesar de la distancia y los conflictos, siempre hemos encontrado la manera de mantenernos en contacto y unidos. Me brindaron un apoyo invaluable en momentos muy difíciles a distancia y son de lo mejor que me ha pasado en la vida.

Agradezco también el apoyo técnico de Víctor M. del Moral Chávez, Alfredo José Hernández Álvarez, Joel Gómez Espíndola e Iván Uhthoff Aguilera del CCG-UNAM, así como el apoyo logístico y moral de Iliana Bahena Arellano.

Finalmente, agradezco a CONACYT por el apoyo financiero otorgado, que ha sido esencial para poder llevar a cabo mis estudios. Su compromiso con el desarrollo de la ciencia y la educación en México ha sido fundamental para mi formación académica y profesional.

Resumen

El objetivo de este estudio fue determinar los costos relacionados con la producción de genoma y de proteoma, con el fin de proponer estrategias óptimas de minimización. Estas estrategias buscan generar diseños de intervenciones genéticas que redirijan recursos celulares a funciones biotecnológicas. Para la comparación de estrategias de minimización se utilizaron datos reportados en la literatura de *Escherichia coli,* de nueve cepas minimizadas por el enfoque de genoma y uno por el enfoque de proteoma. También se utilizaron datos publicados de proteoma de *E. coli* en 22 condiciones y datos de contribución a la adecuación de 3789 mutantes (también de *E. coli*) en 166 condiciones. Para cuantificar la diferencia de costos, además de los datos mencionados, se utilizó un modelo de metabolismo y expresión de *E. coli.* Nuestros hallazgos indicaron que los costos relacionados a la producción de proteoma son mayores que los relacionados a la producción de genoma. También encontramos que los blancos que más recursos relacionados al proteoma pueden liberar suelen estar asociados a una afectación negativa a la adecuación. Asimismo, observamos que tan solo seis genes predecían una liberación de recursos relacionados a proteoma equivalentes a la cepa minimizada MDS12, que tiene 423 genes eliminados. Además de que las cepas que solo se enfocan en minimización genómica, como la Δ16, presentan problemas que afectarían su utilidad como cepas de producción. Por lo que un enfoque de minimización de proteoma racional, junto con un análisis de su contribución a la adecuación, nos puede permitir la identificación de pocos genes blanco que liberen una mayor cantidad de recursos celulares que podrán ser utilizados en funciones biotecnológicas sin afectar negativamente la adecuación de la célula.

Abstract

The aim of the study was to determine the costs associated with genome and proteome production, in order to propose optimal minimization strategies. These strategies aim to generate designs of genetic interventions that redirect cellular resources to biotechnological functions. For the comparison of minimization strategies, we used data reported in the *Escherichia coli* literature of nine strains minimized by the genome approach and one by the proteome approach. Published proteomic data in 22 conditions from *E. coli* and fitness contribution data from 3,789 mutants (also from *E. coli)* in 166 conditions were also used. To quantify the cost difference, besides the data mentioned above, a metabolism and expression model of *E. coli* was also used. Our findings indicated that the costs related to proteome production are higher than those related to genome production. We also found that targets that can release more proteome-related resources are often associated with a negative impact on fitness. Furthermore, we observed that only six genes predicted a proteome release equivalent to the minimized MDS12 strain, which has 423 genes deleted. In addition to the fact that strains that only focus on genomic minimization, such as Δ16 present problems that would affect their use as production strains. Therefore, a rational proteome minimization approach, together with an analysis of its contribution to fitness, can allow us to identify a few target genes that release a greater amount of cellular resources that can be used in biotechnological functions without negatively affecting the fitness of the cell.

Índice de contenido

Índice de figuras

Índice de tablas

Glosario

Asignación de recursos celulares: Proceso dinámico de la célula por el cual asigna cualquier tipo de recursos internos, como maquinaria celular, información genética, membranas y espacio intracelular

Carga proteómica: Cantidad total de proteínas expresadas en una célula en un momento determinado.

CBMs: Modelos matemáticos de metabolismo celular que utilizan restricciones para predecir el comportamiento de la célula.

Circuito sintético biológico: Componente dentro de la célula el cual, en analogía a los circuitos electrónicos, está diseñado para realizar funciones lógicas.

DMs: Modelos matemáticos de metabolismo celular que tienen en cuenta el tiempo y predicen el comportamiento de la célula en situaciones transitorias.

fg: Unidad de medida de masa que equivale a $1x10^{-15}$ gramos.

fgPC: Femtogramos de proteína por célula.

Funciones biotecnológicas: Funciones relacionadas con el desarrollo o manipulación de procesos o productos de interés humano.

Funciones de cobertura: Funciones celulares que permiten la supervivencia de la célula en condiciones cambiantes.

Funciones sintéticas: Funciones derivadas de circuitos de genes sintéticos.

Modelos grano fino: Modelos detallados de metabolismo celular que tienen en cuenta reacciones individuales y su cinética.

Modelos grano grueso: Modelos simplificados de metabolismo celular que agrupan reacciones similares y no tienen en cuenta la cinética de cada reacción.

Modelos M: Modelos de metabolismo a escala genómica.

Modelos ME: Modelos de metabolismo y expresión a escala genómica.

Proteoma: Totalidad de proteínas expresadas en una célula bajo una condición y tiempo determinado.

Proteoma de cobertura: Fracción del proteoma conformado por proteínas de cobertura.

Proteínas de cobertura: Proteínas que en una condición dada no se utilizan para crecimiento celular o están presentes en exceso y, por ende, operando debajo de su capacidad máxima.

RA: Reacciones asociadas.

UPF: Fracción de proteína sin modelar en el modelo ME. Sirve para representar el proteoma de cobertura.

Introducción y antecedentes

Biología sintética y de sistemas

La biología de sistemas es una rama de la bioinformática que estudia e integra las interacciones y comportamiento de una célula u organismo dentro de una misma escala y también dentro de diferentes escalas como lo son células y tejidos (Ideker et al., 2001; Kirschner, 2005; Tavassoly et al., 2018). Por otro lado, la biología sintética combina biología e ingeniería con el objetivo de crear nuevos sistemas biológicos o rediseñar completamente los sistemas existentes (Polizzi, 2013; Shapira et al., 2017). Por lo tanto, la combinación de la biología sintética con la biología de sistemas tiene gran potencial para permitirnos el estudio e ingeniería de organismos (D. Liu et al., 2013). El comprender la relación genotipo-fenotipo es una de las preguntas más relevantes para ambas ramas (O'Brien et al., 2013). Los sistemas biológicos son extremadamente complejos, por lo que para comprender la relación genotipo-fenotipo es necesario estudiar desde procesos moleculares hasta celulares. Los enfoques reduccionistas, que dividen los sistemas hasta sus componentes mínimos, han identificado exitosamente las bases bioquímicas de muchos procesos biológicos (O'Brien et al., 2013; Van Regenmortel, 2004). Sin embargo, muchas propiedades no pueden ser comprendidas estudiando sus partes individuales.

Con el advenimiento de las tecnologías ómicas, ahora tenemos acceso a una gran cantidad de información que nos brinda una visión cuantitativa y detallada de cambios en los procesos biológicos, además de muchas otras aplicaciones (Aebersold & Mann, 2003; De Godoy et al., 2008). Sin embargo, a pesar de tener acceso a toda esa información, muy poca ha sido analizada y/o integrada (Carrera et al., 2014; Ebrahim et al., 2016). En consecuencia, se necesitan métodos para analizar, comprender e integrar la gran cantidad de datos ómicos disponibles. En esta tesis, aprovechamos datos ómicos de la literatura para diseñar intervenciones genéticas.

Uno de los fines de conocer el funcionamiento de los organismos y poder modelarlos, es predecir cómo responderían a cambios en ellos mismos o en su ambiente (Balikó et al., 2018). Los circuitos sintéticos biológicos son componentes dentro de una célula que, en analogía a los circuitos electrónicos, están diseñados para realizar funciones lógicas (Kobayashi et al., 2004). Si deseamos agregar un circuito sintético biológico, debemos recordar que estamos hablando en el contexto de una célula viva y, como se mencionó, son sistemas extremadamente complejos. Lo ideal sería que estas nuevas funciones fueran predecibles y pudiéramos utilizar sus productos fácilmente, sin mucho procesamiento de los mismos. Sin embargo, las células suelen ser un ambiente impredecible para los circuitos sintéticos (Balikó et al., 2018). Debido a estas complicaciones, la creación de una célula simple que redujera nuestras variables y, por lo tanto, la interferencia, facilitaría la comprensión de la misma así como las predicciones a cómo respondería ante su ingeniería. Para la creación de estas células más simples, se ha hecho uso de modelos celulares (Balikó et al., 2018; Hirokawa et al., 2013; Karcagi et al., 2016; Kolisnychenko et al., 2002; Mizoguchi et al., 2008). En este trabajo, también se hizo uso de un modelo celular, el cual será descrito más adelante.

Modelos celulares

Una de las aplicaciones de la biología sintética y de sistemas es la creación de fábricas celulares (Patra et al., 2021). El modelado matemático de procesos celulares es cada vez más importante para el diseño de estas fábricas (Hafner et al., 2021). En particular, el modelado metabólico ha permitido crear diseños racionales y la integración de grandes conjuntos de datos para obtener redes biológicas, lo que ha llevado a nuevos conocimientos sobre los sistemas biológicos y una reducción del trabajo experimental en el desarrollo de fábricas celulares (Chen et al., 2016; Richelle et al., 2020). Los modelos basados en restricciones (*constraint-based models* o CBMs por sus siglas en inglés) y los modelos dinámicos (*dynamic models o* DMs por sus siglas en inglés), son los dos tipos principales de modelado metabólico. La principal diferencia entre los CBM y los DM es que los CBM asumen un estado estacionario, mientras que los DM brindan información sobre comportamientos transitorios (Gudmundsson & Nogales, 2021). Los CBM y DM son parte de las herramientas que pueden ayudarnos a analizar, comprender e integrar la cantidad de datos ómicos que ahora están disponibles (Richelle et al., 2020). Debido a esto, en este trabajo centramos nuestra atención en ellos, más específicamente en los modelos M y los modelos ME (detallados en la siguiente sección) para la integración de datos ómicos.

Modelos M

Los modelos de metabolismo a escala genómica (modelos M por **m**etabolismo), que en su mayoría son parte de los CBMs, representan las reacciones bioquímicas de producción y consumo de la red metabólica de un organismo a escala genómica que han sido determinadas experimentalemte con anterioridad (Baart & Martens, 2012; Lerman et al., 2012). Con esta información, han logrado predecir diversas funciones celulares (Bordbar et al., 2014; Lerman et al., 2012; Lewis et al., 2012; McCloskey et al., 2013; O'Brien et al., 2015). Sin embargo, los modelos M no toman en cuenta la producción de las enzimas ni otros procesos importantes para la célula como lo son la transcripción y la traducción. Por lo que su integración con datos ómicos está limitada por ejemplo a integrar datos ómicos como restricciones enzimáticas para los flujos metabólicos. Por ejemplo, abundancia de proteína y número de recambio (Åkesson et al., 2004; Becker & Palsson, 2008; Colijn et al., 2009; Lloyd et al., 2017; Shlomi et al., 2008). Debido a que estas limitaciones pueden dar como resultado predicciones biológicamente no posibles, se han realizado esfuerzos para que los modelos M sean más precisos, siendo uno de estos el que dio como resultado a los modelos ME que serán descritos a continuación (Feist & Palsson, 2008; Lloyd et al., 2017; Reed & Palsson, 2004; Schellenberger, Lewis, et al., 2011).

Modelos ME

Para aumentar las capacidades de los modelos M, se incluyó el proceso de expresión génica y se generaron los modelos de metabolismo y expresión (denominados modelos ME por **m**etabolismo y **e**xpresión) (Lerman et al., 2012; O'Brien et al., 2013; Thiele et al., 2012). Los modelos ME como los modelos M, también pueden simular los flujos metabólicos óptimos para un determinado estado, pero a diferencia de los modelos M, pueden simular la distribución óptima de proteoma que mantenga el fenotipo metabólico. Esto permite abordar nuevas preguntas biológicas, como lo son los cálculos de distribución de proteoma (LaCroix et al., 2015), cómo se usan las vías metabólicas y las implicaciones de las restricciones de membrana y volumen (J. K. Liu et al., 2014). Además, como

contempla abundacias óptimas de proteoma, es posible integrar datos transcriptómicos y proteómicos en ellos(Lloyd et al., 2017).

Los modelos ME conllevan varios retos, por eso hasta ahora solo hay para tres organismos, *Thermotoga maritima* (Lerman et al., 2012), *Escherichia coli* K12 MG1655 (J. K. Liu et al., 2014; O'Brien et al., 2013; Thiele et al., 2009, 2012) y más recientemente *Clostridium Ljungdahlii* (J. K. Liu et al., 2019). Entre los retos se encuentra la velocidad de resolución, ya que para encontrar los flujos óptimos para las condiciones dadas, los modelos ME pueden llegar a ser hasta cinco órdenes de magnitud más lentos (Lloyd et al., 2017). Otro reto involucra herramientas de software generalizadas, existentes en los modelos M (Agren et al., 2013; Ebrahim et al., 2013; Gelius-Dietrich et al., 2013; King et al., 2015; Schellenberger, Que, et al., 2011), mientras que los primeros modelos ME han utilizado su propia estructura de base de datos y módulos de código, lo cual ha obstaculizando su progreso y extrapolación (Lloyd et al., 2017). Se han logrado avances recientes en esta área, ya que tanto el modelo ME de *Escherichia coli* K12 MG1655 como el de *Clostridium Ljungdahlii* utilizan el software COBRAme, el cual se explicará con más detalle en la siguiente sección (J. K. Liu et al., 2019; Lloyd et al., 2017). Finalmente, debido a que los modelos ME son tan grandes, complejos y requieren una gran cantidad de datos, crearlos y analizarlos ha sido un trabajo de años y mucho esfuerzo (Lloyd et al., 2017). Como una solución para que la creación de los modelos ME fuera más accesible, se generó el programa COBRAme.

COBRAme

Usar COBRAme (Lloyd et al., 2017) para los modelos ME es similar a usar el programa COBRApy (Ebrahim et al., 2013) para los modelos M, ya que ambos son herramientas que simplifican la creación, manipulación, simulación, comparación e interpretación de modelos metabólicos, pero COBRAme permite agregar el proceso de expresión génica. COBRAme es un programa informático desarrollado para crear, manipular, simular e interpretar los resultados de los modelos ME. El modelo iJL1678b-ME se creó con ayuda de COBRAme, así como información de la expresión de genes de *Escherichia coli* (*E. coli*) y usando como base el modelo metabólico iJO1366 de *Escherichia coli* K12 MG1655 (Lloyd et al., 2017). COBRAme se puede usar para generar modelos para cualquier organismo con un modelo M y encuentra los flujos óptimos más rápido, con órdenes de magnitud, que los modelos ME anteriores (O'Brien et al., 2013). Al tratar varios de los problemas que se presentaban en la biología sintética y de sistemas, COBRAme demostró ser una herramienta de mucha utilidad para abordar preguntas fundamentales para la biología (Brenner, 2010), ser guía para diseñar cepas industriales (Otero & Nielsen, 2010) con aplicaciones como la producción de metabolitos de interés (fármacos, cosméticos, productos para industria alimenticia, etc.), biorremediación, generación de energía renovable, entre otras (Durot et al., 2009; Janssen et al., 2005; Peng et al., 2008; Rittmann, 2008; Ro et al., 2006); y proporcionar una perspectiva de sistemas para el análisis de la expansión de los datos ómicos (Palsson & Zengler, 2010). Aunque no hay que olvidar que obtener suficiente información metabólica y de expresión sigue siendo un gran problema al hacer este tipo de modelos. Debido a lo descrito anteriormente, optamos por enfocarnos en los modelos ME dentro de este trabajo para poder analizar datos de flujos de transcripción, traducción y metabólicos así como integrar datos ómicos .

Distribución del proteoma

Distribución del proteoma en *E. coli*

Hui *et al*. 2015 estudiaron cómo era la respuesta de la expresión génica de *E. coli* ante tres limitaciones metabólicas en crecimiento exponencial. Las limitaciones fueron en cuellos de botella cruciales de la red metabólica: catabolismo (controlando el consumo de lactosa en células creciendo en lactosa), anabolismo (cambiando la capacidad de asimilación de amonio como única fuente de nitrógeno) y traducción de los ribosomas (agregando cantidades subletales del inhibidor de traducción, cloranfenicol) (Hui et al., 2015; You et al., 2013). Aunque la red de regulación es compleja, encontraron que el proteoma bacteriano podría estar representado por un modelo de grano grueso, que es un modelo que agrupa niveles de organización celular con funciones similares en una pequeña cantidad de sectores (Bond et al., 2007; Doan et al., 2022). En este caso, su modelo contó con seis sectores. Dependiendo de cómo variaba la concentración de proteínas individuales ante las limitaciones, se les asignó un sector. También demostraron que la partición del proteoma en sectores está fuertemente relacionado con la división de flujos metabólicos. Esto sugiere un principio sobre cómo son asignados los recursos a nivel proteoma en la célula (economía del proteoma de la célula). Las estrategias de grano grueso de regulación global de genes son modelos útiles y más sencillos que sirven como base para futuros estudios sobre la expresión de genes y la construcción de circuitos biológicos sintéticos (Doan et al., 2022; Hui et al., 2015). Una parte importante del proteoma es la fracción de proteína no utilizada, que será detallada a continuación.

Fracción de proteína no utilizada

Los costos y beneficios de la síntesis de proteínas del organismo estructuran la regulación de su expresión y son equilibrados a través de la evolución (Dekel & Alon, 2005; O'Brien et al., 2016). La velocidad específica de crecimiento celular suele ser un componente determinante para la adecuación de las bacterias (Neidhardt, 1999), mientras que gran parte del proteoma es responsable del crecimiento en diversas condiciones (Li et al., 2014). Los costos y beneficios de las proteínas a menudo son cuantificados por sus efectos en el crecimiento celular (Dekel & Alon, 2005; Li et al., 2014; Neidhardt, 1999; Scott et al., 2010). Y se ha observado que la síntesis de proteínas no utilizadas o de cobertura (aquellas que en el entorno actual no se utilizan para crecimiento celular o están presentes en exceso y, por ende, operando debajo de su capacidad óptima) resulta en una reducción en las tasas de crecimiento que además se ha reportado como predecible y lineal (Bienick et al., 2014; O'Brien et al., 2016; Scott et al., 2010).

Las formas más comunes de caracterizar cómo se organiza el proteoma se basan en anotaciones de funciones de proteínas o los objetivos reguladores de la transcripción (O'Brien et al., 2016). Para comprender la magnitud y la variabilidad de la expresión de proteínas de cobertura en ambientes naturales, O'Brien *et al*., 2016 desarrollaron un enfoque novedoso basado en el uso de datos proteómicos absolutos y globales, así como el modelo de asignación de proteoma a escala genómica de *E. coli*, el modelo iJL1678-ME (O'Brien et al., 2013), ya que calcula cuánto cuesta cada gen en el proteoma y qué tan útil es en el medio especificado. Además, el modelo tiene un parámetro llamado "fracción de proteína sin modelar" o UPF por sus siglas en inglés (*Unmodeled Protein Fraction*), el cual sirve para representar aquellos procesos que no se utilizan para crecimiento celular (procesos que no se encuentran dentro del modelo). Más adelante se ahondará sobre la UPF en la sección "*Modelo ME de E. coli".* El estudio fue realizado en 16 de las 22

condiciones distintas que reportaron Schmidt *et al*., 2016 con datos absolutos de proteínas por célula (femtogramos por célula, siendo un femtogramo 10^{-15} gramos) para *E. coli* K12 BW25113 (Tabla 1). Solo se utilizaron 16, ya que los datos de medios enriquecidos que no están definidos y de fase estacionaria no se pueden simular fácilmente con el modelo ME. Sin embargo, como será explicado más adelante, nosotros integramos la información de las 22 condiciones en este trabajo al utilizar estos datos como complemento a la información entregada por el modelo ME.

En general, las cepas en las condiciones de Schmidt et al, 2016 estuvieron en cultivos por lote en matraces Erlenmeyer de 500 ml a una temperatura de 37°C con una agitación orbital de 300 rpm con un pH de siete y se tomaron muestras al llegar a fase estacionaria. Los casos que difieren son la condición de estrés por temperatura (se llevó a 42°C), estrés osmótico (se agregó 50 mM de NaCl), estrés por pH (se llevó a un pH de 6 con ácido clorhídrico fumante), estrés por hambruna (se mantuvieron en agitación de uno a tres días después de alcanzar la fase estacionaria) y la condición de quimiostato, donde se mantuvieron velocidades específicas de crecimiento de 0.12, 0.2, 0.35 y 0.5 h^{-1}.

Tabla 1. Condiciones y velocidades específicas en los que se midieron los datos proteómicos de Schmidt et al., 2016 en E. col K12 BW25113 y las velocidades específicas de crecimiento. (*) Condiciones no utilizadas en el estudio de O'brien et al., 2016.

Condiciones	Velocidad específica de crecimiento. (h^{-1})
LB	1.90
*Glicerol + AA	1.27
42°C + glucosa	0.66
*Fructosa	0.65
pH6 + glucosa	0.63
Glucosa	0.58
*Xilosa	0.55
Estrés-osmótico + glucosa	0.55
Quimiostato μ=0.5	0.5
Glicerol	0.47
*Manosa	0.47
Glucosamina	0.46
Succinato	0.44
Fumarato	0.42
Piruvato	0.4
Quimiostato μ=0.35	0.35
Acetato	0.3
Galactosa	0.26
Quimiostato μ=0.20	0.2
Quimiostato μ=0.12	0.12
*Fase estacionaria 1 día	0
*Fase estacionaria 3 días	0

Con este enfoque, O'Brien *et al*., 2016 encontraron que, en algunas condiciones, casi la mitad de la masa del proteoma (cuya abundancia fue reportada en número de copia por célula) entra en la categoría que definieron como proteína que no se utiliza (proteína que entorno especificado, no se utiliza para el crecimiento celular). Las condiciones con alrededor del 40% de masa de proteoma de cobertura fueron las de quimiostato con μ=0.12 y μ=0.20. Además, descubrieron que el costo de producir estas proteínas de cobertura explica más del 95% de la variación en las tasas de crecimiento de *E. coli* en las 16 condiciones distintas de Schmidt *et al*., 2016. Esto indica que la fracción de proteína de

cobertura puede variar significativamente dependiendo del entorno. Por otro lado, como el modelo iJL1678-ME de *Escherichia coli* ya había sido utilizado con los datos proteómicos de Schmidt *et al.,* 2016, decidimos agregar este conjunto de datos a nuestro análisis.

Además, a partir de experimentos de evolución adaptativa (ALE) en ocho cepas, también encontraron que los cambios en la asignación del proteoma a las fracciones usadas frente a las de cobertura son un mecanismo común para aumentar las velocidades específicas de crecimiento celular. Por lo tanto, estos cambios en asignación de proteoma pueden explicar la variabilidad en las velocidades específicas de crecimiento observadas en las 16 condiciones.

Un ejemplo de compensación ecológica para *E. coli* es el uso de una fuente de carbono diferente a glucosa, como la galactosa, la cual es menos común y está asociada a condiciones estresantes para la célula. Esto debido a que dará como resultado diferencias en las velocidades específicas de crecimiento a pesar de ser fuentes de carbono parecidas ya que su estrategia ecológica se ve reflejada en los patrones regulatorios (Mitchell et al., 2009; O'Brien et al., 2016; Perkins & Swain, 2009; Tagkopoulos et al., 2008). Como resultado, hay grandes costos a la adecuación asociados con la expresión de proteínas de cobertura, pero pueden ayudar a los organismos a adaptarse a condiciones cambiantes (O'Brien et al., 2016).

Modelo ME de *E. coli*

Como se mencionó, el modelo iJL1678b-ME de *Escherichia coli* K12 MG1655 se construyó utilizando la herramienta COBRAme. El modelo cuenta con reacciones de transcripción, traducción, formación de complejos y reacciones metabólicas. Además, este modelo ME fue un 87.5% exacto en sus predicciones de genes esenciales al compararlas contra un estudio de detección de genes esenciales en todo el genoma (al nivel de inactivación de un solo gen) en medio mínimo M9 adicionado con glucosa (Lloyd et al., 2017; Monk et al., 2017). Asimismo, este modelo ME puede encontrar los flujos que den una solución óptima a un ambiente dado en un tiempo de 10 min (Lloyd et al., 2017).

Si en el modelo ME fijamos un valor de consumo en los flujos de carbono (para simular la disponibilidad de glucosa), y cambiamos ese valor en otras simulaciones, el modelo ME (al igual que en el modelo de grano grueso de Basan *et al*. 2015) predice cambios en el metabolismo central del carbono. Esto es consistente con los cambios mostrados en conjuntos de datos fluxómicos de otras investigaciones (Nanchen et al., 2006; O'Brien et al., 2013; Schuetz et al., 2007, 2012). Por otra parte, al hacer la transición de las condiciones de carbono limitado a exceso de carbono en el modelo, donde el proteoma es la limitante, la mayoría de los cambios en el metabolismo central del carbono se deben al cambio a un catabolismo de carbono de rendimiento más bajo (que requiere una cantidad menor de proteoma). Mientras que a bajas velocidades específicas de crecimiento bajo la limitación de fuente de carbono, los cambios son debido a un cambio en la demanda de ATP. El modelo ME fue capaz de hacer la predicción tanto de la condición donde el proteoma es limitado y cuando hay cambios en la demanda de ATP (O'Brien et al., 2013).

Teniendo en cuenta todos los puntos anteriores, debería ser posible usar un modelo a escala genómica de metabolismo y expresión genética (ME) como una aproximación para

la identificación de intervenciones genéticas para la redirección de recursos celulares a funciones que vengan de circuitos de genes sintéticos (Beitz et al., 2022; Kobayashi et al., 2004) a las que llamaremos funciones sintéticas, o que también se redirijan a funciones biotecnológicas (desarrollo o manipulación de procesos o productos de interés humano) (Clarke & Kitney, 2020). En este estudio, se ha utilizado el modelo iJL1678b-ME de *Escherichia coli* para tal fin.

Proteína *dummy*

No todas las reacciones metabólicas modeladas en iJL1678b-ME están anotadas con el complejo enzimático conocido que cataliza la conversión metabólica . Por ende, para evitar que una reacción que no esté explícitamente anotada como espontánea no tenga un costo de operación, se agregó el complejo "*dummy*". El complejo *dummy* está compuesto por una proteína *dummy*, cuya composición de codones es representativa de un gen promedio en *E. coli*.

Otro conflicto que enfrenta el modelo ME es que no contiene todas las proteínas codificadas o expresadas en *E. coli K12* MG1655. Para lidiar con eso, se utiliza el parámetro UPF (el coeficiente o fracción de proteína no modelada). Esto asegura que se produzcan todas las proteínas que no llevan a cabo reacciones enzimáticas en el modelo. También permite que las proteínas que no están modeladas, son subutilizadas, o no son utlizadas en el modelo ME, estén representadas. La UPF requiere la producción de la cantidad apropiada de proteína *dummy*. Esto tiene gran relevancia fisiológica, pues no todos los procesos celulares están abarcados en el modelo ME, especialmente los procesos relacionados a defensa de situaciones de estrés, regulación, producción del flagelo, entre muchos otros, pues no tienen una relación estequiométrica con la formación de biomasa. Similar a lo que ocurre con la energía, en la que se usa un coeficiente de mantenimiento para poder cuantificar toda la energía producida pero no consumida en un proceso biológico modelado (Pirt, 1965; Utrilla et al., 2016)**,** en los modelos ME se usa la UPF para cuantificar toda la proteína que se puede llegar a producir en una cierta condición pero no está abarcada por el modelo. Como resultado, la UPF es un parámetro muy importante en la calibración del modelo, que se puede modificar (Utrilla et al., 2016) y se convierte en un blanco de diseño para cepas de producción. En esta tesis se utilizó la UPF en las calibraciones y predicciones de la distribución de recursos proteómicos, ya que un cambio en la UPF puede simular un cambio en la fracción de proteína de cobertura.

Enfoques de reducción de la complejidad

Teniendo en cuenta un enfoque biotecnológico, en el cual se planea colocar a la célula en un ambiente lo más controlado posible, se puede aplicar la eliminación de características que no le son útiles en esa condición (no esenciales) para reducir la complejidad del sistema y disminuir el desvío de recursos a funciones prescindibles (Balikó et al., 2018; Lastiri-Pancardo et al., 2020). En general, la creación de una célula simple se ha abordado de dos maneras, por un lado, está la opción de generar genomas desde cero, lo cual ha demostrado ser una tarea complicada ya que aún no conocemos del todo bien ni siquiera a los organismos más sencillos (Gibson et al., 2010). Por otro lado, la simplificación y optimización de células que ya conocemos es un método menos complejo y más rápido (Acevedo-Rocha et al., 2013; Esvelt & Wang, 2013).

Reducción de genoma

Las minimizaciones que se han realizado en células ya conocidas, se han enfocado en la reducción del genoma y número de genes (Balikó et al., 2018; Hutchison et al., 2016). Esto sin considerar de manera precisa cómo se realiza la asignación de recursos celulares (distribución dinámica de la célula de recursos internos, como ribosomas, enzimas, información genética, membranas y espacio intracelular, entre otros) (Zeng et al., 2021) una vez que se realizó la modificación genética. Además, se ha observado que una mayor reducción en el genoma que hasta el momento está clasificado como accesorio, puede conllevar una reducción en la velocidad específica de crecimiento (Kurokawa et al., 2016; Nishimura et al., 2017).

Reducción de proteoma

Utilizando el modelo iJL1678b-ME mencionado, sumado con conjuntos de datos experimentales, Lastiri-Pancardo *et al.* 2019 fueron capaces de identificar un número mínimo de intervenciones genéticas que maximizara el ahorro de recursos celulares que regularmente serían usados para expresar genes no esenciales. Una vez identificadas las intervenciones, se realizaron tres mutaciones de factores transcripcionales en la cepa *Escherichia coli* K12 BW25113 y se obtuvo la cepa llamada PFC, la cual, en teoría, redujo un 0.5% su proteoma (Lastiri-Pancardo et al., 2020).

Para analizar las diferencias entre la cepa PFC y su parental, se les agregó a ambas un plásmido con la vía de violaceína (Darlington et al., 2018) y, aunque PFC solo contenía tres mutaciones, incrementó su producción de proteína recombinante en un 18%. Estas tres mutaciones fueron en factores de transcripción (*ΔphoB* – sistema de eliminación de fosfato, *ΔflhC* – regulador maestro de flagelos, *ΔcueR*- sistema de salida de cobre), que llevan a la producción de proteoma de cobertura (proteoma que en el entorno actual no se utiliza para crecimiento celular o está presente en exceso y, por ende, operando debajo de su capacidad máxima), mostrando que la minimización de proteoma de cobertura puede llevar a una redirección de los recursos celulares a la producción de proteínas recombinantes.

Sin embargo, aún no se tiene una formalización sobre los costos de producción de proteínas y genes individuales, como lo puede ser el cálculo de su consumo de energía en términos de ATP. Esto permitiría hacer una comparación cuantitativa entre la reducción de genoma y reducción de proteoma para elegir la estrategia de minimización más adecuada. Por esta razón, en este trabajo se busca hacer esta formalización de costos, como se describirá en detalle más adelante.

Comparación de cepas con genoma minimizado

Las cepas MDS (*Multiple-Deletion Series*) provienen de *E. coli* K12 MG1655 (Figura 1). Hicieron comparaciones genómicas para encontrar segmentos presentes en K12 pero ausentes en otras *E. coli.* Estos segmentos fueron eliminados. Se logró reducir el genoma un 20%. También se buscó eliminar islas con sitios de inserción, elementos móviles y pseudogenes. Karcagi et al. 2016 decidieron comparar las cepas MDS12, MDS42, Y MDS69 con un ensayo de competición directa. Esto significa que cultivaron cada una de las cepas mutantes junto con *E. coli* K12 MG1655 a una relación de volumen 1:1, las transfirieron a un nuevo medio cada día y después de cinco días contaron las células para determinar su aptitud competitiva. La aptitud competitiva nos habla de si se tuvo ventaja o

desventaja en términos de volumen ceular. Si la aptitud competitiva de una cepa fuera 1 en relación a la parental, querría decir que ninguna tiene ventaja sobre otra. Mientras que si la aptitud competitiva de la cepa fuera de 0.8 con relación a la parental, querría decir que tiene una desventaja del 20%. Las cepas MDS tuvieron un peor desempeño que la *E. coli* K12 MG1655 en los ensayos de competición directa (la MDS12 tiene una aptitud competitiva de 0.96, la MDS42 de 0.77 y la MDS69 de 0.7). Además de tener un rendimiento de biomasa reducido, así como niveles elevados de *rpoS* (RpoS es un regulador maestro de respuesta de estrés), lo que significa que la pérdida de los segmentos de las cepas MDS puede inducir respues de estrés (Balikó et al., 2018; Karcagi et al., 2016; Kolisnychenko et al., 2002; Pósfai et al., 2006).

