Universidad Pablo de Olavide

Programa de Doctorado en Biotecnología, Ingeniería y Tecnología Química (R.D.99/2011)

Tesis Doctoral

Búsqueda y caracterización de señales de codificación de proteínas basado en similitudes no significativas

Memoria para optar al grado de doctor presentada por

Carlos Sánchez Casimiro-Soriguer

Directores

Antonio J. Pérez Pulido

Juan Jiménez Martínez

Sevilla, 2020

Los Dres. ANTONIO JESÚS PÉREZ PULIDO y JUAN JIMÉNEZ MARTÍNEZ, profesor titular y catedrático del Departamento de Biología Molecular e Ingeniería Bioquímica de la Universidad Pablo de Olavide,

CERTIFICAN: que el presente trabajo sobre “Búsqueda y caracterización de señales de codificación de proteínas basado en similitudes no significativas” ha sido realizado en los laboratorios de dicho Departamento por el Licenciado D. Carlos Federico Sánchez Casimiro-Soriguer, bajo nuestra dirección y supervisión, reuniendo los requisitos de originalidad y calidad científica para optar al grado de Doctor por la Universidad Pablo de Olavide de Sevilla; dentro del Programa de Doctorado “Biotecnología, Ingeniería y Tecnología Química”.

Sevilla, Mayo de 2020

LOS DIRECTORES DE LA MEMORIA

Fdo.: Antonio J. Pérez Pulido Fdo.: Juan Jiménez Martínez

Trabajo presentado por Carlos Federico Sánchez Casimiro-Soriguer para optar al Grado de Doctor por la Universidad Pablo de Olavide, de Sevilla.

Sevilla, Mayo de 2020

Fdo. Carlos Federico Sánchez Casimiro-Soriguer

«La única posibilidad de descubrir los límites de lo posible es aventurarse un poco más allá de ellos, hacia lo imposible.»

Arthur C. Clark

Índice

Agradecimientos

El trabajo desarrollado durante una tesis, siendo un periodo de aprendizaje y de iniciación a la investigación científica, no puede ser el resultado del trabajo de una única persona y normalmente involucra a una gran cantidad de personas. Por lo tanto es de justicia reconocer todo este trabajo y justifica la existencia de este apartado. A su vez es difícil expresar con palabras el compendio de esfuerzos que supone este tipo de trabajos.

En primer lugar me gustaría agradecer su trabajo y dedicación a mis directores. A Antonio por haber confiado en mi y abrirme las puertas a este mundo fantástico que es la bioinformática. Llegados a este punto, después de haber compartido tantas horas dentro del despacho y algunas más fuera de el, más que un director es un referente y un amigo. A Juan por haber estado siempre dispuesto a ayudar, a pesar de su ajetreada agenda, y enseñarme que la ciencia solo es posible gracias a la conjunción del trabajo y conocimiento de muchas personas.

A mis compañeros del CABD por recibirme siempre con una sonrisa y estar siempre dispuestos a responder a mis dudas y consultas, a pesar de ser uno de los “frikis” bioinformáticos de la planta baja. Sois un grupo genial, y me hubiese gustado compartir con todos vosotros más barbacoas, cervecitas y campeonatos de padel. Además, me gustaría agradecer aquí, por su participación directa en los trabajos involucrados en esta tesis a Andrés, Manolo, Ana María y Pablo. Y en especial a Alejandro y María del Mar por permitirme usar nuestros trabajos como parte de esta tesis. Sin vosotros esta tesis no hubiese sido posible.

Otro apartado fundamental en el desarrollo de una tesis son los apoyos familiares. En este apartado debo confesar que soy muy afortunado, pues creo que es difícil encontrar una familia donde la adquisición de conocimiento y su puesta en práctica sea tan importante. Esos valores que aprendí en mi juventud, que me han acompañado durante toda mi vida han hecho posible esta tesis y me han ayudado en los momentos no tan buenos. En especial a mi abuela Pepa por enseñarme muchas cosas pero sobre todo por

enseñarme que para ser feliz en esta vida hay que aprender cosas nuevas hasta el último momento. Se que le hubiese encantado poder ver como defiendo esta tesis.

A mi madre por ser la red de seguridad que todo hijo desearía, por tu tiempo, esfuerzo y enseñanzas. A mi padre por enseñarme lo que es la dedicación, el esfuerzo y el trabajo bien hecho. A mi hermano por ser quien es, compartir conmigo su vida y dotes artísticas.

Por último y por eso más importante es mi agradecimiento a mi pareja María por su apoyo incondicional, por ser un punto de inflexión en mi vida, por haberme impulsado a salir de mi zona de confort y porque cada minuto que he dedicado a esta tesis fuera del horario laboral ha ido acompañado de un minuto de esfuerzo por su parte. A mis hijos Manuel y Pablo. Se pueden considerar que son producto de esta tesis ya que nacieron durante la misma. Siempre llevaré conmigo vuestras caras de ilusión y alegría al verme entrar por la puerta de casa a pesar de haber echado de menos mi presencia por el necesario tiempo que he tenido que dedicar a esta tesis. Os quiero.

Resumen

Desde que se secuenciaron las primeras secuencias genómicas con las denominadas tecnologías de primera generación hasta el momento actual donde se están implantando las tecnologías de secuenciación de tercera generación se ha producido un abaratamiento del coste que supone secuenciar el genoma completo de un organismo. Esto ha supuesto un gran aumento de los datos genómicos disponibles, generando la necesidad de la automatización del análisis para poder sacar el mayor provecho a toda esta información. En este sentido se han desarrollado aplicaciones capaces de localizar los genes presentes en secuencias biológicas. Estas aplicaciones incluso en genomas de procariotas, que poseen una menor complejidad con respecto a organismos eucariotas, no encuentran todas las secuencias codificantes de proteínas, y en especial si estas se corresponden con sORF, pseudogenes o genes no canónicos.

Con el objetivo de solventar este problema, y permitir completar así el conjunto de genes de un genoma, se desarrolló AnABlast. Este algoritmo es capaz de localizar regiones codificantes de proteínas mediante el acúmulo de alineamientos no significativos, descartados habitualmente. En esta tesis se presenta el desarrollo y aplicación de AnABlast sobre genomas completos. Se han estudiado y optimizado los parámetros para mejorar la sensibilidad y especificidad utilizando como base el genoma de *Drosophila melanogaster* y se han validado los resultados de posibles nuevos genes codificantes de proteínas, especialmente pequeños ORFs, en *Caenorhabditis elegans* mediante la técnica de RNA interferente. Asimismo, se ha desarrollado la segunda versión del anotador funcional Sma3s, que permite asignar funciones tanto a proteomas como a transcriptomas completos. En esta segunda versión se han reducido las dependencias, simplificado su uso, mejorado la sensibilidad y especificidad y reduciendo el tiempo de ejecución y los costes computacionales. Finalmente estas dos herramientas (AnaBlast y Sma3s) se han combinado en la aplicación web de AnABlast, facilitando de esta forma su uso y, por lo tanto, permitiendo a grupos experimentales completar la anotación de sus genomas de estudio.

1. Introducción

1.1 Motivación

1.1.1 Secuenciación

A mediados del siglo XX dio comienzo una de las grandes revoluciones de la biología actual. Severo Ochoa y su discípulo Kornberg, en la Universidad de Nueva York, secuenciaron en laboratorio las primeras moléculas de RNA y DNA respectivamente. Eran simples cadenas de poly-A y poly-T, pero pronto dieron paso a la secuenciación de cadenas más complejas, lo que dio lugar, poco después, a los trabajos de Marshall W. Nirenberg y J. Heinrich Matthaei, que mostraban que a partir de una cadena de uracilo se obtenían péptidos de fenilalanina, y así pudieron concluir que el triplete UUU de RNA codificaba para este aminoácido (Nirenberg and Matthaei 1961). Este descubrimiento desencadenó pronto una carrera por descifrar los aminoácidos codificados por los diferentes tripletes de nucleótidos del DNA, dando lugar al código genético, tal y como lo conocemos actualmente, con sus 64 posibles codones y los 22 aminoácidos que codifican.

Este descubrimiento tardó en dar paso a la secuenciación de DNA de organismos, ya que primero se tuvieron que desarrollar tecnologías adecuadas, como la de Gilbert y Maxan (Maxam and Gilbert 1977) que marcaba el DNA con nucleótidos radiactivos y lo rompía químicamente en nucleótidos concretos para determinar su orden, y la de Sanger y Coulson (Sanger, Nicklen, and Coulson 1977) que en lugar de romper, interrumpía la síntesis de nucleótidos concretos. Esta última se acabó imponiendo por su mayor sencillez, denominándose actualmente como tecnología de secuenciación de primera generación, y permitió la secuenciación por primera vez de un genoma completo, la secuencia del fago PhiX174 (Sanger et al. 1977).

En sucesivos años se fueron secuenciando los genomas completos de diferentes organismos. En 1995 se realizó la primera secuenciación de un genoma bacteriano completo, *Haemophilus influenzae*, (White et al. 1995), en 1996 se secuenció el primer organismo eucariota, *Saccharomyces cerevisiae*, (A. Goffeau et al., 1996) , en 1999 se

secuenció el primer organismo pluricelular *Caenorhabditis elegans (Wilson 1999)*, y un año más tarde se terminó de secuenciar el genoma de *Drosophila melanogaster* (Adams 2000). Estos genomas han servido como modelos para numerosas enfermedades y, con más de dos décadas de estudio a sus espaldas, tienen un elevado nivel de conocimiento con numerosas evidencias moleculares derivadas de numerosos estudios tanto *in silico* como de laboratorio.

Estos genomas iniciales dieron paso a un nuevo hito en esta revolución de la biología, que ha dado paso a la era genómica. En abril de 2003 finalizó el Proyecto Genoma Humano, con la secuenciación del 99% del genoma de *H. sapiens*, con una precisión del 99.9%. Se tardaron 13 años en secuenciar este primer genoma humano, con un coste aproximado de unos 3.000 millones de dólares y con la contribución de múltiples centros internacionales (Collins 2003).

Pero este hito pronto se quedaría pequeño, y sólo unos pocos años más tarde, en 2007, comenzó el proyecto de los 1000 genomas que, gracias a la aparición de nuevas tecnologías de secuenciación como *Illumina*, *Roche Diagnostics* (454) y *Life Technologies* (SOLiD), denominadas *Next Generation Sequencing* (NGS), permitieron en 2012 culminar su primera fase, en la que se secuenciaron 1092 genomas provenientes de 14 poblaciones humanas distintas. La información y datos provenientes de este proyecto actualmente son una herramienta imprescindible en muchos proyectos de investigación, sirviendo como referencia y control de la variabilidad genómica humana, ya que se logró encontrar el 98% de los SNP con una frecuencia mayor del 1 % en la población. En número de bases secuenciadas, hablamos de aproximadamente de 6.000.000 millones de bases, cuyo manejo y análisis supusieron un gran reto, incluso para las infraestructuras de datos que tenían que almacenar esta información (The 1000 Genomes Project Consortium, Clarke, et al. 2012) (The 1000 Genomes Project Consortium, McVean, et al. 2012).

Pronto se vio que esta enorme cantidad de información se quedaba corta a la hora de afrontar retos futuros de salud (Heath et al. 2008), como la búsqueda de asociaciones a enfermedades raras y variantes genéticas debidas a diferencias entre poblaciones. Esto a dado lugar a que, en la actualidad, se esté llevando a cabo el proyecto del millón de

genomas por parte de 13 países de la Unión Europea. En este proyecto se esperan recopilar más de un millón de genomas y tenerlos disponibles para el año 2022.

Todo este número de genomas secuenciados no hubiera sido posible sin el abaratamiento de los costes de secuenciación que han supuesto las tecnologías NGS (Figura 1), lo que ha permitido que, en la actualidad, la secuenciación de genomas completos esté al alcance de cualquier grupo (Loman and Pallen 2015).

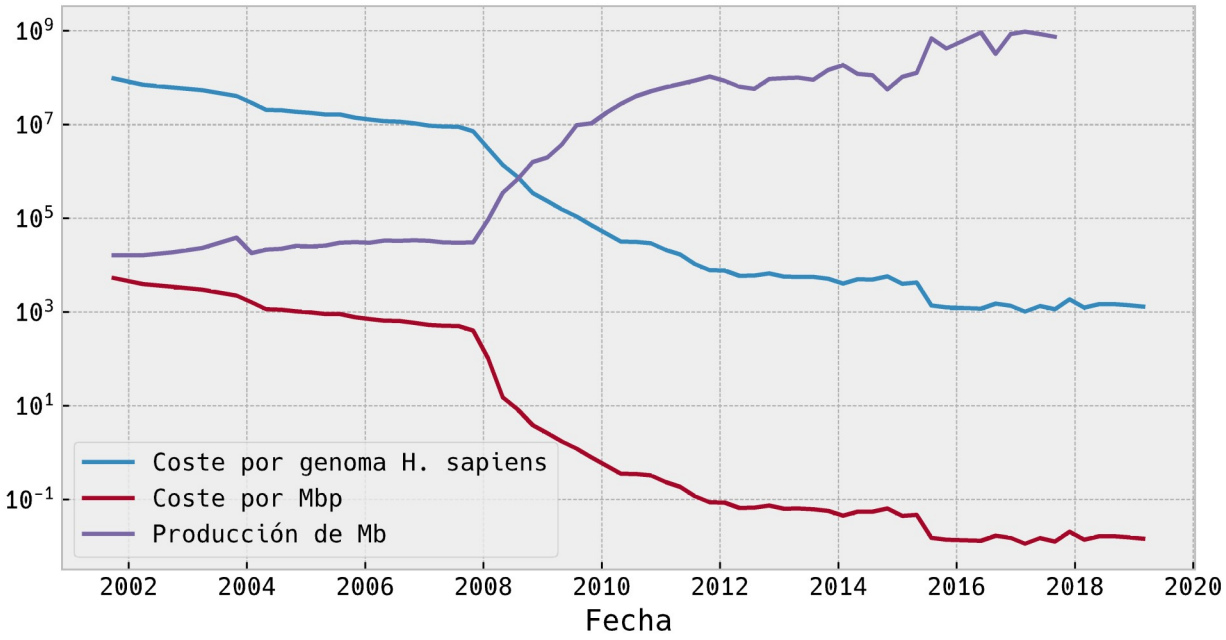


Figura 1: Coste de secuenciación de un genoma humano completo (dólares), coste de secuenciación por Mbp (dólares) y producción de secuencias en Mb, National Human Genome Research Institute, NIH.

En la actualidad existen más de 370.000 organismos secuenciados, según la base de datos GOLD (Genomes Onlines Database) (Mukherjee et al. 2019), muchos de los cuales tienen secuenciadas varias variedades, cepas o individuos, por lo que el número total de genomas únicos puede multiplicar varias veces esta cifra (Figura 2).

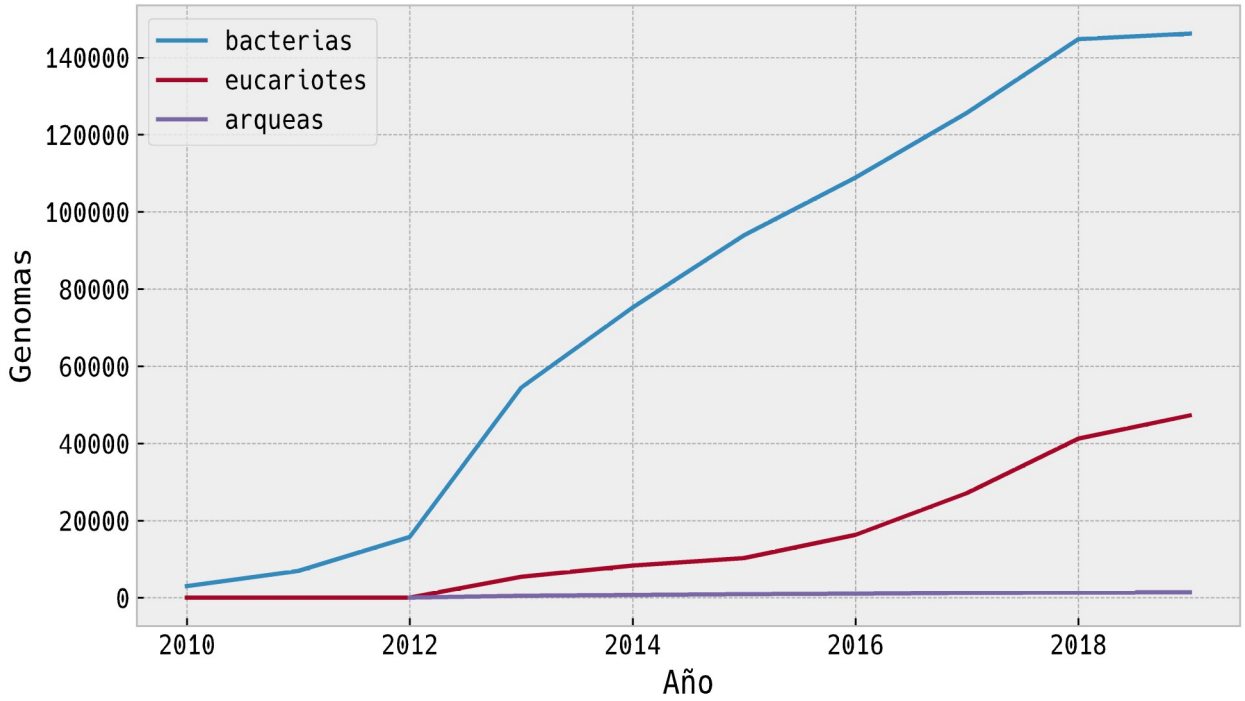


Figura 2: Nº total de proyectos en GOLD (Genomes Online Database, Joint Genome Institute) por año separados por grupo taxonómico.

Por otra parte las tecnologías de secuenciación siguen avanzando, y actualmente las antiguas NGS han pasado a denominarse tecnologías de secuenciación de segunda generación, ya que se están asentando las denominadas tecnologías de tercera generación como Oxford Nanopore (Deamer, Akeson, and Branton 2016) y PacBio (Rhoads and Au 2015). Estas tecnologías se centran en obtener la secuencia de nucleótidos a partir de una sola molécula de DNA, lo que evita los sesgos provocados por el uso de PCR y la sincronización necesaria en las tecnologías de segunda generación (Schadt, Turner, and Kasarskis 2010). Las tecnologías de 3ª generación son capaces de secuenciar cadenas muy largas de DNA, de varios miles de pares de bases, aunque actualmente con el hándicap de una mayor tasa de error que las tecnologías de 1ª generación, lo cual suele afrontarse complementando con datos obtenidos de las tecnologías anteriores.

A pesar de toda la información que se puede producir mediante la secuenciación de un organismo, este paso solo es el inicio, y la base para los subsiguientes análisis que se suelen realizar sobre los datos genómicos. Una vez se ha conseguido secuenciar el genoma de un organismo, el siguiente paso consiste en el ensamblado de las lecturas para formar secuencias más largas. Sobre el ensamblado se realiza la anotación o

búsqueda de elementos funcionales, siendo habitual la localización de los DNA codificantes de proteínas, así como las regiones correspondientes a los genes no codificantes de proteínas (Kersey et al. 2016). Una vez localizados los genes, el siguiente paso consiste en asignar una función a las proteínas que codifican, pudiendo incluir en su anotación características como la localización celular y una estructura (Figura 3). Estos primeros pasos en el análisis de un genoma son fundamentales, ya que de la calidad de esta información dependerán los subsiguientes análisis y estudios.

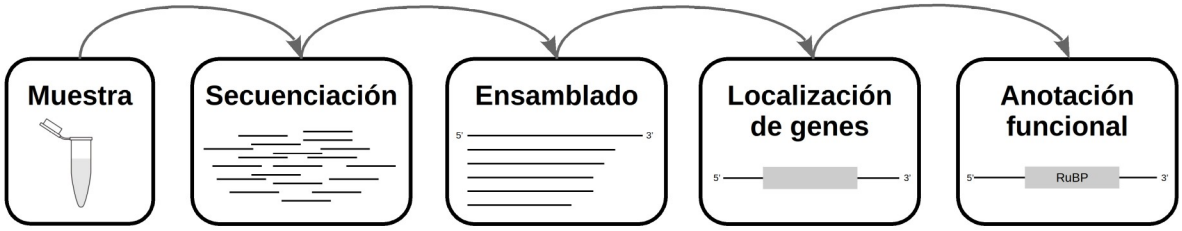


Figura 3: Esquema de análisis *de novo* de una secuencia nucleotídica

La necesidad de análisis y gestión de toda esta enorme cantidad de información, dio lugar ya en los años 60 a la bioinformática, con la pionera Margaret Dayhoff. Esta disciplina científica se encuentra en la intersección de la biología molecular y la informática. Su dominio lo constituye, esencialmente, el desarrollo y aplicación de herramientas computacionales para la recolección, almacenamiento, manejo, análisis e interpretación de la información evolutiva, estructural y funcional contenida en la secuencia de las biomoléculas, con el objetivo final de tener aplicación directa en estudios relacionados con los sistemas biológicos (Mount, D.W. 2001).

1.1.2 Ensamblado

Las tecnologías de secuenciación no producen directamente la secuencia del genoma de un organismo. El resultado de estas tecnologías son múltiples (repetidas) pequeñas secuencias que proceden de las diferentes regiones a lo largo de todo el genoma. Estas pequeñas secuencias se denominan lecturas (del ingles “reads”). Esto requiere que se haga necesario un siguiente paso de ensamblado de las lecturas, para montar el genoma del organismo de interés.

En la actualidad, según el tipo de lecturas que ensamblan, podemos encontrar por un lado herramientas bioinformáticas capaces de trabajar con lecturas cortas (aproximadamente de unas 300 pb) provenientes de tecnologías de secuenciación de segunda generación como: SOAPdenovo (Luo et al. 2012), Velvet (Zerbino and Birney 2008), SPAdes (Bankevich et al. 2012). Por otro lado más recientemente, y acompañando a las tecnologías de secuenciación de tercera generación, encontramos ensambladores que pueden trabajar con lecturas largas (hasta varios miles de pb). Dentro de este grupo están: Canu (Koren et al. 2017), miniasm (Li 2016) y FALCON (Chin et al. 2016).

No obstante, recientemente se ha mostrado que la combinación de estas cadenas largas (Oxford nanopore y PacBio) junto con las lecturas más cortas de Illunima, pueden mejorar el ensamblado de genomas muy bien estudiados, como es el caso de *Caenorhabditis elegans*, (Yoshimura et al. 2019). También se ha empleado esta combinación para mejorar otro tipo de ensamblados, como los de las lecturas de ARN usadas en experimentos de transcriptómica denominados RNA-seq (Au et al. 2012). Normalmente, el ensamblado se optimiza con la corrección de la mayor tasa de error de las secuencias largas, mediante lecturas de secuencias cortas, con menor tasa de error. Esto ha puesto en evidencia los numerosos errores existentes en secuencias que se creían bien conocidas como en el caso de *C. elegans*, donde con la nueva secuenciación se han encontrado 53 posibles nuevos genes, o el de *Acinetobacter baumannii,* donde 400 regiones codificantes de proteínas se predijeron con un tamaño diferente y otras 200 fueron eliminadas (Hamidian et al. 2019).

Estas mejoras se deben sobre todo a que las secuencias largas evitan los problemas que tienen las secuencias cortas en zonas de DNA repetidas y de baja complejidad (Schadt, Turner, and Kasarskis 2010). Estas ventajas han hecho que versiones recientes de ensambladores añadan la opción de poder combinar estos dos tipos de lecturas de forma automática, como puede hacer actualmente SPAdes (Antipov et al. 2016).

1.2 Localización de genes codificantes de proteínas

La localización de genes ha sido un reto desde los inicios de la secuenciación de genomas y en especial de los genes que codifican para secuencias de aminoácidos y son traducidos a proteínas (genes codificantes de proteínas). Aunque ésta puede ser una definición canónica, no siempre se ajusta a la realidad y los límites entre genes codificantes y no codificantes de proteínas no siempre están claros (Gerstein et al. 2007).

1.2.1 Procariotas

En organismos procariotas, inicialmente se confiaba en la búsqueda de similitud entre secuencias para localizar las nuevas regiones codificantes, de tal forma que si la región del nuevo genoma se parecía lo suficiente a algún gen codificante conocido, se consideraba que en esa nueva región había un gen codificante. Para realizar estas comparaciones, en un principio, se utilizaron algoritmos de programación dinámica (Needleman and Wunsch 1970; Smith and Waterman 1981) que son más exhaustivos, pero se acabaron imponiendo algoritmos heurísticos, que realizan una aproximación a la solución óptima como BLAST (S. F. Altschul et al. 1990) y FASTA (Pearson and Lipman 1988), que son lo que se siguen utilizando en la actualidad. Todos estos algoritmos se basan en la búsqueda de similitudes significativas entre secuencias. Para valorar esta semejanza emplean matrices de puntuación asignando un valor o penalización para cada sustitución en las cadenas de DNA o aminoácidos. Entre estas matrices de puntuación se encuentran las PAM (Dayhoff, M. O 1979), basadas en un modelo evolutivo de sustitución, y las más recientes y utilizadas BLOSUM (Henikoff and Henikoff 1992), las cuales dan más peso a sustituciones frecuentes encontradas en motivos o dominios de familias proteicas.

Estas técnicas de búsqueda de similitud entre secuencias tienen el problema de que, debido al enorme aumento de en el número de secuencias que se han incorporado a las bases de datos en los últimos años, la búsqueda de similitud entre secuencias supone un problema en términos de tiempos de computación, si se quiere utilizar una base de datos actualizada. Además tienen el hándicap de que si en la base de datos de referencia no existe una proteína suficientemente similar, no se detectará esa región codificante.

Para solventar estos problemas se desarrollaron métodos *ab initio, los cuales* buscan señales de secuencia que marcan el inicio y fin de las regiones codificantes. Inicialmente se crearon algoritmos basados en Modelos Ocultos de Markov como GeneMark (Borodovsky and McIninch 1993) y GLIMMER (Salzberg et al. 1998) (A. Delcher 1999) (A. L. Delcher et al. 2007). Más adelante, para solventar la falta de precisión a la hora de detectar la posición de inicio en genomas con un alto contenido en GC, se desarrolló Prodigal (Hyatt et al. 2010) que es capaz de usar los sitios de unión al ribosoma (RBS en sus siglas en ingles) existentes en el genoma, mejorando de esta forma la precisión a la hora de localizar los genes codificantes de proteínas.

Con estos programas se pueden llegar a predecir correctamente más del 95% de los genes codificantes de proteínas, pero también predicen erróneamente algunas regiones no codificantes, como sucede por ejemplo con las regiones ribosómicas. Para solucionar estos falsos positivos, se han incorporado estas herramientas a flujos de trabajo automatizados de localización de genes, que tienen en cuenta tanto los genes codificantes como los no codificantes, como RAST (Aziz et al. 2008) y Prokka (Seemann 2014).

1.2.2 Eucariotas

En organismos eucariotas la predicción de regiones codificantes se complica debido a la mayor complejidad en la estructura de su genoma (Figura 4), sobre todo por la presencia de intrones, exones, sitios de splicing enhancers/potenciadores y las grandes regiones no codificantes que contienen, y que separan a las regiones codificantes que pueden ser muy cortas, siendo incluso de la longitud de una base (Abebrese et al. 2017).

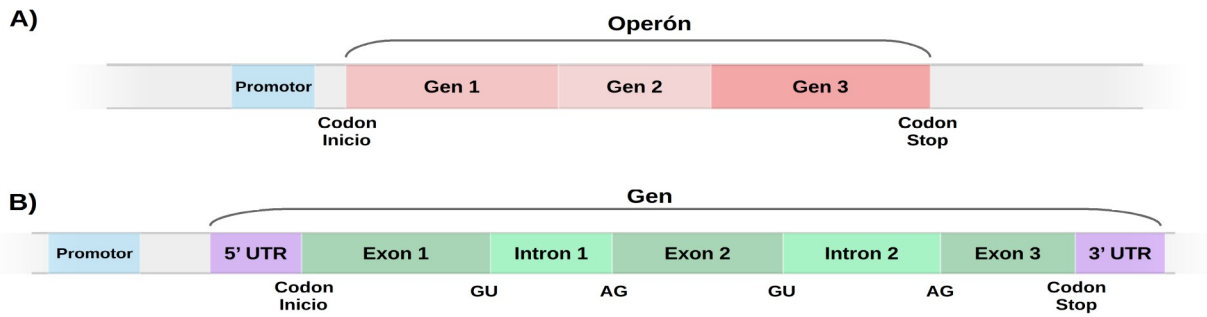


Figura 4: Estructura general de los genes de procariotas (A) y eucariotas (B)

Inicialmente también se usaban algoritmos basados en la búsqueda de similitud de secuencias, pero se pasó rápidamente a los métodos *ab initio* debido al aumento del tamaño de las bases de datos y a que la longitud de los genomas de eucariotas hacía que estos métodos fueran impracticables.

Entre estos métodos *ab initio* se encuentran geneID (Guigó et al. 1992), que se basa en detección de señales y reglas jerarquizadas, GLIMMER HMM (Majoros, Pertea, and Salzberg 2004) que utiliza modelos ocultos de Markov para detectar señales (HMM) y AUGUSTUS (Stanke et al. 2004), que aplica un método híbrido entre búsqueda de similitud entre secuencias y modelos ocultos de Markov.

En el mejor de los casos la sensibilidad de estos algoritmos no supera el 85% a nivel de exon, y si lo llevamos a nivel de gen la sensibilidad cae por debajo del 70% con precisiones aún más bajas, por lo que el reto de localizar *in silico* genes codificantes de proteínas en genomas de eucariotas está lejos de estar resuelto (Goodswen, Kennedy, and Ellis 2012).

Cuando los organismos han sido muy estudiados, como es el caso de los organismos modelo *S. pombe*, *D. melanogaster*, *C. elegans,* esta falta de sensibilidad y especificidad de los algoritmos se ha solventado parcialmente. Estos organismos se han estudiado de forma exhaustiva por diferentes métodos y por lo tanto se conocen la mayoría de genes presentes en sus genomas, existiendo bases de datos específicas para ellos como WormBase (Harris et al. 2010), FlyBase (Tweedie et al. 2009) y PomBase (Wood et al. 2012). El hecho de que estos genomas estén muy bien estudiados no significa que se hayan encontrado todos sus genes (Figura 5). Por ejemplo, el número de genes

codificantes de proteínas en *C. elegans* ha ido variando a lo largo de los años, pasando de unos 19000 detectados en los análisis iniciales, a más de 21000 predichos en la última versión del genoma de *C. elegans* (WS272) disponible en WomBase (Dubaj Price and Hurd 2019).

Los genes codificantes de proteínas (una vez procesados los intrones) se localizan dentro de marcos abiertos de lectura (ORFs en sus siglas en ingles). Estos ORFs son regiones del genoma que comienzan con un codón de inicio y terminan con un codón de stop y no presentan ningún codón de stop entre ambos. Esto indica que esta región puede ser traducida a aminoácidos.

Se puede afirmar que de forma generalizada existe una subestimación del número de genes codificantes de proteínas localizados en marcos abiertos de lectura (ORFs) no canónicos, de pequeño tamaño, pseudogenes o con codones de inicio alternativos. Estas proteínas además pueden jugar un papel importante en sistemas clave para el organismo como la respuesta inmune (Jackson et al. 2018). Véase por ejemplo el gran aumento en el número de genes de C. elegans a partir de 2018 (Fig. 5b).

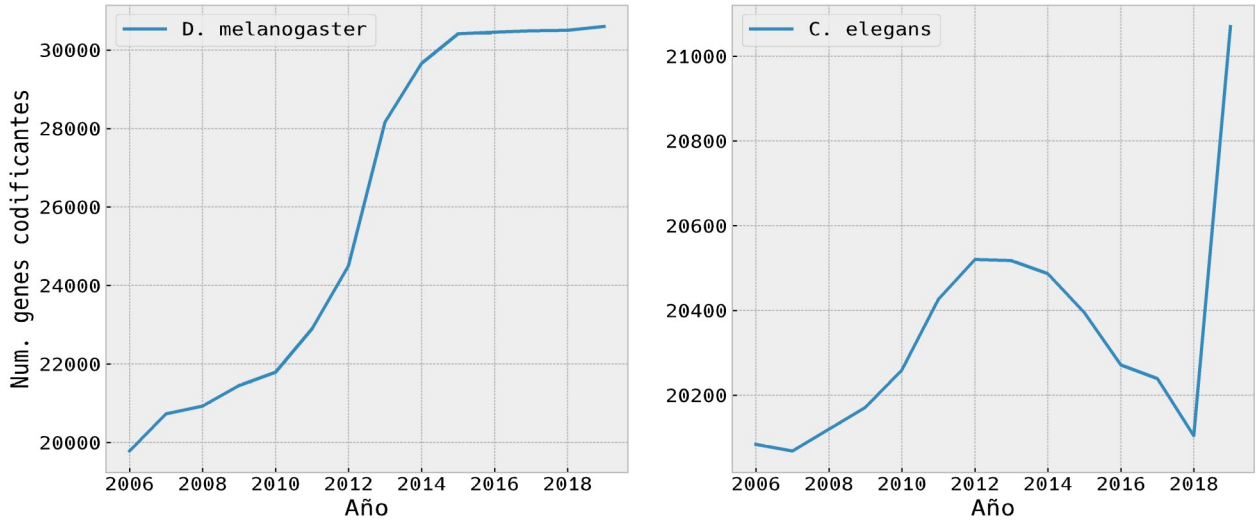


Figura 5: Evolución del número de genes codificantes de proteínas en las bases de datos de FlyBase y WormBase para *D. melanogaster* y *C. elegans* respectivamente. Datos obtenidos del histórico procedente del ftp de FlyBase y WormBase respectivamente.

1.2.3 Marcos abiertos de lectura cortos

Trabajos recientes han mostrado un nuevo componente de los genomas, los marcos abiertos de lectura cortos, o en sus siglas en inglés sORFs (Small Open Reading Frames). Se considera un ORF como pequeño, cuando tiene 100 codones o menos, y se ha visto que pueden codificar péptidos biólogicamente activos que suelen tener un papel regulatorio en la expresión de genes canónicos (Andrews and Rothnagel 2014), o en la regulación de procesos fundamentales como el transporte de calcio en *D. melanogaster*, llegando a afectar a la contracción del corazón en estos organismos (Magny et al. 2013).

Los sORF pueden ser intergénicos, pero en su mayoría se encuentran en las regiones 5'UTR, 3'UTR, dentro del propio gen que regula, pero en un marco de lectura diferente, en regiones intrónicas, e incluso dentro de RNAs no codificantes (ncRNA) (Figura 6). Además los sORF suelen tener como inicio codones alternativos al AUG (Hellens et al. 2016). La función biológica de gran mayoría de estos sORF tiene que ser investigada y debido al gran número en que se encuentran en los genomas de eucariotas, es imprescindible discernir entre aquellos sORF que pueden codificar péptidos y los que no.