Otras cepas también provenientes de *E. coli* K12 MG1655 son la Δ16 (Hashimoto et al., 2005) y la MS56 (Park et al., 2014). La cepa Δ16 fue construida con el objetivo de minimizar al máximo el genoma de *E. coli* hasta obtener los componentes esenciales para mantener las actividades celulares. Las modificaciones fueron enfocadas en eliminar genes no esenciales que estuvieran reportados en la literatura. Como fenotipo, se reportó un crecimiento deficiente y localización anormal de nucleoides. La reducción fue del 30% (Figura 1). Por su parte, la cepa MS56, que tuvo una reducción de genoma del 24%, fue construida pensando en obtener células simples y altamente controlables (pensando en que un genoma reducido habría mayor comprensión de los genes e interaciones, lo cual mejoraría el control sobre rutas introducidas a estas células) sin regiones no esenciales (Juhas et al., 2011; Kurokawa & Ying, 2020).

Por otro lado, las cepas MGF (*Minimum Genome Factory)*, derivadas de *E. coli* K12 W3110 (Figura 1), surgieron de eliminar, además de regiones no compartidas por hibridación genómica comparativa, genes con función desconocida y regiones eliminadas en otras cepas de genoma reducido. Alcanzó un máximo de reducción de genoma del 25%. Se usaron deleciones de cepas de crecimiento normal para construir una cepa de genoma reducido. La cepa MGF-01 muestra una velocidad específica de crecimiento comparable a la silvestre (*E. coli* K12 W3110), mientras que la MGF-02 presenta una velocidad específica de crecimiento más alta (Hirokawa et al., 2013; Mizoguchi et al., 2008). A su vez, a partir de las cepas MGF se crearon las cepas DGF (*Design Genome Factory),* que incluyeron la eliminación de secuencias de inserción, sistemas toxina antitoxina y la restauración de regiones anteriormente eliminadas (alta osmolalidad, auxotrofía para uracilo). Estas alcanzaron una reducción máxima de genoma del 35% (Hirokawa et al., 2013).

Adecuación de cepas minimizadas

Como se mencionó, las cepas con un enfoque de reducción de genoma tienen problemas relacionados con la gran cantidad de genoma eliminado y el desconocimiento de las funciones de este. Mientras que, en el enfoque de reducción de proteoma, al tener una menor cantidad de mutaciones, es mucho menos probable que las cepas tengan problemas que comprometan su utilidad para producir funciones sintéticas o funciones biotecnológicas. Sin embargo, lo mejor sería que antes de realizar eliminación en cualquiera de los dos enfoques se tuviera conocimiento de cómo afectarán a la célula.

La base de datos de adecuación *Fitness Browser* (Price et al., 2018; Wetmore et al., 2015) es una base de datos en la cual están recopilados datos de adecuación de diferentes mutantes de varias cepas de diversos organismos, incluyendo *Escherichia coli K12*

BW25113. Para definir la adecuación, se crearon mutantes a las cuales les agregaron transposones con códigos de barra aleatorios, insertados en ubicaciones aleatorias del genoma que posteriormente fueron rastreados. Después, se crecen todas juntas en diferentes ambientes y se compara la abundancia de los códigos de barras utilizando PCR. Al final, se tienen datos que describen el cambio en la abundancia de las mutantes en los diferentes genes utilizados durante el experimento. Una adecuación de 0 significa que las mutantes tuvieron la misma abundancia que otras mutantes y probablemente tan bien como la parental. Una adecuación menor a 0 significa que el gen era importante para la adecuación, haciendo que la abundancia de las mutantes disminuyera (-1 significa que las mutantes fueron la mitad de abundantes al final del experimento que al principio). Una adecuación mayor a 0 significa que la ausencia de ese gen mejora la adecuación y las mutantes fueron más abundantes. Mencionan que un valor de adecuación de -1 a 1 es un fenotipo sutil o neutral, mientras que un gen que es esencial en ciertas condiciones tendrá un valor de -4 a -8. Para *E. coli K12 BW25113, Fitness Browser* contiene información de 166 condiciones diferentes, entre las que hay variación de fuentes de carbono, variación de fuentes de nitrógeno y condiciones de estrés. Por lo tanto, podríamos tomar la información de esta base de datos para, con su definición de adecuación, elegir mutaciones que no sean esenciales o que no tengan un efecto negativo al fenotipo que nos interese lograr.

Trabajo presente

Este trabajo aborda la necesidad de optimizar el diseño de intervenciones genéticas mediante la formalización de los costos de producción y la comparación cuantitativa de los enfoques de reducción de genoma y proteoma. Se describen las metodologías utilizadas para predecir genes objetivo tomando en cuenta los diferentes ambientes que podrían ser de interés. Para este fin, el análisis fue completamente computacional, haciendo uso de datos de la literatura, así como un modelo ME de *E. coli* con el que se realizaron diversas simulaciones. Además, se discuten las limitaciones de la investigación y se proponen posibles direcciones para futuras investigaciones.

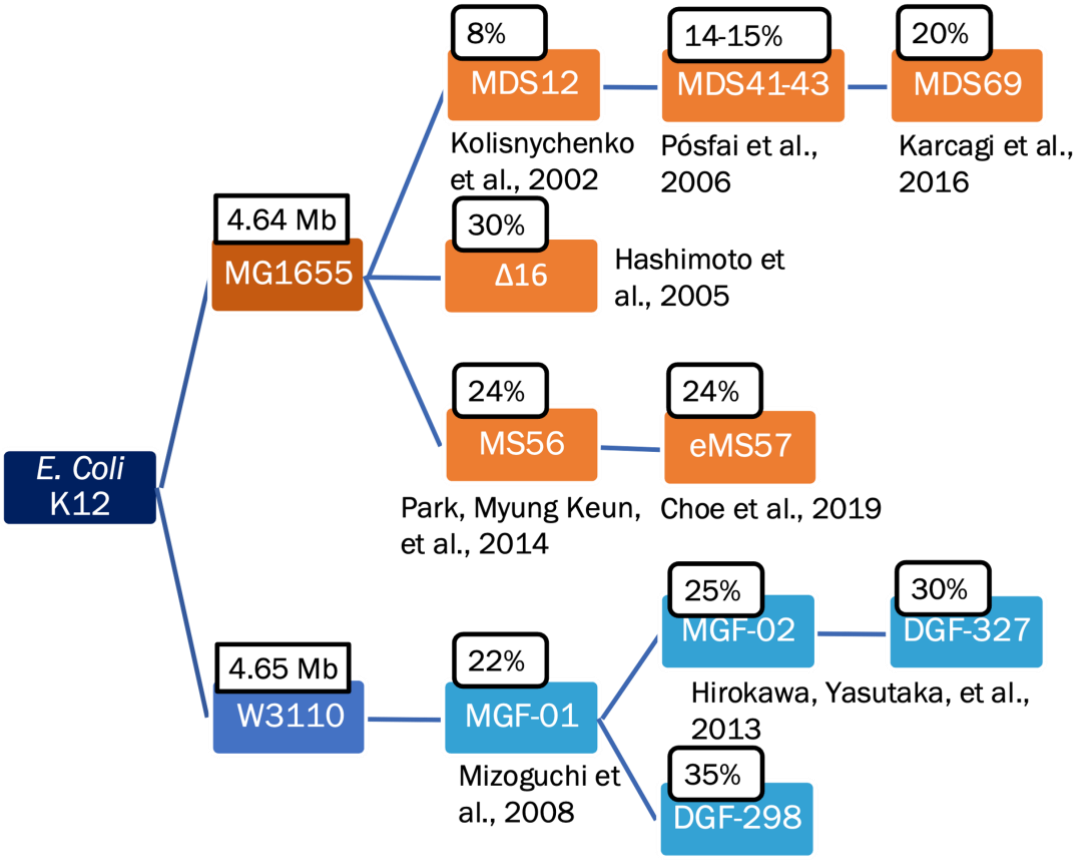

Figura 1. Mapa de relación de cepas minimizadas con el porcentage de minimización de genoma alcanzado y el estudio del que surgieron.

Hipótesis

La determinación de costos, basándonos en el cálculo del consumo de energía en términos de ATP de diferentes funciones, permitirá la identificación de genes cuya eliminación reducirán el proteoma de cobertura y llevarán a la redirección de recursos celulares a funciones biotecnológicas o funciones sintéticas.

Objetivos

Objetivo General

Identificar una estrategia para reducir la expresión de funciones de cobertura para generar diseños de intervenciones genéticas.

Objetivos específicos

a. Comparar los enfoques de reducción de complejidad (reducción de genoma y reducción de proteoma) así como sus efectos fenotípicos reportados en *Escherichia coli*. Esto mediante una revisión de la literatura científica.
b. Evaluar la eficacia de los enfoques de reducción de complejidad para la redirección de recursos celulares hacia funciones biotecnológicas o funciones sintéticas, mediante el cálculo de los costos de producción basados en el consumo de energía en términos de ATP de diferentes funciones según datos proteómicos y simulaciones con el modelo iJL1678b-ME.
c. Combinar el enfoque de reducción elegido con los datos de adecuación de la base de datos *Fitness Browser* de *Escherichia coli* en los 166 distintos ambientes que reportan, para encontrar genes blanco de modificación para la reducción de la complejidad celular.
d. Comparar los resultados obtenidos en los objetivos anteriores para determinar la estrategia más adecuada para reducir la complejidad celular y generar diseños de intervenciones genéticas efectivas.

Metodología

El análisis en este proyecto fue completamente bioinformático. Combinamos datos de proteómica, genómica, simulaciones obtenidas de un modelo ME así como datos de adecuación. Todos ellos provenientes de *E. coli* y detallados a continuación. Se emplearon los datos de proteómica antes mencionados provenientes de Schmidt *et al*., 2016, los cuales también fueron utilizados por O'Brien *et al*., 2016 para analizar el proteoma de *E. coli*. Sin embargo, O'Brien *et al*., 2016 utilizaron 16 condiciones de las 22 disponibles (Tabla 1) ya que eran las compatibles con su metodología, mientras que en este proyecto se integraron todas las condiciones (crecimiento en medios mínimos con un exceso de diferentes fuentes de carbono y energía, crecimiento en cultivos de quimiostato con glucosa limitada con diferentes velocidades específicas de crecimiento, crecimiento en exceso de glucosa con diferentes condiciones de estrés y crecimiento en medio complejo).

Otros conjunto de datos utilizados comprenden once de las ya mencionadas de cepas con genoma minimizado: MDS12, MDS41, MDS42, MDS43 y MDS69 (Kolisnychenko et al., 2002; Pósfai et al., 2006); Δ16 (Hashimoto et al., 2005), MS56 (Hall, 2004), MGF-01, MGF-02, DGF-327, y DGF-298 (Hirokawa et al., 2013; Mizoguchi et al., 2008). Así como datos obtenidos de la cepa con proteoma minimizado: PFC [7]. Se compararon enfoques de proteómica y genómica para determinar cuál era el más efectivo minimizando la complejidad y revisando en la literatura qué efectos había reportados sobre el fenotipo.

También se utilizó el modelo de metabolismo y expresión de *Escherichia coli* K12 MG1655 [2,16] llamado iJL1678b-ME para calcular el consumo de ATP en términos de $mmol \bullet gDW^{-1} \bullet h^{-1}$ de ATP de los genes presentes en el modelo. Además, también se realizaron simulaciones de replicación, transcripción y producción de proteína para, junto con datos proteómicos y tamaño de los genes, determinar cuál era el proceso más costoso y establecer si era más conveniente realizar una minimización enfocada a genoma o enfocada a proteoma.

Por último, al determinar cuál fue el proceso más costoso (traducción), se integraron datos ómicos relacionados a ese proceso con datos de adecuación de de *Escherichia coli* K12 BW25113 tomados de la base de datos de adecuación *Fitness Browser* (Price et al., 2018; Wetmore et al., 2015). Este enfoque nos permitió elegir los genes blanco más apropiados para diseñar intervenciones genéticas que nos permitan reducir la complejidad celular al mismo tiempo que disminuimos la probabilidad de presentar fenotipos negativos.

Procesamiento de datos de las cepas provenientes de *E. coli* K12 MG1655

Para obtener los identificadores de los genes eliminados en las cepas MDS se utilizaron los artículos más recientes del grupo de Pósfai. Para la cepa MDS12 se utilizó la información de Pósfai, György, *et al*, 2006 (Pósfai et al., 2006). Mientras que para la cepa MDS42 y MDS69 se utilizó el artículo de Karcagi, Ildikó, *et al*, 2016 (Karcagi et al., 2016). En un inicio se planeaba mapear los segmentos eliminados reportados por si algún gen anteriormente reportado con función desconocida ahora ya era conocido. Sin embargo, al mapear los segmentos no coinciden con los genes reportados, así como hay genes extra

reportados en los artículos más recientes del mismo grupo, por lo que sospechamos que hay un error con el reporte inicial de segmentos eliminados, así que optamos por utilizar la lista reportada con los identificadores de los genes eliminados en los artículos más recientes. Debido a esto, es probable que algunos genes eliminados no hayan sido incluidos en este estudio. Por consiguiente, se optó por considerar únicamente la cepa MDS42 como representante de las cepas MDS41 y MDS43 ya que cuenta con información más actualizada. Además, la cantidad de genes de diferencia reportados en el artículo donde se mencionan a las tres cepas (Pósfai et al., 2006) no nos parece significativa (MDS41 tiene 703 genes, MDS42 tiene 704 y la MDS43 tiene 743). En contraste, el artículo más reciente sobre MDS42 (Karcagi et al., 2016) ya informa de 725 genes, mientras que las cepas MDS41 y MDS43 no son mencionadas.

Los nombres de los genes de la cepa MS56 ya venían reportados en el artículo donde la presentan (Park et al., 2014) por lo que esos fueron los identificadores utilizados.

Para obtener los genes de las cepas Δ16 se utilizaron los intervalos reportados en la literatura y se mapearon contra la información presente en la base de datos de NCBI (Brown et al., 2015) bajo el nombre de *Escherichia coli* str. K12 substr. MG1655 versión NC_000913.3.

Procesamiento de datos de las cepas provenientes de *E. coli* K12 W3110

Para obtener los genes de las cepas MGF-01, MGF-02, DGF-298, y DGF-298 se utilizaron los intervalos reportados en la literatura y se mapearon contra la información presente en la base de datos de NCBI (Brown et al., 2015) bajo el nombre de *Escherichia coli* str. K12 substr. W3110 versión AP009048.1.

Análisis de consumo de recursos en términos de ATP por reacción con el modelo ME

Para el cálculo de consumo de recursos en términos de ATP utilizando el modelo iJL1678b-ME, se analizaron todas las reacciones, seleccionando aquellas que incluían ATP en sus reactivos (indicando que la reacción consumía ATP). Una vez que identificadas las reacciones que consumían ATP, se multiplicó el valor del flujo de la reacción por su coeficiente estequiométrico de ATP, obteniendo así el valor de flujo de consumo de ATP.

Análisis de consumo de recursos en términos de ATP por gen con el modelo ME

Para obtener el consumo de recursos en términos de ATP para cada gen utilizando el modelo iJL1678b-ME primero se buscaron las reacciones de las categorías metabólica, traducción y transcripción las cuales representan las tres reacciones que abarcan la mayor parte del consumo de recursos en términos de ATP. A partir de las reacciones se calculó la aportación al flujo de cada gen involucrado según la estequiometría de la reacción.

Análisis de consumo de recursos en términos de ATP por gen con el modelo ME y datos ómicos

Para obtener costos de replicación, transcripción y producción de proteína, primero se utilizó el modelo iJL1678b-ME para simular cambios en el porcentaje de ADN en la célula, cambios en el número de genes transcritos y cambios en la producción de proteína para obtener los costos de ATP según el modelo para cada uno de estos cambios. Posteriormente, ya que estos costos dependían del tamaño del gen así como el porcentaje de proteoma que representaban, se calculó el costo de replicación, transcripción y producción de proteína utilizando datos del tamaño del gen obtenidos de NCBI (Brown et al., 2015) y de los datos de proteína obtenidos de Schmidt *et al*., 2016.

Búsqueda de genes blanco uniendo información proteómica y de adecuación

Se utilizaron los datos de adecuación de *Fitness Browser* (Price et al., 2018; Wetmore et al., 2015) en el cual están recopilados datos de adecuación de diferentes mutantes de diversos organismos, incluyendo *Escherichia coli K12 BW25113*. Al momento del estudio se contaban con 166 diferentes condiciones para *E. coli K12 BW25113.* Se buscaron aquellos genes con datos de adecuación en las 166 condiciones que además estuvieran dentro del conjunto de datos proteómicos de Schmidt *et al*., 2016. Como se mencionó, un valor de adecuación de -1 a 1 lo interpretaron un fenotipo sutil o neutral, mientras que un gen que es esencial en ciertas condiciones tendrá un valor de -4 a -8. Teniendo esto en cuenta, se clasificaron los genes como “neutrales” aquellos en el rango de -1 a 1. Si tenían un valor menor a -1, se clasificaron como genes con aportación negativa a la adecuación. Dentro de los genes con aportación negativa también estaban aquellos con valores de entre -4 a -8, que marcamos como “esenciales”. Si por el contrario tenían valores mayores a 1, se clasificaron como genes con aportación positiva. Por otro lado, los genes que no estaban en las listas de *Fitness Browser* fueron catalogados como los tomaremos como “posibles esenciales” ya que por cómo se obtuvieron los datos, lo más probable es que no fue posible obtener una mutante ya que eran esenciales.

Resultados

Liberación predicha de proteoma en cepas provenientes de *E. coli* K12 MG1655

Utilizando la información proteómica del trabajo de Schmidt *et al*., 2016 calculamos qué porción del proteoma es representada por las proteínas codificadas en los genes eliminados de cada cepa de la línea MDS en cada una de las 22 condiciones (Tabla 1). Para este fin, se cuantificó qué porcentaje de los genes eliminados tenía información proteómica en alguna de las 22 condiciones. Esto con respecto al total de los genes eliminados de cada cepa y al conjunto total de genes presentes en Schmidt *et al*., 2016. Se estimó cuánta carga proteómica representaban las proteínas codificadas en genes eliminados con respecto al proteoma total de cada una de las condiciones de Schmidt *et al*., 2016.). La carga proteómica está representada por la cantidad de proteoma en femtogramos (fg o 10^{-15} g) por célula que proporcionarían los genes eliminados de la cepas en cada una de las condiciones. En este estudio, utilizamos la abreviatura 'fgPC' para referirnos a femtogramos de proteína por célula. Esta carga proteómica la tomaremos como el proteoma predicho liberado (ya que esos genes fueron eliminados).

La cepa con más genes eliminados, Δ16, es también de la cual tenemos mayor información proteómica. El 36.67% de los genes eliminados de Δ16 fueron encontrados entre los reportados por Schmidt et al., 2016., los cuales solo representan un 18.19% del total de genes con información proteómica en el conjunto de datos de Schmidt et al., 2016 (Tabla 2). Mientras que la cepa con menos genes eliminados, MDS12, es de la cual se encontró menos información proteómica, con solo el 18.44% de sus genes eliminados estando presentes en la información de Schmidt et al., 2016 y representando un 3.31% con respecto al conjunto de genes de Schmidt et al., 2016. A pesar de que la cepa MDS69 tiene aproximadamente el doble de genes eliminados que la cepa MDS12, el porcentaje de genes con información proteómica con respecto al conjunto de datos de Schmidt et al., 2016 se triplica (9.83% contra 3.31%). Mientras que la cepa Δ16 tiene casi el triple de genes eliminados que la cepa MDS12, pero tiene más de cinco veces el porcentaje de genes con información proteómica con respecto al conjunto de datos de Schmidt et al., 2016 (18.19% contra 3.31%).

Tabla 2. Porcentaje de los genes eliminados de las cepas provenientes de E. coli K12 MG1655 con información en el conjunto de datos proteómicos.

	Genes eliminados	Genes con información proteómica		
		Número de genes	Con respecto a los genes eliminados	Con respecto al conjunto de genes presentes en Schmidt *et al.*, 2016
MDS12	423	78	18.44%	3.31%
MDS42	725	157	21.66%	6.66%
MDS69	966	232	24.02%	9.83%
MS56	1005	286	28.46%	12.12%
Δ16	1170	429	36.67%	18.19%

Una vez estimada la carga proteómica de los genes eliminados, a la cual podemos ver como el proteoma estimado liberado de las cepas, esta fue graficada para cada cepa en cada condición (Figura 2). Las condiciones fueron acomodadas de mayor a menor velocidad específica de crecimiento (Tabla 1).

En la figura 2a podemos observar que la cepa que en todas las condiciones tiene una mayor cantidad de proteoma estimado liberado es la cepa Δ16 con un máximo de ~29.83 fgPC en medio LB, después viene la cepa MS56 con ~18.73 fgPC en la condición de glicerol más aminoácidos, posteriormente la MDS69 con ~13.79 fgPC en la condición de glicerol más aminoácidos, seguida por la MDS42 con ~7.32 fgPC en la condición en acetato y, finalmente, la MDS12 con ~2.6 fgPC en la condición de 42°C en glucosa.

Sin embargo, el proteoma total en fgPC para cada una de las 22 condiciones de Schmidt et al., 2016 es diferente, por lo que una mayor cantidad de fgPC liberados no significa un mayor porcentaje de proteoma estimado liberado. Por lo tanto, se hizo la misma gráfica para porcentaje de proteoma estimado liberado (Figura 2b). Del mismo modo, podemos observar en la figura 2b que la cepa que mayor porcentaje de proteoma estimado liberó fue la cepa Δ16 con un máximo de ~13.13% en la condición en acetato, después viene la cepa MS56 con ~7.37% en la condición en fase estacionaria de un día, posteriormente la MDS69 con ~4.64% en la condición en acetato, seguida por la cepa MDS42 con ~3.73% en la condición en acetato y, finalmente, ~1.02% en la condición de 42° en glucosa. En general, se puede apreciar en la figura 2b que hay una tendencia en la que las condiciones donde hay un mayor porcentaje de proteoma estimado liberado, son aquellas donde hay una menor velocidad específica de crecimiento.

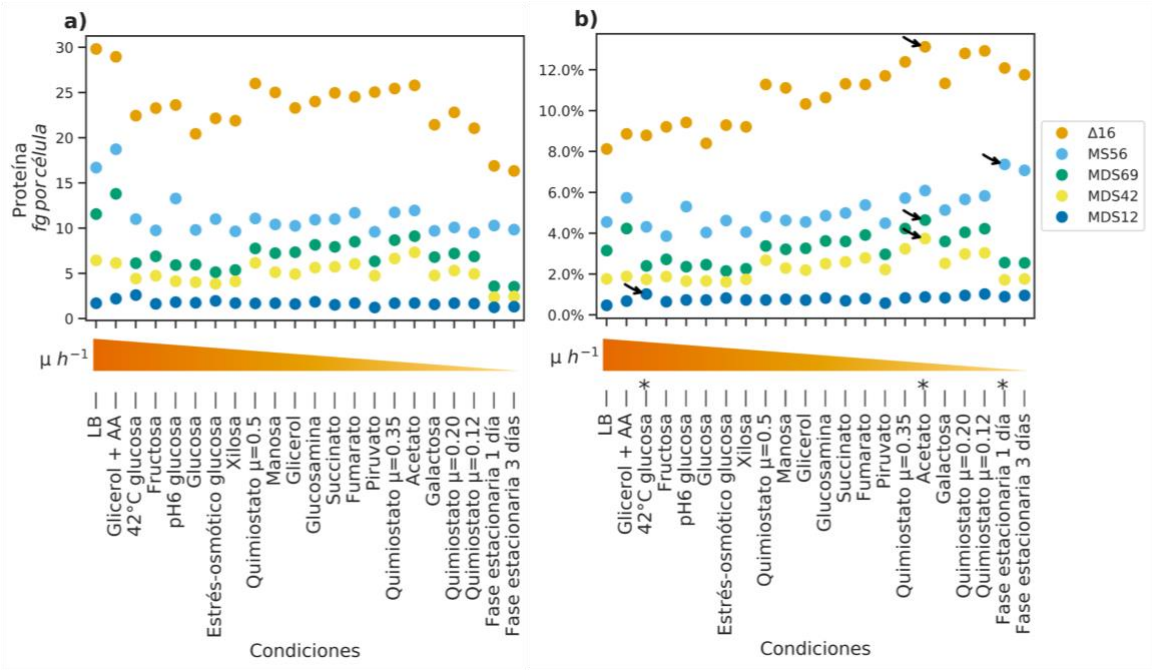

Figura 2. Proteoma estimado liberado en las cepas MDS12, MDS42, MDS69, MS56 y Δ16. a) Proteoma estimado liberado en femtogramos por célula. b) Proteoma estimado liberado en porcentaje. (*) Condición en la que una o más cepas obtuvieron el porcentaje máximo de proteoma estimado liberado. (→) Punto máximo de porcentaje de proteoma estimado liberado para cada cepa.

Liberación predicha de proteoma en cepas provenientes de *E. coli* K12 W3110

Al igual que para las cepas MDS, calculamos la liberación predicha de proteoma para cada cepa de las líneas MGF y DGF en cada una de las 22 condiciones de Schmidt *et al*., 2016. En primer lugar se obtuvieron los porcentajes de los genes eliminados de los cuales se tenía información proteómica en alguna de las 22 condiciones de modo que obtuvimos valores con respecto al total de los genes eliminados de cada cepa y con respecto al conjunto total de genes presentes en Schmidt *et al*., 2016. Una vez que teníamos los genes con información proteómica, se calculó la carga proteómica que representaban las proteínas codificadas en los genes eliminados con respecto al proteoma total de cada una de las condiciones de Schmidt *et al*., 2016. En general, todas las cepas tienen el mismo porcentaje de genes eliminados con información proteómica (32.80% en promedio). Pero la cepa con más genes eliminados, DGF298, es de la cual tenemos mayor información proteómica con respecto a todos los genes del conjunto de Schmidt *et al*., 2016 con un 23.53% (Tabla 3).

Tabla 3. Porcentaje de los genes eliminados de las cepas provenientes de de E.coli K12 W3110 con información en el conjunto de datos proteómicos.

	Genes eliminados	Genes con información proteómica		
		Número de genes	Con respecto a los genes eliminados	Con respecto al conjunto de genes presentes en Schmidt *et al*., 2016
MGF01	1035	339	32.75%	14.37%
MGF02	1176	388	32.99%	16.45%
DGF327	1392	464	33.33%	19.67%
DGF298	1727	555	32.14%	23.53%

Posteriormente se graficó la carga proteómica, o proteoma estimado liberado (llamado así ya que los genes cuyas proteínas codificadas aportaban esa carga fueron eliminados), para cada cepa en cada condición (Figura 3). Las condiciones fueron acomodadas de mayor a menor velocidad específica de crecimiento (Tabla 1).

Podemos observar en la figura 3a que la cepa que en todas las condiciones tiene una mayor cantidad de proteoma estimado liberado en fgPC es la cepa DGF-298 con un máximo de ~31.15 fgPC en la condición en glicerol con aminoácidos, seguida por la DGF-327 con ~27.16 fgPC en la condición en acetato, luego la MGF-02 con ~16.32 fgPC en la condición de quimiostato con μ=0.35 h^{-1} y, finalmente, la MGF-01 con ~15.86 fgPC en la condición de quimiostato con μ=0.35 h^{-1}.

Sin embargo, como se mencionó, una mayor cantidad de proteoma estimado liberado en fgPC no significa un mayor porcentaje de proteoma estimado liberado así que se realizó el mismo análisis para porcentaje de proteoma estimado liberado (Figura 3b). Del mismo modo, la cepa que mayor porcentaje liberó fue DGF-298 con un máximo de ~14.95%, seguida por la DGF-327 con ~10.4%, luego la MGF-02 con ~8.02% y, finalmente, la MGF-01 con ~7.77%. Todas alcanzando su máximo en la condición de quimiostato con μ=0.12 h^{-1}. Se aprecia la misma tendencia que con las cepas derivadas de MG1655 en la que las

condiciones donde hay un mayor porcentaje de proteoma estimado liberado, son aquellas donde hay una menor velocidad específica de crecimiento.

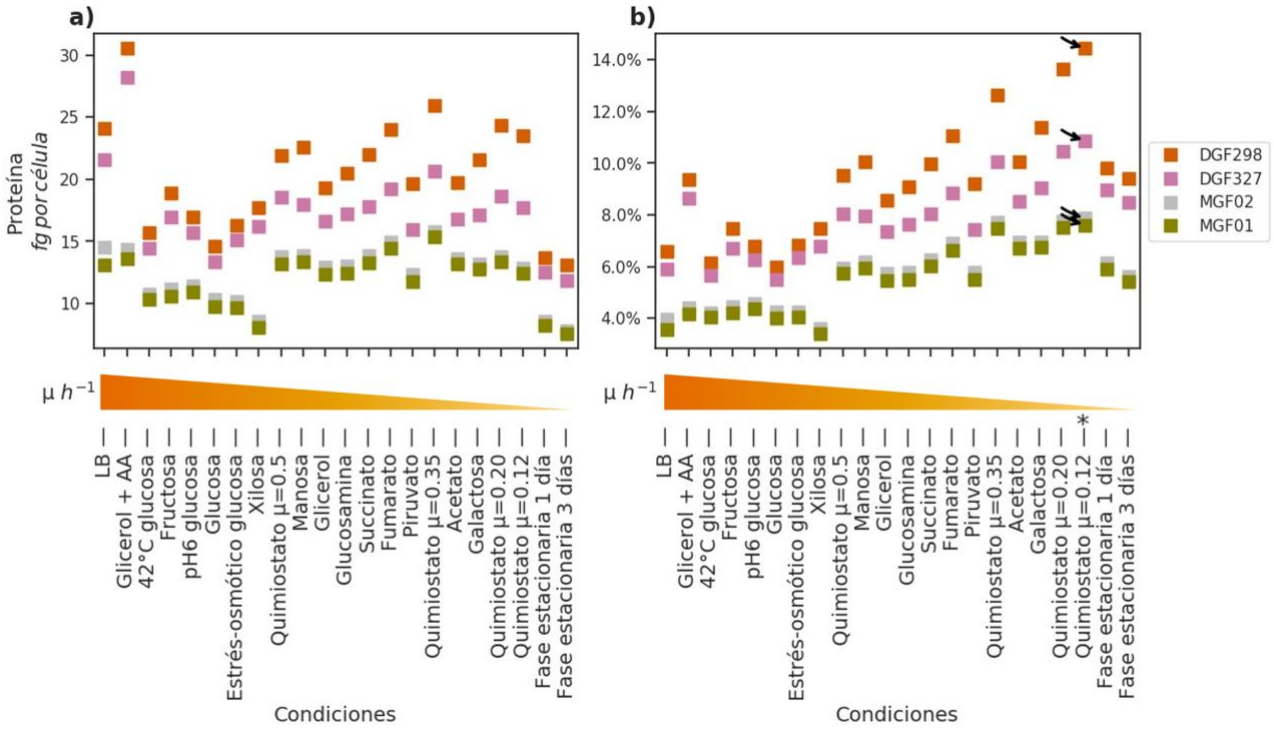

Figura 3. Proteoma estimado liberado en las cepas MGF01, MGF02, DGF298 y DGF327. a) Proteoma estimado liberado en femtogramos por célula. b) Proteoma estimado liberado en porcentaje. (*) Condición en la que una o más cepas obtuvieron el porcentaje máximo de proteoma estimado liberado. (→) Punto máximo de porcentaje de proteoma estimado liberado para cada cepa.

Distribución del proteoma

Proteoma promedio de las 22 condiciones

Para observar la distribución del promedio de proteoma de las 22 condiciones se utilizó la información de cuánta proteína aportada, en fgPC, estaba relacionada a cada gen en cada una de las condiciones. Se obtuvo el valor promedio relacionado a cada proteína codificada por cada gen así como su desviación estándar. En total, había información de 2359 proteínas. Una vez que teníamos el promedio individual de producción de proteína, tomamos su suma (224.52 fgPC ± 114.11) como un proteoma total promedio de las 22 condiciones y se analizó su distribución. Sin embargo, al haber valores de proteínas entre 10^{-6} fgPC y ~30 fgPC, y más del 90% de los valores se encontraban debajo de 2.5 fgPC, se optó por obtener el logaritmo base 10 para observar más fácilmente la distribución. Solo un 14.6% (o 344) de los genes codificaban para proteínas cuya aportación estaba por encima de los 0.1 fgPC (Figura 4).