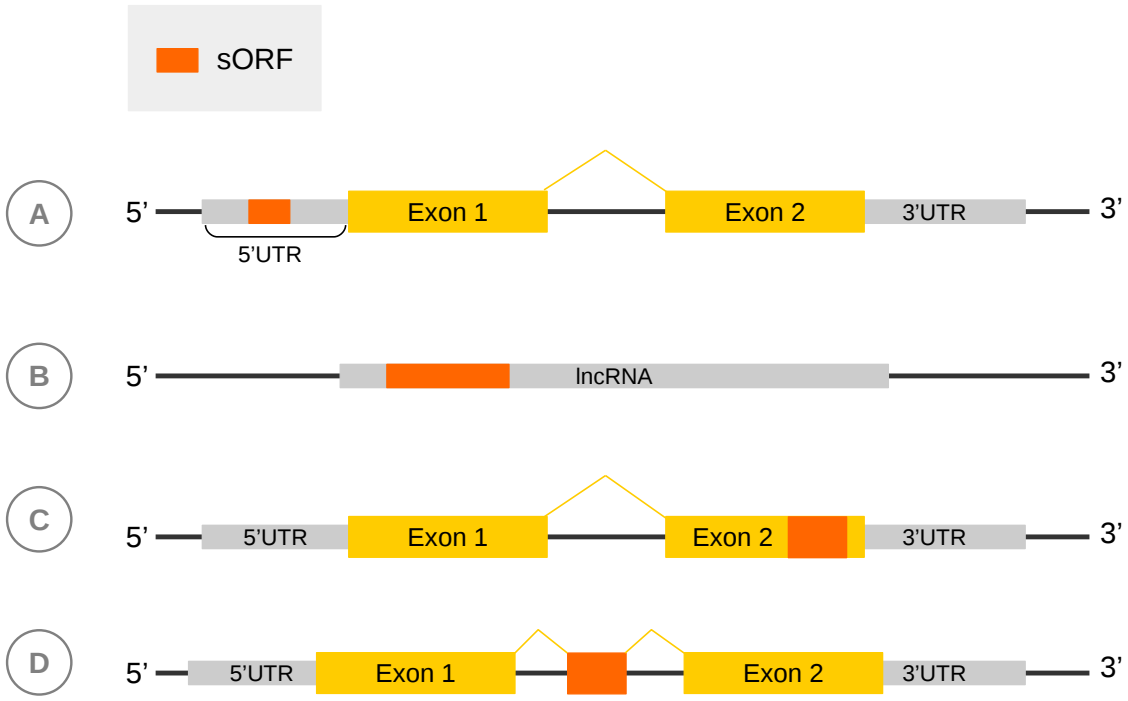


Figura 6: Posibles localizaciones de sORF que pueden traducirse a aminoácidos. (A) sORF en las regiones no traducidas de genes canónicos. (B) sORF dentro de lncRNA. (C) sORF dentro de exones canónicos pero en diferente marcos de lectura al correspondiente a la proteína canónica. (D) sORF dentro de regiones intrónicas de genes canónicos que se pueden traducir mediante splicing alternativo.

Dentro de un genoma complejo como el de *D. melanogaste*r se proponen que pueden existir más de 550.000 sORF de al menos 10 codones de longitud, de los cuales no todos tienen una función biológica (Pueyo, Magny, and Couso 2016). Por lo tanto discernir entre aquellos que se traducen y codifican un péptido con función y los que no, es una tarea primordial y abrumadora, ya que el número de péptidos activos que forman parte de los organismos se puede multiplicar varias veces. Los algoritmos que identifican genes canónicos no tienen en cuenta estas secuencias tan cortas (<= 100 aa), ya que se considera que la mayoría de genes codificantes lo hacen para proteínas más largas, pero si se dieran por válidos estos sORFs provenientes de predicciones automáticas, las bases de datos se inundarían de falsos positivos, a menos que se comprobaran experimentalmente los resultados (Pueyo, Magny, and Couso 2016). Todo esto hace que se abra un nuevo campo en la genómica, ya que existen evidencias de que una gran parte de estos pequeños péptidos pueden tener funciones importantes en los genomas.

Los primeros algoritmos computacionales que se usaron para predecir los sORFs, eran programas basados en parte en la similitud entre secuencias, y en características intrínsecas de las mismas, como puede ser la composición de hexámeros, la frecuencia de dicodones y la proporción entre sustituciones de aminoácidos sinónimos y no sinónimos, como son Coding Region Identification Tool Invoking Comparative Analysis (CRITICA) (Badger and Olsen 1999), Coding Potential Calculator (CPC) (Kong et al. 2007) y sORF Finder (Hanada et al. 2010). Otro grupo de algoritmos se basan en estudios filogenéticos de las secuencias, ya sea mediante filogenética basada en modelos ocultos de Markov (phylo-HMM), como PhastCons (Siepel 2005), o basados en modelos filogenéticos de sustitución de codones como PhyloCSF (Lin, Jungreis, and Kellis 2011).

Actualmente estos algoritmos no son lo suficientemente precisos, siendo necesario combinarlos con otros métodos experimentales de detección de traducción como son Ribo-Seq y espectrometría de masas. Pero aún combinando estos métodos puede ser difícil diferenciar los resultados del ruido de fondo de traducciones aleatorias (Pueyo, Magny, and Couso 2016).

1.3 Anotación funcional de genes codificantes de proteínas

Una vez localizados los genes del genoma, otro paso fundamental es predecir la función de éstos. El enorme crecimiento en el número de secuencias genómicas ha provocado un incremento similar en el número de secuencias proteicas. Esto ha provocado que la gran mayoría de proteínas estén anotadas de forma computacional, sin ninguna validación experimental (Figura 7). Lo que hace que la precisión de los algoritmos encargados de asignar funciones a las proteínas sea fundamental para el conocimiento general de los genomas y para disciplinas más aplicadas como la biomedicina y la farmacéutica (Radivojac et al. 2013).

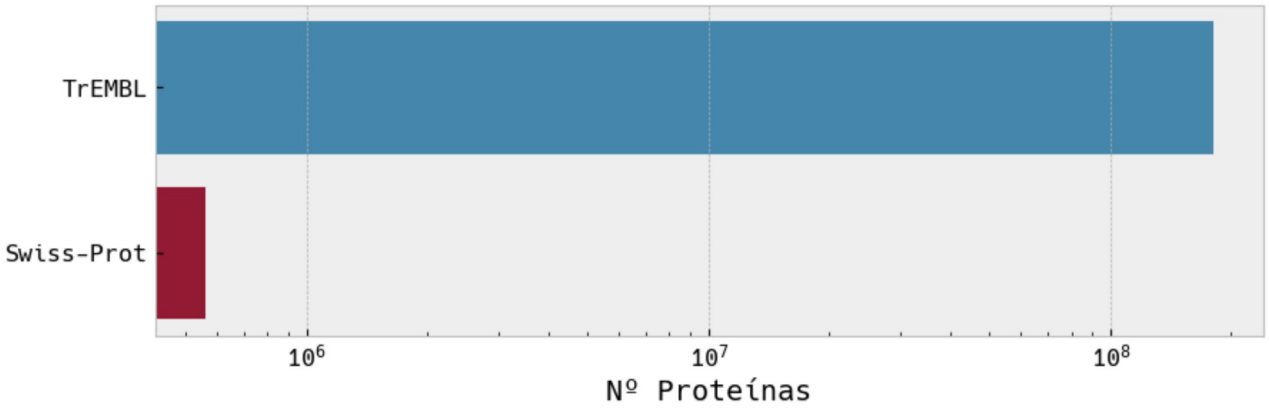


Figura 7: Número de proteínas en UniProtKB revisadas y anotadas manualmente (Swiss-Prot) y anotadas automáticamente y no revisadas (TrEMBL)

En un principio se usaban algoritmos como BLAST (BLASTP para proteínas), para la búsqueda de similitud entre secuencias. Si las proteínas eran suficientemente similares,la proteína problema recibía las anotaciones de la proteína similar procedente de la base de datos. Como evolución de BLAST, que usa matrices de peso por posición para puntuar la similitud entre 2 secuencias, apareció PSI-BLAST que se basa en perfiles de secuencia, con un número de secuencias de una base de datos que se consideran similares a la proteina problema (S. Altschul 1997). Estos algoritmos realizan búsquedas de similitud frente a bases de datos de proteínas como UniProt Knowledgebase (The UniProt Consortium 2017), o bases de datos de secuencias nucleotídicas como

GenBank (Benson et al. 2012). Estas bases de datos intentan ser un compendio de todas las secuencias conocidas. Surgen de esfuerzos internacionales y de la integración de diferentes bases de datos que cruzan periódicamente su información para mantenerse actualizadas. UniProtKB surge de la colaboración entre NCBI (National Center for Biotechnology Information), EBI (European Bioiformatic Institute) y SIB (Swiss Instiure of Bioinformatic). GenBank pertenece al NCBI y al mismo tiempo forma parte de la "International Nucleotide Sequence Database Collaboration" donde participan las bases de datos DDBJ (DNA DataBank of Japan) y ENA (European Nucleotide Archive).

En un siguiente paso, junto a los programas de búsqueda de similitud que usaban BLAST, se empezaron a usar programas basados en Modelos Ocultos de Markov, como hace InterProScan (Jones et al. 2014), el cual emplea los algoritmos de TMHMM y HMMER3, ademas de BLAST, para buscar en un gran número de bases de datos como son: Pfam (El-Gebali et al. 2019) y TIGRFAMs (Haft et al. 2012), constituidas por familias de proteínas generadas por alineamientos múltiples y modelos ocultos de Markov; SMART (Letunic and Bork 2018), Prosite (Sigrist et al. 2012) y PRINTS (Attwood 2000), que son bases de datos de dominios y motivos proteicos; además de PIRSF (Wu 2004), Panther (Mi et al. 2019) y HAMAP (Pedruzzi et al. 2015), que son bases de datos de familias de proteínas revisadas manualmente; y finalmente ProDom (Bru 2004) y CATH-Gene3D (Dawson et al. 2017), que son bases de datos de proteínas que incluyen información 3D, al igual que SUPERFAMILY (Pandurangan et al. 2019). Todas estas bases de datos se han integrado en InterPro (Mitchell et al. 2015).

Más recientemente aparecieron nuevos algoritmos como DIAMOND (Buchfink, Xie, and Huson 2015). Este algoritmo, realizando un BLASTX, es decir buscando la traducción de los 6 marcos de lectura de una secuencia nucleotídica frente a una base de datos de proteínas, puede ser hasta 20.000 veces más rápido que BLAST alineando secuencias, pero a costa de tener una menor sensibilidad. Para conseguir este incremento se basa en estrategias como la reducción del alfabeto, espaciado e indexación de semillas.

DIAMOND se emplea en aplicaciones como eggNOG-mapper (Huerta-Cepas et al. 2017), para asignar funciones realizando búsquedas de similitud en la base de datos de ortólogos eggNOG (Huerta-Cepas et al. 2019). Se considera que dos genes son ortólogos cuando se han originado a partir de un gen ancestral procedente del genoma del último

antepasado común. Esta base de datos proporciona información funcional y filogenética de los grupos de ortólogos que contiene.

En la actualidad se siguen usando programas que se basan en BLAST para alinear secuencias, sobre todo cuando realizan un BLASTP (alineamiento de proteínas frente a bases de datos de proteínas), ya que en estos casos la rapidez de DIAMOND no sobrepasa a la de BLAST para valores de igual sensibilidad.

Existen otros programas de anotación funcional pero suelen tener otras limitaciones como es el caso de Argot2.5 (Lavezzo et al. 2016), ESG/PFP (Khan et al. 2015), BAR-PLUS (Piovesan et al. 2011) y FastAnnotator (Chen et al. 2012), que solo están disponibles mediante una aplicación web donde el número de proteínas a anotar está limitado. Trinotate (Bryant et al. 2017) es otro programa que sólo permite anotar transcriptomas, y Blast2GO (Conesa et al. 2005) que necesita una licencia de pago para acceder a todas sus funcionalidades.

A estas limitaciones hay que sumarle el uso que hacen algunos programas de grandes bases de datos con millones de proteínas, lo que hace que los tiempos de alineamiento aumenten año tras año conforme estas bases de datos aumentan. Un ejemplo de esto lo tenemos en BAR-PLUS que hace uso de UniProt completa, y por tanto teóricamente su tiempo de ejecución crecería exponencialmente, en paralelo al crecimiento actual de esta base de datos.

Todos estos programas se basan por tanto, de algún modo u otro, en la búsqueda de similitud de las secuencias problemas frente a diferentes bases de datos, con lo que volvemos al problema del crecimiento y actualización de las bases de datos y los problemas que provoca el aumento del tiempo de cálculo y desactualización de las anotaciones presentes en las mismas.

Sma3s v1 & AnABlast

Previo al desarrollo de esta tesis se desarrolló el anotador funcional de proteínas Sma3s, por *Sequence Massive Annotation using 3 Modules* (Munoz-Merida et al. 2014). Esta aplicación permite la anotación automatizada de grandes cantidades de secuencias biológicas (nucleótidos o aminoácidos). Para realizar la anotación se basa en tres módulos que de forma sucesiva anotan la secuencia problema. En el primer módulo se extraen las anotaciones de secuencias muy similares, el segundo módulo obtiene las anotaciones mediante la búsqueda inicial de ortólogos y el tercero usa los términos enriquecidos procedentes de grupos de secuencias similares a la secuencia problema. Como base de datos de referencia puede usar tanto UniprotKB completa como cada uno de sus secciones por separado, Swiss-Prot y TrEMBL, o incluso sólo secuencias de grupos taxonómicos concretos.

Un año más tarde se desarrolló el algoritmo AnABlast para la búsqueda de regiones codificantes de proteínas en secuencias genómicas (Jimenez et al. 2015), realizándose las primeras pruebas con regiones intergénicas de la levadura *Schizosaccharomyces pombe.* Esta primera versión del algoritmo se desarrolló con el objetivo de probar que la acumulación de alineamientos no significativos de BLAST de proteínas sobre secuencias genómicas no tenían una distribución azarosa. En estas primeras versiones se realizaba un BLASTX, es decir se alineaban las secuencias de los seis marcos de lectura posibles de las regiones intergénicas frente a la base de datos de proteínas UniRef50 (The UniProt Consortium 2017). Esta base de datos proviene de la agrupación, mediante el programa CD-HIT (Fu et al. 2012), de las proteínas de UniProtKB que tienen una identidad mayor o igual al 50% y un solapamiento mayor o igual al 80%. De esta forma, esta base de datos elimina en gran medida la redundancia y disminuye el posible sesgo debido a la sobrerrepresentación en UniProtKB de grupos de secuencias de la misma familia que pueden dar lugar acumulaciones de alineamientos anormalmente altos.

Como se ha descrito anteriormente, la versión inicial de AnABlast es capaz de localizar posibles secuencias codificantes de proteínas en pequeñas regiones genómicas (como son las regiones intergénicas de *S. pombe*). Por otro lado, la primera versión de Sma3s es

capaz de anotar estas secuencias. Por lo tanto, la combinación de estas dos herramientas podría ser empleada en los pasos posteriores a la secuenciación y ensamblado de secuencias genómicas y esta posibilidad ha sido explorada y desarrollada durante esta tesis.

2. Objetivos

2.1 Objetivos

El objetivo planteado en esta tesis se centra por un lado en el desarrollo y mejora del algoritmo AnABlast para la localización de genes codificantes de proteínas en genomas completos, así como la puesta en práctica del mismo sobre genomas de diferentes organismos eucariotas. Y por otro lado se trata de mejorar la anotación masiva de secuencias biológicas mediante la herramienta computacional Sm3s, así como el desarrollo de una aplicación web que permita el uso combinado de estas dos aplicaciones de forma rápida y abierta para toda la comunidad científica. Este objetivo general se puede dividir en los siguientes apartados:

1. Puesta a punto de la búsqueda de acúmulos de alineamientos no significativos de BLAST en genomas completos y validación del protocolo comparando frente a los resultados ya publicados de las regiones intergénicas de *Schizosaccharomyces pombe.*

2. Mejora de la anotación funcional de secuencias biológicas (nucleótidos y aminoácidos) realizada mediante la aplicación Sma3s, incluyendo la mejora de la sensibilidad y especificidad de las anotaciones, así como el tiempo de ejecución, eliminación de dependencias del programa, facilidad de uso y mejora en las salidas de la aplicación.

3. Optimización de parámetros para localizar los acúmulos de alineamientos no significativos de BLAST para la búsqueda de ORFs en el genoma de *Drosophila melanogaster*.

4. Aplicación de la búsqueda de acúmulos de alineamientos no significativos de BLAST sobre el genoma de *Caenorhabditis elegans,* para buscar y validar genes codificantes de proteínas descartados por otros algoritmos, haciendo hincapié en pequeños ORFs.

5. Desarrollo de una aplicación web que permita el uso fácil, abierto y combinado de AnABlast y Sma3s.

3. Marco teórico

3.1 Marco teórico de las publicaciones

En este apartado se describirá el desarrollo de la tesis y se resumirán brevemente los resultados más importantes de los artículos científicos implicados en la misma. En los artículos que se encuentran adjuntos más abajo, se describen en más detalle todos los desarrollos y resultados obtenidos durante la tesis (Figura 8).

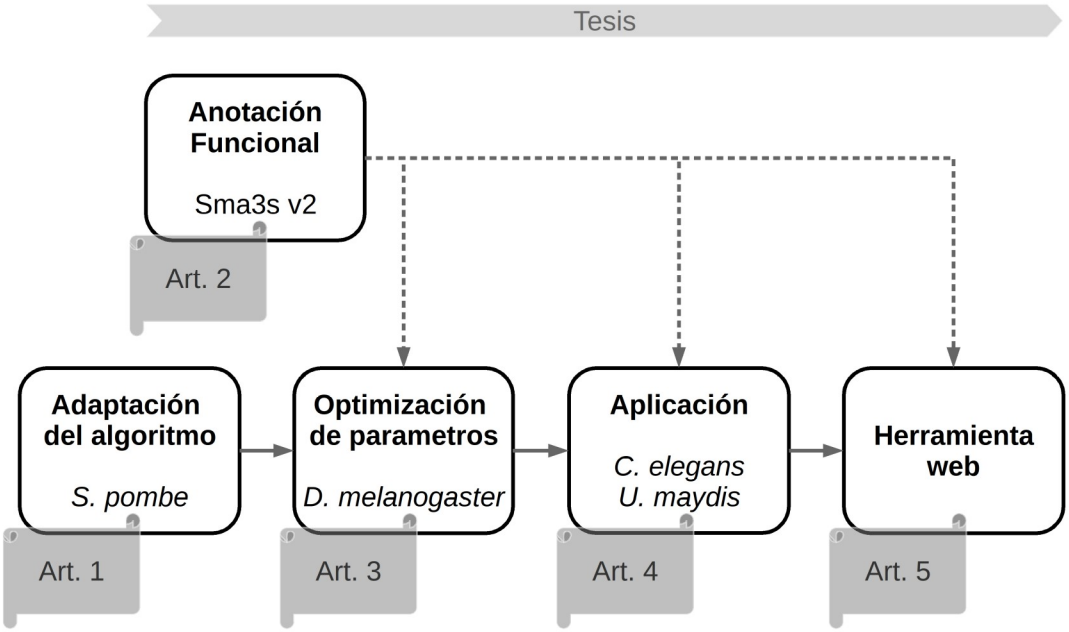


Figura 8: Esquema de los trabajos realizados en la tesis. Art. 1: AnABlast: a new in silico strategy for the genome-wide search of novel genes and fossil regions (previo a esta tesis) . Art.2: Sma3s: A universal tool for easy functional annotation of proteomes and transcriptomes (https://doi.org/10.1002/pmic.201700071). Art. 3 *Drosophila melanogaster* (https://doi.org/10.1186/s12864-020-6632-y). Art. 4: *Caenorhabditis elegans* (enviado a revista). Art. *5: Aplicación* web (DOI https://doi.org/10.1007/978-1-4939-9173-0_12 *).*

El desempeño de este trabajo se inició tras la publicación de los artículos científicos de Sma3s v1 (Munoz-Merida et al. 2014) y AnABlast (Jimenez et al. 2015). En este último artículo se describen las bases del algoritmo AnABlast, que sirve como base para el inicio de esta tesis.