Al organizar los genes según la aportación de sus proteínas codificadas al proteoma promedio, identificamos los cinco genes principales cuyas proteínas codificadas tienen una mayor aportación. Posteriormente buscamos sus funciones (Tabla 4). Tan solo la proteína

codificada por el gen *tufA* aporta el 6.26% ± 2.68 del total del proteoma. Con esto podemos observar que son pocos los genes cuyas proteínas codificadas tienen una aportación grande al proteoma promedio. Por lo tanto, si eliminámaramos un gen al azar cuya proteína codificada sí sea traducida, lo más probable es que sería un gen que aporte menos de 0.1 fgPC al proteoma promedio, o lo que sería equivalente, un ~0.05% del proteoma promedio.

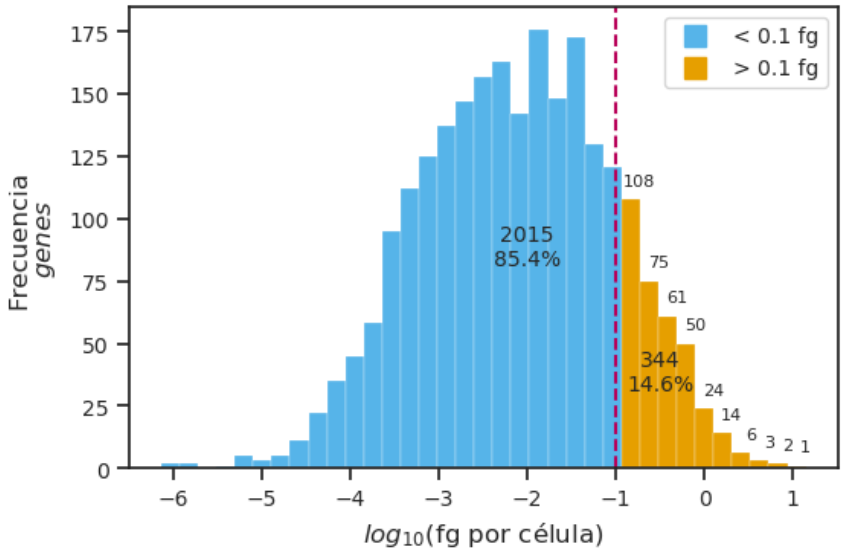

Figura 4. Distribución de aportación al proteoma promedio de las 22 condiciones por gen, en términos de su proteína codificada en fgPC.

Tabla 4. Primeros cinco genes cuyas proteínas codificadas aportan más al proteoma promedio.

Gen	Promedio fg/célula	σ fg/célula	Promedio	σ	Descripción
tufA	14.05	±6.03	6.26%	±2.68%	Factor de elongación Tu 1
aceA	7.38	±5.91	3.29%	±2.63%	Isocitrato liasa
ompA	6.36	±1.66	2.83%	±0.74%	Porina A de la membrana externa
fusA	3.95	±2.13	1.76%	±0.95%	Factor de elongación G
cysK	3.67	±1.39	1.63%	±0.62%	O-acetilserina sulfhidrilasa A

A pesar de que podría sonar tentador tomar como genes blanco para minimización aquellos cuyas proteínas codificadas más aportan al proteoma promedio, probablemente sean importantes para su crecimiento en ciertas condiciones o que sean incapaces de sobrevivir sin ellos. Para revisar si la eliminación de los genes podría tener efectos negativos en las cepas se consultó la base de datos *Fitness Browser* (Price et al., 2018; Wetmore et al., 2015) que contiene datos de adecuación de *E. coli* en 166 condiciones (detallado en la sección de métodos). Si la mutante tenía una adecuación menor a -1, lo que significa una menor abundancia de la mutante, lo catalogamos como que esa mutante tendría defectos en la adecuación. Por lo que quitarlo presentaría una desventaja para la cepa en esa condición. Si la mutante tenía una adecuación menor a -4, se catalogó como que era un gen esencial para esa condición (como es mencionado también por la base de datos de *Fitness Browser)* y su eliminación tendría como consecuencia que la cepa no pueda sobrevivir en dicha condición (Tabla 5).

En los casos donde no hubo información en la base de datos de *Fitness Browser,* se consultaron las bases de datos EcoCyc (Karp et al., 2002; Keseler et al., 2017) y *Profiling of Escherichia coli Chromosome (PEC)* (Yamazaki et al., 2007). Por ejemplo, *tufA* no tiene datos reportados en la base de datos de *Fitness Browser*. Pero sí hay reportes en la base de datos EcoCyc de esencialidad en medio LB enriquecido (con *tufA* mutado, la cepa no presenta crecimiento en ese medio). Por la parte de *aceA, ompA* y *cysK* sí hay datos en *Fitness Browser*, por lo que pudimos clasificar aquellos cuyas mutantes tienen valores de adecuación menores a -1 en alguna de las 166 condiciones en la categoría de "defectos de adecuación en la mutante" y aquellos menores a -4 como "esenciales". Podemos observar que para todos los genes, hay reportes de que sus mutantes afectan negativamente a la adecuación, siendo *cysK* el que en más condiciones se ve afectado (90 de las 166 reportadas en *Fitness Browser*). Finalmente, *fusA* tampoco está presente en la base de datos de *Fitness Browser*, pero sí es reportado como un gen esencial (hay evidencia de que la mutación del gen es letal) para *Escherichia coli* por la base de datos PEC. Por lo tanto, el eliminar alguno de estos genes sin tomar en consideración cómo afectan a la adecuación, puede llevar a que la cepa no sea viable o sea viable condicionalmente; Además de que presente fenotipos no deseados para la producción de funciones sintéticas.

Tabla 5. Cantidad de condiciones con afectación a la adecuación de los cinco genes cuyas proteínas codificadas aportan más al proteoma promedio. (**) Genes que no se encontraron en la base de datos Fitness Browser, pero sí en la base de datos EcoCyc. (†) Genes que no se encontraron en la base de datos Fitness Browser, pero sí en PEC.

Gen	Esencial	Número de condiciones con defectos de adecuación en la mutante
tufA	1†	
aceA	2	2
ompA	1	51
fusA	Sí **	
cysK	62	90

Además, no debemos olvidar que el que algunos genes no aparezcan en las listas de *Fitness Browser* podría indicar que sean "posibles esenciales" puesto que no fue posible obtener una mutante por medio de transposones porque probablemente eran esenciales y no se pudo obtener una mutante. Además de estar presentes como esenciales en otras bases de datos.

Proteoma total en medio mínimo con glucosa

El proteoma promedio de las 22 condiciones nos da una visión general de los genes cuyas proteínas codificadas aportan más al proteoma. Sin embargo, esto no quiere decir que las mismas proteínas constituyan una carga mayor en todas las condiciones. Por lo tanto, los genes blanco pueden variar según la condición de interés. Por esta razón decidimos analizar la distribución del proteoma en una condición en específico. La condición elegida fue glucosa en medio mínimo (medio mínimo M9 con 5 g/ L de glucosa) ya que el medio M9 con glucosa se destaca por su atractivo industrial y su costo accesible, siendo de especial interés para cepas de producción con relevancia en la industria (Rugbjerg et al., 2018).

Para observar la distribución del proteoma en la condición de medio mínimo con glucosa también se graficó y analizó la distribución de la aportación al proteoma por proteína codificada de cada gen y se sacó el logaritmo base 10 para observar más fácilmente la distribución. En este caso solo tomamos en cuenta 2356 genes, a diferencia de las 2359 del proteoma promedio ya que no se detectaron tres proteínas en este medio. Solo un 14.4% (o 340) de los genes codificaban para proteínas cuya aportación era mayor a 0.1 fgPC (Figura 4, Figura 5). Al organizar los genes según la aportación de sus proteínas codificadas al proteoma en la condición de glucosa, podemos ver los primeros cinco genes cuyas proteínas codificadas tienen una mayor aportación. A partir de esta lista, buscamos sus funciones (Tabla 6). De nuevo, la proteían codificada por *tufA* es la que mayor carga presenta con un 7.45% de todo el proteoma en la condición de glucosa. También observamos que, al igual que en la distribución del proteoma promedio, son pocos los genes cuyas proteínas codificadas aportan una gran cantidad de proteoma en esa condición. Finalmente, los genes cuyas proteínas codificadas más aportan pueden ir variando según la condición que analicemos, por ejemplo, entre los primeros cinco genes cuyas proteínas codificadas más aportan al proteoma promedio se encuentra *aceA*. Mientras que en los primeros cinco genes cuyas proteínas codificadas más aportan al proteoma en glucosa no se encuentra *aceA* y en su lugar podemos observar a *metE.*

El gen *aceA* codifica una de las proteínas que permiten a *E. coli* evitar una parte del ciclo del ácido cítrico, pasando más bien por el ciclo del glioxilato (Fischer & Sauer, 2003). Esta ruta la usa con frecuencia cuando presenta velocidades específicas de crecimiento bajas (Peebo et al., 2015), así como en crecimiento teniendo como fuente de carbono acetato (Zarembinski et al., 1991). En el contexto de los datos que constituyen nuestro proteoma promedio, aproximadamente el 60% de las condiciones tienen velocidades específicas de crecimiento menores a 0.5 h^{-1}. Además, solo dos condiciones tienen valores superiores a 0.66 h^{-1} (1.27 y 1.9 h^{-1}). Por último, el 25% de las condiciones tienen valores inferiores a 0.3 h^{-1}. En consecuencia, esto puede explicar por qué *aceA* aparece entre los cinco genes principales cuyas proteínas codificadas contribuyen más al proteoma promedio.

Por otro lado, la proteína codificada por *metE* es una enzima crucial en la biosíntesis de metionina (Gonzalez et al., 1992). Se ha reportado que en medio mínimo con glucosa puede representar el 5% del proteoma (Russell & Berry, 1986). Además, se sabe que el acetato inhibe la biosíntesis de metionina, concretamente la actividad de la enzima MetE (Pinhal et al., 2019). Debido a lo anterior, la contribución al proteoma de las proteínas codificadas por *aceA* y *metE* a una determinada condición puede variar. Por lo tanto, a pesar de que el proteoma promedio nos da una idea general de qué proteínas aportan más, si tenemos en mente una condición particular, sería ideal examinar la contribución al proteoma en esa condición.

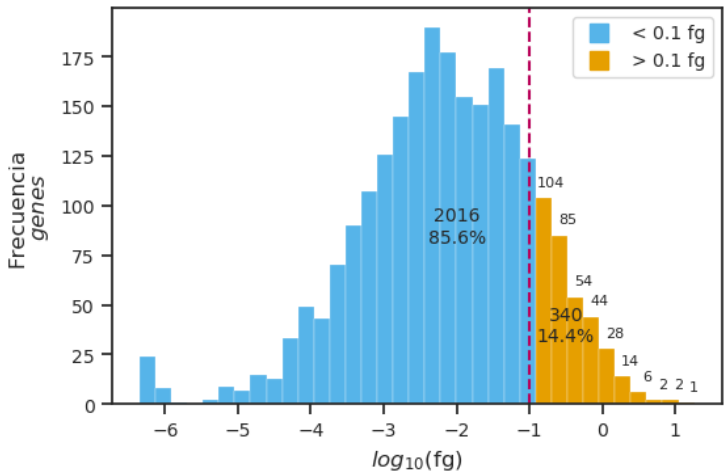

Figura 5. Distribución de aportación al proteoma en la condición de glucosa por gen, en términos de su proteína codificada en fgPC.

Tabla 6. Primeros cinco genes cuyas proteínas codificadas aportan más al proteoma en la condición en glucosa. Los genes que no están presentes en la lista del proteoma promedio están marcados con un asterisco (*).

Gen	fg/célula	%	Descripción
tufA	18.13	7.45	Factor de elongación Tu 1
ompA	7.38	3.04	Porina A de la membrana externa
**metE*	7.25	2.98	Homocisteína transmetilasa independiente de cobalamina
fusA	5.02	2.07	Factor de elongación G
cysK	4.19	1.72	O-acetilserina sulfhidrilasa A

De igual manera que con el proteoma promedio, se analizaron los genes cuyas mutaciones afectaran negativamente a la adecuación en algunas condiciones o fueran genes esenciales (según las bases de datos *Fitness Browser,* EcoCyc y PEC) (Tabla 7). El único gen que no aparece en la lista del proteoma promedio es *metE*, y este gen presenta aún más condiciones en las que su ausencia afecta negativamente a la adecuación. Esto nos demuestra de nuevo que el eliminar alguno de estos genes sin tomar en consideración cómo afectan a la adecuación puede llevar a que la cepa no sea viable o sea viable condicionalmente. Además de que pueda presentar fenotipos no deseados para la producción de funciones sintéticas.

De nuevo, no hay que olvidar que el que algunos genes no aparezcan en las listas de *Fitness Browser* podría indicar que sean “posibles esenciales” ya que no fue posible obtener una mutante por medio de transposones.

Tabla 7. Cantidad de condiciones con afectación a la adecuación de los cinco genes cuyas proteínas codificadas aportan más al proteoma en la condición de glucosa. (*) Genes que no están presentes en la lista del proteoma promedio. (**) Genes que no se encontraron en la base de datos Fitness Browser, pero sí en la base de datos EcoCyc. (†) Genes que no se encontraron en la base de datos Fitness Browser, pero sí en PEC.

Gen	Esencial	Número de condiciones con defectos de adecuación en la mutante
tufA	1†	
ompA	1	51
**metE*	70	94
fusA	Sí **	
cysK	62	90

Proteoma estimado de cepas provenientes de *E. coli* K12 MG1655

Al igual que con el proteoma promedio total, para observar la distribución del proteoma promedio estimado liberado, resultado de la eliminación de los genes de las cepas provenientes de *E. coli* K12 MG1655. Se graficó y analizó la distribución de la aportación al proteoma por proteína codificada de cada gen eliminado en las cepas en cada una de las condiciones. De manera que para cada cepa se analizó la distribución para una cantidad de genes que iban de entre 429 y 78. Para esto, se obtuvo el promedio de aportación de cada proteína codificada por los genes eliminados de cada cepa en las 22 condiciones y se sacó el logaritmo base 10 para observar más fácilmente la distribución (Figura 6). La cepa Δ16 fue la que más genes cuyas proteínas tenían valores mayores a 0.1 fgPC, teniendo 52 genes que representaban un 12.1% de sus genes eliminados, y la que menos tuvo fue la cepa MDS12, con tres genes que representaban un 3.8% de sus genes eliminados. Esto indica que en este subconjunto de genes se mantiene la tendencia a que hay un menor número de genes cuyas proteínas codificadas tienen una alta carga proteómica.

Haciendo la unión del conjunto de todos los genes eliminados de las cepas provenientes de *E. coli* K12 MG1655, se realizaron tablas con los cinco primeros genes cuyas proteínas codificadas más aportaban al proteoma (Tabla 8). Además, se agregó una primera columna '#' que nos indica la posición que tenía ese gen dentro de los genes cuyas proteínas codificada más carga proteómica representaban dentro del promedio de la información proteómica (la posición entre los 2359 genes del promedio de las 22 condiciones) (Schmidt et al., 2016). El gen cuya proteína más carga representaba dentro de los genes eliminados de las cepas provenientes de *E. coli* K12 MG1655 es *gltA.* Este gen representa un 0.71% ± 0.30 del promedio de proteoma total y es el gen número 22 de 2359 con más carga con respecto al proteoma promedio total.

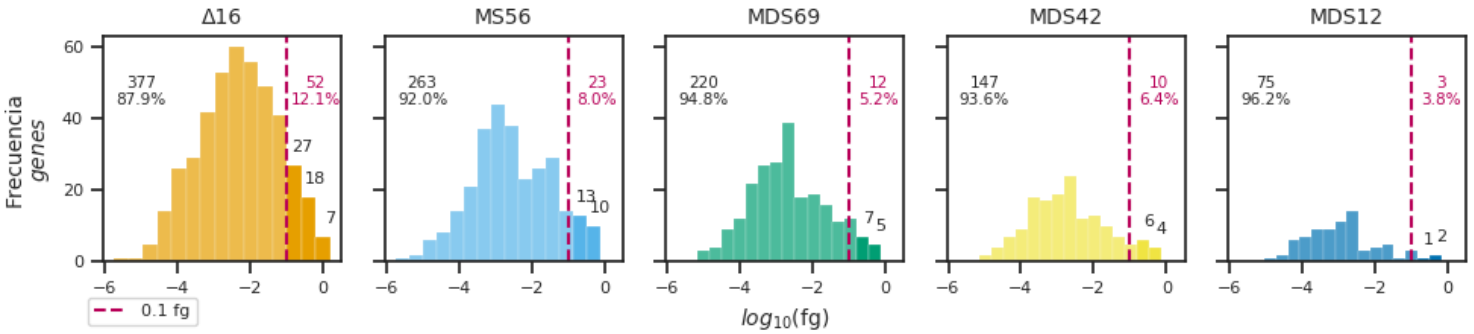

Figura 6. Distribución de aportación al proteoma promedio de las proteínas codificadas en los genes eliminados de las cepas provenientes de E. coli K12 MG1655 en fgPC.

Tabla 8. Primeros cinco genes cuyas proteínas codificadas aportan más al proteoma de los genes eliminados de las cepas provenientes de E. coli K12 MG1655 en fgPC y porcentajes con sus respectivas desviaciones estándar.

#	Gen	Promedio fg/célula	σ fg/célula	Promedio	σ	Descripción
22	*gltA*	1.60	±0.67	0.71%	±0.30%	Citrato sintasa
23	*aldA*	1.44	±1.07	0.64%	±0.48%	Aldehído deshidrogenasa A
53	*tpx*	0.78	±0.19	0.35%	±0.09%	Tiol peroxidasa
54	*sucC*	0.78	±0.36	0.35%	±0.16%	Succinil-CoA sintetasa subunidad β
55	*gnd*	0.78	±0.37	0.35%	±0.16%	6-fosfogluconato deshidrogenasa

Una lista de esencialidad de los genes cuyas proteínas codificadas más aportan al proteoma de los genes eliminados puede darnos información de genes que podrían ser de nuestro interés como genes blanco para minimización. Pero también pueden darnos pistas sobre por qué las cepas minimizadas presentan algunos fenotipos desfavorables al observar si existe evidencia de que la eliminación de los genes que han eliminado tiene un efecto negativo en la adecuación (Tabla 9). Por ejemplo, *gltA* es esencial en 23 condiciones y su ausencia afecta negativamente a la adecuación en 93 de 166 condiciones reportadas. Pero, por otro lado, quizás podríamos tomar *aldA* o *tpx* como genes blanco si las condiciones en las que planeamos tener a las cepas no se encuentran dentro de las condiciones en las que afecta negativamente a la adecuación (estos genes no serán considerados como blancos en nuestro estudio ya que nuestros criterios eliminaron aquellos que afecten negativamente a la adecuación en al menos una condición).

Tabla 9. Cantidad de condiciones con afectación a la adecuación de los cinco genes cuyas proteínas codificadas aportan más al proteoma de los genes eliminados de las cepas provenientes de E. coli K12 MG1655.

Gen	Esencial	Número de condiciones con defectos de adecuación en la mutante
gltA	23	93
aldA	0	2
tpx	0	4
sucC	6	22
gnd	0	18

Proteoma estimado de cepas provenientes de *E. coli* K12 MG1655 en glucosa

Al igual que en el análisis del proteoma total, además de utilizar el promedio, también se estudió el proteoma estimado liberado en la condición de glucosa resultado la eliminación de los genes de las cepas provenientes de *E. coli* K12 MG1655 (Figura 7). Como mencionamos en el análisis del proteoma total, los genes cuyas proteínas codificadas más aportan pueden variar dependiendo de la condición, y en este caso, al organizar los genes según la aportación al proteoma de sus proteínas codificadas (Tabla 10) encontramos tres genes que no estaban presentes en el mismo análisis pero para el proteoma promedio (Tabla 8): *gnd, ompT* y *ompX.* Además, en este caso el gen cuya proteína codificada más aporta es *gnd* con un 0.5%. También podemos observar que, esta condición con este subconjunto de genes, la distribución se mantiene y son pocos los genes cuyas proteínas codificadas tienen una mayor aportación al proteoma.

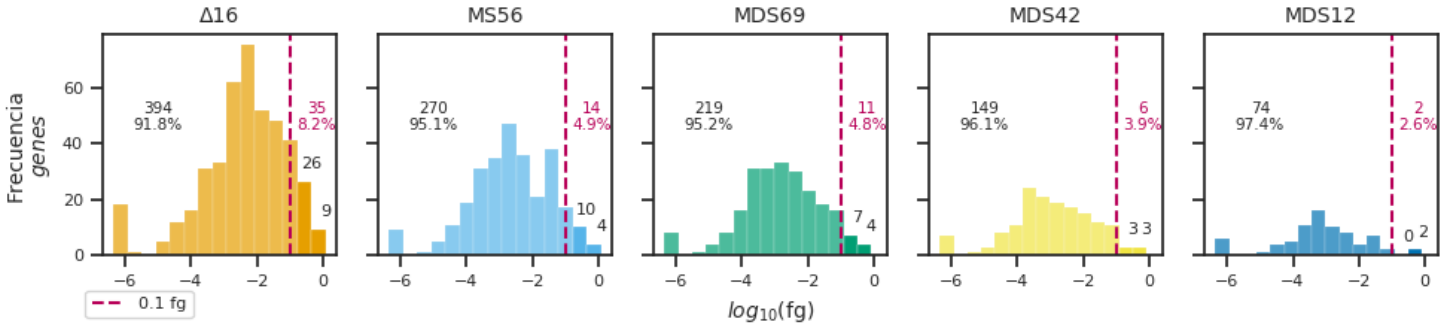

Figura 7. Distribución de aportación al proteoma en la condición de glucosa de las proteínas codificadas en los genes eliminados de las cepas provenientes de E. coli K12 MG1655 en fgPC.

Tabla 10. Primeros cinco genes cuyas proteínas codificadas aportan más al proteoma en la condición de glucosa de los genes eliminados de las cepas provenientes de E. coli K12 MG1655 en fgPC y porcentajes con sus respectivas desviaciones estándar. (*) Genes que no están presentes en la lista de la aportación promedio.

#	Gen	fg/célula	%	Descripción
31	**gnd*	1.20	0.50	6-fosfogluconato deshidrogenasa
63	**ompT*	0.82	0.34	Proteasa VII de la membrana externa
68	*gltA*	0.77	0.32	Citrato sintasa
70	*tpx*	0.73	0.30	Tiol peroxidasa
75	**ompX*	0.69	0.28	Proteína X de la membrana externa

Una vez más, al revisar cómo afectan a la adecuación estos genes, podemos observar cuáles podrían ser los responsables de fenotipos donde la adecuación se ve afectada negativamente y también cuáles podrían ser posibles blancos de minimización. En este caso, *ompT* no afecta negativamente a ninguna de las 166 condiciones reportadas por *Fitness Browser* por lo que podría ser un blanco de minimización.

Tabla 11. Cantidad de condiciones con afectación a la adecuación de los cinco genes cuyas proteínas codificadas aportan más al proteoma en la condición de glucosa de los genes eliminados de las cepas provenientes de E. coli K12 MG1655.

Gen	Esencial	Número de condiciones con defectos de adecuación en la mutante
**gnd*	0	18
**ompT*	0	0
gltA	23	93
tpx	0	4
**ompX*	0	6

Proteoma promedio estimado de cepas de *E. coli* K12 W3110

Así como se hizo con el proteoma total, se representó gráficamente y analizó la distribución del proteoma estimado originado por las proteínas codificadas por los genes eliminados en las cepas MGF y DGF. En este análisis se graficó y analizó la distribución de la aportación al proteoma promedio de cada proteína codificada por gen en cada una de las condiciones, abarcando todas las cepas (Figura 8). También se le aplicó el logaritmo base 10 al promedio de aportación para observar mejor la distribución. En general, todas las cepas tuvieron alrededor de 24 genes que pasaban los 0.1 fgPC, variando ligeramente el porcentaje. Todas las cepas mantuvieron la tendencia de, con este subconjunto de genes, tener pocos genes cuyas proteínas codificadas aportan mucho a la carga proteómica.

En la Tabla 12 se ordenaron los cinco genes cuyas porteínas codificadas, en promedio, más aportaban al proteoma. El primero fue *aldA,* el cual ocupa el lugar 23 de 2359 como gen cuya proteína más aporta al proteoma promedio total con 0.64% ± 0.48.

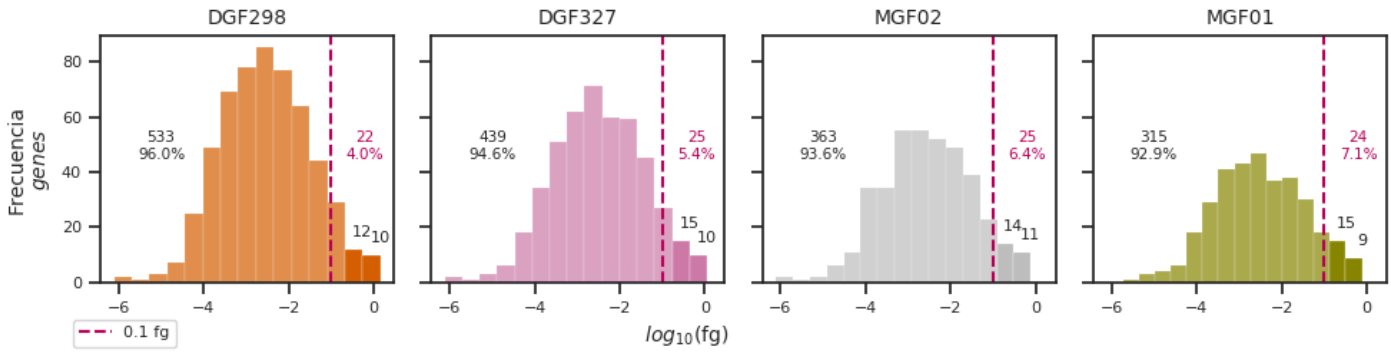

Figura 8. Distribución de aportación al proteoma promedio de las proteínas codificadas en los genes eliminados de las cepas MGF y DGF en fgPC.

Tabla 12. Primeros cinco genes cuyas proteínas codificadas aportan más al proteoma promedio de los genes eliminados de las cepas provenientes de E. coli K12 W3110 en fgPC y porcentajes con sus respectivas desviaciones estándar.

#	Gen	Promedio fg/célula	σ fg/célula	Promedio	σ	Descripción
23	*aldA*	1.44	±1.07	0.64%	±0.48%	Aldehído deshidrogenasa A
29	*adhE*	1.13	±0.55	0.50%	±0.24%	Aldehído-alcohol deshidrogenasa
44	*dppA*	0.86	±0.32	0.38%	±0.14%	Proteína de fijación de dipéptido
58	*rbsB*	0.77	±0.41	0.34%	±0.18%	Proteína de transporte de ribosa
61	*acs*	0.76	±0.56	0.34%	±0.25%	Acetil-CoA sintetasa

De nuevo, al revisar cómo afectan a la adecuación la eliminación de estos genes, podemos observar cuáles podrían ser los responsables de fenotipos negativos y también cuáles podrían ser posibles blancos dependiendo de la condición de nuestro interés. En este caso, *rbsB* no afecta negativamente a ninguna de las 166 condiciones reportadas por *Fitness Browser* por lo que podría ser un blanco de minimización.

Tabla 13. Cantidad de condiciones con afectación a la adecuación de los cinco genes cuyas proteínas codificadas aportan más al proteoma de los genes eliminados de las cepas provenientes de E. coli K12 W3110.

Gen	Esencial	Número de condiciones con defectos de adecuación en la mutante
aldA	0	2
adhE	0	3
dppA	0	4
rbsB	0	0
acs	0	3

Proteoma estimado de cepas provenientes de *E. coli* K12 W3110 en glucosa

Como en los casos anteriores, además de utilizar el promedio, también se estudió el proteoma estimado liberado en la condición de glucosa, siendo este proteoma estimado liberado proveniente de las proteínas codificadas por los genes eliminados en las cepas derivadas de *E. coli* K12 W3110 (Figura 9). Como mencionamos en el análisis del proteoma total, los genes cuyas proteínas codificadas más aportan pueden variar, y en este caso, al organizar los genes según la aportación al proteoma de sus proteínas codificadas (Tabla 14) encontramos cuatro genes que no estaban presentes en el mismo análisis pero para el proteoma promedio (Tabla 6): *ompT*, *cusF, oppA* y *tcyJ.* En este caso, el gen cuya proteína codificada más aporta es *ompT* con un 0.34%. También podemos observar que, en esta condición y en este subconjunto de genes, la distribución se mantiene y son pocos los genes cuyas proteínas codificadas tienen una mayor aportación al proteoma.

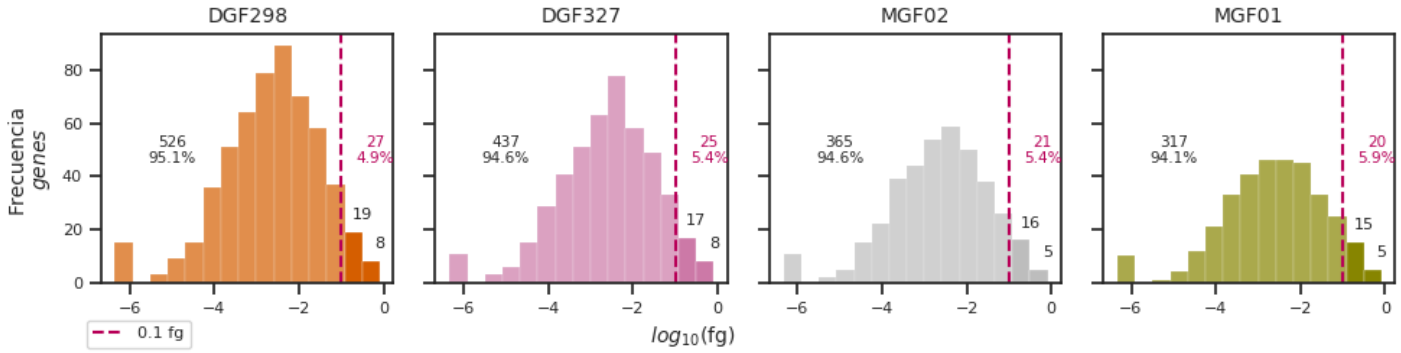

Figura 9. Distribución de aportación al proteoma en medio de glucosa de las proteínas codificadas en los genes eliminados de las cepas provenientes de E. coli K12 W3110 en fgPC

Tabla 14. Primeros cinco genes cuyas proteínas codificadas aportan más al proteoma en medio de glucosa de los genes eliminados de las cepas provenientes de E. coli K12 W3110 en porcentaje.

#	Gen	fg/célula	%	Descripción
63	**ompT*	0.82	0.34	Proteasa VII de la membrana externa
66	*adhE*	0.77	0.32	Aldehído-alcohol deshidrogenasa
87	**cusF*	0.60	0.25	Sistema de exportación de cobre / plata
90	**oppA*	0.58	0.24	Proteína de enlace del oligopéptido periplásmico
97	**tcyJ*	0.55	0.23	Proteína de unión a L-cistina

Al igual que en los casos anteriores, al revisar cómo afectan a la adecuación estos genes, podemos observar cuáles podrían ser los responsables de fenotipos negativos y también cuáles podrían ser posibles blancos dependiendo de la condición de nuestro interés (Tabla 15). En este caso, tenemos tres genes que no afectan negativamente a ninguna de las 166 condiciones reportadas por *Fitness Browser: ompT, cusF* y *oppA.* Por lo tanto, estos podrían ser candidatos a blancos de minimización.

Tabla 15. Cantidad de condiciones con afectación a la adecuación de los cinco genes cuyas proteínas codificadas aportan más al proteoma en la condición de glucosa de los genes eliminados de las cepas provenientes de E. coli K12 W3110.

Gen	Esencial	Número de condiciones con defectos de adecuación en la mutante
**ompT*	0	0
adhE	0	3
**cusF*	0	0
**oppA*	0	0
**tcyJ*	0	4

Consumo de recursos en términos de ATP del modelo ME

Para analizar el consumo de recursos en términos de ATP, se utilizó el modelo iJL1678b-ME. Este modelo proporciona información sobre los flujos metabólicos, de traducción y de transcripción. Sin embargo, en este modelo no se pueden simular diferencias en temperatura o pH. Pero sí podemos establecer diferencias en consumo, por ejemplo, de fuentes de carbono. Elegimos el medio mínimo con glucosa como punto de partida ya que, como se mencionó en la introducción, se ha demostrado que este modelo tiene una precisión del 87.5 % en sus predicciones de genes esenciales en comparación con un estudio de detección de genes esenciales en medio mínimo con glucosa (Lloyd et al., 2017; Monk et al., 2017).