Durante las etapas iniciales de la tesis se rediseñó el algoritmo para su aplicación en genomas completos, ya que previamente solo se había aplicado sobre las regiones intergénicas de *Schizosaccharomyces pombe.* Durante este proceso se pasó a realizar

un TBLASTN en lugar de un BLASTX, ambos provenientes del paquete BLAST+ (Camacho et al. 2009). El segundo algoritmo (utilizado para S. pombe) realiza alineamientos de los seis marcos de lectura de una secuencia nucleotídica frente a una base de proteínas y el primero (implementado en esta tesis) permite alinear una proteína frente a una base de datos de secuencias nucleotídicas, traduciendo sus seis posibles marcos de lectura. Este cambio permitió comparar la base de datos de proteínas UniRef50 frente a genomas completos (siendo las proteínas de UniRef, en este caso, las secuencias problema), ya que de forma inversa (siendo las secuencias genómicas las secuencias problema) el algoritmo no funcionaría debido a limitaciones del mismo por la longitud de las secuencias problema.

Como resultado de este cambio, se emplearon como secuencias problema las más de 24 millones de proteínas existentes en UniRef50, y como base de datos el genoma completo de *Schizosaccharomyces pombe.* Una vez obtenidos los resultados del análisis del genoma completo, se validaron con los obtenidos en las regiones intergénicas en Jiménez et al, mostrando resultados muy similares (Figura 9).

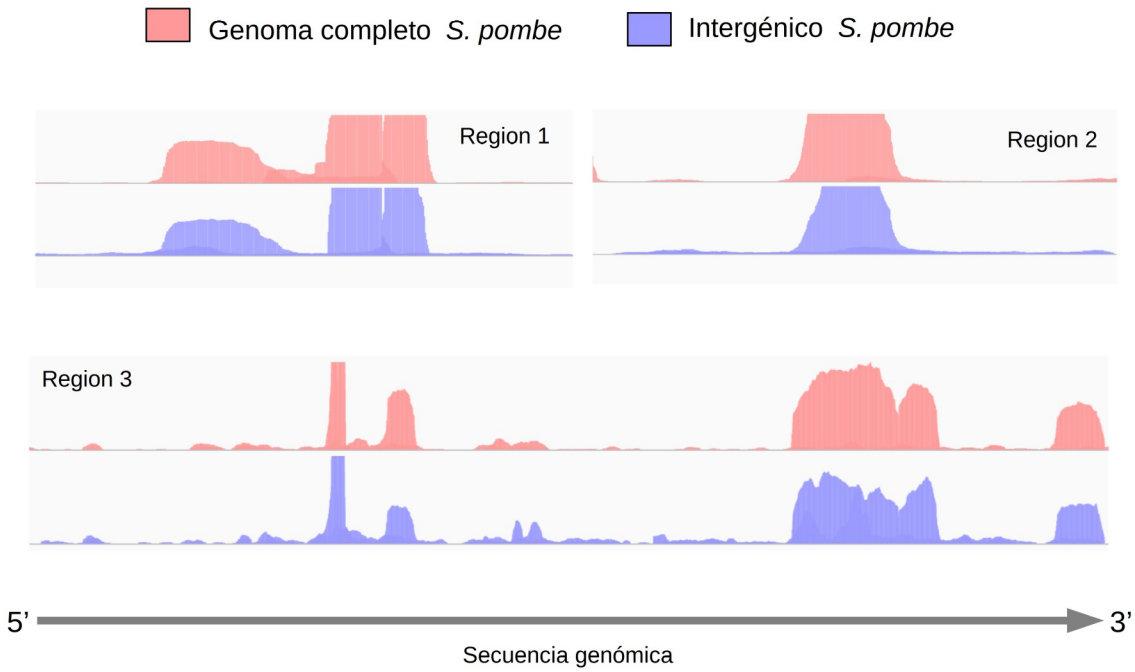


Figura 9: Comparación en tres regiones intergénicas de *S. pombe* de los picos de AnABlast obtenidos mediante la aplicación de este algoritmo sobre el genoma completo (rosa) o exclusivamente sobre las regiones intergénicas (azul).

En resumen, el algoritmo para el análisis de genomas completos consiste en alinear los aminoácidos predichos a partir de los seis marcos de lectura de una secuencia genómica contra la base de datos de proteínas no redundante UniRef50. Cada una de las secuencias representantes de UniRef50 se emplea como secuencia problema y mediante un TBLASTN se alinea contra los 6 marcos de lectura de la secuencia genómica traducidos a aminoácidos. De todos los alineamientos posibles nos quedamos con aquellos que tienen un bitscore de 30, independientemente del p-valor que tengan (protomotivos). Esto supone que la gran mayoría de alineamientos son no significativos (p-value mayor de 10e-5). A partir de estos alineamientos se genera una puntuación para cada posición de cada uno de los seis marcos de lectura, de tal forma que esta puntuación se corresponde con el sumatorio de los alineamientos que se superponen en cada posición. Esta puntuación se traduce en una intensidad de señal o altura y es la agrupación de las alturas de cada posición lo que genera las señales o picos de AnABlast (Figura 10).

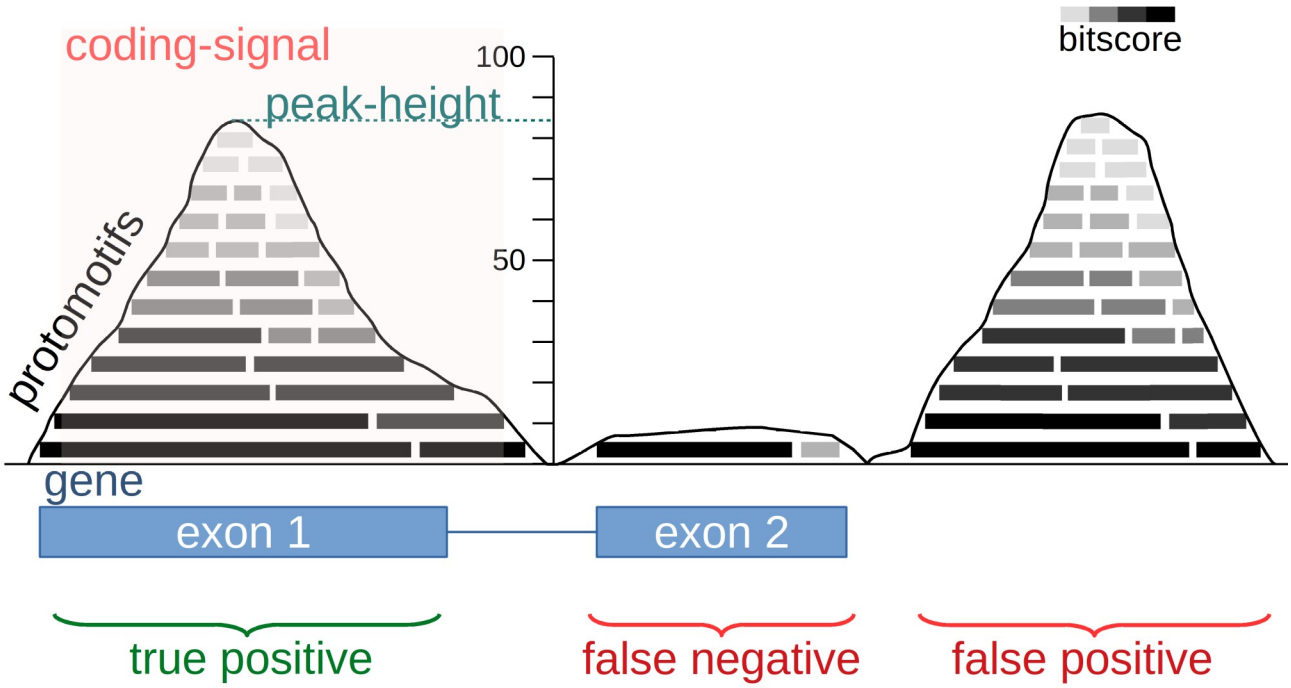


Figura 10: Esquema de los perfiles de AnABlast obtenidos en teórico genoma con un gen con dos exones. El tamaño de los picos es la máxima acumulación de protomotivos en una posición específica del genoma (Alineamientos de BLAST incluidos los que tienen bitscore bajos). Los picos con una acumulación de protomotivos por encima de un cierto límite son considerados como posibles regiones codificantes de proteínas (señales codificantes). Los picos significativos que coinciden con un exon conocido se consideran verdaderos positivos, mientras que aquellos que coinciden con regiones genómicas sin exones conocidos son considerados como falsos positivos. Los exones conocidos que no coinciden con acumulaciones significativas de protomotivos constituyen los falsos negativos.

Tras realizar estos cambios en el algoritmo se procedió a actualizar la aplicación de anotación funcional Sma3s, obteniendo su segunda versión, la cual ofrecía una mejor precisión y un menor tiempo de ejecución.

Para comprobar si los parámetros aplicados en el análisis inicial, de las regiones intergénicas, eran los óptimos para analizar genomas completos se decidió analizar el genoma de un eucariota complejo como es *Drosophila melanogaster.*

Una vez determinados cuales eran estos parámetros óptimos se paso a la aplicación en el genoma de *Caenorhabditis elegans* obteniendo resultados positivos validados *in vivo* con fenotipos obtenidos mediante RNA interferente.

Por último se decidió facilitar el uso de AnABlast a los miembros de la comunidad científica sin conocimientos de bioinformática, mediante el desarrollo de una aplicación web programada en Perl.

3.1.1 Anotación funcional (Sma3s v2)

Como se ha comentado previamente, el objetivo principal de esta tesis consistía en la localización de nuevas regiones codificantes de proteínas. Pero una vez obtenidas estas regiones, su anotación funcional podría constituir un paso fundamental y un apoyo a la posibilidad de que estás regiones realmente son codificantes.

Para realizar esta anotación funcional desarrollamos la segunda versión de la aplicación Sma3s (Munoz-Merida et al. 2014). Esta aplicación se basa en alineamientos locales, mediante BLAST, y enriquecimiento biológico que se combinan mediante tres módulos. Permitiendo un uso sencillo y una alta precisión en las anotaciones.

En su segunda versión, realizada durante esta tesis, se redujeron el número de dependencias de software, hasta precisar sólo la instalación previa de los paquetes BLAST+ y Perl. Además se redujo el tiempo de anotación medio por secuencia problema mediante el uso de una base de datos de proteínas no redundante Uniref90 (The UniProt Consortium 2017). Esta base de datos proviene de la agrupación de todas

las proteínas de UniProtKB usando un umbral mínimo de 90% de identidad y 80% de cobertura (The UniProt Consortium 2017), lo que se consigue mediante la aplicación CD-HIT (Fu et al. 2012). Además, se mejoró la precisión de sus anotaciones, permitiendo asignar nombres de proteínas y descripciones informativas, las cuales son de gran utilidad para proyectos futuros. Por último, para mejorar el análisis posterior, se mejoró la salida de resultados, los cuales ahora se pueden abrir desde un porgrama de hoja de cálculo y permiten analizar clases funcionales de los proteomas o transcriptomas anotados. Todo esto, junto a una mejora en la precisión de los resultados debida a mejoras en la integración de los módulos y el uso de las etiquetas de calidad de las anotaciones presentes en la base de datos. El primero de los módulos realiza una búsqueda de secuencias muy similares y el segundo módulo una búsqueda de ortólogos. A partir de estos dos módulos se asignan, principalmente, los nombres de los genes y las descripciones de las secuencias, evitando la asignación de nombres espurios, mediante un filtrado de los mismos a través de unas sencillas reglas como son la longitud del nombre o la presencia de caracteres extraños. El tercer módulo emplea todos los alineamientos significativos que ha generado la secuencia problema para obtener las anotaciones compartidas. Para la asignación de las anotaciones se emplean los 3 módulos, lo que mejora la sensibilidad de las mismas.

La herramienta bionformática resultante permite realizar la anotación de proteomas o transcriptomas completos a cualquier investigador, dando un resultado en forma de anotaciones funcionales por secuencia, muy fácil de revisar y útil para la realización de análisis posteriores. Para mostrar las mejoras de la versión 2 se anotaron 52 proteomas del genero *Bacillus* de bacterias. Utilizando la hoja de clases funcionales de resultado de cada anotación, se pudieron comparar funciones generales entre las diferentes especies (Figura 11). Estos procesos biológicos provienen de los GO Slim (Gene Ontology Consortium 2004), y representan una visión de los niveles superiores procedentes de las tres ontologías de términos GO presentes en la base de datos Gene Ontology (http://geneontology.org/). Estas tres ontologías son Función Molecular, que describe la actividad, Procesos Biológicos, que describe los objetivos por una o más conjuntos ordenados de funciones moleculares y Componente Celular, que describe la localización, a niveles de componentes subcelulares y complejos macromoleculares.

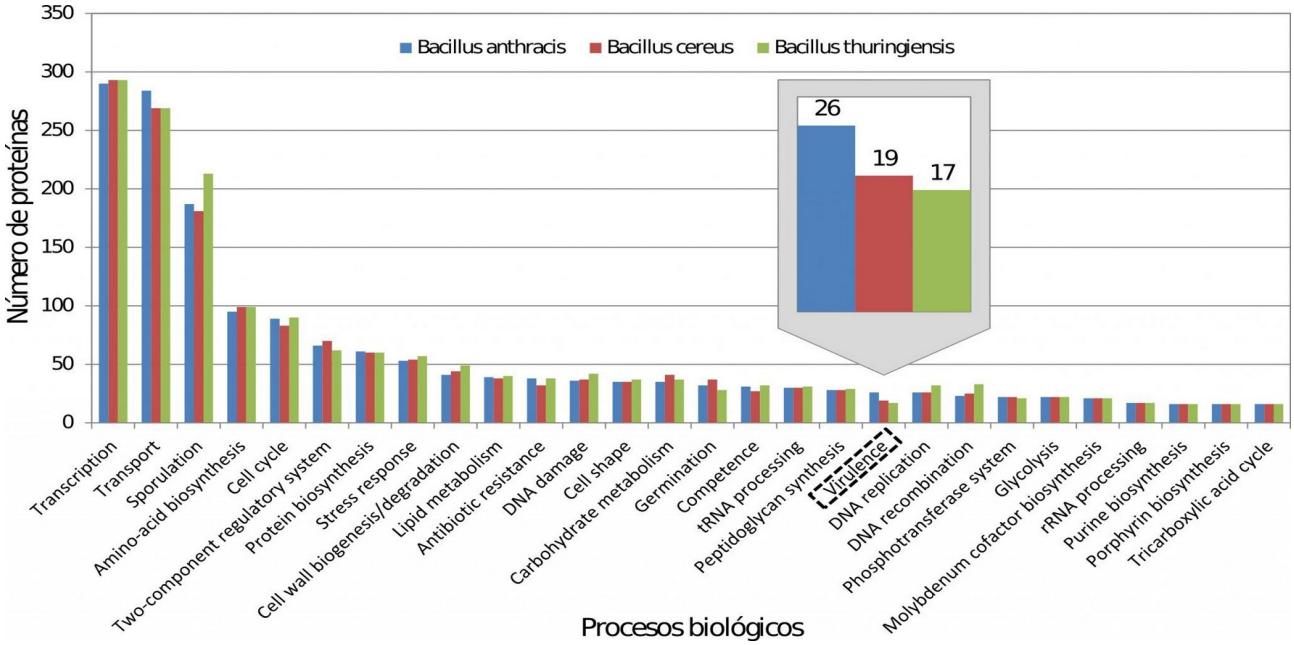


Figura 11: Número de proteínas anotadas por Sma3s en diferentes procesos biológicos para tres proteomas del género *Bacillus*. Se destaca el número de proteínas relacionadas con virulencia, el cual es mayor en *B. anthracis* (agente etiológico del antrax), algo menor en *B. cereus* (patógeno oportunista causante de intoxicaciones alimentarias), y menor aún en *B. thuringiensis*. Es esta última especie destaca el elevado número de proteínas relacionadas con esporulación, seguramente debido a proteínas formadoras de paraesporas denominadas cristales, las cuales presenta específicamente esta especie.

3.1.2 Optimización de parámetros (*Drosophila melanogaster*)

Para ajustar los parámetros del algoritmo y comprobar si los alineamientos de baja puntuación que se obtenian con AnABlast eran debidos al azar, se aplico el algoritmo sobre el genoma de *Drosophila melanogaster.* En un principio se habían empleado todos los alineamientos obtenidos al comparar la base de datos Uniref50 frente al genoma problema, los cuales tuvieran un bitscore >= 30.

Si tomamos como referencia un p-valor considerado como significativo de 10e-5, obtendremos que la mediana de los bitscore para ese valor está cercana a 51 (Figura 12). Para evaluar la precisión de este límite se obtuvieron los picos de AnABlast aplicando un límite de bitscore 31 y 29, siendo en el primer caso más restrictivos que el segundo. Además del bitscore, otro parámetro clave de AnABlast es la altura a partir de la cual se considera que un pico destaca la presencia de una secuencia codificante de proteínas.

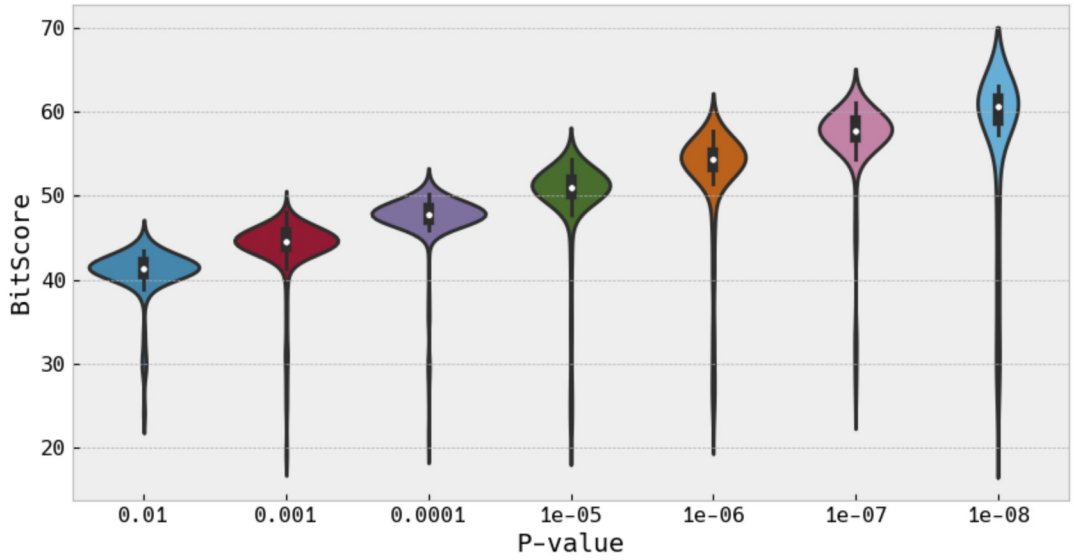


Figura 12: Distibución de los bitscores para diferentes p-valores obtenidos de los 100.000 primeros alineamientos de blast procedentes de los resultados de AnABlast de *D. melanogaster*

Se evaluó tanto la precisión como la sensibilidad de ambos parámetros utilizando dos bases de datos de referencia de diferentes años, 2012 y 2017 (Figura 13) . La especificidad aumenta conforme aumenta el bitscore o la altura mínima de pico, ya que son parámetros más restrictivos, pero en este caso la sensibilidad disminuye. De manera similar la sensibilidad aumentaba cuando se empleaba la base de datos de referencia más actual (2017). Teniendo en cuenta estás variaciones se determinó que los mejores parámetros son un bitscore de 30 y una altura mínima de pico de 70.

Para mostrar los resultados de AnABlast sobre *D. melanogaster* se creó un visualizador genómico usando la herramienta JBrowse (Buels et al. 2016). En este visualizador, además de los picos generados por AnABlast, se pueden activar y desactivar diferentes pistas (tracks) con información de utilidad sobre elementos genómicos como: Genes presentes en FlyBase y Expressed Sequence Tags (EST). Se puede acceder a los resultados de AnABlast con *D. melanogaster*desde la siguiente URL:
http://www.bioinfocabd.upo.es/drome

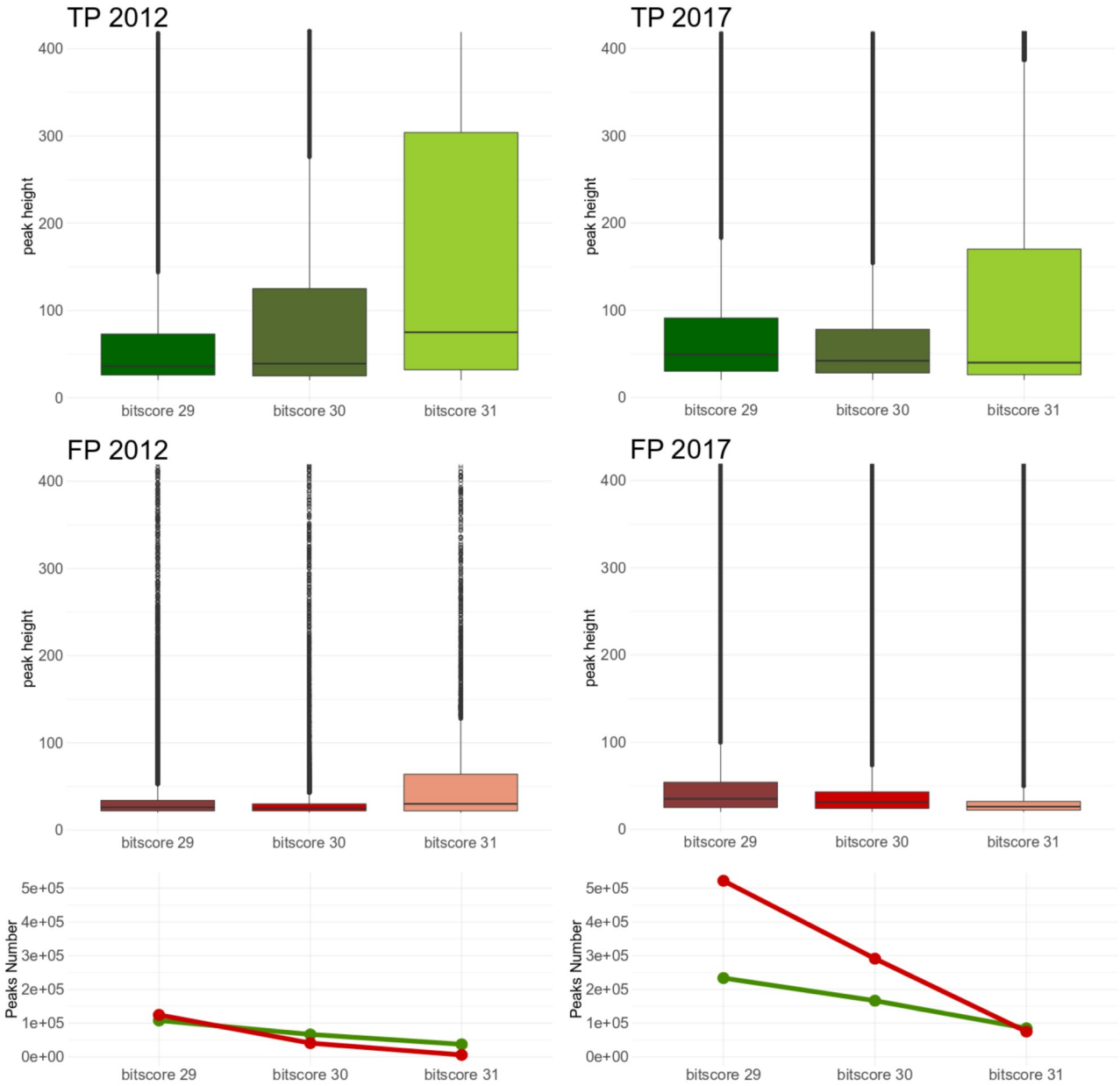


Figura 13: Distribución de la altura de picos y número de señales codificantes encontradas a diferentes valores de bitscore. La distribución de altura de pico esta separada por a) verdaderos positivos y b) falsos positivos, y se muestran por cada distribución de la base de datos (2012 y 2017) y según tres límites de bitscore. Los valores atípicos se muestran como una cadena de puntos por encima de las cajas. c) El número de señales codificantes de verdaderos y falsos positivos para cualquier altura de pico según los correspondientes límites de bitscore (Se muestran el número de picos a partir de una altura mínima de 20).

3.1.3 Aplicación de AnABlast (*Caenorhabditis elegans*)

El principal organismo sobre el que se utilizó AnABlast es *Caenorhabditis elegans.* El objetivo de este análisis era localizar nuevos genes codificantes de proteínas potenciales, así como nuevos exones pertenecientes a genes conocidos, con el valor de poder comprobar posteriormente en laboratorio los resultados. Para llevar a cabo este objetivo,

en un principio, se generaron los perfiles de AnABlast para todo el genoma, se extrajeron los picos significativos y se realizó un filtrado de los mismos. Este filtrado consistió en descartar:

- Picos que solapaban con genes o pseudogenes conocidos (Se descartan los genes conocidos ya que estamos buscando secuencias codificantes nuevas)
- Picos que solapaban con predicciones de otros programas (Se descartan predicciones de otros programas ya que nos interesan genes que solo AnABlast puede localizar).

Después del filtrado se obtuvieron 92 picos correspondientes a posibles nuevas regiones codificantes. Estos 92 picos se dividieron en dos grupos, dependiendo de su proximidad a genes conocidos, con la intención de separar aquellos posibles nuevos genes de los posibles nuevos exones: 82 picos, a más de 500 nucleótidos de genes conocidos y 10 picos a menos de 500.

Posteriormente se realizó un análisis funcional en laboratorio mediante RNAi, sobre los 82 picos predichos como nuevos genes codificantes de proteínas. En este análisis se realizó un muestreo de posibles nuevos fenotipos, encontrando 3 picos de AnABlast que presentaban fenotipo (G71, G98 y G107) (Figura 14). Teniendo en cuenta que no todos los genes silenciados mediante este método generan fenotipo, es destacable haber encontrado tres candidatos AnABlast correspondientes a secuencias que al silenciarlas generan fenotipo.

Al igual que en el caso de *D. melanogaster,* para *C. elegans* también se creó un visualizador genómico para revisar los resultados de AnABlast, donde se pueden activar y desactivar diferentes apartados que contienen información de: Genes presentes en WormBase, transposones, evidencias de transcripción como puede ser polisomas, espectrografía de masas y RNAseq, y predicciones realizadas por diferentes programas como son GeneFinder, GeneMarkHMM, mGene, mSplicer. Se puede acceder a esta información en la siguiente URL:
http://www.bioinfocabd.upo.es/celegans

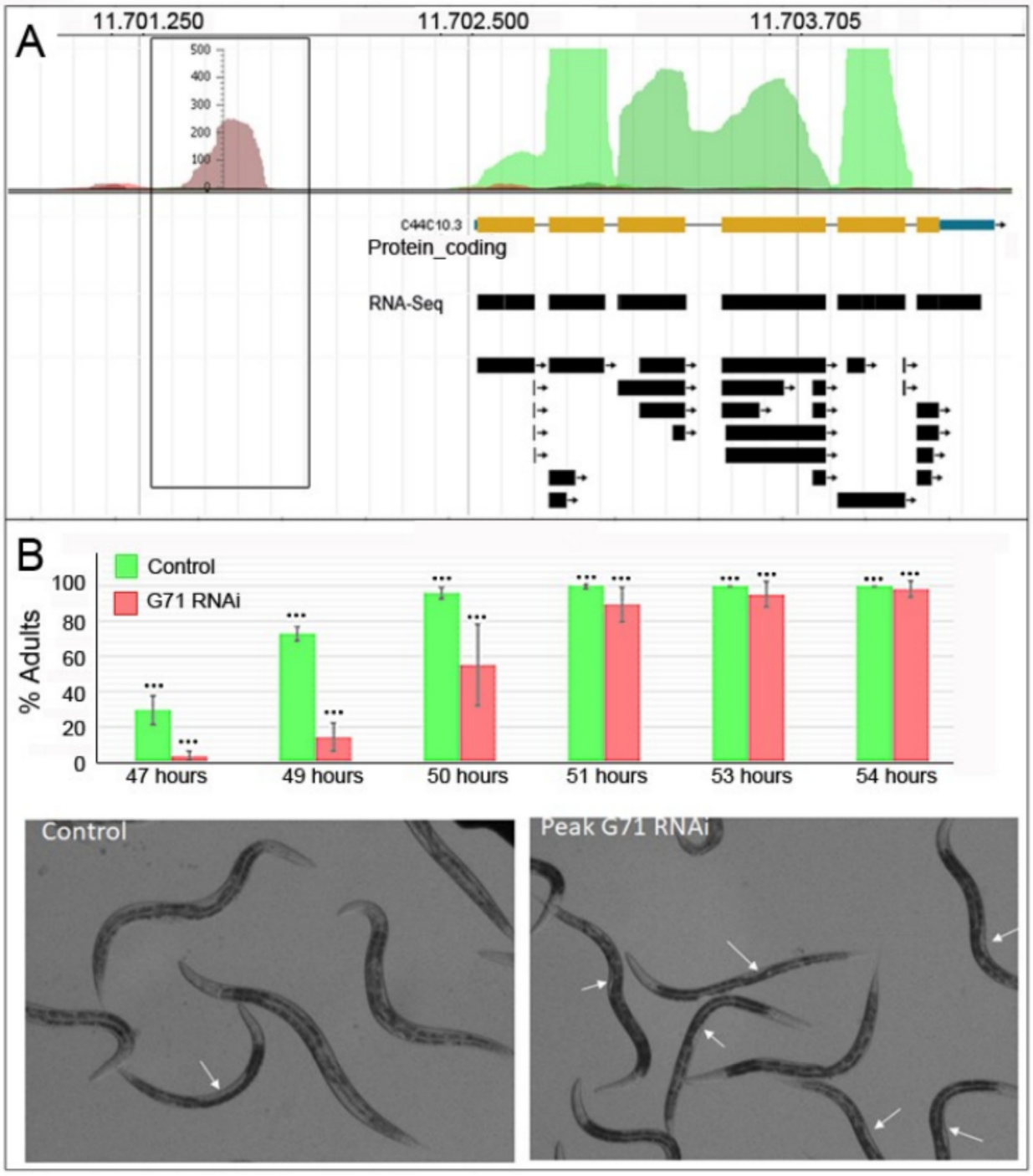


Figura 14: Análisis del pico de AnABlast G71 (X: 11701412-11701730). A) Perfiles de AnABlast mostrando el pico G71 (caja caudrada) y las señales adyacentes que coinciden con los exones de la proteína C44C10.3. Se muestran los datos de expresión de RNA (RNA-seq). B) Media de gusanos (%) que alcanzan el estado adulto desde la L1 (tiempo 0 horas) en gusanos sometidos al RNAi E71 (rosa) con respecto a los controles (verde). Las barra de desviación estándar son mostradas. Se muestran fotografías del fenotipo causado por G71 RNAi con respecto a los controles a las 49 horas (paneles inferiores). Las flechas indican la vulva no madura de los animales en L4.

3.1.3 Aplicación web

Una vez demostrada la utilidad del algoritmo y testados los parámetros más adecuados para su aplicación, se decidió ampliar el número posible de usuarios, mediante la simplificación de su uso. Para alcanzar este objetivo se desarrolló un aplicación web.

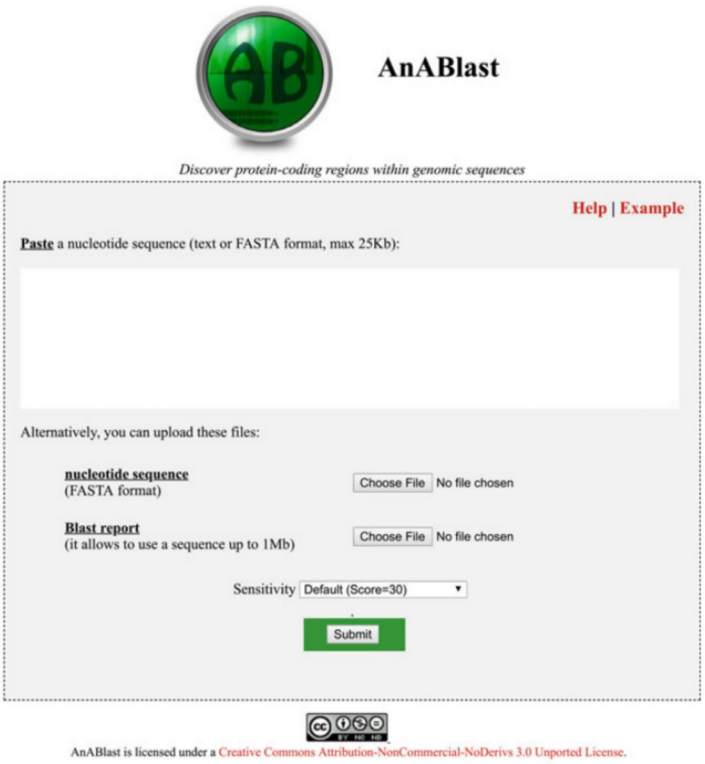


Figura 15: Imagen inicial de la aplicación web de AnABlast

Esta aplicación se programó en lenguaje Perl mediante CGI (Common Gateway Interface), que es un protocolo para ejecutar programas informáticos vía peticiones web (“CGI” 2019). Mediante este protocolo se aceptan las secuencias problemas y se lanzan los programas necesarios para realizar el control de calidad de las mismas y ejecutar AnABlast, así como Sma3s para anotar las predicciones, y otros predictores como Prodigal y AUGUSTUS para comparar resultados. Por último, para una mejor visualización de los resultados, se puso en marcha un visualizador de genomas basado en JBrowse (Buels et al. 2016), el cual esta programado en JavaScript (“JavaScript” 2019) y HMTL5 y permite una rápida visualización de genomas e información asociada como la posición de los picos de AnABlast los cuales están almacenados en 6 ficheros en formato BigWig (Kent et al. 2010), uno para cada marco de lectura, que se muestran en JBrowse

en un solo track mediante el módulo multiBigWig (“Elsiklab, Github” 2019) facilitando de esta forma su visualización, comparación y análisis posterior.