Cuando se simula el crecimiento en medio mínimo con glucosa, el modelo iJL1678b-ME produce ~69 mmol•gDW^{-1}•h^{-1} de ATP, de los cuales se analizó el consumo de todas las categorías de reacciones del modelo individualmente (Figura 10 y Tabla 16). Primero observamos que las reacciones tipo "variable resumen" son las que mayor consumo de ATP tienen, representando el 43.7% del ATP producido. Hablaremos de dos de estas reacciones más adelante, pero en general son reacciones que imponen restricciones globales al modelo. Por ejemplo, la variable resumen "dilución de biomasa", obliga a que la tasa de producción de biomasa de diversos metabolitos sea igual a la tasa de su dilución en las células hijas durante el crecimiento. Después las reacciones metabólicas representan el 34.7% y en tercer lugar las reacciones de traducción con un 6.8%. Un 14.5% del ATP producido no es utilizado dentro de las reacciones del modelo (Otras reacciones), pero está ahí para representar a la energía que utiliza la célula para diferentes funciones, pero que no están modeladas. Esta energía no consumida en el modelo sería el sobrante de la energía producida por la reacción GAM (energía de mantenimiento asociada a crecimiento) sobre la cual ahondaremos a continuación.

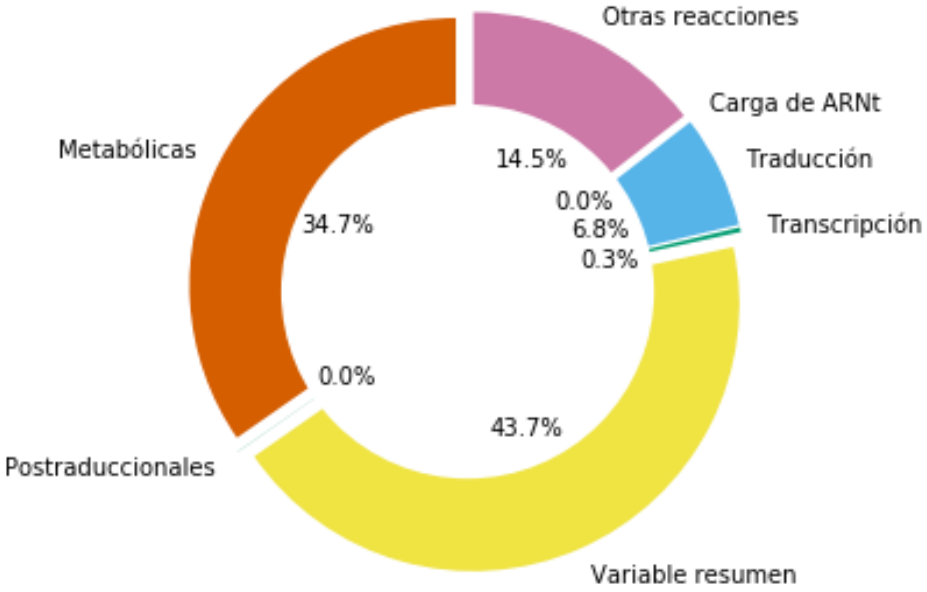

Figura 10. Porcentajes del ATP producido en el modelo iJL1678b-ME que es utilizado por las reacciones del mismo modelo.

Tabla 16. Consumo de recursos en términos de ATP de cada tipo de reacción del modelo ME

Tipo de reacción	Consumo de ATP $mmol \bullet gDW^{-1} \bullet h^{-1}$
Variable resumen	30.12
Metabólica	23.87
Otras	9.99
Traducción	4.70
Transcripción	0.19
Postraduccional	0.013
Carga de ARNt	0.0006
Total	68.88

Dentro de la categoría de "variable resumen" solo dos reacciones son las que consumen ATP: GAM (energía de mantenimiento asociada a crecimiento) y ATPM (función de mantenimiento de ATP) (Tabla 17). La reacción GAM representa la energía en forma de ATP necesaria para replicar una célula, incluida la síntesis de macromoléculas (secuencias extensas y repetitivas de una estructura química básica conocida como monómero) como ADN, ARN y proteínas (Thiele & Palsson, 2010). En cuanto a la reacción ATPM es una reacción de hidrólisis de ATP equilibrada que se utiliza para simular demandas de energía no asociadas con crecimiento, como lo es mantener el potencial de membrana, reparación celular y motilidad (Varma & Palsson, 1995). En el contexto del modelo, es "la energía que se requiere para mantener la homeostasis de una unidad de biomasa durante un intervalo de tiempo unitario" (Iizuka, 2016). En modelos a escala genómica es común encontrar coeficientes de mantenimiento que representan toda aquella energía que no es tomada en

cuenta en reacciones de biosíntesis y se usa para cerrar el balance energético. Las demás reacciones con mayor consumo entran en la categoría de reacciones metabólicas.

Tabla 17. Primeras 10 reacciones que más consumen recursos en términos de ATP.

Reacción	Tipo	Consumo de ATP $mmol \bullet gDW^{-1} \bullet h^{-1}$
**GAM*	Variale resumen	29.124
Adenilato quinasa	Metabólica	6.872
Fosfofructoquinasa	Metabólica	4.320
translation_dummy	Traducción	1.671
Glutamina sintetasa	Metabólica	1.573
Hexoquinasa (D-glucosa: ATP)	Metabólica	1.467
**ATPM*	Variale resumen	1.000
Aspartato quinasa	Metabólica	0.929
Fosforribosilpirofosfato sintetasa	Metabólica	0.819
Carbamato quinasa	Metabólica	0.626

Liberación teórica de recursos en términos de ATP en las cepas provenientes de *E. coli* K12 MG1655

Una vez que tenemos los valores de flujo de consumo de ATP para cada reacción, podemos determinar qué genes están involucrados en cada una y cuánto contribuyen (ver la sección de métodos). Esto nos permite determinar la cantidad de ATP consumida por las reacciones asociadas a un gen. Utilizaremos la abreviatura 'RA' para hacer referencia a estas reacciones asociadas a lo largo de este informe.

En promedio, 15% de los genes eliminados en las cepas MDS están representados en el modelo ME (Tabla 18). Debido a esto, solo podremos calcular el flujo de consumo de ATP según el modelo ME que tendrían las RA a ese ~15% de genes si no hubiesen sido eliminados. Dicho de otro modo, podemos calcular su flujo de consumo de ATP equivalente en la cepa sin modificaciones del modelo ME (*E. coli* K12 MG1655). Como estos genes fueron eliminados en las cepas minimizadas, tomaremos la cantidad de ATP que hubiesen consumido sus RA como la liberación teórica de recursos en términos de ATP. La cepa proveniente de *E. coli* K12 MG1655 con mayor representación en el modelo ME sería la Δ16 con 25% (Tabla 18).

Tabla 18. Porcentaje de cobertura de los genes eliminados de las cepas provenientes de E. coli K12 MG1655 con los genes del modelo ME

MDS12	12%
MDS42	15%
MDS69	17%
MS56	21%
Δ16	25%

La liberación teórica de recursos en términos de ATP en las cepas MDS y MS56 fue en promedio de 0.16 mmol•gDW^{-1}•h^{-1} de ATP o 0.23% del ATP producido, siendo ligeramente más alta la MS56. Finalmente, la Δ16 llega a 0.7 mmol•gDW^{-1}•h^{-1} de ATP o 1.13% de todo el ATP producido, siendo mucho mayor a todas las demás cepas (Figura 11).

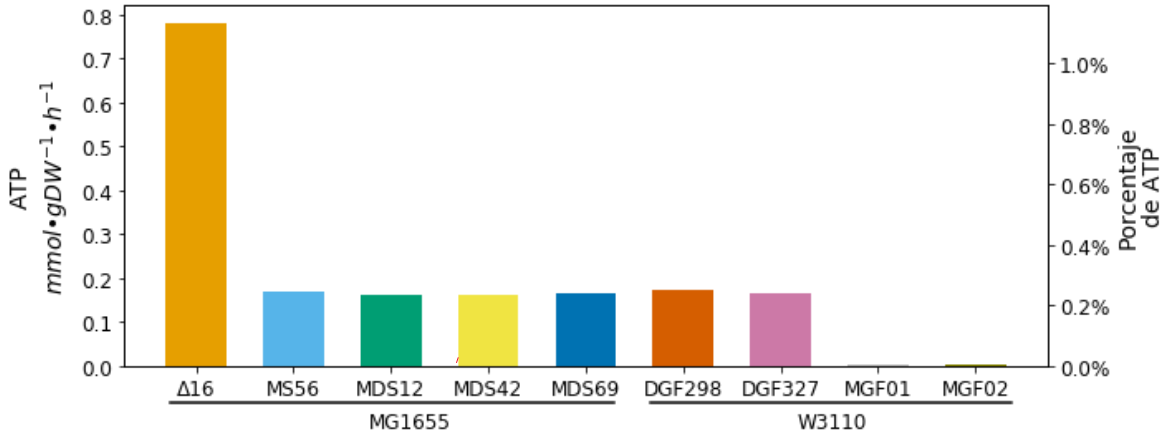

Figura 11. Flujo de consumo de ATP equivalente en la silvestre aportado por las reacciones relacionadas a los genes que se eliminaron de las cepas minimizadas en mmol•gDW^{-1}•h^{-1} de ATP en el eje y izquierdo. En el eje y derecho se encuentra el valor en porcentaje tomando como total el ATP producido por el modelo ME.

Liberación teórica de recursos en términos de ATP en las cepas provenientes de *E. coli* K12 W3110

Al igual que para las cepas provenientes de *E. coli K12 MG1655,* se determinó qué genes estaban involucrados en cada una de las reacciones y cuánto contribuían (ver la sección de métodos). De esta manera, obtuvimos la cantidad de ATP consumida por las RA a un gen.

En promedio, 25.78% de los genes eliminados en las cepas MGF y DGF están representados en el modelo ME. Debido a esto, únicamente conoceremos lo que pasa en ~25.78% de los genes y cómo es es el flujo de consumo de ATP de sus RA. Siendo este consumo el equivalente en la silvestre del modelo ME (*E. coli* K12 MG1655). Como este es el consumo aportado por las RA a los genes que se eliminaron de las cepas minimizadas, lo tomaremos como la liberación teórica de recursos en términos de ATP. Las cepas cuyos genes están más representados en el modelo ME serían las DGF con 27% (Tabla 19).

Tabla 19. Porcentaje de cobertura de los genes eliminados de las cepas provenientes de E.coli K12 W3110 con los genes del modelo ME

MGF01	24%
MGF02	26%
DGF298	27%
DGF327	27%

La liberación teórica de recursos en términos de ATP en las cepas MGF (~0.004 mmol•gDW^{-1}•h^{-1} de ATP o un 0.006% del ATP producido) fue más bajo que las DGF (~0.17 mmol•gDW^{-1}•h^{-1} de ATP o un 0.24% del ATP producido) (Figura 11). Entre las cepas DGF la DGF-298 liberó un 0.008 más de mmol•gDW^{-1}•h^{-1} de ATP que la DGF-327.

Distribución del consumo de recursos en términos de ATP

Consumo total de recursos en términos de ATP

Para observar la distribución del consumo de recursos en términos de ATP del modelo iJL1678b-ME, se graficó y analizó la aportación al flujo de consumo de ATP por las reacciones relacionadas a cada gen. En promedio, apenas el 13.6% de las RA a los genes alcanzaron un consumo de más de 0.01 mmol•gDW^{-1}•h^{-1} de ATP, lo que sugiere que son pocos los genes cuyas RA tienen un mayor flujo de consumo de ATP (Figura 12). En la Tabla 20 se puede observar los cinco primeros genes cuyas RA tienen una mayor aportación en porcentaje al consumo de ATP y sus funciones. Tan solo las RA a *accC* consumen el 32.42% de todo el ATP producido. El gen accC codifica para la biotina carboxilasa, la cual es un dominio de la acetil-CoA carboxilasa. La acetil-CoA carboxilasa es responsable de catalizar el paso regulado en la síntesis de ácidos grasos y es de vital importancia ya que estos son empleados para la biogénesis de la membrana en bacterias (Broussard et al., 2013; Cronan & Waldrop, 2002). Además, de acuerdo a la información proteómica en medio con glucosa de Schmidt *et al*., 2016, es del ~14% de genes cuyas proteínas asociadas tienen un valor mayor a 0.1 fgPC de aportación al proteoma.

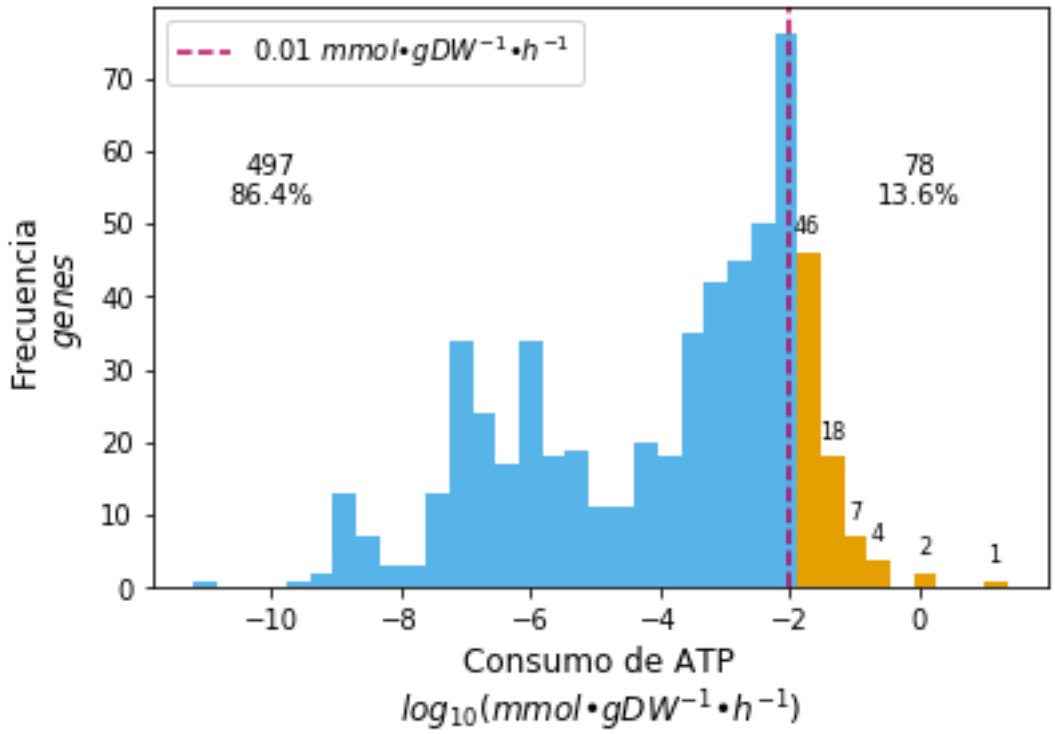

Figura 12. Distribución del consumo de recursos en términos de ATP por gen en el modelo ME.

Tabla 20. Primeros cinco genes (sin contar el gen dummy) cuyas RA consumen más ATP en el modelo ME.

Gen	Consumo de ATP mmol•gDW^{-1}•h^{-1}	%	Descripción
accC	22.208	32.24	Biotina carboxilasa
dummy	1.678	2.44	
cysD	0.971	1.41	Sulfato adenililtransferasa subunidad 2
lpp	0.231	0.34	Lipoproteína de mureína
sucB	0.193	0.28	Dihidrolipoil transuccinilasa
lpd	0.180	0.26	Lipoamida deshidrogenasa

Al igual que con el proteoma, el eliminar los genes cuyas RA consumen más ATP podría sonar como buenos blancos para minimización, pero al revisar cómo afectan a la adecuación, podemos ver que *accC* es esencial de acuerdo a la base de datos PEC (Tabla 21). Esto no es sorpresa, ya que como se mencionó, este gen está involucrado en procesos esenciales. Por parte de *cysD* y *lpp*, según la información de la base de datos *Fitness Browser*, son esenciales en algunas condiciones y su ausencia afecta negativamente muchas otras. Finalmente, *sucB* y *lpd* no se encuentran en la base de datos *Fitness Browser,* pero sí se menciona en la base de datos EcoCyc que son esenciales en algunas condiciones. Esto apoya la idea de que la minimización de recursos no debería realizarse sin considerar los efectos a la adecuación para evitar perder viabilidad o presentar fenotipos que puedan afectar negativamente la producción de metabolitos de interés.

Tabla 21. Cantidad de condiciones con afectación a la adecuación de los primeros cinco genes (sin contar el gen dummy) que consumen más ATP en el modelo ME. (**) Sin información de adecuación, datos obtenidos de Profiling of Escherichia coli Chromosome (PEC). (†) Sin información de adecuación, datos obtenidos de Escherichia coli K12 substr. MG1655 reference genome (EcoCyc)

Gen	Esencial	Número de condiciones con defectos de adecuación en la mutante
accC	Sí**	
cysD	14	85
lpp	1	66
sucB	2†	
lpd	3†	

Una vez más, no debemos olvidar que el que algunos genes estén ausentes en las listas de Fitness Browser puede llegar a indicar que sean "posibles esenciales" ya que no fue posible obtener una mutante por medio de transposones ya que probablemente eran esenciales.

Consumo teórico de recursos en términos de ATP en cepas provenientes de *E. coli* K12 MG1655

Al igual que con el consumo total de recursos en términos de ATP, para observar la distribución del flujo de consumo de ATP equivalente en la silvestre, aportado por las RA a los genes eliminados de las cepas provenientes de *E. coli* K12 MG1655, se graficó y analizó el flujo de consumo de ATP por cada gen y todas las cepas. En promedio, únicamente el 6.7% de los genes llegaba a aportar más de 0.01 mmol•gDW^{-1}•h^{-1} de ATP por lo que son pocos los genes cuyas RA tienen un mayor consumo de ATP del modelo ME (Figura 13). En la Tabla 22 se puede observar los cinco primeros genes cuyas RA tienen un mayor consumo de ATP en porcentaje (de entre los genes eliminados), sus funciones y la posición que representan en el consumo total. Estos cinco genes están entre los 30 que cuyas RA consumen más ATP, siendo *sucB* el primero de esta lista de cinco con un 0.28% del ATP producido.

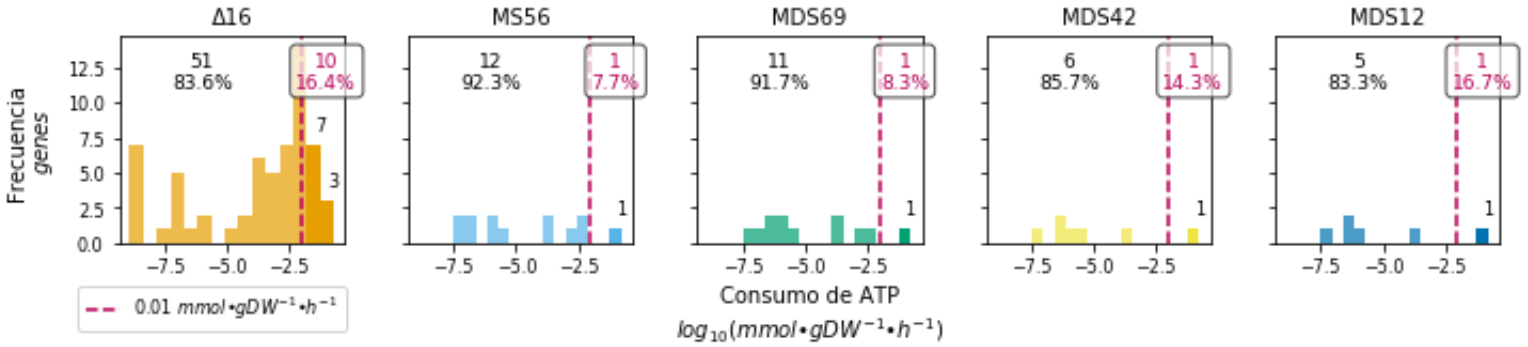

Figura 13. Distribución del flujo de consumo de ATP equivalente en la silvestre por los genes que se eliminaron de las cepas derivadas de E. coli K12 MG1655 en mmol•gDW^{-1}•h^{-1} de ATP.

Tabla 22. Primeros cinco genes cuyas proteínas codificadas aportan más al consumo de ATP equivalente en la silvestre de los genes eliminados de las cepas derivadas de E. coli K12 MG1655 en porcentaje

#	Gen	Consumo de ATP mmol•gDW^{-1}•h^{-1}	%	Descripción
5	*sucB*	0.193	0.280	Dihidrolipoil transuccinilasa
7	*ompN*	0.161	0.234	Porina N de la membrana externa
11	*sucA*	0.099	0.143	Complejo 2-oxoglutarato deshidrogenasa
27	*hisD*	0.034	0.050	Histidinol deshidrogenasa
30	*trpD*	0.032	0.047	Subunidad de antranilato sintasa

Consumo teórico de recursos en términos de ATP en cepas de *E. coli* K12 W3110Tabla 23 se puede observar los cinco primeros genes con mayor consumo de ATP en porcentaje (de entre los genes eliminados), sus funciones y la posición que representan en el consumo total.Estos cinco genes están entre los 257 que más consumen ATP. El número uno de las cepas MGF y DGF, y el número siete en consumo total, es el gen *ompN* el cual representa el 0.23% del ATP producido.

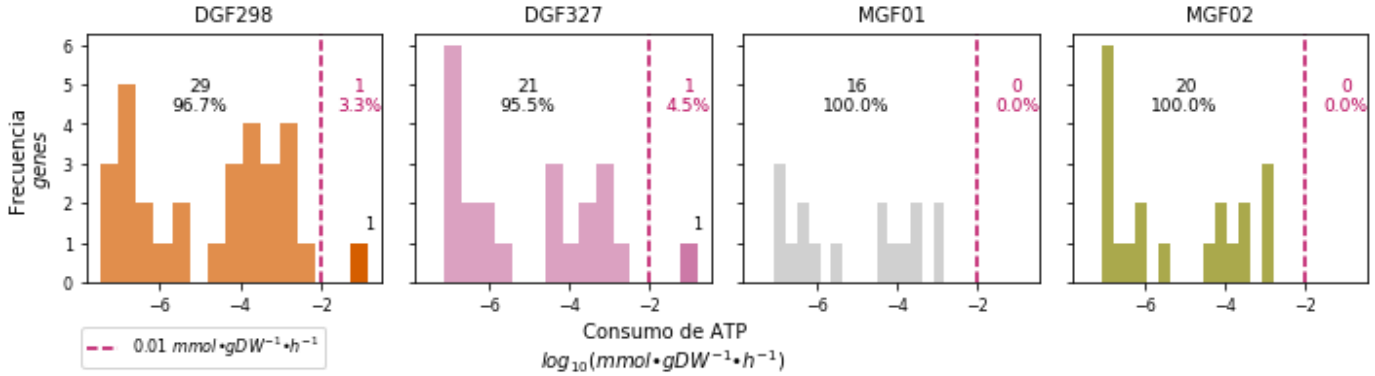

Figura 14. Distribución del flujo de consumo de ATP equivalente en la silvestre por los genes eliminados de las cepas MGF y DGF en mmol•gDW-1•h-1 de ATP.

Tabla 23. Primeros cinco genes que más ATP consumen equivalente en la silvestre por los genes eliminados de las cepas MGF y DGF en porcentaje.

#	Gen	Consumo de ATP $mmol \bullet gDW^{-1} \bullet h^{-1}$	%	Descripción
7	*ompN*	0.161	0.2340	Porina N de la membrana externa
200	*rpiB*	0.003	0.0040	Ribosa-5-fosfato isomerasa B
232	*yqeA*	0.002	0.0022	Aminoácido quinasa (putativa) YqeA
242	*queF*	0.001	0.0018	7-ciano-7-deazaguanina reductasa
257	*cynT*	0.001	0.0014	Anhidrasa carbónica 1

Costo de ATP en megabases (Mb)

A pesar de haber obtenido valores de costo de ATP en los puntos anteriores, estamos limitados a los genes que se encuentren en los datos proteómicos o en los datos del modelo ME. Por lo que a continuación se utilizó el modelo ME y los datos proteómicos para poder generalizar costos de un gen en relación a sus Mb.

Costo de ATP por replicación según modelo ME

Con el objetivo de modelar el costo de ATP de la replicación de ADN, se modificó el porcentaje de ADN por célula en el modelo iJL1678b-ME. Para obtener la cantidad de ADN en Mb por célula en las velocidades específicas de cremiento modeladas, se utilizaron datos basados en estudios de Bremer & Dennis, 1996 (Bremer & Dennis, 1987; Dennis & Bremer, 1974) que describen el cambio observado en la cantidad de ADN dependiendo de la velocidad específica de crecimiento (Davison, 1966). Los parámetros de porcentaje de ADN por célula que se encuentran el modelo, basados en valores experimentales de Bremer & Dennis, 1996 (Bremer & Dennis, 1987; Dennis & Bremer, 1974), se multiplicaron por 1.125, 1.25 y 1.5 para obtener nuevos valores. De manera que si se multiplica por 1.25 quiere decir que se aumentó el porcentaje de ADN en un 25%. Después se calculó cuál sería el valor respectivo de equivalentes de genoma de ADN por célula (número promedio de genomas presentes en una célula) de acuerdo a la velocidad específica de crecimiento así como el porcentaje de ADN para obtener el aumento en Mb. Para esto, se hizo un ajuste de curva para los datos proporcionados por Bremer & Dennis, 1996 (Bremer & Dennis, 1987; Dennis & Bremer, 1974) para *E. coli* y determinar el equivalente en genomas para las simulaciones hechas a una velocidad específica de crecimiento de 0.69 h^{-1}. El ajuste con una R^2=0.99 arrojó que había un equivalente de ~1.63 genomas a una μ de 0.69 h^{-1} (Figura 15).

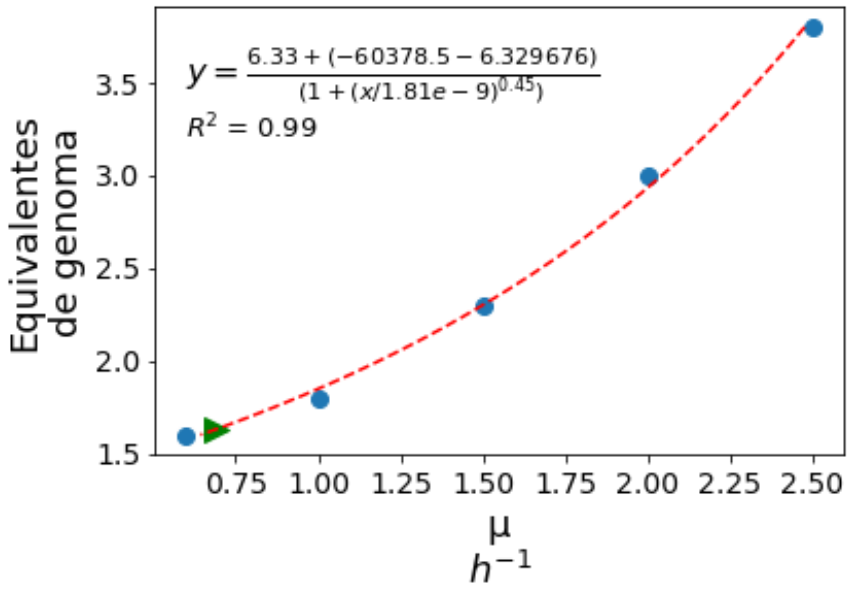

Figura 15. Predicción de los equivalentes en genomas para la velocidad específica elegida. Los puntos azules corresponed a los datos proporcionados por Bremer & Dennis. El triángulo verde hace referencia al valor determinado de equivalentes de genoma para la velocidad específica de crecimiento de las condiciones modeladas.

Una vez obtenido el equivalente en genomas para nuestras simulaciones, pudimos calcular el aumento del costo de ATP por Mb. Para esto multiplicamos el tamaño del genoma de *Escherichia coli* str. K12 substr. MG1655 por el equivalente en genomas y después por el porcentaje de ADN aumentado en cada una de las simulaciones. Después se calculó el flujo de consumo de ATP para cada valor y obtuvimos la Figura 16. Siendo que, si no aumentamos el porcentaje de ADN a replicar, tenemos un consumo menor a 48.835. Sin embargo, al aumentar el tamaño del genoma, y, por lo tanto, la cantidad de ADN a replicar, el flujo de consumo de ATP aumenta con una razón de cambio de ~0.017 mmol•gDW^{-1}•h^{-1} por Mb.

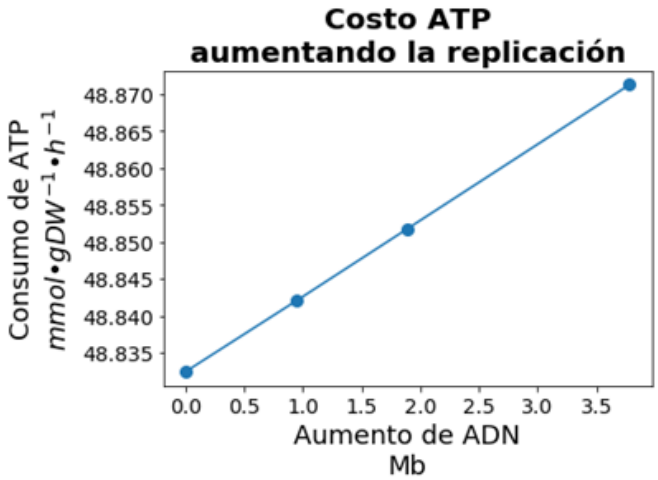

Figura 16. Variación en el flujo de consumo de ATP al aumentar la cantidad de genes a replicar.

Costo de ATP por transcripción según modelo ME

Una vez analizados los costos de replicación, buscamos analizar el costo de aquellos genes que no solo son replicados, sino también transcritos, ya que en la simulación anterior solo no se agregó un flujo de transcripción a la cantidad de ADN añadida. Para analizar los costos de ATP por agregar genes que se transcriben también se utilizó el modelo iJL1678b-ME. Tomando como modelo base aquel con los valores por defecto UPF=0.36. Se calculó el valor promedio del flujo de transcripción de los genes, de manera los genes agregados al modelo no fueran solo replicados, sino que también cumplieran con dicho flujo de transcripción, así como el tamaño promedio de un gen (~1 kb). Este aumento de genes fue modelado cuatro veces con valores diferentes de aumento de genes, los cuales eran 230, 460, 690 y 920 genes. ya que representan un aumento de aproximadamente 5, 10, 15 y 20% de los 4600 genes de *E. coli K12* (*Escherichia Coli K-12 Substr. MG1655 All-Genes*, n.d.; Keseler et al., 2017).

Al inicio los cálculos se realizaron sin fijar la velocidad específica de crecimiento, ni valores predeterminados de flujos de consumo de oxígeno y glucosa. Esto permitió que estos parámetros variaran entre simulaciones, lo cual imposibilitó utilizar las simulaciones para comparar la demanda de ATP entre ellas. Esto dado que un cambio en la velocidad específica de crecimiento implica cambios en la producción de biomasa, así como en costos de replicación. Debido a esto, se realizaron las simulaciones donde, además de los parámetros ya establecidos, se agregaron restricción para que el modelo mantuviera una μ de 0.69 h^{-1}, un consumo de glucosa de 10 mmol•gDW^{-1}•h^{-1} y un consumo de oxígeno de 18 mmol•gDW^{-1}•h^{-1}. Una vez aplicados estos parámetros, pudimos comparar el cambio en la demanda de ATP de manera global (tomando en cuenta todos los procesos del modelo) (Figura 17).

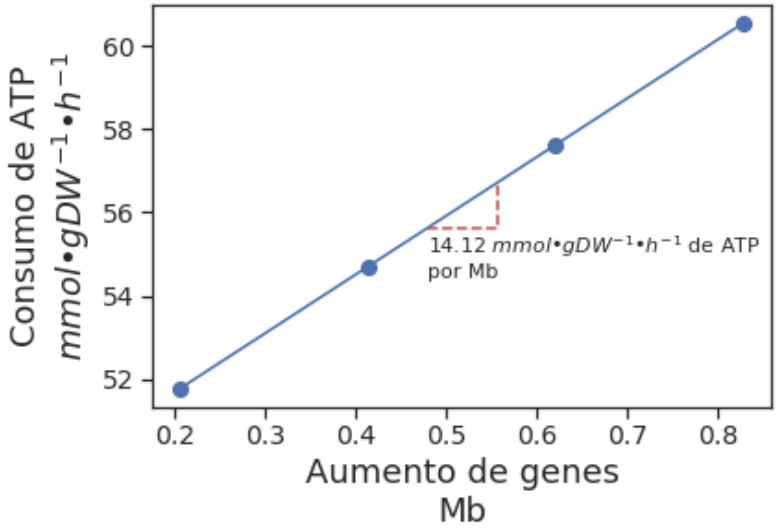

Figura 17. Variación en el flujo de consumo de ATP de manera de los flujos de transcripción, traducción y metabólicos al aumentar la cantidad de genes con flujos de transcripción.

Podemos observar que, al aumentar el número de genes transcritos, y, en consecuencia, la cantidad de ADN a transcribir, el flujo de consumo de ATP aumenta con una razón de cambio de ~14.12 mmol•gDW^{-1}•h^{-1} por Mb.

Sin embargo, esto toma en cuenta los flujos de traducción y flujos metabólicos, no solamente los flujos de transcripción. Si contamos solamente aquellos flujos destinados a transcripción, obtenemos la Figura 18, donde el flujo de consumo de ATP aumenta con una razón de cambio de ~ 0.29 mmol•gDW^{-1}•h^{-1} por Mb.

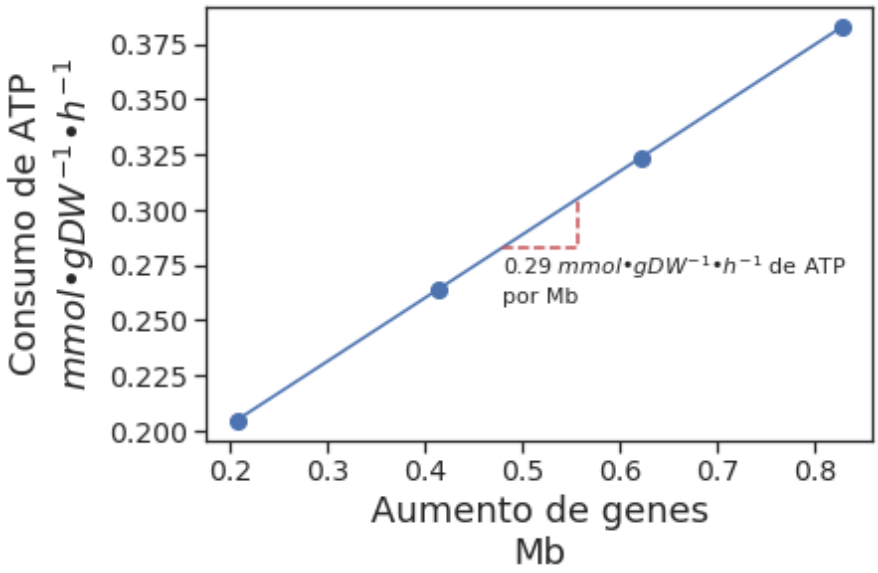

Figura 18. Variación en el flujo de consumo de ATP solo tomando en cuenta los flujos de transcripción, al aumentar la cantidad de genes a transcribir.

Cálculo manual de costo de ATP por replicación y transcripción

De acuerdo con la ecuación de Lynch & Marinov, 2015., el costo de la replicación de ADN en bacterias, en unidades de hidrólisis de ATP por célula, se puede calcular con la siguiente fórmula (Lynch & Marinov, 2015):

$$101L_g$$

Donde L_g representa el tamaño del gen en nucleótidos. Mientras que el costo en la transcripción en bacterias se calcula con la siguiente fórmula:

$$2N_r L_g(23 + \delta_r t)$$

Siendo N_r el número promedio permanente de ARNm maduros para el gen dentro la célula, L_g el tamaño del gen en nucleótidos, δ_r la tasa de degradación por transcrito en copias por segundo y t la velocidad específica de crecimiento en horas.

Con estos datos calcularon que, para un gen promedio de 950 bases, el costo para la replicación es de ~1.36x10^5 unidades de hidrólisis de ATP por célula mientras que para la transcripción es de ~1.5x10^5 unidades de hidrólisis de ATP por célula. Esto representaría un costo de ~1.43x10^5 unidades de hidrólisis de ATP por célula por Mb para la replicación y ~1.57x10^5 unidades de hidrólisis de ATP por célula por Mb para la transcripción.

Para compararlo con los costos obtenidos por los modelos, cambiamos las unidades de mmol•gDW^{-1}•h^{-1} de ATP por Mb en unidades de hidrólisis de ATP por célula. Con esto obtuvimos que, de acuerdo al modelo ME se tenía un costo de ~2.62 x10^{6} unidades de hidrólisis de ATP por célula por Mb para la replicación. Mientras que la transcripción tiene un costo de ~3.60x10^{9} unidades de hidrólisis de ATP por célula por Mb de manera global, y de ~7.3x10^{7} unidades de hidrólisis de ATP por célula por Mb contando solo flujos de transcripción. Las diferencias entre las magnitudes pueden deberse a que Lynch & Marinov, 2015 asumen un reciclaje rápido y eficiente de los ribonucleótidos (Lynch & Marinov, 2015).

Costos de ATP por modificación de proteoma

Costo de ATP por proteoma

Para el cálculo de costo de proteoma, se utilizaron los valores de UPF de 0.27, 0.30, 0.33 y 0.36 siendo 0.36 (36%) el valor basal del modelo. También se fijó la μ a 0.69 h^{-1}. Este cálculo representa la predicción del modelo ante una reducción del proteoma del 0.03 o 3% entre cada valor.

La función objetivo ATPM (mantenimiento de ATP) puede utilizarse para maximizar la producción de ATP (De Jong & Giordano, 2020). En este caso había diferencias en los flujos de consumo de glucosa y oxígeno, por lo que se fijaron para que los cambios en el flujo de consumo de ATP no fueran resultado de los cambios de consumo de oxígeno y glucosa. La función ATMP tiene como ventaja es que es una reacción de *SummaryVariable* (reacciones que imponen restricciones al modelo). Por lo tanto, no es una reacción reversible, de manera que evitamos artificios de parte del modelo en los cuales para contender con el aumento de una reacción solo aumenta el flujo de la reacción reversa.

Debido a que, como ya mencionamos, el valor de consumo de oxígeno y glucosa variaban los fijamos. Primero fijamos el valor de consumo de oxígeno a 18 mmol•gDW^{-1}•h^{-1} tomando en cuenta que no fuera un valor restrictivo para el flujo de consumo de ATP tomando como base los resultados de la simulación anterior. Sin embargo, ese valor de oxígeno seguía permitiendo variación en el consumo de glucosa por lo que también se fijó el valor de consumo de glucosa a 10 mmol•gDW^{-1}•h^{-1}, de lo cual obtuvimos la Figura 19.

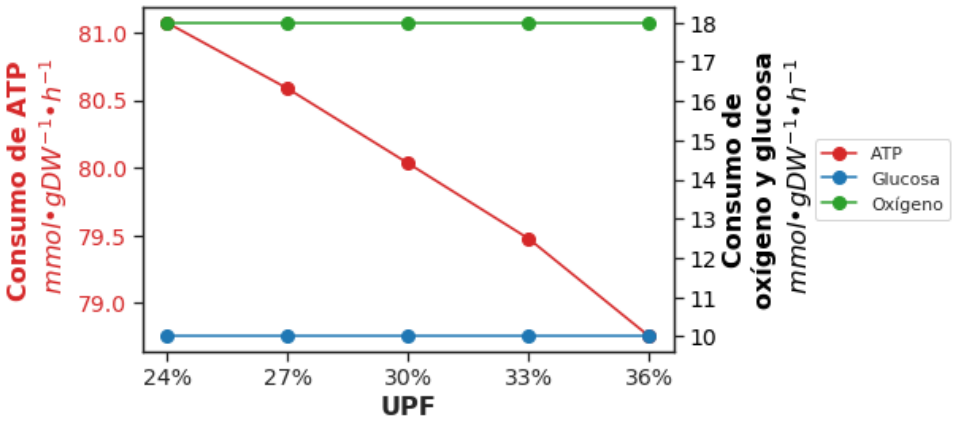

Figura 19. Variaciones en el flujo de consumo total de ATP en diferentes valores de UPF, con valores de flujo de consumo de oxígeno y glucosa fijos, y teniendo como objetivo la reacción ATPM.

Consumo celular de ATP por proteoma

Restamos el flujo de consumo de ATP de la función objetivo, ATPM, del consumo total para conocer el consumo celular (Figura 20). El costo de ATP aumenta con una razón de cambio de ~0.16 mmol•gDW^{-1}•h^{-1} por 1% de UPF.

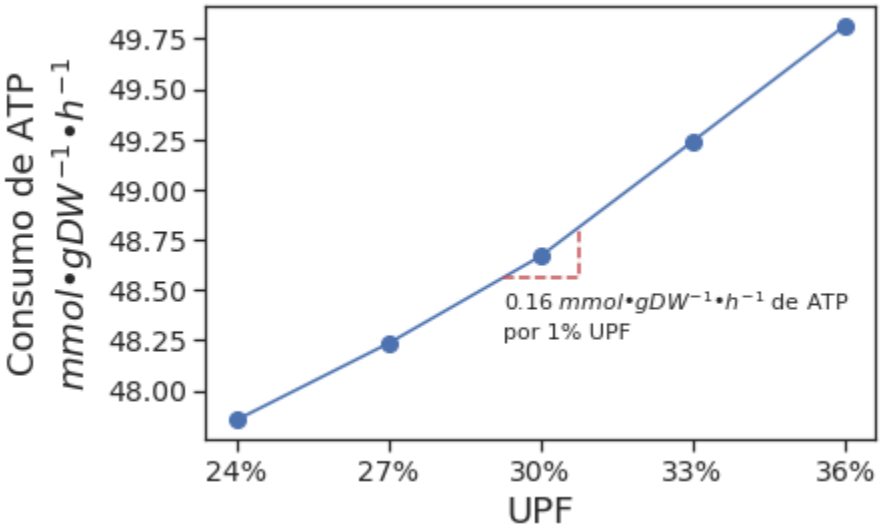

Figura 20. Variaciones en el consumo celular de ATP en diferentes valores de UPF, con valores de flujos de consumo de oxígeno y glucosa fijos, y teniendo como objetivo la reacción ATPM.

Costo de ATP por proteína recombinante

ATPM como objetivo y variando los valores de GFP

Para observar el costo de ATP en el caso de producción de una proteína recombinante, se agregó la secuencia nucleotídica de la proteína GFP en el modelo, así como los procesos de transcripción y traducción. En esta simulación. se fijó la μ a 0.69 h^{-1}, la UPF se mantuvo en el valor basal de 36% y los valores del flujo de producción de la proteína GFP fueron 0.00025, 0.005, 0.0075 mmol•gDW^{-1}•h^{-1}. Por lo que, por ejemplo, nuestro modelo simula que nuestra célula produce 0.0025 mmol•gDW^{-1}•h^{-1} de GFP.

En este caso estamos viendo el efecto de aumentar la producción de proteína recombinante sin un cambio en el proteoma. Al igual que en los puntos anteriores, para lograr obtener flujos comparables y observar el efecto del aumento de GFP, pusimos atención en que los flujos de consumo de glucosa y oxígeno se mantuvieran fijos, estableciendo el flujo de consumo de oxígeno a 18 mmol•gDW^{-1}•h^{-1} y el flujo de de consumo de glucosa de 9 mmol•gDW^{-1}•h^{-1} (Figura 21).

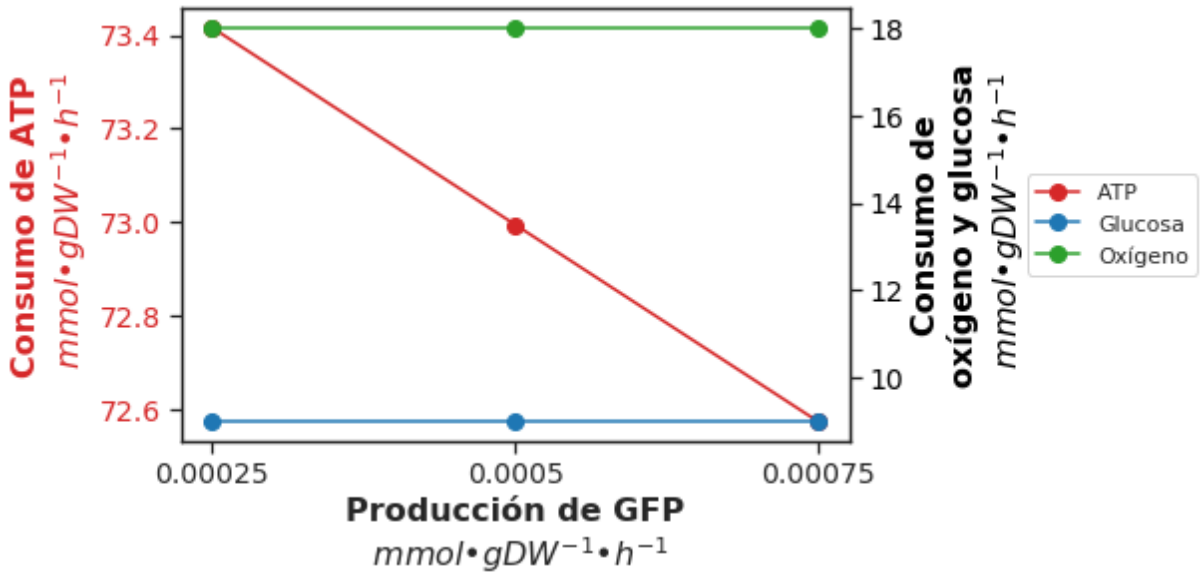

Figura 21. Variaciones en el flujo de consumo total de ATP en diferentes valores de producción de GFP, con valores de flujos de consumo de oxígeno y glucosa fijos, y teniendo como objetivo la reacción ATPM.

Consumo celular de ATP por producción de proteína recombinante

Restamos el flujo de consumo de ATP de la función ATPM del consumo total para conocer el consumo celular (Figura 22). El costo de ATP aumenta con una razón de cambio de ~16.45 mmol•gDW^{-1}•h^{-1} por mmol•gDW^{-1}•h^{-1} de GFP.

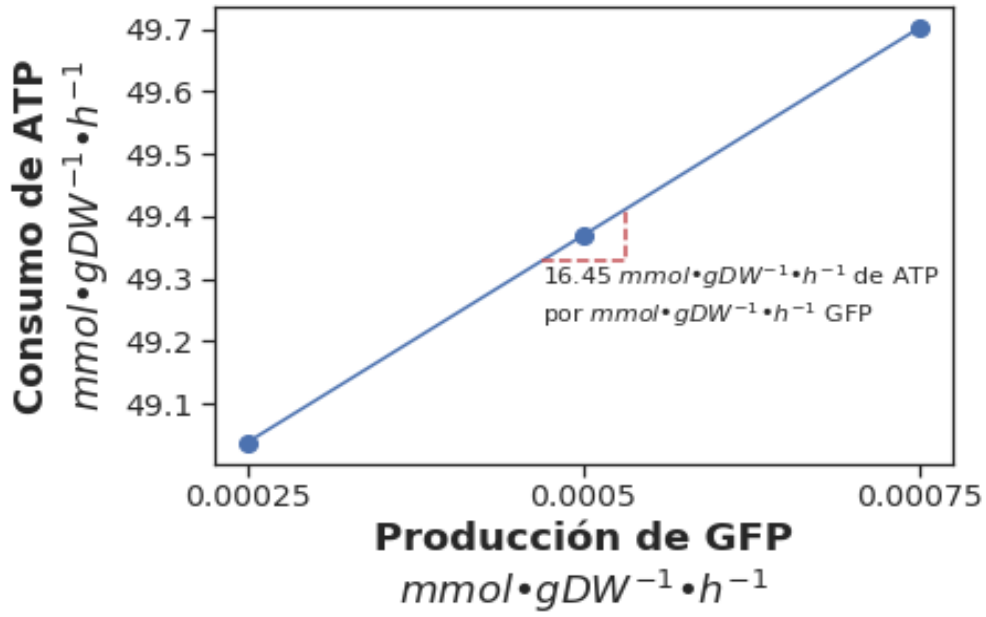

Figura 22. Variaciones en el consumo celular de ATP en diferentes valores de producción de GFP, con valores de flujos de consumo de oxígeno y glucosa fijos, y teniendo como objetivo la reacción ATPM.

GFP como objetivo y variando los valores de UPF

Para analizar cómo afectaba el flujo de consumo de ATP cuando utilizamos a la GFP como objetivo y variamos la UPF, se fijó la μ a 0.69 h^{-1}. Se utilizaron los valores de UPF de 27, 30, 33 y 36% recordando que 36% es valor basal del modelo. En este caso, lo que veríamos es el cambio del consumo de recursos en términos de ATP cuando liberamos proteoma y tenemos producción de una proteína recombinante.

A diferencia de los demás casos, el poner como objetivo a GFP fue suficiente para que el consumo de oxígeno y de glucosa se mantuvieran estables, por lo que no fue necesario restringir ningún flujo extra (Figura 23).

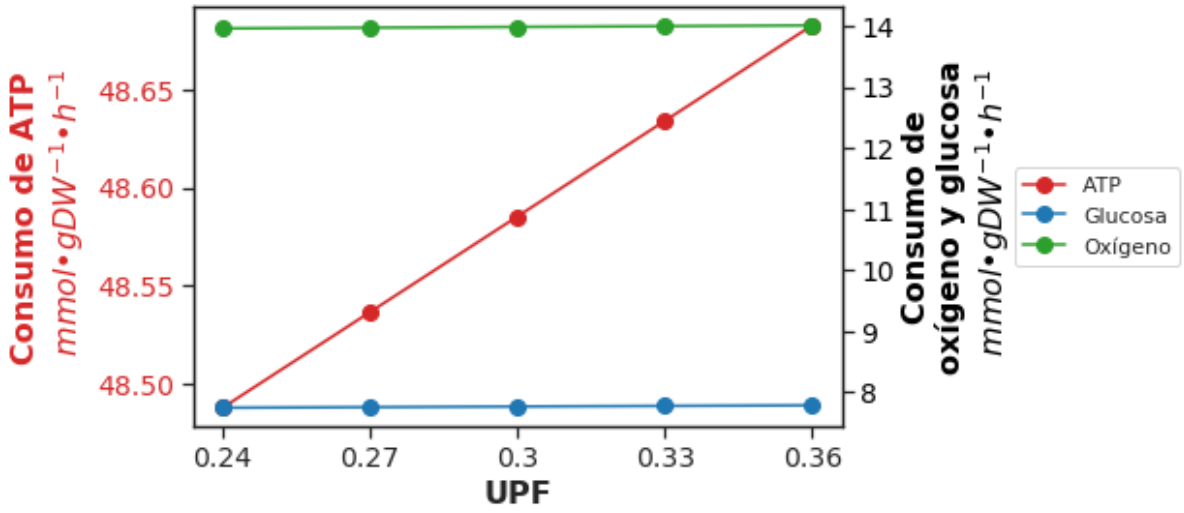

Figura 23. Variaciones en el flujo de consumo total de ATP en diferentes valores de UPF, con flujos de consumo de oxígeno y glucosa fijados a un valor, y teniendo como objetivo la producción de GFP.

Consumo celular

Restamos el el flujo de consumo de ATP de la función ATPM del consumo total para conocer el consumo celular (Figura 24). El costo de ATP aumenta con una razón de cambio de ~1.63x10^{-2} mmol•gDW^{-1}•h^{-1} por porcentaje de UPF. Mientras que el flujo de producción de GFP disminuyó con una razón de cambio de 1.6x10^{-4} mmol•gDW^{-1}•h^{-1} por 1% de UPF.

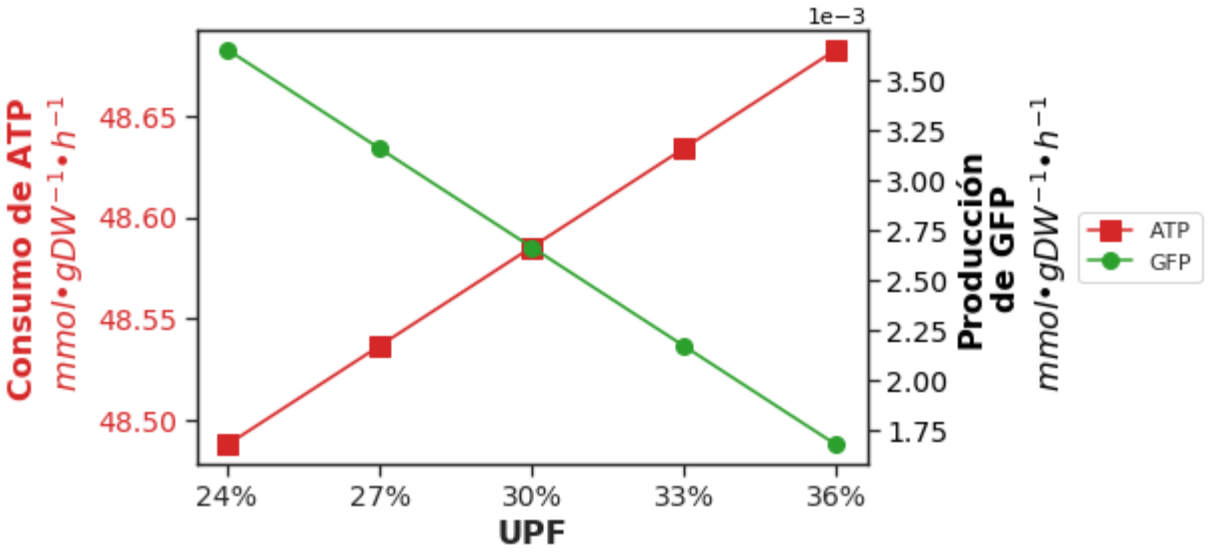

Figura 24. Variaciones en el consumo celular de ATP en diferentes valores de UPF, con flujos de consumo de de oxígeno y glucosa fijados a un valor,, y teniendo como objetivo la producción de GFP.

Cálculo manual de costo de ATP por traducción

Lynch & Marinov, 2015., también hacen el cálculo del costo de la producción de proteínas en bacterias en unidades de hidrólisis de ATP por célula, con la siguiente fórmula (Lynch & Marinov, 2015):

$$N_p L_p \left[(\bar{c}_{AA} - 1) + 5\delta_p t \right]$$

Donde N_p es la abundancia celular de las proteínas, L_p representa la longitud de la proteína en aminoácidos, $\bar{c}_{AA}$ es el costo promedio de síntesis por residuo de aminoácido en una proteína, δ_p es la tasa de degradación de las proteínas en copias por segundo y t es la velocidad específica de crecimiento en horas.

Con esta fórmula y datos para *E. coli* (Lynch & Marinov, 2015) calcularon que, para un gen promedio de 950 bases, el costo para la traducción es de ~4.11x10^6 unidades de hidrólisis de ATP por célula. Esto representa un costo de ~4.2x10^6 unidades de hidrólisis de ATP por célula por Mb para la traducción.

Nuevo cálculo con costos por replicación, transcripción y producción de proteína según el modelo ME

Una vez que conocemos los costos de replicación, transcripción y producción de proteína, podemos calcular los costos para cada uno de los genes eliminados en las cepas minimizadas tomando en cuenta el tamaño del gen y su aportación al proteoma (Figura 25). Las cepas que más recursos en términos de ATP liberaron son la Δ16 y la DGF298, siendo la Δ16 la que más ATP liberó con 2.186 mmol•gDW^{-1}•h^{-1} de ATP lo cual equivale a un ~3% del ATP producido en el modelo ME (Figura 25 a).

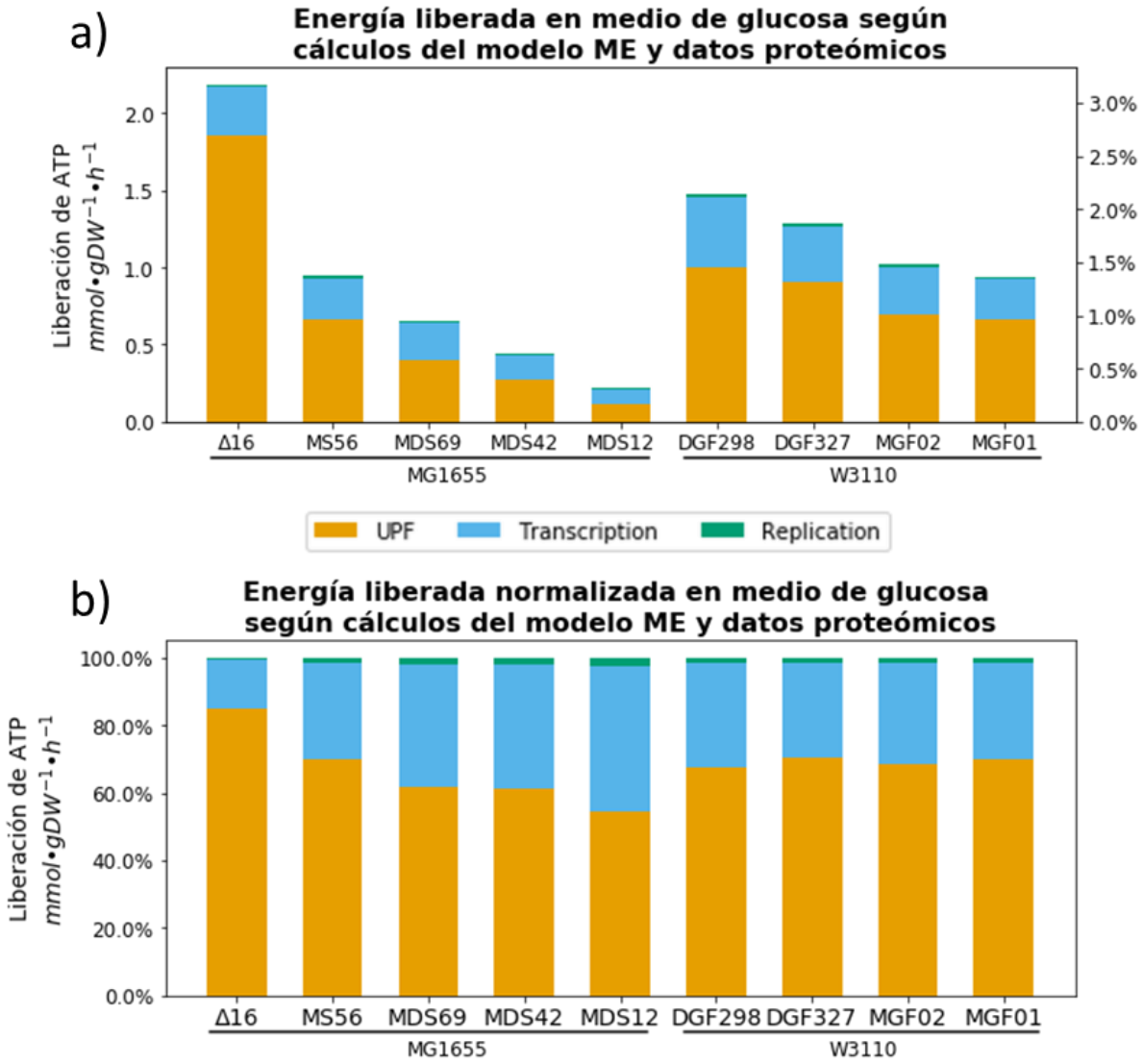

Figura 25. Recursos en términos de ATP liberados en medio de glucosa según cálculos del modelo ME, tamaño del gen y datos proteómicos

En general, la mayor parte del flujo de consumo de ATP proviene de los procesos de traducción (Figura 25 b) que representa entre el 55 y el 85% del total, mientras que la

transcripción y replicación son los menos representando la transcripción entre el 14 y 42% y la replicación entre el 0.8 y 2.5%.

Búsqueda de genes blanco con información proteómica y de adecuación

Para la creación de listas con genes blanco, buscamos aquellos genes cuyas mutantes tuvieran datos de adecuación en las 166 condiciones presentadas en *Fitness Browser* (Price et al., 2018; Wetmore et al., 2015) que además estuvieran dentro del conjunto de datos proteómicos de Schmidt *et al*., 2016. Después filtramos aquellos que tuvieran un efecto negativo en la adecuación y nos quedamos con aquellos que tuvieran ya sea un efecto neutral o un efecto positivo en la adecuación. Finalmente, para cada una de las 22 condiciones de Schmidt *et al*., 2016 obtuvimos los 20 genes que representaban una mayor carga proteómica y buscamos aquellos presentes en las 22 listas o la mayoría de ellas. En este caso obtuvimos dos listas, una donde el requisito era que no hubiera efectos negativos hacia la adecuación (Tabla 24.a) y otra donde el requisito era que tuvieran mínimo un efecto positivo hacia la adecuación (Tabla 24.b).

Tabla 24. Genes compartidos entre el top 20 de cada una de las condiciones.

a) Sin afectación a la adecuación

	19	20	22	21	19
Fructosa					
Succinato					
Galactosa					
Mannosa					
Xylosa					
pH6 glucosa					
42°C glucosa					
Estrés-osmótico glucosa					
Fase estacionaria 3 días					
Fase estacionaria 1 día					
Quimiostato μ=0.12					
Quimiostato μ=0.20					
Quimiostato μ=0.35					
Quimiostato μ=0.5					
Piruvate					
Glicerol					
Glucosamina					
Fumarato					
Acetato					
Glicerol + AA					
LB					
Glucosa					
	1	1	2	1	1

Condiciones

Genes

b) Mejora a la adecuación

	20	21	22	21	20
Fructosa					
Succinato					
Galactosa					
Mannosa					
Xylosa					
pH6 glucosa					
42°C glucosa					
Estrés-osmótico glucosa					
Fase estacionaria 3 días					
Fase estacionaria 1 día					
Quimiostato μ=0.12					
Quimiostato μ=0.20					
Quimiostato μ=0.35					
Quimiostato μ=0.5					
Piruvate					
Glicerol					
Glucosamina					
Fumarato					
Acetato					
Glicerol + AA					
LB					
Glucosa					
	2	1	6	2	3

Condiciones

Genes

Para la lista donde se requería no presentar efectos negativos a la adecuación buscamos aquellos genes que estuvieran compartidos en al menos 19 condiciones. Encontramos que había dos genes compartidos entre todas 22 condiciones, un gen compartido en 21, un gen compartido en 20, un gen compartido en 19 condiciones y otro gen compartido en otras 19 condiciones. Para la selección de genes blanco y su búsqueda en literatura decidimos tomar en cuenta todos estos seis genes y ver su carga proteómica en la condición de glucosa

(Tabla 25). La mitad de estos genes presentan mejoría en algunas condiciones, pero el más alto, *ompT*. se mantiene neutral en las condiciones. En general los seis genes, según información recopilada de la base de datos *Profiling of Escherichia coli Chromosome (PEC)* (Yamazaki et al., 2007), tienen mutantes que están reportadas como no esenciales, además de presentar una morfología celular normal.

Teóricamente, si eliminásemos estos seis genes, obtendríamos una liberación de proteoma del 1.163% tomando en cuenta los valores de proteoma en glucosa.

Tabla 25. Lista de genes blanco en condiciones sin afectación a la adecuación y su carga proteómica en la condición de glucosa.

#	*Gen*	Descripción	fg/célula	%	Neutral	Mejora
29	*ompT*	Proteasa VII de la membrana externa	0.820	0.337	166	0
46	*oppA*	Proteína de enlace del oligopéptido periplásmico	0.581	0.239	165	1
62	*yjgK*	Proteína de biopelícula de toxina-antitoxina TabA	0.443	0.182	166	0
63	*aceB*	Malato sintasa A	0.439	0.181	159	7
92	*argT*	Proteína periplásmica de unión a lisina / arginina / ornitina	0.286	0.118	166	0
103	*rbsB*	Proteína de transporte de ribosa	0.259	0.106	163	3

Para la lista donde debían mostrar una mejora en la adecuación también buscamos aquellos genes que estuvieran compartidos en al menos 19 condiciones. Encontramos que había seis genes compartidos entre todas 22 condiciones, por lo que tomamos estos seis genes y su carga proteómica en la condición de glucosa para generar nuestra lista (Tabla 26). El más alto, *oppA*. presenta mejora en solo una condición, mientras que el gen siguiente, *aceB*, presenta mejora en siete condiciones. En general los seis genes, según información recopilada de la base de datos *Profiling of Escherichia coli Chromosome (PEC)* (Yamazaki et al., 2007), presentan mutantes reportadas como no esenciales, además de presentar una morfología celular normal.

Teóricamente, si eliminásemos estos seis genes, obtendríamos una liberación de proteoma del 0.787% tomando en cuenta los valores de proteoma en glucosa.

Tabla 26. Lista de genes blanco en condiciones con mejora a la adecuación y su carga proteómica en la condición de glucosa.

#	*Gen*	Descripción	fg/célula	%	Neutral	Mejora
46	*oppA*	Proteína de enlace del oligopéptido periplásmico	0.581	0.239	165	1
63	*aceB*	Malato sintasa A	0.439	0.181	159	7
97	*fkpA*	Peptidil-prolil cis-trans isomerasa FkpA	0.272	0.112	164	2
103	*rbsB*	Proteína de transporte de ribosa	0.259	0.106	163	3
107	*gatY*	D-tagatosa-1,6-bisfosfato aldolasa subunidad GatY	0.252	0.104	158	8
222	*ydgA*	Proteína hipotética YdgA	0.110	0.045	165	1

Discusión

Comparación de cepas

Comparación cepas MG1655 y W3110

En general, la cepa Δ16 tuvo una mayor liberación predicha de proteoma (Figura 26) y energética (Figura 11 y Figura 25). Sin embargo, hay que tener en cuenta que los enfoques de minimización entre la Δ16 y las demás cepas eran diferentes, ya que el proyecto del que surgió la cepa Δ16 tenía como objetivo una cepa con la menor cantidad de genoma, mientras que las demás buscaban cepas estables para producción. A pesar de que la cepa Δ16 tenía una mayor cantidad de recursos disponibles, también tenía fenotipos reportados que afectaban gravemente a la célula como lo son crecimiento deficiente y localización anormal de nucleoides. Por lo que es importante realizar las mutaciones de una manera racional para lograr obtener una cepa estable que nos sea útil en la producción de metabolitos de interés. Por ejemplo, al consultar el impacto de dichas mutación en bases de datos como PEC.