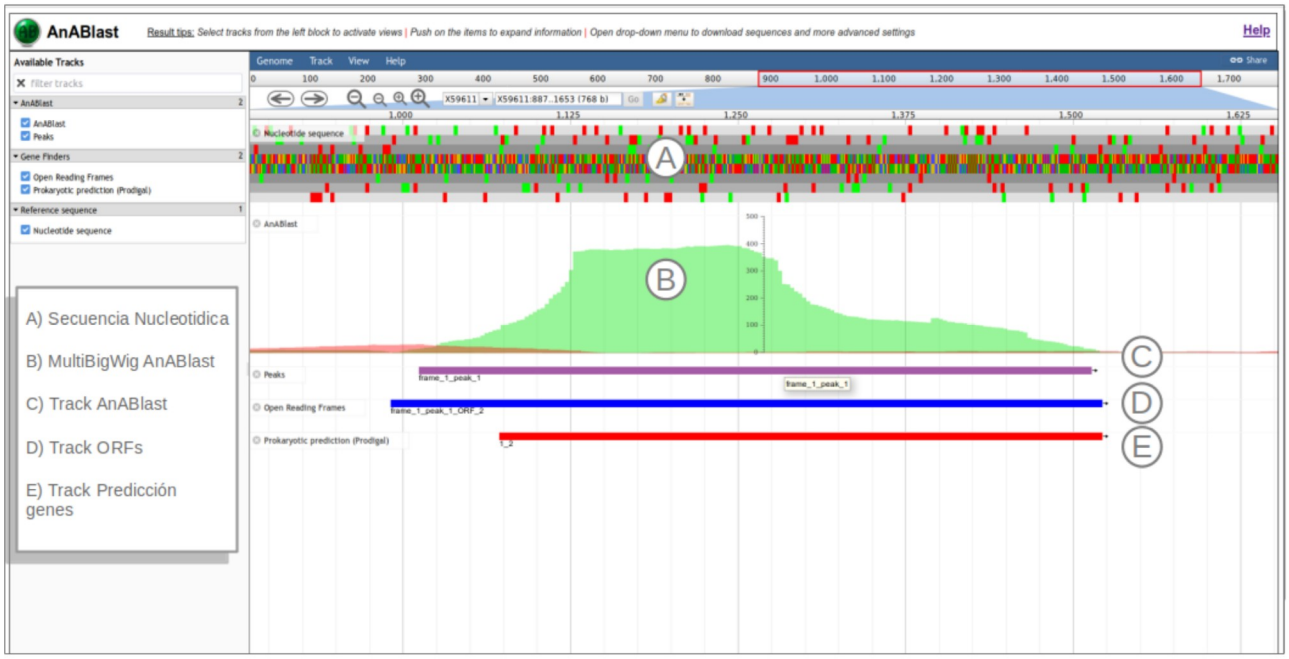


Figura 16: Ejemplo de los resultados mostrados por la aplicación web de AnABlast. A) Secuencia genómica con su traducción a aminoácidos en sus 6 marcos de lectura. Las dos líneas centrales representan los nucleótidos (un color para cada tipo) y las líneas grises arriba y abajo de la secuencia nucleotídica representan los 6 marcos de lectura donde se marcan en rojo los codones de parada y en verde los codones de inicio. B) Acumulación de alineamientos producida por AnABlast (pico). C) Track de los nucleótidos que están dentro de la señal de AnABlast (morado). D) Track de los ORF solapantes con la señal de AnABlast (azul). E) Track de las predicciones realizadas por otros algoritmos, en este caso Prodigal (Rojo).

Este visualizador de genomas (Figura 16) es el último paso de la aplicación web y en él se muestran todos los resultados de forma visual y accesible, pudiendo navegar por toda la secuencia problema. Además todos los ficheros generados por los análisis están disponibles para su descarga posibilitando su posterior análisis e integración con otras herramientas.

Se puede acceder a la aplicación web de AnABlast desde la siguiente URL:
http://www.bioinfocabd.upo.es/anablast/

4. Capítulo 1: Sma3s v2

4.1 Sma3s: a universal tool for easy functional annotation of proteomes and transcriptomes

Carlos S. Casimiro-Soriguer[1], Antonio Muñoz-Mérida[2], Antonio J. Pérez-Pulido[1]

[1]Centro Andaluz de Biología del Desarrollo (CABD-CSIC), Universidad Pablo de Olavide, Ctra. Utrera, Km. 1, 41013 Sevilla, Spain; [2]CIBIO-InBIO, Research Network in Biodiversity and Evolutionary Biology, Universidade do Porto, Campus Agrário de Vairão, 4485-661 Vairão, Portugal

4.1.1 Abstract

The current cheapening of next-generation sequencing has led to an enormous growth in the number of sequenced genomes and transcriptomes, allowing wet labs to get the sequences from their organisms of study. To make the most of these data, one of the first things that should be done is the functional annotation of the protein-coding genes. But it used to be a slow and tedious step that can involve the characterization of thousands of sequences.

Sma3s is an accurate computational tool for annotating proteins in an unattended way. Now, we have developed a completely new version, which includes functionalities that will be of utility for fundamental and applied science. Currently, the results provide functional categories such as biological processes, which become useful for both characterizing particular sequence datasets and comparing results from different projects. But one of the most important implemented innovations is that it has now low computational requirements, and the complete annotation of a simple proteome or transcriptome usually takes around 24 hours in a personal computer.

Sma3s has been tested with a large amount of complete proteomes and transcriptomes, and it has demonstrated its potential in health science and other specific projects.

Keywords: Functional annotation; Bioinformatic tool; Proteome; Transcriptome

4.1.2 Introduction

The current omics era is generating an enormous amount of available proteomes that we need to functionally annotate if we want to make the best of them. In fact, the annotation of all the protein-coding genes from an organism accelerates its knowledge at the molecular level [1,2].

Whereas the transcriptome allows both assessing the gene expression in a particular condition and knowing the specific protein-coding genes expressed by an organism, the genome allows knowing all the genes of this organism. It includes genes encoding for proteins predicted by a gene finder utility or proteogenomics strategy [3]. But, both of them need the functional annotation of protein-coding sequences prior to further analysis.

Functional annotation of a dataset coming from a whole genome or transcriptome can become a slow and tedious job, mainly when researchers have no knowledge on bioinformatics. In this context, the annotation of large sequence datasets is usually performed after a default similarity search followed by the assignment of functional terms from the best hits, without any more analysis [4,5]. Furthermore, the automatic annotation of large datasets, usually based on sequence similarity, is something that cannot be easily managed by the current computational tools. Most of available automatic annotators only allow the use of web applications with limitations in the number of query sequences to analyze. This is the case of the best scored tools arisen from the first *critical assessment of protein function annotation experiment* (CAFA), such as FFPred, Argot, PANNZER, ESG/PFP, or BAR-PLUS [6]. Only two of them allow the annotation of large datasets using specific programming scripts (FFPred [7], PANNZER [8]), but their standalone versions have a lot of requirements, including large databases and specialized software, all of which are elusive for experimental researchers. Blast2GO also belongs to this category [9]. It is a widely used annotation tool that can be difficult to use, especially with large datasets of sequences and without a commercial subscription. All of these tools annotate mainly amino acid sequences, but there are other methods for specifically annotating

protein-coding sequences, useful for analyzing transcriptomic data, which have the same weaknesses [10–12].

4.1.3 Results and discussion

To overcome the challenges of the current functional annotators we developed Sma3s, which is a standalone tool that has already shown a high accuracy with large sequences datasets, both of proteins and protein-coding nucleotide sequences [13]. From its first version it has been used in projects with organisms coming from heterogeneous taxonomic divisions such as bacteria [14], fungus [15–17], invertebrates [18], plants [19,20], and animals [21–23], and it provided useful results in all cases. Despite being easy to use, Sma3s had some computational requirements that we have now overcome (Supporting information: sections S1.1). Currently, it only depends on the installation of the standard BLAST software, and it allows using Sma3s on any operating system while hardly using computational resources. Additionally, it has now a low requirement of time thanks to the use of a shorter non-redundant database providing the precomputated reference annotations. As mentioned above, Sma3s is being used to annotate large datasets, but detailed analysis of the annotations is sometimes performed by other software [15,21], which is not able to exploit the full potential of the results. Thus, Sma3s now provides elaborated results, including functional categories which allow the user the easy creation of charts and the further analysis of the obtained annotations. Finally, Sma3s accuracy has been improved, thanks to the integration of its three previously independent modules (Fig. S1), and by the use of quality tags from the annotations, such as the evidence codes coming from Gene Ontology [24].

Initially, the new Sma3s has been compared with two of the best tools to annotate proteins, Argot and Blast2GO, and Sma3s obtains the best results with a benchmark of currently well-annotated sequences where self-annotation was avoided using an early release of the reference database (Supporting information: section S1.2).

To annotate a proteome the default mode can be used, though it is recommended to add two special parameters for improving the annotation with GO terms (-go and -goslim). The Sma3s result offers a report with different annotation types, including the most probable

gene name and protein description, EC numbers for enzymes, GO terms, and UniProt keywords and pathways. But one of the most demanded usability for massive annotators is the possibility of giving summaries with information about functional categories. This could be useful, for example, to compare annotations of different organisms in the same project. To enable this functionality, Sma3s now reports a summary with different categories and the number of sequences belonging to each category. To present this new functionality, we selected *Bacillus subtilis* as a complete proteome to annotate. From the 3,940 proteins from *B. subtilis*, Sma3s is able to assign annotations to 3,583 of them (91%), with GO and keywords as the most abundant annotated terms (Suppl. file 2). From the results, the different biological processes of this bacterium can be studied from GO Slim as well as from UniProt keywords categories (Fig. 1). The coverage from the former is higher (it offers an average of 2 terms by each protein), but keywords offer novel terms which can be very useful, as it can be checked in the present example with the group *Sporulation*. Hence, Sma3s found 255 proteins related to *Sporulation* in this well-known sporulated bacterium [25]. These different functional categories sources offer complementary results. For example, whereas the number of *transport* proteins is much higher in keyword than in GO Slim category, the number of proteins involved in *carbohydrate metabolism* is lower (Fig. 1). All of this allows a more complete collection of the biological processes characteristics of the organism to annotate.

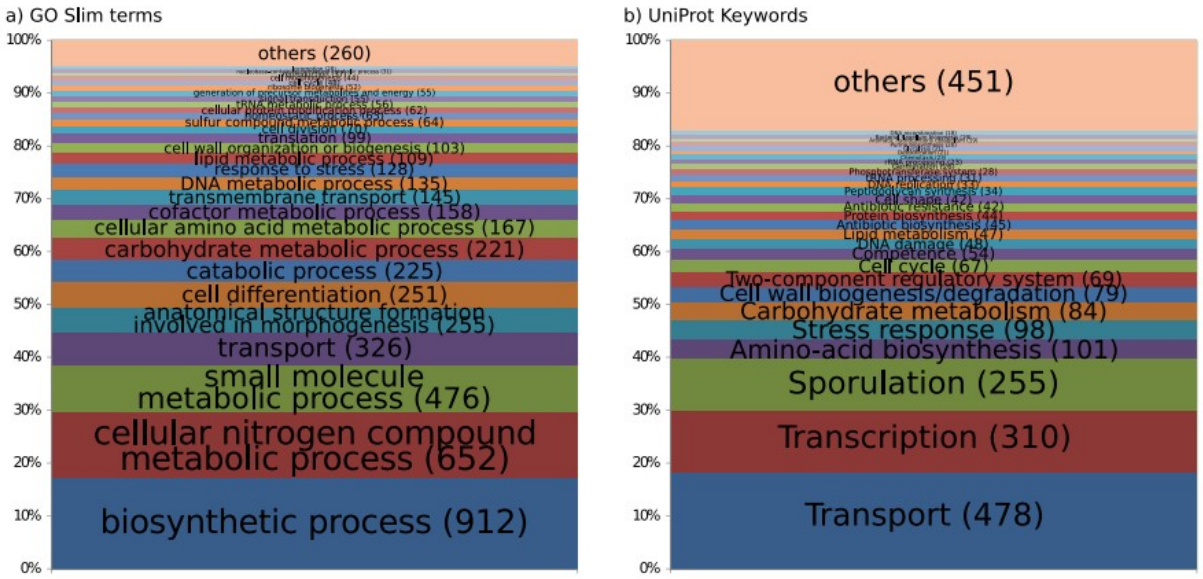


Figure 1. Number of proteins predicted to be involved in different biological processes. The figure represents the percentage of proteins belonging to different biological processes from a) GO Slim and b) UniProt keyword categories. The number of proteins predicted to each process is shown in brackets.

The Sma3s reports can be also used to compare different functional annotations coming from different organisms. To show this utility, we annotate 52 proteomes of the *Bacillus* genus, and compare 2 representative annotations from both biological process categories: *Cell motility* from GO Slim, and *Antibiotic biosynthesis* from UniProt keywords (Fig. 2). All annotated proteomes have a similar number of proteins involved in cell motility, with the exception of *B. gaemokensis* [26], and *B. mycoides*, which is one of the rare *Bacillus* that has been previously reported to lack of motility [27]. On the other hand, several species highlight if we check the other biological process, *antibiotic biosynthesis*. For *B. subtilis*, the representative species of the genus, Sma3s finds 45 proteins related with this annotation, and it is known that this species produces more than two dozen of different antibiotics of a great variety of types [25]. Further, we have been able to support this interesting ability in two more species, *B. megaterium*, and different strains of *B. amyloliquefaciens* [28–30]. All together shows a practical example of a discovering of bacterial strains of interest in a specific field, with for example application in medicine.

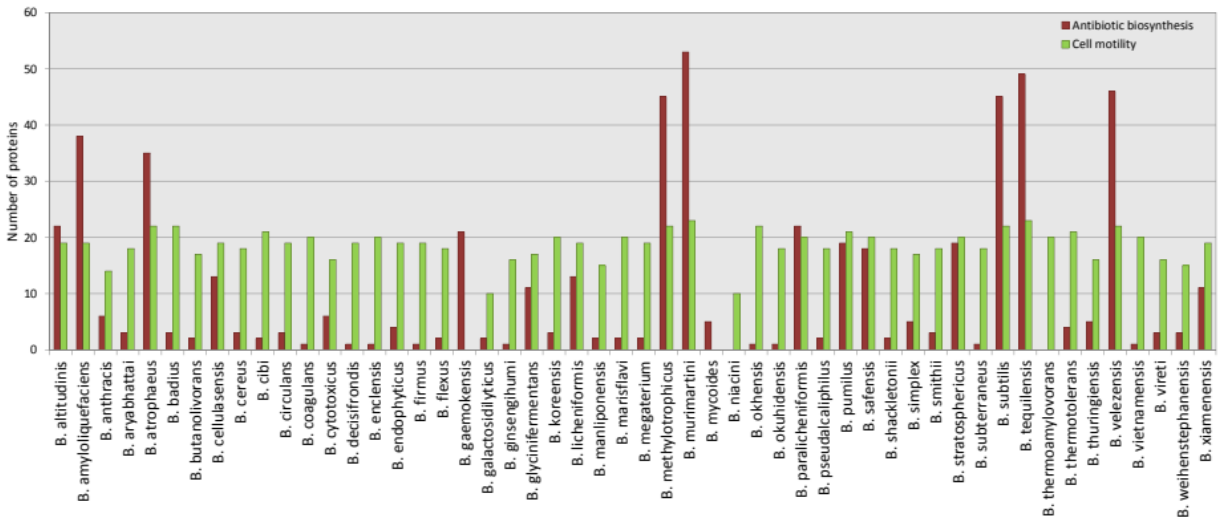


Figure 2. Number of proteins predicted to be involved in antibiotic biosynthesis and motility from different species of Bacillus. Proteins annotated in the antibiotic biosynthesis were extracted from UniProt keyword category, and those annotated in cell motility were extracted from GO Slim category.

So far we have showed Sma3s functionalities with well annotated proteomes. But we have also showed its great utility in heterogeneous annotation projects by annotating a non-model species proteome (Supporting information: section S1.3), and a transcriptome from a polyploid plant (Supporting information: section S1.4). The latter has been compared to the results of the widely used tool Trinotate [10], and Sma3s showed a great performance.

In conclusion, here we show that Sma3s is a useful computational tool for annotate proteomes and transcriptomes in a short time and with minimal requirements. Furthermore, the most important characteristic of Sma3s is that it can be used to annotate complete sequence datasets by any user without computational skills, which will allow increasing the knowledge about their organisms in study. All of them, as far as we know, will leave Sma3s as virtually one of the easier and faster, keeping its accuracy, functional annotator of proteins and protein-coding sequences currently available.

Funding

This research was supported by the Ministry of Economy and Competitiveness of the Spanish Government grant BFU2013-46923-P. Authors declare no competing interest.

Acknowledgements

We would like to thank C3UPO and CICA for the HPC support.

4.1.4 References

[1] Rougon-Cardoso, A., Flores-Ponce, M., Ramos-Aboites, H.E., Martínez-Guerrero, C.E., et al., The genome, transcriptome, and proteome of the nematode Steinernema carpocapsae: evolutionary signatures of a pathogenic lifestyle. *Sci. Rep.* 2016, 6, 37536.

[2] Castro, J.C., Maddox, J.D., Cobos, M., Requena, D., et al., De novo assembly and functional annotation of Myrciaria dubia fruit transcriptome reveals multiple metabolic pathways for L-ascorbic acid biosynthesis. *BMC Genomics* 2015, 16, 997.

[3] Renuse, S., Chaerkady, R., Pandey, A., Proteogenomics. *Proteomics* 2011, 11, 620–630.

[4] Prochnik, S.E., Umen, J., Nedelcu, A.M., Hallmann, A., et al., Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri. *Science* 2010, 329, 223–226.

[5] Venturini, L., Ferrarini, A., Zenoni, S., Tornielli, G.B., et al., De novo transcriptome characterization of Vitis vinifera cv. Corvina unveils varietal diversity. *BMC Genomics* 2013, 14, 41.

[6] Radivojac, P., Clark, W.T., Oron, T.R., Schnoes, A.M., et al., A large-scale evaluation of computational protein function prediction. *Nat. Methods* 2013, 10, 221–227.

[7] Minneci, F., Piovesan, D., Cozzetto, D., Jones, D.T., FFPred 2.0: improved homology-independent prediction of gene ontology terms for eukaryotic protein sequences. *PloS One* 2013, 8, e63754.