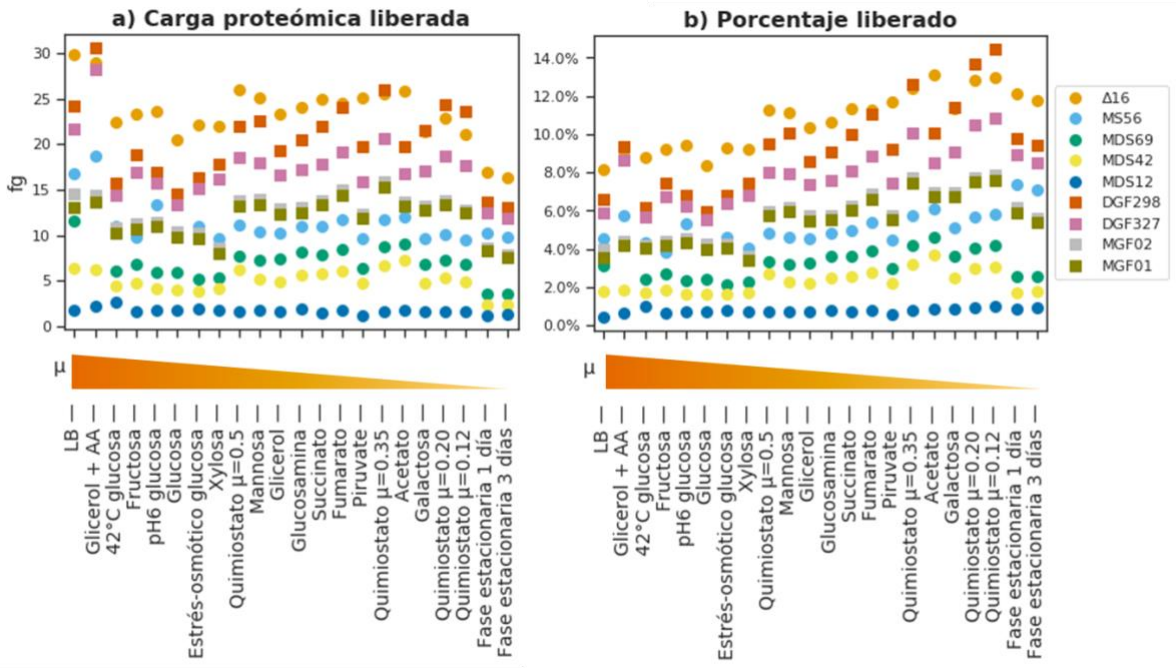

Figura 26. Proteoma estimado liberado en las cepas MDS12, MDS42, MDS69, MS56 y Δ16. a) proteoma estimado liberado en femtogramos por célula. b) Proteoma estimado liberado en porcentaje.

Comparación cepas genoma y proteoma

La relación entre la eliminación de genoma y proteoma no es lineal. Por ejemplo, al graficar esta relación para todas las cepas en el ambiente de glucosa, la cepa Δ16 tiene una menor cantidad de genoma eliminado que la DGF-298.

Sin embargo, tiene el doble de proteoma estimado liberado (Figura 27). Por lo que una mayor eliminación de genoma no garantiza una eliminación equivalente de proteoma.

En el caso contrario, a la PFC se le eliminaron solo tres factores de transcripción de su genoma y, en teoría, liberó un 0.5% de proteoma (Lastiri-Pancardo et al., 2020), mostrando que se puede obtener una mayor liberación de proteoma con una cantidad mínima de eliminación de genes.

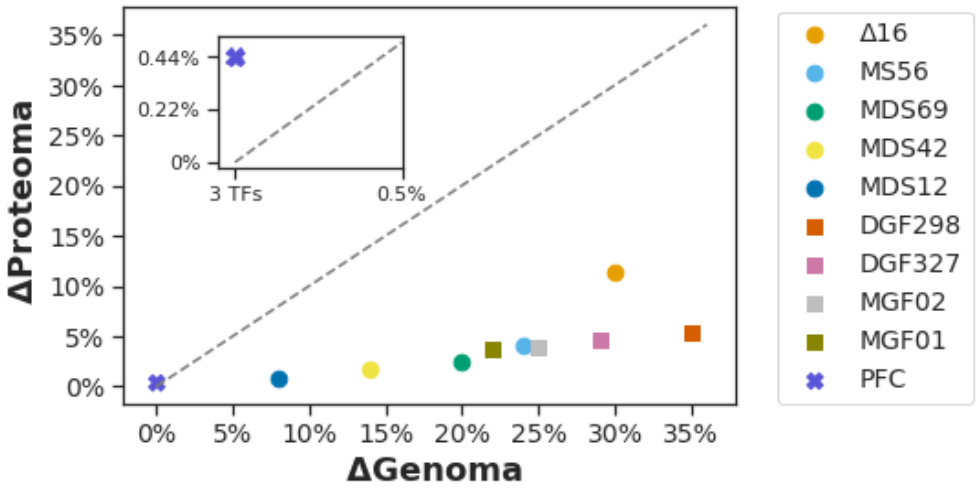

Figura 27. Relación entre eliminación de genoma y proteoma.

Distribución de proteoma y distribución de recursos en términos de ATP

Hay una menor cantidad de genes cuyas proteínas codificadas contribuyen significativamente al proteoma. Por ende, al realizar una eliminación enfocada en minimizar genoma, sin tomar en cuenta la carga al proteoma o de recursos en términos de ATP, es poco probable que eliminemos genes con alta carga proteómica o alto costo de recursos en términos de ATP.

Comparación de costos

Manual

Los costos obtenidos a partir de las simulaciones del modelo ME de replicación y transcripción son bastante parecidos para los cálculos de Lynch & Marinov, 2015 para *E. coli* (Lynch & Marinov, 2015). Sin embargo, son mínimos al compararlos con los de traducción de un gen promedio al ser un orden de magnitud más pequeños. Por lo que, en términos energéticos, es más caro producir una proteína que producir un transcrito o replicar un gen. Sin embargo, nos hace falta hacer el cálculo para comparar las magnitudes de costo para la traducción entre Lynch & Marinov,2015 y el modelo ME.

Modelo ME

Tabla 27. Costos de a partir de las simulaciones del modelo ME.

mmol•gDW^{-1}•h^{-1} de **ATP** por **Mb**			mmol•gDW^{-1}•h^{-1} de **ATP** por 1% de **UPF**		mmol•gDW^{-1}•h^{-1} de **GFP** por 1% **UPF**	mmol•gDW^{-1}•h^{-1} de **ATP** por mmol•gDW^{-1}•h^{-1} de **GFP**
Replicación	Transcripción / Flujos globales	Transcripción / Flujos de transcription	UPF / objetivo: ATPM	UPF / objetivo: GFP	UPF / objetivo: GFP	GFP / objetivo: ATPM
0.017	14.12	0.29	0.16	0.016	0.00016	16.45

Gen promedio: 950 bases y 0.103 fg de proteoma en condición de glucosa

mmol•gDW^{-1}•h^{-1} de **ATP** por **Mb**			mmol•gDW^{-1}•h^{-1} de **ATP** por **Mb**		mmol•gDW^{-1}•h^{-1} de **GFP** por **Mb**	mmol•gDW^{-1}•h^{-1} de **ATP** por 7.13x10^{-3} mmol•gDW^{-1}•h^{-1} de **GFP**
Replicación	Transcripción / Flujos globales	Transcripción / Flujos de transcription	UPF / objetivo: ATPM	UPF / objetivo: GFP	UPF / objetivo: GFP	GFP / objetivo: ATPM
0.017	14.12	0.29	7.36	0.73	0.0071	1.17

En este caso, en las comparaciones de mmol•gDW^{-1}•h^{-1} de ATP por Mb, la transcripción es un proceso más costoso que la replicación (Tabla 27). Mientras que un aumento de 1% de UPF, tiene un costo de 1.65x10^{-1} mmol•gDW^{-1}•h^{-1} de ATP. Sin embargo, no hay equivalencia exacta de un aumento de 1% de UPF, pero podemos hacer un aproximado en el caso de el gen de tamaño promedio (~1 kb) y con aportación promedio al proteoma en la condición de glucosa (0.103 fgPC) según los datos de Schmidt *et al*., 2016 (Schmidt et al., 2016).

Este gen promedio tendría un tamaño de 950 bases (como en el análisis de Lynch y Marinov, 2015) y un porcentaje de proteoma en la condición de glucosa de 4.24x10^{-2}%. Esto querría decir que necesitaríamos ~1052 veces este gen para obtener una Mb del mismo, lo cual representaría el 44.60% del proteoma en la condición de glucosa. Este aumento de proteoma representaría un costo de ~7.35 mmol•gDW^{-1}•h^{-1} de ATP por aumentar 1 Mb de gen promedio, mientras que la replicación y transcripción tendrían costos de ~1.68x10^{-2} y de ~2.87x10^{-1} mmol•gDW^{-1}•h^{-1} de ATP. Por lo que, para un gen promedio el costo de producción de una proteína es mucho mayor al costo de transcripción y replicación.

En los casos donde se agregó al modelo la producción de la proteína GFP, tenemos dos casos; Uno donde se varió la UPF y se puso como objetivo la producción de GFP y otro donde se varió la producción GFP y el objetivo fue ATPM. El primer caso nos representaría los cambios si tenemos producción de una proteína recombinante y liberamos o aumentamos el proteoma, mientras que el segundo caso son los costos energéticos si solo forzáramos un aumento en la producción.

En el primer caso, el costo de ATP aumenta con una razón de cambio de ~1.63x10^{-2} mmol•gDW^{-1}•h^{-1} por 1% de UPF. Mientras que el costo de producción de GFP aumentó con una razón de cambio de -1.6x10^{-4} mmol•gDW^{-1}•h^{-1} por 1% de UPF.

De nuevo utilizando el ejemplo de un gen promedio, donde una Mb equivale a un aumento de ~44.6% de proteoma en la condición de glucosa, aumentar una Mb nos costaría 7.27x10^{-1} mmol•gDW^{-1}•h^{-1} de ATP y disminuiría la producción de GFP en 7.13x10^{-3} mmol•gDW^{-1}•h^{-1}. Por el lado contrario, liberar esa Mb aumentaría la producción de GFP en la misma medida.

En el segundo caso tenemos un costo de ~16.453 mmol•gDW^{-1}•h^{-1} de ATP por mmol•gDW^{-1}•h^{-1} de GFP. Por lo tanto, para lograr un aumento de GFP de 7.13x10^{-3} mmol•gDW^{-1}•h^{-1} (como el que obtuvimos en el primer caso), necesitaríamos 1.17x10^{-1} mmol•gDW^{-1}•h^{-1} de ATP. Mientras que en el primer caso para mejorar la producción de GFP en 7.13x10^{-3} mmol•gDW^{-1}•h^{-1} nos costó liberar 7.27x10^{-1} mmol•gDW^{-1}•h^{-1} de ATP, en este caso necesitamos invertir 1.17x10^{-1} mmol•gDW^{-1}•h^{-1} de ATP para lograr el mismo aumento. En ordenes de magnitud, es equivalente el eliminar Mb para mejorar la producción de GFP a forzar su flujo.

En resumen, las magnitudes de costos entre el modelo ME y los cálculos de Lynch & Marinov son bastante similares, sobre todo en términos de la etapa de traducción, ya que representa la mayor parte del costo de un gen promedio (Figura 28). Esto apoya la idea de que un blanco para la liberación de recursos es el eliminar proteoma.

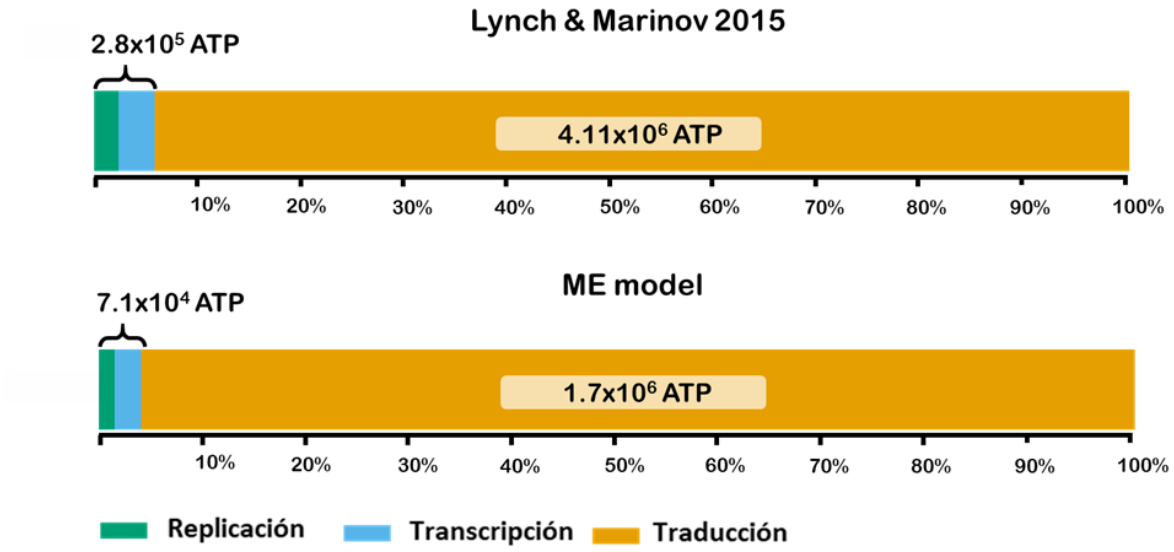

Figura 28. Diagrama de comparación de costos de un gen promedio en Escherichia coli. Elaboración propia con base en la figura suplementaria 1 de Lynch y Marinov, 2015.

Adecuación

En términos energéticos y proteómicos, podemos observar que los genes con costos energéticos más altos suelen tener un costo en términos de adecuación y ser considerados como esenciales o esenciales en ciertas condiciones. Debido a esto, no podemos solamente eliminar aquellos genes que tengan una mayor liberación de recursos, sino también pensar en qué afectaciones conllevan y si éstas afectan el uso que planemos darle a nuestra bacteria.

Por parte de las cepas minimizadas, si no tomamos en cuenta su carga proteómica, podemos organizar la información según su aportación a la adecuación para ver qué tanta afectación negativa presentan. Para esto podemos hacer dos clasificaciones. La primera es ver cuántos de los genes no aparecen en las listas de *Fitness Browser* ya que, por cómo se obtuvieron los datos, aquellos genes no presentes los tomaremos como “posibles

esenciales" ya que no fue posible obtener una mutante probablemente porque eran esenciales. Por otra parte, de los genes que sí tenemos información en *Fitness Browser,* podemos ver cuántos de ellos tienen aportaciones negativas.

Tabla 28. Porcentaje de los genes eliminados que pueden mostrar una afectación negativa a la adecuación.

Cepa	Posibles esenciales %	Esenciales en al menos 1 condición (Fitness Browser) %	Afectacion negativa en al menos 1 condición %	Afectación negativa en al menos 20% de las condiciones %	Afectación negativa en al menos 50% de las condiciones %
Δ16	30.427	5.726	34.017	4.103	2.735
MS56	14.030	5.075	30.348	0.896	0.398
MDS69	12.940	5.797	31.781	1.346	0.311
MDS42	16.690	6.621	34.897	1.241	0.414
MDS12	17.494	2.128	31.206	1.182	0.236
DGF298	15.981	5.038	29.184	0.926	0.290
DGF327	13.147	4.382	29.167	0.934	0.287
MGF02	10.799	4.082	29.082	0.935	0.340
MGF01	11.304	3.961	28.599	0.870	0.290

Podemos ver que la menor cantidad de posibles esenciales eliminados, es de ~10% (Tabla 28) y que la cepa más afectada es la Δ16 con un ~30%. Siendo poco menos del doble de la segunda cepa más afectada (la MDS12 con ~17.5%). Estos porcentajes pueden ayudarnos a explicar el porqué vemos fenotipos con afectaciones negativas en la adecuación en las cepas minimizadas además de mostrarnos la diferencia entre los enfoques de minimización de genoma. Por ejemplo, la cepa Δ16 que fue creada con el objetivo de lograr el mínimo de genoma es aquella la que tiene que mayor porcentaje de de eliminación de genes posiblemente esenciales, así como afectaciones negativas según los datos de *Fitness Browser,* siendo la más alta por mucho tanto en los posibles esenciales como en las afectaciones negativas en al menos 20% y 50% de las condiciones. Esto a diferencia de las cepas que fueron creadas para producción. Sin embargo, todas siguen mostrando un porcentaje importante de posibles esenciales y afectaciones negativas a la adecuación, por lo que además de una métrica de proteoma o genoma, sería conveniente tener en cuenta las afectaciones de la adecuación.

Una vez elegida la métrica de proteoma como minimización ya que la mayor parte del costo energético viene de la producción de proteína y junto con información de afectación a la adecuación de diferentes genes, podemos seleccionar aquellos cuyas proteínas codificadas tienen una mayor aportación al proteoma y menos afectan en la adecuación o que incluso presenten mejorías según bases de datos como *Fitness Browser*. De esta manera, tendremos una mayor liberación de recursos minimizando efectos negativos en la célula y que nos permitan mejorar la producción de las cepas. Por ejemplo, la lista de blancos que presentan mejoría en la adecuación tiene una liberación de proteoma teórica comparable a la cepa MDS12 con la eliminación de solo seis genes. Por otro lado, los que vienen de la lista donde la mutación se mantiene neutral, presentan una liberación de proteoma teórica más alta que la cepa MDS12.

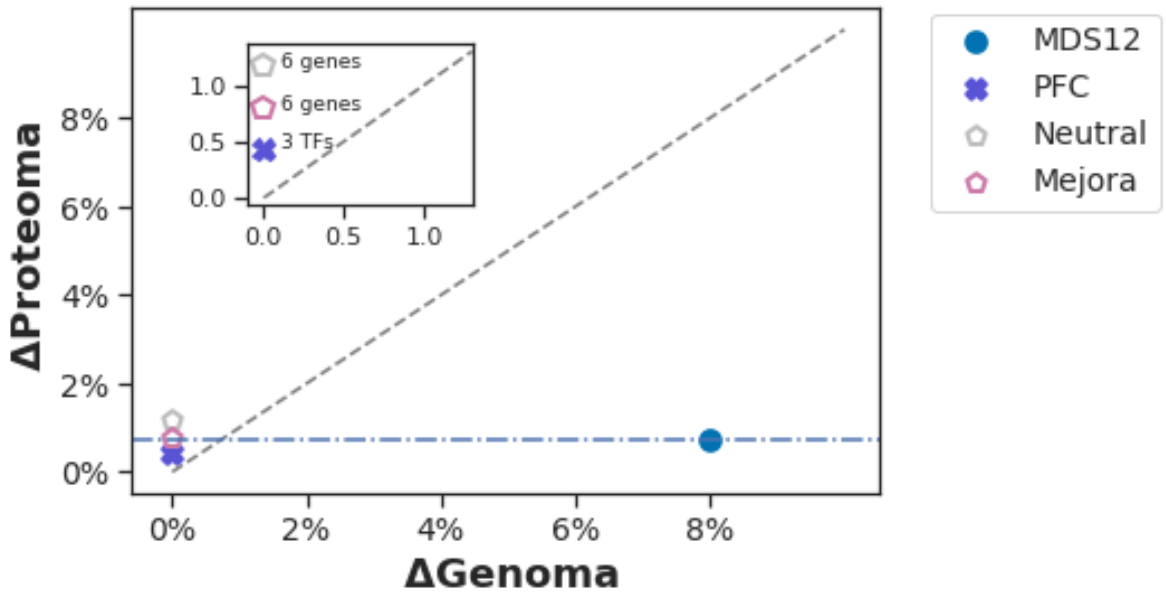

Figura 29. Comparación de la predicción de liberación de proteoma de cepas minimizadas y de genes blanco obtenidos en este estudio.

Conclusiones

La eliminación de una mayor cantidad de genoma no indica una eliminación equivalente de proteoma y puede comprometer su supervivencia en diferentes ambientes. En general, la comparación entre las cepas MG1655 y W3110 reveló que la cepa Δ16 tuvo una mayor liberación predicha de proteoma y energía. Es importante destacar que aunque la cepa Δ16 tenía más recursos disponibles, también presentaba algunos fenotipos graves, como un crecimiento deficiente y una localización anormal de nucleoides. Por tanto, es crucial llevar a cabo las mutaciones de una manera racional para obtener una cepa estable que sea útil en la producción de metabolitos de interés.

En cuanto a la relación entre la eliminación de genoma y proteoma, se observó que no es lineal, y una mayor eliminación de genoma no garantiza una eliminación equivalente de proteoma. La cepa Δ16 tiene una menor cantidad de genoma eliminado que la DGF-298, pero presenta el doble de proteoma estimado liberado. Además, la cepa PFC solo eliminó tres factores de transcripción de su genoma y liberó un 0,5% de proteoma, lo que muestra que se puede obtener una mayor liberación de proteoma con una cantidad mínima de mutaciones.

También se encontró que hay una menor cantidad de genes que tienen una mayor contribución al proteoma, por lo que al hacer una eliminación enfocada en el genoma, sin tomar en cuenta la carga proteómica o de recursos en términos de ATP, es poco probable que se eliminen genes con alta carga proteómica o alto costo de recursos en términos de ATP.

Como es indicado por las simulaciones del modelo ME y las estimaciones de proteoma de acuerdo a los valores de Schmidt *et al*., 2016, en general es más costoso en términos de ATP la producción de proteínas que la replicación o la transcripción. En cuanto a los costos, se encontró que, en términos energéticos, es más costoso producir una proteína que producir un transcrito o replicar un gen. Los costos obtenidos a partir de las simulaciones del modelo ME de replicación y transcripción son bastante similares a los cálculos de Lynch y Marinov (2015) para *E. coli*. Sin embargo, estos son mínimos al compararlos con los de traducción de un gen promedio, ya que son un orden de magnitud más pequeños, lo que puede tener implicaciones importantes en la producción de proteínas recombinantes y la ingeniería metabólica.

Además, se demostró que la eliminación de genes basada en la carga proteómica o de recursos en términos de ATP es más efectiva que la eliminación enfocada en el genoma. El enfoque de minimización de proteoma puede permitirnos liberar una gran cantidad de recursos con pocas mutaciones que además puede ser comparable a otras cepas minimizadas que cuentan con una mayor cantidad de genoma eliminado.

En resumen, los resultados obtenidos en este estudio pueden ayudar a comprender mejor la relación entre la eliminación del genoma y la liberación de proteoma y energía, así como a optimizar las mutaciones de una manera racional para obtener cepas estables y útiles en la producción de metabolitos de interés. Enfoques de minimización como el presentado en

este estudio pueden llegar a permitirnos crear cepas de producción eficientes y estables con . Además, el tomar en cuenta datos de adecuación junto con el enfoque de minimización nos puede permitir liberar una gran cantidad de recursos y que sea menos probable el eliminar genes esenciales o genes condicionalmente esenciales que puedan ser de interés.

Perspectivas

Para probar los genes blanco de minimización predichos podríamos utilizar la misma cepa parental que PFC. Elegir tres genes blanco ideales para tres condiciones diferentes (por ejemplo, LB, fructosa y glucosa). Y una vez que tengamos nuestras tres cepas mutantes, hacer el mismo análisis que Lastiri-Pancardo et al., 2019 para la producción de proteína recombinante en PFC. De manera que podamos analizar si hubo reducción en el proteoma y cómo afecta esto en la producción de proteína recombinante comparada con la cepa parental en las diferentes condiciones.

También sería interesante mejorar la información sobre la degradación del ARNm al modelo, ya que en también es un elemento clave en cómo se utilizan los recursos celulares (Roux et al., 2022). Además se ha demostrado que es un blanco que puede mejorar la producción de proteínas sin competir por los recursos celulares del organismo huesped (Mao & Inouye, 2012; Venturelli et al., 2017; Wu et al., 2020).

Referencias

Acevedo-Rocha, C. G., Fang, G., Schmidt, M., Ussery, D. W., & Danchin, A. (2013). From essential to persistent genes: A functional approach to constructing synthetic life. *Trends in Genetics*, *29*(5), 273–279. https://doi.org/10.1016/j.tig.2012.11.001

Aebersold, R., & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, *422*(6928), 198–207. https://doi.org/10.1038/nature01511

Agren, R., Liu, L., Shoaie, S., Vongsangnak, W., Nookaew, I., & Nielsen, J. (2013). The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for Penicillium chrysogenum. *PLoS Computational Biology*, *9*(3). https://doi.org/10.1371/journal.pcbi.1002980

Åkesson, M., Förster, J., & Nielsen, J. (2004). Integration of gene expression data into genome-scale metabolic models. *Metabolic Engineering*, *6*(4), 285–293. https://doi.org/10.1016/j.ymben.2003.12.002

Baart, G. J. E., & Martens, D. E. (2012). Genome-scale metabolic models: Reconstruction and analysis. *Methods in Molecular Biology*, *799*, 107–126. https://doi.org/10.1007/978-1-61779-346-2_7

Balikó, G., Vernyik, V., Karcagi, I., Györfy, Z., Draskovits, G., Fehér, T., & Pósfai, G. (2018). Rational efforts to streamline the escherichia coli genome. In *Synthetic Biology: Parts, Devices and Applications* (pp. 49–80). https://doi.org/10.1002/9783527688104.ch4

Becker, S. A., & Palsson, B. O. (2008). Context-specific metabolic networks are consistent with experiments. *PLoS Computational Biology*, *4*(5). https://doi.org/10.1371/journal.pcbi.1000082

Beitz, A. M., Oakes, C. G., & Galloway, K. E. (2022). Synthetic gene circuits as tools for drug discovery. *Trends in Biotechnology*, *40*(2), 210–225. https://doi.org/10.1016/j.tibtech.2021.06.007

Bienick, M. S., Young, K. W., Klesmith, J. R., Detwiler, E. E., Tomek, K. J., & Whitehead, T. A. (2014). The interrelationship between promoter strength, gene expression, and growth rate. *PLoS ONE*, *9*(10). https://doi.org/10.1371/journal.pone.0109105

Bond, P. J., Holyoake, J., Ivetac, A., Khalid, S., & Sansom, M. S. P. (2007). Coarse-grained molecular dynamics simulations of membrane proteins and peptides. *Journal of Structural Biology*, *157*(3), 593–605. https://doi.org/10.1016/j.jsb.2006.10.004

Bordbar, A., Monk, J. M., King, Z. A., & Palsson, B. O. (2014). Constraint-based models predict metabolic and associated cellular functions. *Nature Reviews Genetics*, *15*(2), 107–120. https://doi.org/10.1038/nrg3643

Bremer, H., & Dennis, P. P. (1987). Modulation of Chemical Composition and Other Parameters of the Cell by Growth Rate. *Escherichia Coli and Salmonella: Cellular and Molecular Biology*, *2*(122), 1527–1542. http://ctbp.ucsd.edu/qbio/beemer96.pdf

Brenner, S. (2010). Sequences and consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1537), 207–212. https://doi.org/10.1098/rstb.2009.0221

Broussard, T. C., Price, A. E., Laborde, S. M., & Waldrop, G. L. (2013). Complex formation and regulation of escherichia coli acetyl-CoA carboxylase. *Biochemistry*, *52*(19), 3346–

3357. https://doi.org/10.1021/bi4000707

Brown, G. R., Hem, V., Katz, K. S., Ovetsky, M., Wallin, C., Ermolaeva, O., Tolstoy, I., Tatusova, T., Pruitt, K. D., Maglott, D. R., & Murphy, T. D. (2015). Gene: A gene-centered information resource at NCBI. *Nucleic Acids Research*, *43*(D1), D36–D42. https://doi.org/10.1093/nar/gku1055

Carrera, J., Estrela, R., Luo, J., Rai, N., Tsoukalas, A., & Tagkopoulos, I. (2014). An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of E scherichia coli . *Molecular Systems Biology*, *10*(7). https://doi.org/10.15252/msb.20145108

Chen, C., Le, H., & Goudar, C. T. (2016). Integration of systems biology in cell line and process development for biopharmaceutical manufacturing. *Biochemical Engineering Journal*, *107*, 11–17. https://doi.org/10.1016/j.bej.2015.11.013

Clarke, L., & Kitney, R. (2020). Developing synthetic biology for industrial biotechnology applications. *Biochemical Society Transactions*, *48*(1), 113–122. https://doi.org/10.1042/BST20190349

Colijn, C., Brandes, A., Zucker, J., Lun, D. S., Weiner, B., Farhat, M. R., Cheng, T. Y., Moody, D. B., Murray, M., & Galagan, J. E. (2009). Interpreting expression data with metabolic flux models: Predicting Mycobacterium tuberculosis mycolic acid production. *PLoS Computational Biology*, *5*(8). https://doi.org/10.1371/journal.pcbi.1000489

Cronan, J. E., & Waldrop, G. L. (2002). Multi-subunit acetyl-CoA carboxylases. *Progress in Lipid Research*, *41*(5), 407–435. https://doi.org/10.1016/S0163-7827(02)00007-3

Darlington, A. P. S., Kim, J., Jiménez, J. I., & Bates, D. G. (2018). Dynamic allocation of orthogonal ribosomes facilitates uncoupling of co-expressed genes. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-02898-6

Davison, P. F. (1966). Molecular Biology: Control of Macromolecular Synthesis: A Study of DNA, RNA, and Protein Synthesis in Bacteria . By Ole Maaløe and Niels O. Kjeldgaard (Benjamin, New York, 1966. 296 pp., illus. $12.50. . *Science*, *154*(3753), 1159–1159. https://doi.org/10.1126/science.154.3753.1159.a

De Godoy, L. M. F., Olsen, J. V., Cox, J., Nielsen, M. L., Hubner, N. C., Fröhlich, F., Walther, T. C., & Mann, M. (2008). Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, *455*(7217), 1251–1254. https://doi.org/10.1038/nature07341

De Jong, H., & Giordano, N. (2020). *Practical Exercises: Flux Balance Analysis using the COnstraint-Based Reconstruction and Analysis (COBRA) Toolbox 1 The COBRA toolbox 1.1 Principle*. http://sbml.org

Dekel, E., & Alon, U. (2005). Optimality and evolutionary tuning of the expression level of a protein. *Nature*, *436*(7050), 588–592. https://doi.org/10.1038/nature03842

Dennis, P. P., & Bremer, H. (1974). Macromolecular composition during steady state growth of Escherichia coli B/r. *Journal of Bacteriology*, *119*(1), 270–281. https://doi.org/10.1128/jb.119.1.270-281.1974

Doan, D. T., Hoang, M. D., Heins, A. L., & Kremling, A. (2022). Applications of Coarse-Grained Models in Metabolic Engineering. *Frontiers in Molecular Biosciences*, *9*. https://doi.org/10.3389/fmolb.2022.806213

Durot, M., Bourguignon, P. Y., & Schachter, V. (2009). Genome-scale models of bacterial metabolism: Reconstruction and applications. *FEMS Microbiology Reviews*, *33*(1), 164–190. https://doi.org/10.1111/j.1574-6976.2008.00146.x

Ebrahim, A., Brunk, E., Tan, J., O'Brien, E. J., Kim, D., Szubin, R., Lerman, J. A., Lechner, A., Sastry, A., Bordbar, A., Feist, A. M., & Palsson, B. O. (2016). Multi-omic data integration enables discovery of hidden biological regularities. *Nature Communications*, *7*, 1–9. https://doi.org/10.1038/ncomms13091

Ebrahim, A., Lerman, J. A., Palsson, B. O., & Hyduke, D. R. (2013). COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Systems Biology*, *7*. https://doi.org/10.1186/1752-0509-7-74

Escherichia coli K-12 substr. MG1655 All-Genes. (n.d.). Retrieved October 20, 2023, from https://biocyc.org/ECOLI/NEW-IMAGE?type=ECOCYC-CLASS&object=All-Genes

Esvelt, K. M., & Wang, H. H. (2013). Genome-scale engineering for systems and synthetic biology. *Molecular Systems Biology*, *9*. https://doi.org/10.1038/msb.2012.66

Feist, A. M., & Palsson, B. (2008). The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nature Biotechnology*, *26*(6), 659–667. https://doi.org/10.1038/nbt1401

Fischer, E., & Sauer, U. (2003). A Novel Metabolic Cycle Catalyzes Glucose Oxidation and Anaplerosis in Hungry Escherichia coli. *Journal of Biological Chemistry*, *278*(47), 46446–46451. https://doi.org/10.1074/jbc.M307968200

Gelius-Dietrich, G., Desouki, A. A., Fritzemeier, C. J., & Lercher, M. J. (2013). Sybil - Efficient constraint-based modelling in R. *BMC Systems Biology*, *7*. https://doi.org/10.1186/1752-0509-7-125

Gibson, D. G., Glass, J. I., Lartigue, C., Noskov, V. N., Chuang, R. Y., Algire, M. A., Benders, G. A., Montague, M. G., Ma, L., Moodie, M. M., Merryman, C., Vashee, S., Krishnakumar, R., Assad-Garcia, N., Andrews-Pfannkoch, C., Denisova, E. A., Young, L., Qi, Z. N., Segall-Shapiro, T. H., … Venter, J. C. (2010). Creation of a bacterial cell controlled by a chemically synthesized genome. *Science*, *329*(5987), 52–56. https://doi.org/10.1126/science.1190719

Gonzalez, J. C., Banerjee, R. V., Huang, S., Sumner, J. S., & Matthews, R. G. (1992). Comparison of Cobalamin-independent and Cobalamin-Dependent Methionine Synthases from Escherichia coli : Two Solutions to the Same Chemical Problem. *Biochemistry*, *31*(26), 6045–6056. https://doi.org/10.1021/bi00141a013

Gudmundsson, S., & Nogales, J. (2021). Recent advances in model-assisted metabolic engineering. *Current Opinion in Systems Biology*, *28*. https://doi.org/10.1016/j.coisb.2021.100392

Hafner, J., Payne, J., MohammadiPeyhani, H., Hatzimanikatis, V., & Smolke, C. (2021). A computational workflow for the expansion of heterologous biosynthetic pathways to natural product derivatives. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-22022-5

Hall, S. (2004). Electronic theses and dissertations: Electronic theses and dissertations. In *Science and Technology Libraries* (Vol. 22, Issues 3–4). https://doi.org/10.1300/J122v22n03_06

Hashimoto, M., Ichimura, T., Mizoguchi, H., Tanaka, K., Fujimitsu, K., Keyamura, K., Ote, T., Yamakawa, T., Yamazaki, Y., Mori, H., Katayama, T., & Kato, J. I. (2005). Cell size and nucleoid organization of engineered Escherichia coli cells with a reduced genome. *Molecular Microbiology*, *55*(1), 137–149. https://doi.org/10.1111/j.1365-2958.2004.04386.x

Hirokawa, Y., Kawano, H., Tanaka-Masuda, K., Nakamura, N., Nakagawa, A., Ito, M., Mori, H., Oshima, T., & Ogasawara, N. (2013). Genetic manipulations restored the growth fitness of reduced-genome Escherichia coli. *Journal of Bioscience and Bioengineering*, *116*(1), 52–58. https://doi.org/10.1016/j.jbiosc.2013.01.010

Hui, S., Silverman, J. M., Chen, S. S., Erickson, D. W., Basan, M., Wang, J., Hwa, T., & Williamson, J. R. (2015). Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria. *Molecular Systems Biology*, *11*(2), e784–e784. https://doi.org/10.15252/msb.20145697

Hutchison, C. A., Chuang, R. Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., Ellisman, M. H., Gill, J., Kannan, K., Karas, B. J., Ma, L., Pelletier, J. F., Qi, Z. Q., Richter, R. A., Strychalski, E. A., Sun, L., Suzuki, Y., Tsvetanova, B., Wise, K. S., Smith, H. O., … Venter, J. C. (2016). Design and synthesis of a minimal bacterial genome. *Science*, *351*(6280). https://doi.org/10.1126/science.aad6253

Ideker, T., Galitski, T., & Hood, L. (2001). A new approach to decoding life: Systems biology. *Annual Review of Genomics and Human Genetics*, *2*, 343–372. https://doi.org/10.1146/annurev.genom.2.1.343

Iizuka, K. (2016). *A novel approach to dynamic flux balance analysis that accounts for the dynamic transfer of information by internal metabolites Kazuki Iizuka University of York Biology December 2016* (Issue December). University of York.

Janssen, D. B., Dinkla, I. J. T., Poelarends, G. J., & Terpstra, P. (2005). Bacterial degradation of xenobiotic compounds: Evolution and distribution of novel enzyme activities. *Environmental Microbiology*, *7*(12), 1868–1882. https://doi.org/10.1111/j.1462-2920.2005.00966.x

Juhas, M., Eberl, L., & Glass, J. I. (2011). Essence of life: Essential genes of minimal genomes. In *Trends in Cell Biology* (Vol. 21, Issue 10, pp. 562–568). https://doi.org/10.1016/j.tcb.2011.07.005

Karcagi, I., Draskovits, G., Umenhoffer, K., Fekete, G., Kovács, K., Méhi, O., Balikó, G., Szappanos, B., Györfy, Z., Fehér, T., Bogos, B., Blattner, F. R., Pál, C., Pósfai, G., & Papp, B. (2016). Indispensability of Horizontally Transferred Genes and Its Impact on Bacterial Genome Streamlining. *Molecular Biology and Evolution*, *33*(5), 1257–1269. https://doi.org/10.1093/molbev/msw009

Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Collado-Vides, J., Paley, S. M., Pellegrini-Toole, A., Bonavides, C., & Gama-Castro, S. (2002). The EcoCyc database. *Nucleic Acids Research*, *30*(1), 56–58. https://doi.org/10.1093/nar/30.1.56

Keseler, I. M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martínez, C., Caspi, R., Fulcher, C., Gama-Castro, S., Kothari, A., Krummenacker, M., Latendresse, M., Muñiz-Rascado, L., Ong, Q., Paley, S., Peralta-Gil, M., Subhraveti, P., Velázquez-Ramírez, D. A., Weaver, D., Collado-Vides, J., … Karp, P. D. (2017). The EcoCyc database: Reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Research*, *45*(D1), D543–D550. https://doi.org/10.1093/nar/gkw1003

King, Z. A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N. E., & Palsson, B. O. (2015). Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. *PLoS Computational Biology*, *11*(8). https://doi.org/10.1371/journal.pcbi.1004321

Kirschner, M. W. (2005). The meaning of systems biology. *Cell*, *121*(4), 503–504. https://doi.org/10.1016/j.cell.2005.05.005

Kobayashi, H., Kærn, M., Araki, M., Chung, K., Gardner, T. S., Cantor, C. R., & Collins, J. J. (2004). Programmable cells: Interfacing natural and engineered gene networks. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(22), 8414–8419. https://doi.org/10.1073/pnas.0402940101

Kolisnychenko, V., Plunkett, G., Herring, C. D., Fehér, T., Pósfai, J., Blattner, F. R., & Pósfai, G. (2002). Engineering a reduced Escherichia coli genome. *Genome Research*, *12*(4), 640–647. https://doi.org/10.1101/gr.217202

Kurokawa, M., Seno, S., Matsuda, H., & Ying, B. W. (2016). Correlation between genome reduction and bacterial growth. *DNA Research*, *23*(6), 517–525. https://doi.org/10.1093/dnares/dsw035

Kurokawa, M., & Ying, B. W. (2020). Experimental challenges for reduced genomes: The cell model escherichia coli. *Microorganisms*, *8*(1). https://doi.org/10.3390/microorganisms8010003

LaCroix, R. A., Sandberg, T. E., O'Brien, E. J., Utrilla, J., Ebrahim, A., Guzman, G. I., Szubin, R., Palsson, B. O., & Feist, A. M. (2015). Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. *Applied and Environmental Microbiology*, *81*(1), 17–30. https://doi.org/10.1128/AEM.02246-14

Lastiri-Pancardo, G., Mercado-Hernández, J. S., Kim, J., Jiménez, J. I., & Utrilla, J. (2020). A quantitative method for proteome reallocation using minimal regulatory interventions. *Nature Chemical Biology*, *16*(9), 1026–1033. https://doi.org/10.1038/s41589-020-0593-y

Lerman, J. A., Hyduke, D. R., Latif, H., Portnoy, V. A., Lewis, N. E., Orth, J. D., Schrimpe-Rutledge, A. C., Smith, R. D., Adkins, J. N., Zengler, K., & Palsson, B. O. (2012). In silico method for modelling metabolism and gene product expression at genome scale. *Nature Communications*, *3*(May), 910–929. https://doi.org/10.1038/ncomms1928

Lewis, N. E., Nagarajan, H., & Palsson, B. O. (2012). Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology*, *10*(4), 291–305. https://doi.org/10.1038/nrmicro2737

Li, G. W., Burkhardt, D., Gross, C., & Weissman, J. S. (2014). Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, *157*(3), 624–635. https://doi.org/10.1016/j.cell.2014.02.033

Liu, D., Hoynes-O'Connor, A., & Zhang, F. (2013). Bridging the gap between systems biology and synthetic biology. *Frontiers in Microbiology*, *4*(JUL). https://doi.org/10.3389/fmicb.2013.00211

Liu, J. K., Lloyd, C., Al-Bassam, M. M., Ebrahim, A., Kim, J. N., Olson, C., Aksenov, A., Dorrestein, P., & Zengler, K. (2019). Predicting proteome allocation, overflow

metabolism, and metal requirements in a model acetogen. *PLoS Computational Biology*, *15*(3). https://doi.org/10.1371/journal.pcbi.1006848

Liu, J. K., O'Brien, E. J., Lerman, J. A., Zengler, K., Palsson, B. O., & Feist, A. M. (2014). Reconstruction and modeling protein translocation and compartmentalization in Escherichia coli at the genome-scale. *BMC Systems Biology*, *8*(1). https://doi.org/10.1186/s12918-014-0110-6

Lloyd, C. J., Ebrahim, A., Yang, L., King, Z. A., Catoiu, E., O\textquoterightBrien, E. J., Liu, J. K., & Palsson, B. O. (2017). COBRAme: A Computational Framework for Building and Manipulating Models of Metabolism and Gene Expression. *BioRxiv*. https://doi.org/10.1101/106559

Lynch, M., & Marinov, G. K. (2015). The bioenergetic costs of a gene. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(51), 15690–15695. https://doi.org/10.1073/pnas.1514974112

Mao, L., & Inouye, M. (2012). Use of E. coli for the production of a single protein. *Methods in Molecular Biology*, *899*, 177–185. https://doi.org/10.1007/978-1-61779-921-1_11

McCloskey, D., Palsson, B., & Feist, A. M. (2013). Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Molecular Systems Biology*, *9*(1). https://doi.org/10.1038/msb.2013.18

Mitchell, A., Romano, G. H., Groisman, B., Yona, A., Dekel, E., Kupiec, M., Dahan, O., & Pilpel, Y. (2009). Adaptive prediction of environmental changes by microorganisms. *Nature*, *460*(7252), 220–224. https://doi.org/10.1038/nature08112

Mizoguchi, H., Sawano, Y., Kato, J. I., & Mori, H. (2008). Superpositioning of deletions promotes growth of escherichia coli with a reduced genome. *DNA Research*, *15*(5), 277–284. https://doi.org/10.1093/dnares/dsn019

Monk, J. M., Lloyd, C. J., Brunk, E., Mih, N., Sastry, A., King, Z., Takeuchi, R., Nomura, W., Zhang, Z., Mori, H., Feist, A. M., & Palsson, B. O. (2017). iML1515, a knowledgebase that computes Escherichia coli traits. *Nature Biotechnology*, *35*(10), 904–908. https://doi.org/10.1038/nbt.3956

Nanchen, A., Schicker, A., & Sauer, U. (2006). Nonlinear dependency of intracellular fluxes on growth rate in miniaturized continuous cultures of Escherichia coli. *Applied and Environmental Microbiology*, *72*(2), 1164–1172. https://doi.org/10.1128/AEM.72.2.1164-1172.2006

Neidhardt, F. C. (1999). Bacterial growth: Constant obsession with dN/dt. *Journal of Bacteriology*, *181*(24), 7405–7408. https://doi.org/10.1128/jb.181.24.7405-7408.1999

Nishimura, I., Kurokawa, M., Liu, L., & Ying, B. W. (2017). Coordinated changes in mutation and growth rates induced by genome reduction. *MBio*, *8*(4). https://doi.org/10.1128/mBio.00676-17

O'Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R., & Palsson, B. (2013). Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Molecular Systems Biology*, *9*(693). https://doi.org/10.1038/msb.2013.52

O'Brien, E. J., Monk, J. M., & Palsson, B. O. (2015). Using genome-scale models to predict biological capabilities. *Cell*, *161*(5), 971–987. https://doi.org/10.1016/j.cell.2015.05.019

O'Brien, E. J., Utrilla, J., & Palsson, B. O. (2016). Quantification and Classification of E. coli Proteome Utilization and Unused Protein Costs across Environments. *PLoS Computational Biology*, *12*(6), 1–22. https://doi.org/10.1371/journal.pcbi.1004998

Otero, J. M., & Nielsen, J. (2010). Industrial systems biology. *Biotechnology and Bioengineering*, *105*(3), 439–460. https://doi.org/10.1002/bit.22592

Palsson, B., & Zengler, K. (2010). The challenges of integrating multi-omic data sets. *Nature Chemical Biology*, *6*(11), 787–789. https://doi.org/10.1038/nchembio.462

Park, M. K., Lee, S. H., Yang, K. S., Jung, S. C., Lee, J. H., & Kim, S. C. (2014). Enhancing recombinant protein production with an Escherichia coli host strain lacking insertion sequences. *Applied Microbiology and Biotechnology*, *98*(15), 6701–6713. https://doi.org/10.1007/s00253-014-5739-y

Patra, P., Das, M., Kundu, P., & Ghosh, A. (2021). Recent advances in systems and synthetic biology approaches for developing novel cell-factories in non-conventional yeasts. *Biotechnology Advances*, *47*. https://doi.org/10.1016/j.biotechadv.2021.107695

Peebo, K., Valgepea, K., Maser, A., Nahku, R., Adamberg, K., & Vilu, R. (2015). Proteome reallocation in Escherichia coli with increasing specific growth rate. *Molecular BioSystems*, *11*(4), 1184–1193. https://doi.org/10.1039/c4mb00721b

Peng, R. H., Xiong, A. S., Xue, Y., Fu, X. Y., Gao, F., Zhao, W., Tian, Y. S., & Yao, Q. H. (2008). Microbial biodegradation of polyaromatic hydrocarbons. *FEMS Microbiology Reviews*, *32*(6), 927–955. https://doi.org/10.1111/j.1574-6976.2008.00127.x

Perkins, T. J., & Swain, P. S. (2009). Strategies for cellular decision-making. *Molecular Systems Biology*, *5*. https://doi.org/10.1038/msb.2009.83

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). https://doi.org/10.1128/JB.00147-19

Pirt, S. J. (1965). The maintenance energy of bacteria in growing cultures. *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character. Royal Society (Great Britain)*, *163*(991), 224–231. https://doi.org/10.1098/rspb.1965.0069

Polizzi, K. M. (2013). What is synthetic biology? *Methods in Molecular Biology*, *1073*, 3–6. https://doi.org/10.1007/978-1-62703-625-2_1

Pósfai, G., Plunkett, G., Fehér, T., Frisch, D., Keil, G. M., Umenhoffer, K., Kolisnychenko, V., Stahl, B., Sharma, S. S., De Arruda, M., Burland, V., Harcum, S. W., & Blattner, F. R. (2006). Emergent properties of reduced-genome Escherichia coli. *Science*, *312*(5776), 1044–1046. https://doi.org/10.1126/science.1126439

Price, M. N., Wetmore, K. M., Waters, R. J., Callaghan, M., Ray, J., Liu, H., Kuehl, J. V., Melnyk, R. A., Lamson, J. S., Suh, Y., Carlson, H. K., Esquivel, Z., Sadeeshkumar, H., Chakraborty, R., Zane, G. M., Rubin, B. E., Wall, J. D., Visel, A., Bristow, J., … Deutschbauer, A. M. (2018). Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature*, *557*(7706), 503–509. https://doi.org/10.1038/s41586-018-0124-0

Reed, J. L., & Palsson, B. (2004). Genome-scale in silico models of E. coli have multiple

equivalent phenotypic states: Assessment of correlated reaction subsets that comprise network states. *Genome Research*, *14*(9), 1797–1805. https://doi.org/10.1101/gr.2546004

Richelle, A., David, B., Demaegd, D., Dewerchin, M., Kinet, R., Morreale, A., Portela, R., Zune, Q., & von Stosch, M. (2020). Towards a widespread adoption of metabolic modeling tools in biopharmaceutical industry: a process systems biology engineering perspective. *Npj Systems Biology and Applications*, *6*(1). https://doi.org/10.1038/s41540-020-0127-y

Rittmann, B. E. (2008). Opportunities for renewable bioenergy using microorganisms. *Biotechnology and Bioengineering*, *100*(2), 203–212. https://doi.org/10.1002/bit.21875

Ro, D. K., Paradise, E. M., Quellet, M., Fisher, K. J., Newman, K. L., Ndungu, J. M., Ho, K. A., Eachus, R. A., Ham, T. S., Kirby, J., Chang, M. C. Y., Withers, S. T., Shiba, Y., Sarpong, R., & Keasling, J. D. (2006). Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, *440*(7086), 940–943. https://doi.org/10.1038/nature04640

Roux, C., Etienne, T. A., Hajnsdorf, E., Ropers, D., Carpousis, A. J., Cocaign-Bousquet, M., & Girbal, L. (2022). The essential role of mRNA degradation in understanding and engineering E. coli metabolism. *Biotechnology Advances*, *54*. https://doi.org/10.1016/j.biotechadv.2021.107805

Rugbjerg, P., Feist, A. M., & Sommer, M. O. A. (2018). Enhanced metabolite productivity of Escherichia coli adapted to glucose M9 minimal medium. *Frontiers in Bioengineering and Biotechnology*, *6*(NOV), 166. https://doi.org/10.3389/fbioe.2018.00166

Russell, G. A., & Berry, P. J. (1986). Approaches to the demonstration of congenital heart disease. In *Journal of Clinical Pathology* (Vol. 39, Issue 5). https://doi.org/10.1136/jcp.39.5.503

Schellenberger, J., Lewis, N. E., & Palsson, B. (2011). Elimination of thermodynamically infeasible loops in steady-state metabolic models. *Biophysical Journal*, *100*(3), 544–553. https://doi.org/10.1016/j.bpj.2010.12.3707

Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., Bordbar, A., Lewis, N. E., Rahmanian, S., Kang, J., Hyduke, D. R., & Palsson, B. (2011). Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox v2.0. *Nature Protocols*, *6*(9), 1290–1307. https://doi.org/10.1038/nprot.2011.308

Schmidt, A., Kochanowski, K., Vedelaar, S., Ahrné, E., Volkmer, B., Callipo, L., Knoops, K., Bauer, M., Aebersold, R., & Heinemann, M. (2016). The quantitative and condition-dependent Escherichia coli proteome. *Nature Biotechnology*, *34*(1), 104–110. https://doi.org/10.1038/nbt.3418

Schuetz, R., Kuepfer, L., & Sauer, U. (2007). Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Molecular Systems Biology*, *3*. https://doi.org/10.1038/msb4100162

Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M., & Sauer, U. (2012). Multidimensional optimality of microbial metabolism. *Science*, *336*(6081), 601–604. https://doi.org/10.1126/science.1216882

Scott, M., Gunderson, C. W., Mateescu, E. M., Zhang, Z., & Hwa, T. (2010). Interdependence of cell growth and gene expression: Origins and consequences. *Science*, *330*(6007), 1099–1102. https://doi.org/10.1126/science.1192588

Shapira, P., Kwon, S., & Youtie, J. (2017). Tracking the emergence of synthetic biology. *Scientometrics*, *112*(3), 1439–1469. https://doi.org/10.1007/s11192-017-2452-5

Shlomi, T., Cabili, M. N., Herrgård, M. J., Palsson, B., & Ruppin, E. (2008). Network-based prediction of human tissue-specific metabolism. *Nature Biotechnology*, *26*(9), 1003–1010. https://doi.org/10.1038/nbt.1487

Tagkopoulos, I., Liu, Y. C., & Tavazoie, S. (2008). Predictive behavior within microbial genetic networks. *Science*, *320*(5881), 1313–1317. https://doi.org/10.1126/science.1154456

Tavassoly, I., Goldfarb, J., & Iyengar, R. (2018). Systems biology primer: The basic methods and approaches. *Essays in Biochemistry*, *62*(4), 487–500. https://doi.org/10.1042/EBC20180003

Thiele, I., Fleming, R. M. T., Que, R., Bordbar, A., Diep, D., & Palsson, B. O. (2012). Multiscale Modeling of Metabolism and Macromolecular Synthesis in E. coli and Its Application to the Evolution of Codon Usage. *PLoS ONE*, *7*(9). https://doi.org/10.1371/journal.pone.0045635

Thiele, I., Jamshidi, N., Fleming, R. M. T., & Palsson, B. O. (2009). Genome-scale reconstruction of escherichia coli's transcriptional and translational machinery: A knowledge base, its mathematical formulation, and its functional characterization. *PLoS Computational Biology*, *5*(3). https://doi.org/10.1371/journal.pcbi.1000312

Thiele, I., & Palsson, B. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, *5*(1), 93–121. https://doi.org/10.1038/nprot.2009.203

Utrilla, J., O'Brien, E. J., Chen, K., McCloskey, D., Cheung, J., Wang, H., Armenta-Medina, D., Feist, A. M., & Palsson, B. O. (2016). Global Rebalancing of Cellular Resources by Pleiotropic Point Mutations Illustrates a Multi-scale Mechanism of Adaptive Evolution. *Cell Systems*, *2*(4), 260–271. https://doi.org/10.1016/j.cels.2016.04.003

Van Regenmortel, M. H. V. (2004). Reductionism and complexity in molecular biology. Scientists now have the tools to unravel biological complexity and overcome the limitations of reductionism. *EMBO Reports*, *5*(11), 1016–1020. https://doi.org/10.1038/sj.embor.7400284

Varma, A., & Palsson, B. O. (1995). Parametric sensitivity of stoichiometric flux balance models applied to wild-type Escherichia coli metabolism. *Biotechnology and Bioengineering*, *45*(1), 69–79. https://doi.org/10.1002/bit.260450110

Venturelli, O. S., Tei, M., Bauer, S., Chan, L. J. G., Petzold, C. J., & Arkin, A. P. (2017). Programming mRNA decay to modulate synthetic circuit resource allocation. *Nature Communications*, *8*. https://doi.org/10.1038/ncomms15128

Wetmore, K. M., Price, M. N., Waters, R. J., Lamson, J. S., He, J., Hoover, C. A., Blow, M. J., Bristow, J., Butland, G., Arkin, A. P., & Deutschbauer, A. (2015). Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. *MBio*, *6*(3), 1–15. https://doi.org/10.1128/mBio.00306-15

Wu, J., Bao, M., Duan, X., Zhou, P., Chen, C., Gao, J., Cheng, S., Zhuang, Q., & Zhao, Z. (2020). Developing a pathway-independent and full-autonomous global resource allocation strategy to dynamically switching phenotypic states. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-19432-2

Yamazaki, Y., Niki, H., & Kato, J. I. (2007). Profiling of Escherichia coli chromosome database. *Methods in Molecular Biology*, *416*, 385–389. https://doi.org/10.1007/978-1-59745-321-9_26

You, C., Okano, H., Hui, S., Zhang, Z., Kim, M., Gunderson, C. W., Wang, Y. P., Lenz, P., Yan, D., & Hwa, T. (2013). Coordination of bacterial proteome with metabolism by cyclic AMP signalling. *Nature*, *500*(7462), 301–306. https://doi.org/10.1038/nature12446

Zarembinski, C. M., Hoyt, J. C., & Reeves, H. C. (1991). Properties of isocitrate lyase from Escherichia coli K12 grown on acetate or glycolate. *Current Microbiology*, *22*(1), 65–68. https://doi.org/10.1007/BF02106215

Zeng, H., Rohani, R., Huang, W. E., & Yang, A. (2021). Understanding and mathematical modelling of cellular resource allocation in microorganisms: a comparative synthesis. *BMC Bioinformatics*, *22*(1). https://doi.org/10.1186/s12859-021-04382-3

Anexos

Código para los cálculos en el repositorio de github: https://github.com/utrillalab/proteomeVSgenome

Publicación derivada:

Marquez-Zavala, E., & Utrilla, J. (2023). Engineering resource allocation in artificially minimized cells: Is genome reduction the best strategy?. *Microbial Biotechnology, 16*(5), 990-999. https://doi.org/10.1111/1751-7915.14233

Received: 11 September 2022 | Accepted: 26 January 2023

DOI: 10.1111/1751-7915.14233

MICROBIAL BIOTECHNOLOGY Applied Microbiology International

RESEARCH ARTICLE

Engineering resource allocation in artificially minimized cells: Is genome reduction the best strategy?

Elisa Marquez-Zavala[1,2] | **Jose Utrilla**[2]

[1]Department of Biotechnology and Food Science, Norwegian University of Science and Technology, Trondheim, Norway

[2]Synthetic Biology Program, Center for Genomic Sciences, Universidad Nacional Autónoma de México, Cuernavaca, Mexico

Correspondence
José Utrilla, Synthetic Biology Program, Center for Genomic Sciences, Universidad Nacional Autónoma de MéxicoAv. Universidad sn Col. Chamilpa, Cuernavaca, Morelos, Mexico.
Email: utrilla@ccg.unam.mx

Funding information
DGAPA- UNAM- PAPIIT, Grant/Award Number: IN213420; CONACyT Ciencia de Frontera, Grant/Award Number: 319352

Abstract

The elimination of the expression of cellular functions that are not needed in a certain well-defined artificial environment, such as those used in industrial production facilities, has been the goal of many cellular minimization projects. The generation of a minimal cell with reduced burden and less host-function interactions has been pursued as a tool to improve microbial production strains. In this work, we analysed two cellular complexity reduction strategies: genome and proteome reduction. With the aid of an absolute proteomics data set and a genome-scale model of metabolism and protein expression (ME-model), we quantitatively assessed the difference of reducing genome to the correspondence of reducing proteome. We compare the approaches in terms of energy consumption, defined in ATP equivalents. We aim to show what is the best strategy for improving resource allocation in minimized cells. Our results show that genome reduction by length is not proportional to reducing resource use. When we normalize calculated energy savings, we show that strains with the larger calculated proteome reduction show the largest resource use reduction. Furthermore, we propose that reducing highly expressed proteins should be the target as the translation of a gene uses most of the energy. The strategies proposed here should guide cell design when the aim of a project is to reduce the maximum amount or cellular resources.

INTRODUCTION

Biological complexity is a challenge to understand and model, many synthetic biology projects are based on the idea that to improve production hosts or chassis we need to reduce the inherent biological complexity. Starting from a simplified organism has enormous potential to reduce detrimental host-function interaction, therefore, there have been many efforts to reduce microbial genomes (Fredens et al., 2019; Michalik et al., 2021) potentially also reducing the complexity of microbial production hosts. The idea of a simple cell with the minimal set of functions needed to grow and to perform a programmed synthetic biology function has been placed as the cornerstone of the establishment of a quasi-universal synthetic biology chassis. Many complexity reduction approaches are mainly based on genome reduction, focused on finding the minimal set of genes that are able to sustain life, however many of such strains have shown growth defects (Choe et al., 2019).

Building bottom-up synthetic genomes should provide us with the ability to better understand and design a minimal cell, however to date those projects have shown to be challenging. In the last iteration of a bottom-up synthetic genome, many genes with unknown functions had to be added to the minimal cell in order to produce a viable cell (Hutchison et al., 2016). Top-down projects have mainly focused on reducing parts of the genome using different approaches, however, the cellular resource consumption of eliminated genes-proteins-functions has not been considered.

In this work, we aim to compare bacterial complexity reduction approaches using *Escherichia coli* as a model organism. We formalize the calculations of the saved

resources of eliminating genes when those genes are transcribed and translated (to a median value). We calculated the theoretical liberated resources in terms of energy and proteome liberation for a defined growth environment. We show that resource reallocation efforts can be optimized if they are focused on a few genes producing highly expressed dispensable proteins. We then propose that resource allocation reduction strategies should be focused on the proteome rather than on the genome.

EXPERIMENTAL PROCEDURES

Data processing of strains

We processed the information of nine strains with a minimized genome (Hashimoto et al., 2005; Hirokawa et al., 2013; Karcagi et al., 2016; Mizoguchi et al., 2008; Park et al., 2014; Pósfai et al., 2006) and a strain with a minimized proteome (Lastiri-Pancardo et al., 2020) (Table 1). We obtained the eliminated genes of the minimized genome strains from their most recent articles. If they reported genome elimination ranges, these coordinates were mapped against the information from their parental strain in the GenBank database (Sayers et al., 2019; the *Escherichia coli* str. K-12 substr. MG1655 genome, version NC_000913.3 and the *Escherichia coli* str. K-12 substr. W3110 genome, version AP009048.1).

Calculation of the proteome reduction of each analysed strain

To calculate the amount of proteome that a certain strain is reducing we assumed that the removed protein-coding genes produce the amount of protein measured in the absolute proteomics data set in the same growth condition (Schmidt et al., 2016). The amount of saved proteome was calculated in femtograms per cell, however, it was expressed in percentage of the total measured proteome to facilitate comparison amongst strains and conditions. The used data set comprises 95% of proteome coverage by mass (around 2300 proteins in each condition). Proteins with no data in a certain condition were not considered by these analyses since their proteome contribution may be considered negligible.

Estimation of replication, transcription and translation costs

To have a normalized basis for comparison amongst different cellular process we performed calculations on a gene with an average length of 950 bp as defined by Lynch & Marinov, 2015. We calculate the energetic cost of the replication, transcription or translation of a gene of 950 pb with the iJLE1678-ME model of *E. coli*. In order to perform these calculations, we simulated the change in biomass composition by increasing the genome, the transcriptome or the proteome size by a unit equivalent of a gene of 950 pb (or its gene product). Once the biomass composition was modified in the model we used fixed the growth rate, glucose uptake rate and oxygen uptake rate to simulate batch growth on a minimal medium and we used ATP maintenance (ATPM) as the objective function to calculate the differences on unaccounted for energy by changing the biomass composition of each cellular process. The calculated change in ATPM is the cost of producing DNA, RNA or protein. As the abundance of proteins can span several orders of magnitude and in order to set the translation level of this average gene, we used the median contribution to the proteome of a gene. This is that a 950 pb gene would be translated to reach 0.042% of the proteome and 22.4 Kbp of the genome are needed to produce 1% of the proteome.

Analysis of the cellular resources in terms of ATP from each eliminated gene with the ME model and proteomic data

To calculate total replication, transcription and protein production costs for each minimized strain, first, we used the amount of genome, transcriptome and calculated proteome corresponding to each reduced strain. We used the cellular composition reported in Bremer and Dennis 1996 (Dennis & Bremer, 2008) to compensate for growth-dependent changes in cellular composition, we also compensated the genome equivalents considering the genome equivalents at a given growth rate (e.g. at 0.69 h^{-1} we estimate 1.63 genomes per cell). The iLE1678-ME model was used by fixing the growth rate, the glucose uptake rate and the oxygen uptake rate, we used the ATP maintenance (ATPM) as the objective function for the proteome simulation, whereas for the other simulations, we maintained the default objective of dummy protein production. Then, changing the amount of DNA in the cell, changes in the number of transcribed genes, and changes in protein production respectively. ME-model simulations calculated the changes in ATP consumption, then it was converted to cost per Mb depending on the length of the gene (data obtained from NCBI) as well as the percentage of proteome they represented in a specific condition from the quantitative proteome data of Schmidt et al., 2016. Detailed computational simulations can be reproduced using the code and data provided in the accompanying github repository.

TABLE 1 Genotype and phenotypic traits of the analysed strains.