[8] Koskinen, P., Törönen, P., Nokso-Koivisto, J., Holm, L., PANNZER: high-throughput functional annotation of uncharacterized proteins in an error-prone environment. *Bioinformatics* 2015, 31, 1544–1552.

[9] Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., et al., Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 2005, 21, 3674–6.

[10] Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., et al., Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 2011, 29, 644–652.

[11] Chen, T.-W., Gan, R.-C.R., Wu, T.H., Huang, P.-J., et al., FastAnnotator--an efficient transcript annotation web tool. *BMC Genomics* 2012, 13 Suppl 7, S9.

[12] Hotz-Wagenblatt, A., Hankeln, T., Ernst, P., Glatting, K.H., et al., ESTAnnotator: A tool for high throughput EST annotation. *Nucleic Acids Res.* 2003, 31, 3716–9.

[13] Muñoz-Mérida, A., Viguera, E., Claros, M.G., Trelles, O., Pérez-Pulido, A.J., Sma3s: a three-step modular annotator for large sequence datasets. *DNA Res.* 2014, 21, 341–353.

[14] García-Romero, I., Pérez-Pulido, A.J., González-Flores, Y.E., Reyes-Ramírez, F., et al., Genomic analysis of the nitrate-respiring Sphingopyxis granuli (formerly Sphingomonas macrogoltabida) strain TFA. *BMC Genomics* 2016, 17, 93.

[15] Mikhailov, K.V., Simdyanov, T.G., Aleoshin, V.V., Genomic survey of a hyperparasitic microsporidian Amphiamblys sp. (Metchnikovellidae). *Genome Biol. Evol.* 2016.

[16] Masuya, H., Manabe, R.-I., Ohkuma, M., Endoh, R., Draft Genome Sequence of Raffaelea quercivora JCM 11526, a Japanese Oak Wilt Pathogen Associated with the Platypodid Beetle, Platypus quercivorus. *Genome Announc.* 2016, 4.

[17] Cho, O., Ichikawa, T., Kurakado, S., Takashima, M., et al., Draft Genome Sequence of the Causative Antigen of Summer-Type Hypersensitivity Pneumonitis, Trichosporon domesticum JCM 9580. *Genome Announc.* 2016, 4.

[18] Perina, A., González-Tizón, A.M., Meilán, I.F., Martínez-Lage, A., De novo transcriptome assembly of shrimp Palaemon serratus. *Genomics Data* 2017, 11, 89–91.

[19] Pascual, J., Alegre, S., Nagler, M., Escandón, M., et al., The variations in the nuclear proteome reveal new transcription factors and mechanisms involved in UV stress response in Pinus radiata. *J. Proteomics* 2016, 143, 390–400.

[20] Carmona, R., Zafra, A., Seoane, P., Castro, A.J., et al., ReprOlive: a database with linked data for the olive tree (Olea europaea L.) reproductive transcriptome. *Front. Plant Sci.* 2015, 6, 625.

[21] Kumar, V., Kutschera, V.E., Nilsson, M.A., Janke, A., Genetic signatures of adaptation revealed from transcriptome sequencing of Arctic and red foxes. *BMC Genomics* 2015, 16, 585.

[22] Benzekri, H., Armesto, P., Cousin, X., Rovira, M., et al., De novo assembly, characterization and functional annotation of Senegalese sole (Solea senegalensis) and common sole (Solea solea) transcriptomes: integration in a database and design of a microarray. *BMC Genomics* 2014, 15, 952.

[23] Nourisson, C., Muñoz-Merida, A., Carneiro, M., Sequeira, F., De novo transcriptome assembly and polymorphism detection in two highly divergent evolutionary units of Bosca's newt (Lissotriton boscai) endemic to the Iberian Peninsula. *Mol. Ecol. Resour.* 2016.

[24] Gene Ontology Consortium, Gene Ontology Consortium: going forward. *Nucleic Acids Res.* 2015, 43, D1049-1056.

[25] Stein, T., Bacillus subtilis antibiotics: structures, syntheses and specific functions. *Mol. Microbiol.* 2005, 56, 845–857.

[26] Nakamura, L.K., Jackson, M.A., Clarification of the Taxonomy of Bacillus mycoides. *Int. J. Syst. Evol. Microbiol.* 1995, 45, 46–49.

[27] Jung, M.Y., Jung, M.-Y., Paek, W.K., Park, I.-S., et al., Bacillus gaemokensis sp. nov., isolated from foreshore tidal flat sediment from the Yellow Sea. *J. Microbiol. Seoul Korea* 2010, 48, 867–871.

[28] Malanicheva, I.A., Kozlov, D.G., Sumarukova, I.G., Efremenkova, O.V., et al., Antimicrobial activity of Bacillus megaterium strains. *Mikrobiologiia* 2012, 81, 196–204.

[29] Jeong, H., Park, S.-H., Choi, S.-K., Genome Sequence of Antibiotic-Producing Bacillus amyloliquefaciens Strain KCTC 13012. *Genome Announc.* 2015, 3.

[30] Arguelles-Arias, A., Ongena, M., Halimi, B., Lara, Y., et al., Bacillus amyloliquefaciens GA1 as a source of potent antibiotics and other secondary metabolites for biocontrol of plant pathogens. *Microb. Cell Factories* 2009, 8, 63.

4.2 Sma3s: a universal tool for easy functional annotation of proteomes and transcriptomes. Supporting information

Carlos S. Casimiro-Soriguer[1], Antonio Muñoz-Mérida[2], Antonio J. Pérez-Pulido[1]

[1]Centro Andaluz de Biología del Desarrollo (CABD-CSIC), Universidad Pablo de Olavide, Ctra. Utrera, Km. 1, 41013 Sevilla, Spain; [2]CIBIO-InBIO, Research Network in Biodiversity and Evolutionary Biology, Universidade do Porto, Campus Agrário de Vairão, 4485-661 Vairão, Portugal

4.2.1 Additional results

4.2.1.1 Sma3 algorithm overview

The Sma3s algorithm has been completely rewritten to simplify its use and offer more useful information in the results, in addition to reduce the entire run time (Fig. S1). Originally, Sma3s was composed of three independent but complementary modules. The first of them searches for very similar sequences and its annotations, the second searches for orthologs, and the third one uses all significant alignments from a Blast search to retrieve shared annotations with significant expected values. In the current version, all the three modules have been merged to both improve the ease to use and complete the results with more accurate assignments.

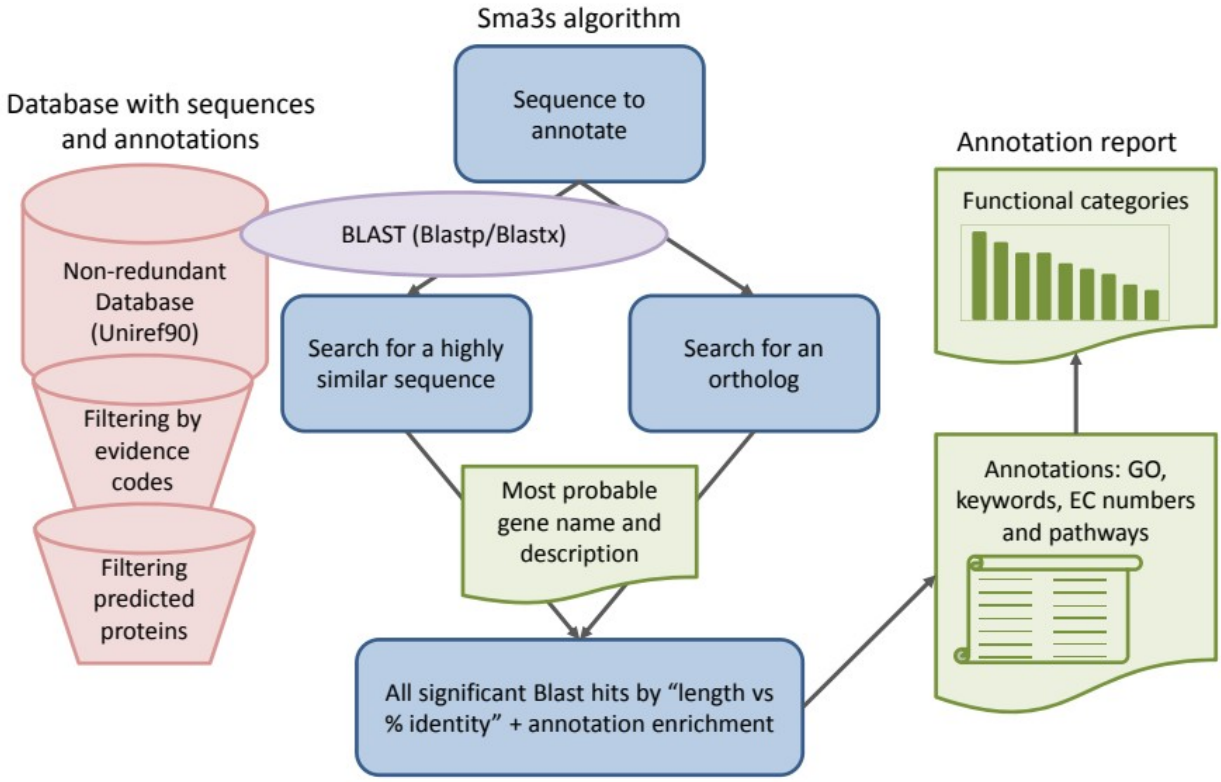


Figure S1. Organigram of the Sma3s pipeline. Every sequence from the query dataset is compared to the reference database using Blast (blastp for proteomes and blastx for transcriptomes). This database can be filtered to taking in account only sequences of expected high quality. Sma3s prioritizes annotations (especially gene name and description) coming from very similar or orthologous proteins. The last module enriches the annotations with proteins sharing short similarity regions but with common annotations. Finally, Sma3s gives the results in two files, one of them with all the annotations for each query sequence, and the other with functional categories that can be used to both create figures and compare different annotations.

Since both of the initial modules are based on more similar sequences, gene names and descriptions are mainly assigned from them rather than the third module. But annotations overall are assigned from all the 3 modules to improve the final sensitivity.

Nevertheless, users can run Sma3s with a single module to check if their sequences are already annotated in the database (using only the module 1), or if there are orthologs in the database that enable the annotation (using only the module 2).

4.2.1.2 Proving the accuracy of Sma3s for annotating protein sequences

To test the accuracy of Sma3s, a dataset of 349 manually curated proteins was used. These proteins entered in Swiss-Prot in 2015 or later. So, to avoid self-annotation Sma3s

was used with Swiss-Prot 2014 as the reference database for the annotation. The result shows that despite of they are new sequences in the public database of proteins, the obtained accuracy is high, with recall around 60% to both keywords and GO terms (Fig. S2).

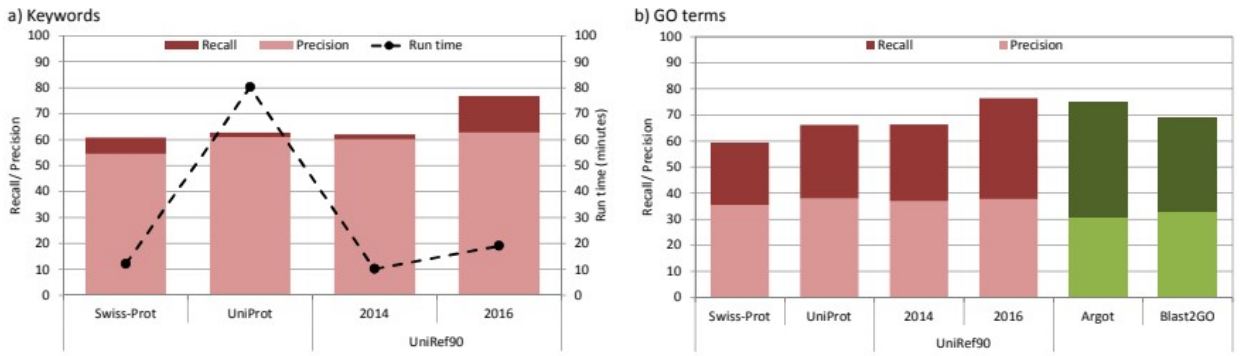


Figure S2. Annotation accuracy when using different reference databases. Recall and precision is represented by percent values, and run times (dashed line) is represented in minutes. The results are shown both a) when UniProt keywords, and b) GO terms are evaluated. In all cases the 2014 versions of the databases were used, except for UniRef90 version 2016. For both Argot and Blast2GO only GO terms were evaluated, and databases from 2014 were used.

Sma3s was initially based on Swiss-Prot as the reference database, but we wanted to increase the accuracy using a greater database. Thus, UniProt database was selected to annotate the same dataset. This is a database with a number of sequences 100 times higher (Table S1). In this case, the results show that though the precision increases slightly, the run time makes it unfeasible, especially if we want to annotate huge sequence datasets.

Table S1. Number of annotated proteins within different databases and number of Blast hits obtained for all the sequences in the test dataset. Two different versions are used for the different databases.

	2014			2016	
	Swiss-Prot	UniProt	UniRef90	UniProt	UniRef90
Number of annotated proteins	524,643	52,924,113	10,809,500	47,715,549	21,330,801
Number of Blast hits	36,000	65,636	63,421	71,995	68,811

In order to decrease the run time, we wanted to use a shorter database but maintaining a more complete collection of protein sequences than with Swiss-Prot. To do that we used UniRef90, where sequence redundancy is reduced using 90% identity as a threshold. In

this case, the number of sequences is 5 times lower than with UniProt (Table S1). So, when we run Sma3s with UniRef90 (2014) as the reference database, the time run is similar to that obtained when we used Swiss-Prot. But also more importantly, the accuracy is almost as high as when using UniProt. In fact, the number of significant Blast alignments is similar to that obtained when we used the complete UniProt database.

The query dataset is composed of new sequences that are currently better known. So, when we use Sma3s with the current release of UniRef90 (2016), the accuracy is now the highest, with a recall close to 80% and a precision of 63%. All of this proposes the last version of UniRef90 as the reference database to annotate proteins and protein-coding sequences with a high accuracy and a short run time.

To compare Sma3s with other protein functional annotators we chose Argot, since it has recently showed a high accuracy [1], and Blast2GO, since it is a widely used computational tool. We annotate the query dataset with both of them and databases from 2014, and the predicted GO terms were collected and compared with those from the Sma3s results. The precision with Sma3s was the highest out of three annotators (37% versus 31% and 27%) (Fig. 2b), though it showed the lowest recall when using UniRef90 2014 (66% versus 75% and 70%), but not when using the 2016 version. But, it is important to note that, even though the recall is higher with Argot, this value changes depending on the type of GO term. The higher recall of Argot seems related to the most generic terms coming from the cellular component ontology (Fig. S3). However the molecular function and biological process ontologies have better results with Sma3s versus Argot, especially in more specific terms. Although cellular localization is important to know the place where a protein exerts its function, molecular functions and even more biological processes are more demanded in projects analyzing gene expression or comparing the annotation of phylogenetically related organisms.

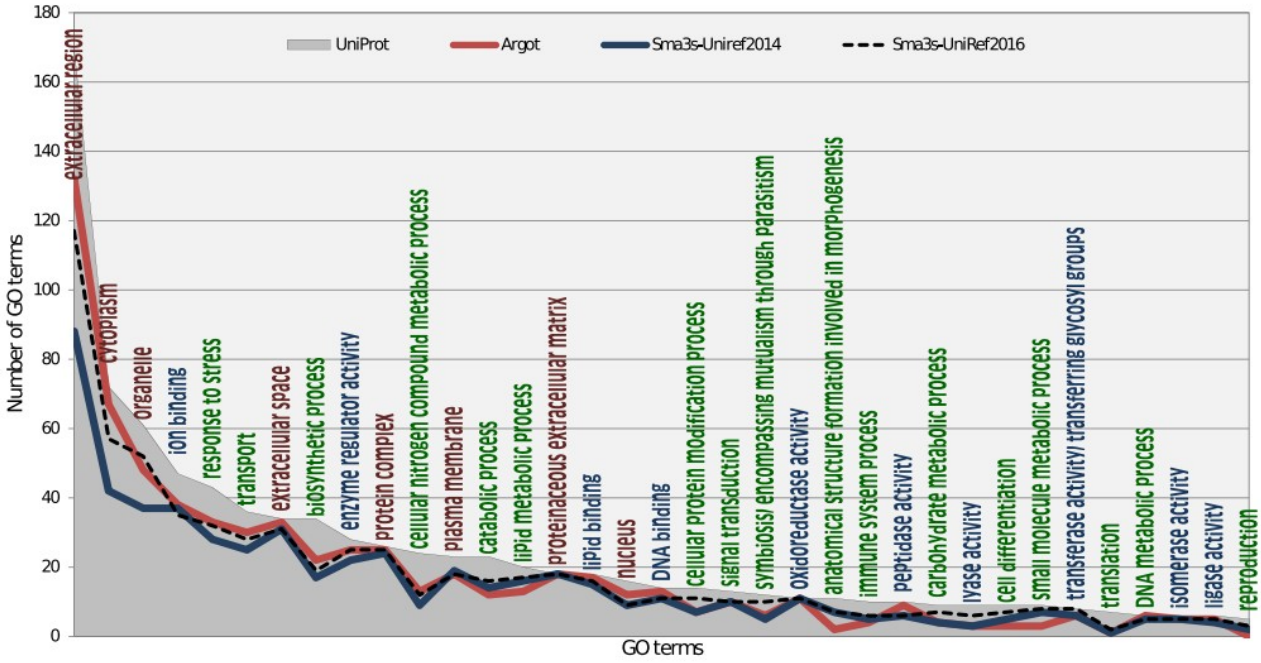


Figure S3. Number of predicted GO terms by different databases and methods. The background in grey color represents the number of specific GO terms in the protein entries from UniProt database. GO terms coming from different ontologies are highlighted in red (Cellular Component), blue (Molecular Function), and green (Biological Process) color.