Strain	Background	Eliminated genome (mb)	Eliminated genome (%)	Eliminated genes (Number)	Phenotype	Reference
Δ16	MG1655	1.3	29.70	1227	• Lower growth rate and changes in cell shape in some media (thicker and elongated) • Increased cell volume likely as a result of deteriorated peptidoglycan integrity and/or by impaired osmotic regulation	Hashimoto et al. (2005)
MS56	MG1655	1.07	23	1072	• Higher stability in the production of recombinant proteins • Possible reduction in the production of the foreign genes (plasmids) • Similar growth rates in rich media, but either faster or slower growth in minimal media • May be sensitive to small differences in cultural conditions • Imbalances in some cofactors and disrupted metabolism in minimal media	Park et al. (2014), Choe et al. (2019)
MDS69	MG1655	0.94	20.30	965	• Growth rate is 17% lower (probably related to large fluctuations in transcriptome caused by genome reduction) • Reduced competitive fitness in LB medium • Decreased nitrogen utilization efficiency • Growth defects in 14.3% of environments where the parental has no issues	Pósfai et al. (2006), Karcagi et al. (2016), Vernyik et al. (2020)
MDS42	MDS/ MG1655	0.66	14.3	704	• Increased L-threonine production by 83% • Higher electroporation efficiency • Similar growth rates in minimal and rich media, with one study showing a 20% lower growth rate • Similar mutation rates and reduced genome stability under minimal media • Growth defects in 8.8% of environments where the parental has no issues	Pósfai et al. (2006), Lee et al. (2009), Ying et al. (2013), Karcagi et al. (2016), Yuan et al. (2017), Ying & Yama (2018). Vernyik et al. (2020)
MDS12	MDS/ MG1655	0.37	8.11	423	• Higher density in stationary phase • Similar growth rates in minimal and rich media • Similar doubling times, electroporation and transformation efficiencies • Marginally outgrown when cocultured	Kolisnychenko et al. (2002), Pósfai et al. (2006), Karcagi et al. (2016), Vernyik et al. (2020)
MGF01	MGF/W3110	1.03	22.2	1081	• Increased L-threonine production (2.4-fold) • Higher glucose consumption (1.44-fold) and acetate accumulation (0.09-fold) • Similar or slightly higher rate in rich media • Either a higher or a lower growth rate and cell density in minimum media • Higher mutation rate	Hashimoto et al. (2005), Mizoguchi et al. (2007, 2008), Hirokawa et al. (2013), Nakayasu et al. (2020)
MGF02		1.1	25	~1200	• Higher growth rate	Hashimoto et al. (2005), Hirokawa et al. (2013), Nakayasu et al. (2020)

TABLE 1 (Continued)

Strain	Background	Eliminated genome (mb)	Eliminated genome (%)	Eliminated genes (Number)	Phenotype	Reference
DGF298	DGF/W3110	1.67	35.80	~1700	• Better growth rate and cell yield in rich medium • More stable genome • No auxotrophy phenotype • Non-productive mutants of protodeoxyviolacein as it was too burdensome	Hashimoto et al. (2005), Hirokawa et al. (2013), Nakayasu et al. (2020)
DGF327		1.38	30.90	~1400	• Better growth rate and cell yield in rich medium • No auxotrophy phenotype	Hashimoto et al. (2005), Hirokawa et al. (2013), Nakayasu et al. (2020)
PFC	BW25113	~1.67 x 10^{-3}	0.004	3	• Higher proteomic budget • Increased production of violacein (18%) and pDNA (33% and 53%)	Lastiri-Pancardo et al. (2020), de la Cruz et al. (2020)

RESULTS

In this work, we aimed to analyse genome or proteome reduction projects with a resource allocation approach. First, to have a basis for comparison we normalized DNA replication, RNA transcription and protein translation into ATP equivalents. Then, as our results confirmed what has been shown before: translation is the main driver of cellular resource consumption (Lynch & Marinov, 2015; Wagner, 2007), we analysed how much of the proteome was reduced assuming that the eliminated genes were expressed at the same magnitude as in the absolute proteomics data set (Schmidt et al., 2016) in the same growth condition (Figure 1).

In order to have a normalized comparison amongst the analysed cellular processes we used two different calculations that rendered a similar result. We used the Lynch & Marinov, 2015 equations, which calculate the energetic cost, measured in units of ATP hydrolyses, associated with the replication, transcription, and translation of an average gene of 950 bp length as previously defined (Lynch & Marinov, 2015). We also simulated the resulting decrease in unaccounted for energy (ATPM) with a ME-model (Lloyd et al., 2018) when we increase the genome, the transcriptome or the proteome by an equivalent of a gene of 950 bp. For calculation basis, we set the median contribution to the proteome by a 950 pb gene to be 0.042% of the proteome (see experimental procedures). Therefore, from ME-model simulations, we obtain (that 1% of proteome represents a cost of 3.76×10^{-4} and of 3.16×10^{-3} mmol ATP gDW h^{-1}) for replication and transcription, respectively. Whereas translation represents 1.65×10^{-1} mmol ATP gDW h^{-1}.

As we show above, the normalized energy cost of the production of the proteome is three and two orders of magnitude higher than the production of the genome or transcriptome respectively. In order to compare the reduction in resource allocation amongst several genome or proteome-reduction approaches we took the information of nine genomes reduced and one proteome-reduced strain of *E. coli* (Table 1). The MDS (Multiple-Deletion Series) set of strains (Karcagi et al., 2016; Pósfai et al., 2006) are derived from MG1655. In that set of strains, genes were eliminated when they are not present in close relatives, mobile elements and also genes with unknown and non-essential functions were eliminated (such as flagella). The Δ16 strain and MS56 were designed to minimize the MG1655 genome as much as possible aiming to obtain a simple and highly controllable cell with non-essential genes and unneeded genome fragments removed (Hashimoto et al., 2005; Park et al., 2014). The MGF (Minimum Genome Factory) strains, derived from W3110, were designed to eliminate genomic regions that do not hybridize with other genomes and regions that were eliminated in other *E. coli* reduction projects. They also eliminated genes with unknown functions

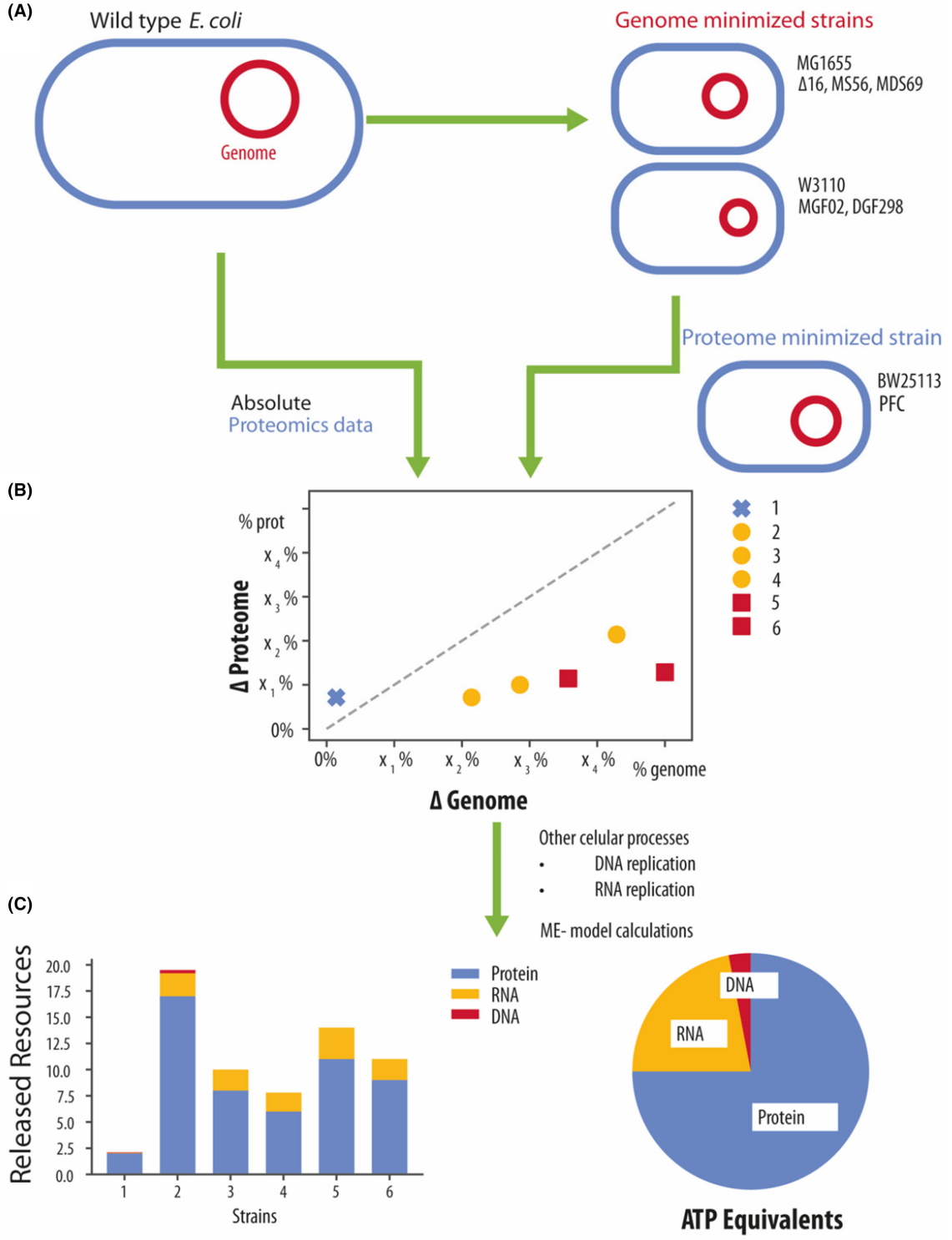

FIGURE 1 Flow chart of the study. (A) The data included for this analysis came from five MG1655 and W3110 derivatives with their genomes minimalized. In addition, data from a proteome-minimized *Escherichia coli* K12 BW25113 derivative was used. (B) Comparison of the amount of proteome that corresponds to the deleted genes in the minimized strains, based on the absolute proteomics data set (Figure 3 for full detail). (C) Calculation of the equivalent ATP consumption in the wild type (WT) provided by the genes that were removed from the minimized genome strains using a metabolism and expression model of *E. coli* and proteomics data. The consumption was evaluated for protein, RNA and DNA production. A pie chart displays the average consumption for each category of the genome-minimized strains.

(Mizoguchi et al., 2008). Based on MGF strains, DGF strains (Designed Genome Factory) were generated by removing insertion sequences and toxin-antitoxin systems. Previously removed regions from MGF strains, causing growth defects such as auxotrophies were restored in this project (Hirokawa et al., 2013). For

better visualization of the results, we selected the strain with the highest genome reduction as a representative of the project the strain with the highest reduction. The MDS69 strain represented the MDS strains, the DGF298 strain represented the DGF strains, and the MGF02 strain represented the MGF strains. Finally, we compare genome-reduced strains to PFC a proteome-reduced strain created by our group (Lastiri-Pancardo et al., 2020). In this project, we applied the ReProMin method, which identifies the minimal set of genetic interventions that maximize the savings in cell resources (Figure 1B).

We used the Schmidt et al., 2016 absolute proteomic data set to calculate the amount of proteome that would be reduced if we assume that the eliminated genes are being expressed at the same magnitude as in the measured proteomes (Figure 2A). Genes with no proteomic information in the data set are not considered since they belong to the 5% of the proteome that is not measured, therefore, we assume that they represent a small contribution to the proteome. In order to set a fixed condition for our comparisons, we used glucose M9 minimal medium as the main analysed growth condition.

All the studied strains from genome reduction projects range from 8 to 35% genome reduction, however, strains were designed and constructed with different approaches, and the per cent of genome reduction is not always proportional to the number of reduced genes (Table 1, Figure 3). As mentioned above, here we focused on the potential resource savings. First, we need to notice the uneven distribution of the contribution of each gene to the total proteome (Figure 2B). The plotted mass contribution of the measured proteins to the proteome in femtograms per cell (fg/cell) spans seven

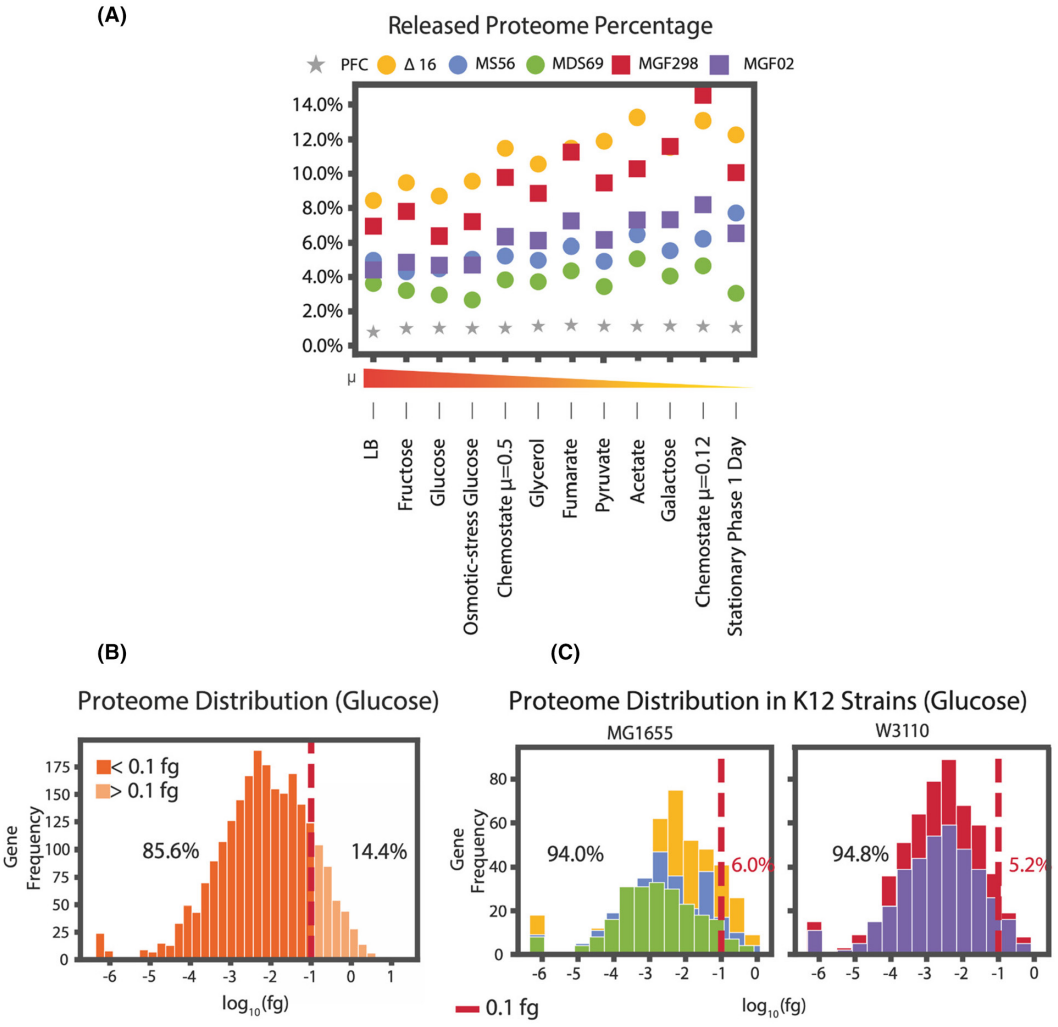

FIGURE 2 Proteome calculations on genome-reduced strains. We used the strains with the smallest genome of each project, and the Schmidt et al., 2016 proteome data set. (A) Percentage of the released proteome in 12 conditions, sorted by specific growth rate. This is the percentage released if we assume that the eliminated genes are being expressed at the same magnitude as in the data set. (B) Proteome mass distribution ($\log_{10}$ fg) per gene on glucose minimal medium growth condition. On the right of the dotted line are the per cent of proteins that have a mass over 0.1 fg. (C) Proteome mass distribution ($\log_{10}$ fg) per from the deleted genes of the genome reduced strains in glucose minimal medium growth condition. On the right of the dotted line are the per cent of proteins that have a mass over 0.1 fg.

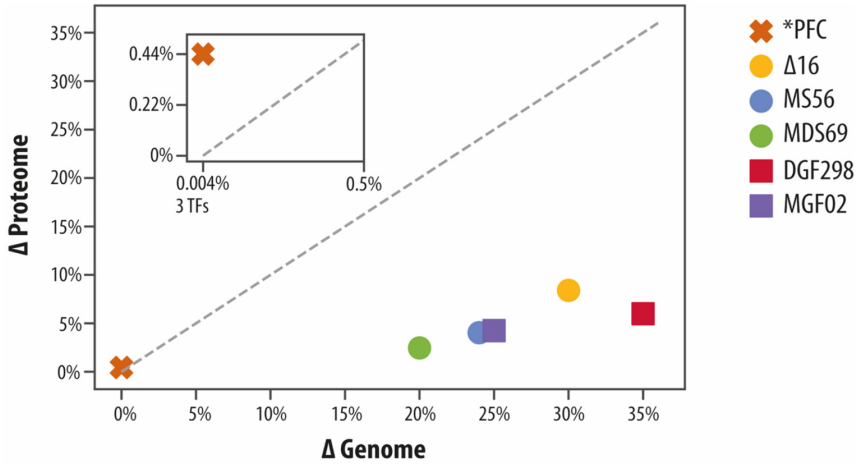

FIGURE 3 Representative strains with their genome reduction in comparison to their calculated proteome reduction in the M9 glucose growth condition. The insert represents the proteome-reduced PFC strain.

orders of magnitude (from 1^{e-6} to 1 fg/cell). Only 14.4% of the measured genes (344) code for a protein that contributes more than 0.1 fg/cell (many of those code for highly expressed essential genes, such as *tufA* or *metE*). This means that few highly expressed genes will be responsible for the greater proteome reduction, therefore it demonstrates that the relation between genome size reduction and potential resource savings is not proportional (Figure 3).

As has been previously stated, our main focus condition is glucose minimal medium, in this condition the calculated proteome reduction of the analysed strains spans from 2.45% to 8.40%. The Δ16 strain showed the largest calculated proteome savings (8.4%) and the MDS69 strain was the one with the smallest calculated proteome savings (2.45%). It is worth mentioning that the Δ16 strain is not the most genome-reduced one, it is the DGF298, however, in that specific condition, the Δ16 strain presents the highest potential proteome reduction. In another remarkable case, for chemostat growth at the 0.12 h^{-1} dilution rate, the DGF298 presents a larger proteome reduction than the Δ16 strain and the largest calculated here (14%, Figure 2A).

We mentioned above that translation (the biosynthesis of the proteome) takes 96% of the energy, when compared on a per gen (950 bp) basis, whereas transcription takes 3.9% and replication just around 0.1% of the total energy (Table 2). Therefore, translation is the main focus of this work. Here, we were also interested in comparing the differences in resource allocation—in terms of normalized flux of ATP in mmol ATP•g DW^{-1} •h—when we calculate the amount of saved energy equivalents. To that end, we used the iLE-1678-ME model to calculate the cost of producing the actual normalized amount of genomic DNA, the transcripts (mRNAs) and the proteins eliminated in each analysed strain. In Figure 4, we show the contribution of each cellular process to the total of saved energy for each strain. These calculations confirmed that translation is the main driver of energy savings ranging from 54 to 84% of saved ATP equivalents. We also showed that

TABLE 2 Calculated ATP cost for each cellular process to produce a protein from a 950 bp gene.

Process	ATP	Per cent
Replication	4180.58	0.23
Transcription	71,359.04	3.93
Translation	1,737,983.23	95.83
Total	1,813,522.86	100

the strains, that according to our calculations, have the greater amount of saved proteome also save the greatest amount of energy. Comparing genome reduction approaches to a proteome reduction-oriented project we show that a strain with only three transcription factors eliminated that resulted in a theoretical proteome reduction of 0.5%, presents a similar calculated resource reduction use than the MSD12 strain with an 8% genome reduction (Figure 4). The proteome-reduced strain showed increases around 20% in the production of fluorescent protein reporters and in violacein from a heterologous metabolic pathway (Lastiri-Pancardo et al., 2020). These results show that if we compare resource savings by the number of removed nucleotides in each project, the proteome reduction approach is more effective than the genome reduction approaches. However, it is also noteworthy that most of the genome-reduced strains have greater savings than the only proteome-reduced strain analysed here.

DISCUSSION

Reduction of cellular complexity is a complicated endeavour that needs to be tackled from many perspectives. The bottom-up approach has proven to be challenging as the definition of minimal gene sets still escapes from our full understanding (Glass et al., 2017). In that regard, top-down approaches may be easier to adopt, therefore more widely used. Considering the

resource use of the genes to eliminate will improve the success of the minimized strains, since reducing less genes with higher contribution to resource use should result in higher budget benefits with less phenotypic defects.

Here, we formalize the calculations of the saved resources of eliminating genes when those genes are transcribed and translated (to a median value). Gene expression is highly variable in terms of transcript and proteins produced per gene, in order to have a basis for calculation we used median values of gene expression. This certainly introduces a large bias in our calculations, however, we provide a coarse estimate of the amount of saved resources for each process. Also, the order of magnitude of change of the resource expenditure amongst the different processes analysed here, highlight that besides those large possible differences in gene expression, the amount of saved resources is much larger for translation than for transcription or DNA replication. Moreover, with less gene deletions, the minimized cell behaviour will become more predictable, and our proteomic data-based calculations should be more accurate, as several gene deletions can dramatically alter protein concentrations. Using a genome-scale model of metabolism and expression (ME-model), we are able to account for "distant" processes such as the need of macromolecular machinery (such as ribosomes) that carry out each process thus incurring a cost.

Many analyses have shown that translation is the major driver of resource consumption of the cell (Kepp, 2020; Lynch & Marinov, 2015; Wagner, 2007). Here, we show that genome reduction does not have a direct correspondence into resource consumption reduction. Our presented findings show that if a cell minimization project aims to reduce the use of resources of the non-used cellular functions then the proteome should be the target. We show that the protein concentration in *E. coli* spans seven orders of magnitude, finding the most expressed non-essential proteins is a pretty simple endeavour that should yield a significant reduction in resource consumption with a less complicated genome editing strategy.

The idea that a minimized genome will yield a minimized cell may not be totally accurate since the mere bearing of a gene does not mean that it will be expressed, and the translation level of a gene can be highly variable. The use of absolute proteomics data sets has proven to be very informative in resource optimization, unfortunately, quantitative absolute proteomics is a quite challenging technique with a large equipment investment cost, therefore those proteome-wide data sets are scarce.

Using a data set with 95% of proteome coverage by mass (2300 genes approximately) leaves many genes outside of our analysis. Although those genes may have a reduced expression and therefore not a great contribution to the total saved resource, those genes may be of great importance to cellular fitness and should not be considered dispensable (Price et al., 2018). In addition, the minimization design should consider the trade-off between strain robustness and resource reduction, by accounting for the fitness contribution of the candidate genes to eliminate, such as the quasi-essential genes on the environment of interest. Genome reduction projects have been carried out in other *E. coli* backgrounds (non-K12), however, the lack of absolute proteomic information limits the application of our method to those strains (Umenhoffer et al., 2017). Absolute proteomic data sets, gene essentiality, cost and fitness contribution of each gene in a genome may provide the data needed to better design minimized strains with a resource allocation approach for *E. coli* strains of different lineage and other microbes of interest.

Improving the understanding of mRNA degradation can play a critical role in optimizing metabolic engineering strategies (Roux et al., 2021). Targeting mRNA degradation has been shown to enhance protein production without competing for cellular resources in the host strain (Mao & Inouye, 2012; Venturelli et al., 2017; Wu et al., 2020). However, due to the variability of mRNA degradation rates between genes, it can be challenging to model this process for an "average gene". Despite this, considering the impact of mRNA degradation on resource utilization is important in order to fully understand the effects of gene modification on cellular behaviour and potentially improve our calculation methods. Overall, mRNA degradation has great potential as a target for metabolic engineering to improve the performance of bacterial cell factories.

Large-scale fermentations generate stressful conditions, such as substrate availability gradients. The regulatory response to those stressful conditions has a cost, as it demands protein expression and it has been shown to increase ATP maintenance demands by 40%–50% (Löffler et al., 2016). A successful approach has been to target those specific stress responses that consume resources. The resulting minimized strains showed a lower maintenance coefficient and increase the production yield of GFP in simulated large-scale bioprocesses (Ziegler et al., 2021).

In this work, we provide a straightforward path to pursue when designing a strain with a higher potential to divert resources to a function of interest whilst performing fewer genetic interventions. Those resource-minimized strains should perform better as microbial production hosts than their non-improved counterparts (de la Cruz et al., 2020; Mizoguchi et al., 2008; Park et al., 2014).

AUTHOR CONTRIBUTIONS

Elisa Marquez-Zavala: Conceptualization (supporting); data curation (lead); investigation (lead); methodology

(equal); writing – original draft (equal); writing – review and editing (supporting). **José Utrilla:** Conceptualization (lead); formal analysis (supporting); funding acquisition (lead); methodology (equal); project administration (lead); supervision (lead); writing – original draft (equal); writing – review and editing (lead).

ACKNOWLEDGMENTS

Funding from DGAPA- UNAM- PAPIIT project IN213420 and CONACyT Ciencia de Frontera (grant no. 319352). Technical support of Jose Alfredo Hernandez and Victor del Moral. EMZ acknowledges the Programa de Maestría y Doctorado en Ciencias Bioquímicas UNAM and the master's scholarship 900742 from CONACyT.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Github repository with the code and source data to reproduce this analysis: https://github.com/utrillalab/Cell_Resource_Minimization_Strategies

ORCID

Elisa Marquez-Zavala https://orcid.org/0000-0001-8096-5280
Jose Utrilla https://orcid.org/0000-0003-3048-9241

REFERENCES

Choe, D., Lee, J.H., Yoo, M., Hwang, S., Sung, B.H., Cho, S. et al. (2019) Adaptive laboratory evolution of a genome-reduced *Escherichia coli*. *Nature Communications*, 10, 935.

de la Cruz, M., Ramírez, E.A., Sigala, J.-C., Utrilla, J. & Lara, A.R. (2020) Plasmid DNA production in proteome-reduced *Escherichia coli*. *Microorganisms*, 8, 1444.

Dennis, P.P. & Bremer, H. (2008) Modulation of chemical composition and other parameters of the cell at different exponential growth rates. *EcoSal Plus*, 3, 1–49.

Fredens, J., Wang, K., de la Torre, D., Funke, L.F.H., Robertson, W.E., Christova, Y. et al. (2019) Total synthesis of *Escherichia coli* with a recoded genome. *Nature*, 569(7757), 514–518.

Glass, J.I., Merryman, C., Wise, K.S., Hutchison, C.A. & Smith, H.O. (2017) Minimal cells—real and imagined. *Cold Spring Harbor Perspectives in Biology*, 9, a023861.

Hashimoto, M., Ichimura, T., Mizoguchi, H., Tanaka, K., Fujimitsu, K., Keyamura, K. et al. (2005) Cell size and nucleoid organization of engineered *Escherichia coli* cells with a reduced genome. *Molecular Microbiology*, 55, 137–149.

Hirokawa, Y., Kawano, H., Tanaka-masuda, K., Nakamura, N., Nakagawa, A., Ito, M. et al. (2013) Genetic manipulations restored the growth fitness of reduced-genome *Escherichia coli*. *Journal of Bioscience and Bioengineering*, 116, 52–58.

Hutchison, C.A., Chuang, R.-Y.R.-Y., Noskov, V.N., Assad-Garcia, N., Deerinck, T.J., Ellisman, M.H. et al. (2016) Design and synthesis of a minimal bacterial genome. *Science*, 351, aad6253.

Karcagi, I., Draskovits, G., Umenhoffer, K., Fekete, G., Kovács, K., Méhi, O. et al. (2016) Indispensability of horizontally transferred genes and its impact on bacterial genome streamlining. *Molecular Biology and Evolution*, 33, 1257–1269.

Kepp, K.P. (2020) Survival of the cheapest: how proteome cost minimization drives evolution. *Quarterly Reviews of Biophysics*, 53, e7.

Kolisnychenko, V., Plunkett, G., Herring, C.D., Fehér, T., Pósfai, J., Blattner, F.R. et al. (2002) Engineering a reduced *Escherichia coli* genome. *Genome Research*, 12, 640–647.

Lastiri-Pancardo, G., Mercado-Hernández, J.S., Kim, J., Jiménez, J.I. & Utrilla, J. (2020) A quantitative method for proteome reallocation using minimal regulatory interventions. *Nature Chemical Biology*, 16, 1026–1033.

Lee, J.H., Sung, B.H., Kim, M.S., Blattner, F.R., Yoon, B.H., Kim, J.H. et al. (2009) Metabolic engineering of a reduced-genome strain of *Escherichia coli* for L-threonine production. *Microbial Cell Factories*, 8, 1–12.

Lloyd, C.J., Ebrahim, A., Yang, L., King, Z.A., Catoiu, E., O'Brien, E.J. et al. (2018) COBRAme: a computational framework for genome-scale models of metabolism and gene expression. *PLoS Computational Biology*, 14, e1006302.

Löffler, M., Simen, J.D., Jäger, G., Schäferhoff, K., Freund, A. & Takors, R. (2016) Engineering E. coli for large-scale production–strategies considering ATP expenses and transcriptional responses. *Metabolic Engineering*, 38, 73–85.

Lynch, M. & Marinov, G.K. (2015) The bioenergetic costs of a gene. *Proceedings of the National Academy of Sciences of the United States of America*, 112, 15690–15695.

Mao, L. & Inouye, M. (2012) Use of *E. coli* for the production of a single protein. *Methods in Molecular Biology*, 899, 177–185.

Michalik, S., Reder, A., Richts, B., Faßhauer, P., Mäder, U., Pedreira, T. et al. (2021) The Bacillus subtilis minimal genome compendium. *ACS Synthetic Biology*, 10, 2767–2771.

Mizoguchi, H., Mori, H. & Fujio, T. (2007) *Escherichia coli* minimum genome factory. *Biotechnology and Applied Biochemistry*, 46, 157–167.

Mizoguchi, H., Sawano, Y., Kato, J.I. & Mori, H. (2008) Superpositioning of deletions promotes growth of *Escherichia coli* with a reduced genome. *DNA Research*, 15, 277–284.

Nakayasu, E.S., Chazin-Gray, A.M., Francis, R.M., Eaton, A.M., Auberry, D.L., Muñoz, N. et al. (2020) Resource reallocation in engineered *Escherichia coli* strains with reduced genomes. *bioRxiv*, 1–35. https://doi.org/10.1101/2020.10.19.346155

Park, M.K., Lee, S.H., Yang, K.S., Jung, S.C., Lee, J.H. & Kim, S.C. (2014) Enhancing recombinant protein production with an *Escherichia coli* host strain lacking insertion sequences. *Applied Microbiology and Biotechnology*, 9815(98), 6701–6713.

Pósfai, G., Plunkett, G., Fehér, T., Frisch, D., Keil, G.M., Umenhoffer, K. et al. (2006) Emergent properties of reduced-genome *Escherichia coli*. *Science*, 312, 1044–1046.

Price, M.N., Wetmore, K.M., Waters, R.J., Callaghan, M., Ray, J., Liu, H. et al. (2018) Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature*, 557, 503–509.

Roux, C., Etienne, T.A., Hajnsdorf, E., Ropers, D., Carpousis, A.J., Cocaign-Bousquet, M. et al. (2021) The essential role of mRNA degradation in understanding and engineering *E. coli* metabolism. *Biotechnology Advances*, 54, 107805.

Sayers, E.W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K.D. & Karsch-Mizrachi, I. (2019) GenBank. *Nucleic Acids Research*, 47, D94–D99.

Schmidt, A., Kochanowski, K., Vedelaar, S., Ahrné, E., Volkmer, B., Callipo, L. et al. (2016) The quantitative and condition-dependent *Escherichia coli* proteome. *Nature Biotechnology*, 34, 104–110.

Umenhoffer, K., Draskovits, G., Nyerges, Á., Karcagi, I., Bogos, B., Tímár, E. et al. (2017) Genome-wide abolishment of Mobile genetic elements using genome shuffling and CRISPR/Cas-assisted MAGE allows the efficient stabilization of a bacterial chassis. *ACS Synthetic Biology*, 6, 1471–1483.

Venturelli, O.S., Tei, M., Bauer, S., Chan, L.J.G., Petzold, C.J. & Arkin, A.P. (2017) Programming mRNA decay to modulate

synthetic circuit resource allocation. *Nature Communications*, 8, 1–11.

Vernyik, V., Karcagi, I., Tímár, E., Nagy, I., Györkei, Á., Papp, B. et al. (2020) Exploring the fitness benefits of genome reduction in *Escherichia coli* by a selection-driven approach. *Scientific Reports*, 10, 1–12.

Wagner, A. (2007) Energy costs constrain the evolution of gene expression. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 308, 322–324.

Wu, J., Bao, M., Duan, X., Zhou, P., Chen, C., Gao, J. et al. (2020) Developing a pathway-independent and full-autonomous global resource allocation strategy to dynamically switching phenotypic states. *Nature Communications*, 11, 1–14.

Ying, B.W., Seno, S., Kaneko, F., Matsuda, H. & Yomo, T. (2013) Multilevel comparative analysis of the contributions of genome reduction and heat shock to the *Escherichia coli* transcriptome. *BMC Genomics*, 14, 1–13.

Ying, B.W. & Yama, K. (2018) Gene expression order attributed to genome reduction and the steady cellular state in *Escherichia coli*. *Frontiers in Microbiology*, 9, 2255.

Yuan, X., Couto, J.M., Glidle, A., Song, Y., Sloan, W. & Yin, H. (2017) Single-cell microfluidics to study the effects of genome deletion on bacterial growth behavior. *ACS Synthetic Biology*, 6, 2219–2227.

Ziegler, M., Zieringer, J., Döring, C.L., Paul, L., Schaal, C. & Takors, R. (2021) Engineering of a robust *Escherichia coli* chassis and exploitation for large-scale production processes. *Metabolic Engineering*, 67, 75–87.

How to cite this article: Marquez-Zavala, E. & Utrilla, J. (2023) Engineering resource allocation in artificially minimized cells: Is genome reduction the best strategy? *Microbial Biotechnology*, 16, 990–999. Available from: https://doi.org/10.1111/1751-7915.14233