4.2.1.3 Sma3s is able to get complete annotations for non-model organisms

To show the performance of Sma3s with non-model organisms we annotate the complete proteome of *Tetranychus urticae*, the red spider mite. This proteome has 17,526 proteins, and Sma3s was able to annotate 7,052. The recall at the level of GO was high (75%), and also at the level of pathways (86%), but lower with keywords (51%). But the prediction assigns a number of annotations much higher than that found in the database UniProt for this proteome. In fact, the precision goes from 20 to 25% with GO and keywords respectively. This latter could be explained due to Sma3s finds new annotations for not well annotated proteomes. It can be showed with the proteins annotated as carotenoid metabolism. Other than certain aphids, *T. urticae* is the only animal known to be able to synthesize carotenoids, which appear to have been acquired through horizontal gene transfer from a fungus [2]. Only 3 proteins of this mite are annotated to participate in carotenoid metabolism in the database UniProt, but Sma3s was able to annotate 7 proteins with a total of 28 annotations related to this process. These results suggest that a great percent of the predicted annotations by Sma3s are really true, contrary to expectation.

4.2.1.4 Sma3s is able to annotate complex transcriptomes

Sma3s can also annotate complete transcriptomes. To show this ability, a transcriptome of the polyploid plant *Festuca pratensis* was annotated. This dataset has been previously annotated by the tool Trinotate and it allows to compare results [3]. Sma3s was able to annotate 35,093 sequences out of the 72,123, versus the 21,324 sequences annotated by Trinotate. The distribution of functional categories was similar between both of the annotations, though Trinotate seems to lose the more specific terms (Fig. S4). For example, though generic localizations such as cytoplasm and plasma membrane have similar numbers from both annotations, specific organelles such as mitochondrion, Golgi apparatus, or the plant characteristic vacuoles are more annotated by Sma3s. And the same happens with biological processes such as anatomical structural development and reproduction. These results again suggest a more specific annotation by Sma3s compared with other current annotators.

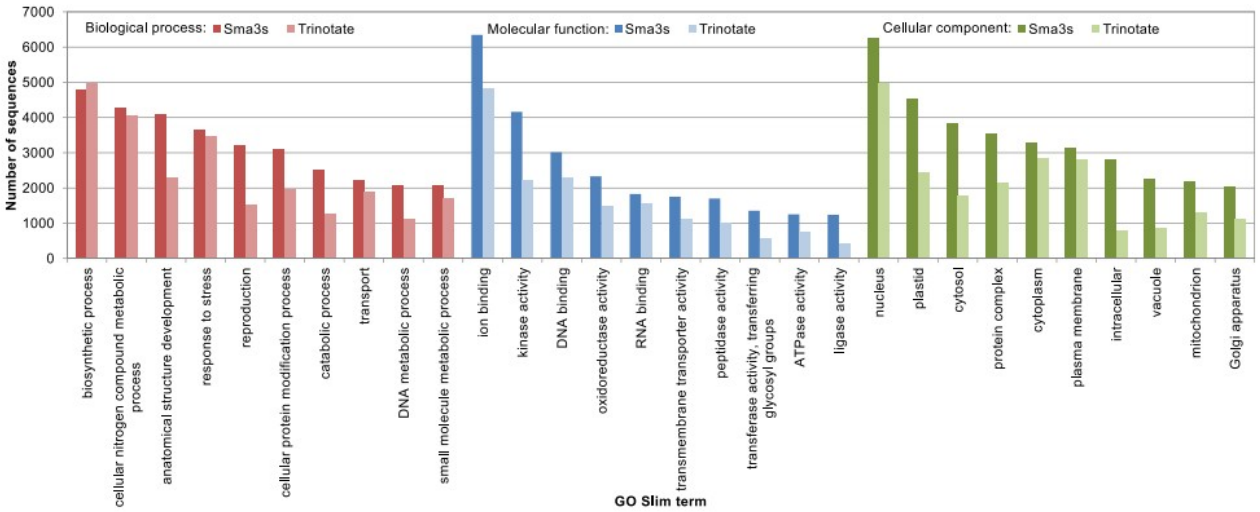


Figure S4. Number of proteins predicted in the different functional categories for the F. pratensis transcriptome. Functional categories coming from GO Slim are shown for both the Sma3s and Trinotate annotation. Only the 10 most frequent terms by ontology are showed.

4.2.2 Materials and methods

4.2.2.1 Improvements in the algorithm

The three independent modules of Sma3s have been now integrated to give a more complete annotation. The first two modules discover significant homologs and they are now used to assign a gene name and a description to each query sequence. But only informative names are used, avoiding those with rare symbols or longer than 6 characters. Then, annotation terms are assigned from either the module 1 or 2, and this preliminary annotation is complemented with the more productive module 3.

One of the principal algorithm improvements concerns to the results. The annotation sources have been extended, and the final annotation report now includes EC numbers for enzymes, together with UniProt keywords and pathways [4], and GO terms [5]. To add functional categories, which are especially useful to undertake great annotation or comparative genomic projects, Sma3s gives GO Slim terms that report more general annotations. The GO Slim terms have been extracted from each GO term in the reference database, using both the Map2Slim script and the Generic GO Slim file from the Gene Ontology web. The results also include four categories used to classify the keywords in UniProt: Biological process, Cellular component, Developmental stage, and Disease.

Due to the exponential growth of the sequence databases, the quality of the annotations is something to keep in mind [6]. Thus, Sma3s allows reporting quality annotations, where only experimental assigned GO terms and keywords will be used. To do this, annotations with the “Inferred Electronic Annotation” evidence code coming from both GO terms (IEA) and UniProt keywords (ECO:0000501, and codes related to this one) can be discarded in the analysis. In addition, non-informative annotations are avoided, as well as database sequences without any GO term and keyword, and predicted sequences from the reference database can be also discarded.

The remaining parameters from Sma3s are now fixed in an automatic way for improving both proteome and transcriptome annotations, and make easier the use of the annotator.

Finally, Sma3s offers the annotations in a text file that can be opened with any spreadsheet program (tab-separated values; TSV), along with a file containing the summary of the results. This latter includes the functional categories, with biological processes and pathways, which allow the easy creation of figures to the end user.

4.2.2.2 Requirements to use Sma3s

The installation of the Blast+ package is the only mandatory requirement for Sma3s (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST), together with the Perl programming language interpreter. Sma3s takes around 24 hours to annotate a simple proteome (around 5000 sequences), and a similar time to annotate a short transcriptome (adding -nucl parameter, which will use blastx instead of blastp). Moreover, Sma3s allows parallelization of the initial Blast step, using high performance computing (HPC) to accelerate the entire process.

The Sma3s script can be found at:
http://www.bioinfocabd.upo.es/sma3s

There, you can see video tutorials about how to use Sma3s in different operating systems. Finally, you can use UniProt files with .dat extension from its website to perform the annotation (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/). But we offer a non-redundant database, coming from UniRef90, which allows shorter times to annotate bigger sequence datasets. This database was used to produce the results presented in this work, and it can be downloaded from our website.

4.2.2.3 Test dataset of new proteins

To annotate proteins with a current minimum quality, but using a previous database version, where the proteins were not annotated, we proceeded with the following steps. We used the manually curated Swiss-Prot section in UniProt database release 2016_07. Then, we collected entries created from 2015, with only one Accession number, thus avoiding any new entry coming from a previous one. But these proteins could come from TrEMBL, the not curated section in UniProt database. To avoid the latter, we checked the

history from each entry. Thus, we selected entries created in UniProt from January 2015 to July 2016 and currently stored in Swiss-Prot. Finally, we remove 41 proteins that lacked GO terms, since they did not allow measuring the accuracy.

Following this strategy, we found 349 proteins from Swiss-Prot release 2016_07, with manually assigned annotations, which did not exist in the release from 2014 (Suppl. file 1). Thus, annotating these proteins using Swiss-Prot release 2014_11, and checking accuracy with Swiss-Prot release 2016_07, constitutes a good procedure to test annotation tools avoiding self-annotation.

To check the prediction, we used GO Slim terms and UniProt keywords from Sma3s, and convert GO annotations found by Argot to GO Slim, using the Map2Slim script by the Gene Ontology Consortium.

We used Sma3s with the following reference databases: Swiss-Prot 2014_11, UniProt 2014_11, UniRef90 2014_11, and UniRef90 2016_07 (taking away the 349 query sequences). And we used Argot release 2.5 from its website (July, 2016), which used databases from 2014, and Blast2GO 4.1 standalone version with the database Swiss-Prot from 2014.

4.2.2.4 Annotation of proteomes and transcriptomes

To annotate different proteomes from *Bacillus* genus we downloaded the protein dataset files from Ensembl bacteria database [7]. We selected only reference species whose names were composed by two words. Thus, we finally downloaded 52 different proteomes. To annotate the non-model proteome from *T. urticae* we downloaded its UniProt entries from the Proteomes section of this database. These entries were removed from the reference database to avoid self-annotation.

Finally, to annotate the plant transcriptome from *F. pratensis* we used the sequences deposited at the Transcriptome Shotgun Assembly database in the NCBI under the accession GBXZ00000000.1 (https://www.ncbi.nlm.nih.gov/Traces/wgs/?val=GBXZ01). The Sma3s annotation was compared with that from the Additional file 1 by Czaban et al. [3].

All of these datasets were used to run Sma3s, using default parameters and a HPC cluster to accelerate the process. The –nucl parameter was activated for the transcriptome annotation.

4.2.3 References

[1] Lavezzo, E., Falda, M., Fontana, P., Bianco, L., Toppo, S., Enhancing protein function prediction with taxonomic constraints--The Argot2.5 web server. *Methods* 2016, 93, 15–23.

[2] Altincicek, B., Kovacs, J.L., Gerardo, N.M., Horizontally transferred fungal carotenoid genes in the two-spotted spider mite Tetranychus urticae. *Biol. Lett.* 2012, 8, 253–257.

[3] Czaban, A., Sharma, S., Byrne, S.L., Spannagl, M., et al., Comparative transcriptome analysis within the Lolium/Festuca species complex reveals high sequence conservation. *BMC Genomics* 2015, 16, 249.

[4] The UniProt Consortium, UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 2017, 45, D158–D169.

[5] Gene Ontology Consortium, Gene Ontology Consortium: going forward. *Nucleic Acids Res.* 2015, 43, D1049-1056.

[6] Holliday, G.L., Davidson, R., Akiva, E., Babbitt, P.C., Evaluating Functional Annotations of Enzymes Using the Gene Ontology. *Methods Mol. Biol. Clifton NJ* 2017, 1446, 111–132.

[7] Kersey, P.J., Allen, J.E., Armean, I., Boddu, S., et al., Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Res.* 2016, 44, D574-580.

5. Capítulo 2: AnABlast

5.1 AnABlast, re-searching for protein-coding sequences in genomic regions

Alejandro Rubio[1], Carlos S. Casimiro-Soriguer[1], Pablo Mier[2], Miguel A. Andrade-Navarro[2], Andrés Garzón[1], Juan Jiménez[1],*, and Antonio J. Pérez-Pulido[1],*

[1]Centro Andaluz de Biologia del Desarrollo (CABD, UPO-CSIC-JA). Facultad de Ciencias Experimentales (Área de Genética), Universidad Pablo de Olavide, Ctra. de Utrera, km.1, 41013, Sevilla, Spain. [2]Faculty of Biology, Johannes Gutenberg University Mainz, Hans-Dieter-Husch-Weg 15, 55128 Mainz, Germany

*Co-corresponding authors jjimmar@upo.es ajperez@upo.es

Running head

Searching protein-coding sequences with AnABlast

5.1.1 Abstract

AnABlast is a computational tool that highlights protein-coding regions within intergenic and intronic DNA sequences which escape detection by standard gene prediction algorithms. DNA sequences with small protein-coding genes or exons, complex intron-containing genes, or degenerated DNA fragments are efficiently targeted by AnABlast. Furthermore, this algorithm is particularly useful in detecting protein-coding sequences with non-significant homologs to sequences in databases. AnABlast can be executed online at http://www.bioinfocabd.upo.es/anablast/.

Keywords: Gene finding, coding DNA sequences, in silico-annotation tool, small genes, fossil DNA sequences.

5.1.2 Introduction

A great number of wet-lab groups are sequencing whole genomes as a common practice, taking advantage of the current burst of the genomics era. In silico analysis of such amount of sequences is essential for accurate annotation tasks [1, 2], but computational tools for predicting genes usually have accuracies of around 90%, or even lower for exons in protein-coding genes coming from eukaryotic organisms [3, 4]. Thus, a significant number of coding sequences escape detection when using currently available genome annotation tools.

The identification of similar proteins through BLAST analysis is one of the most useful strategies in genome annotation. Finding significant alignments facilitates the assignment of putative functions to query amino acid sequences through the identification of related proteins in sequence databases. Non-significant alignments, those below the significant score threshold and, therefore, discarded by in silico gene finders, are often found in conventional similarity searches. In hypothetical polypeptide sequences obtained from the electronic translation of non-coding DNA, such alignments occur by chance at a very low frequency. Protein-coding sequences, however, mostly arise from previous ones during evolution, and non-significant alignments can include both functional and random evolutionary footprints coming from common ancestors [5, 6]. Thus, in polypeptide sequences computationally translated from coding DNA, in addition to random matches, footprints of ancestral common sequences increase the frequency of non-significant alignments, and consequently, the accumulation of non-significant alignments efficiently discriminates putative coding from non-coding DNA sequences (lacking such footprints). Even in the case of highly divergent genes, their coding sequences may contain footprints of common ancestral proteins to be found, among the millions of proteins available in databases, by using this strategy [5, 6]. Therefore, alignments accumulated in predicted amino acid sequences provide a method to discern coding from non-coding DNA, a new strategy used by AnABlast (Ancestral-patterns search through A BLAST-based strategy) to overcome limitations of current in silico algorithms in order to identify putative coding DNA sequences [7].

Here, we present the AnABlast web application, a novel computational tool which allows the identification of putative coding regions in recent genomic sequences, or in intergenic and intron sequences of annotated genomes.

5.1.3 Short introduction to the web interface

AnABlast can be executed from a web application and only needs a genomic sequence of up to 25 Kb to start its execution. Alternatively a BLAST report can be provided if, for example, the user wants to run the slow similarity search in a computer cluster (Fig. 1). In this latter case, the genomic sequence can be as long as 1 Mb.

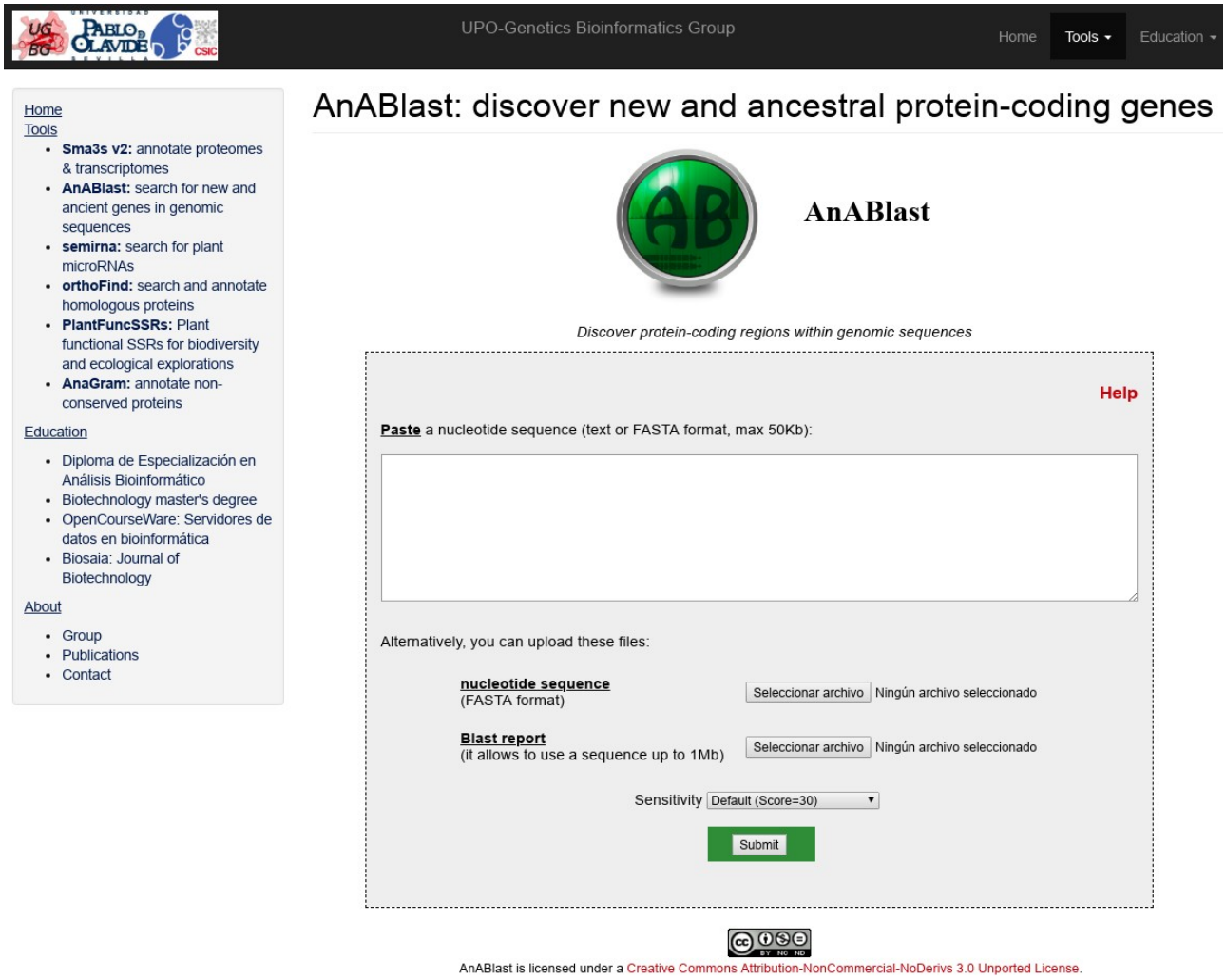


Fig. 1 Screenshot of the AnABlast Web page at http://www.bioinfocabd.upo.es/anablast/

Briefly, AnABlast compares the predicted amino acid sequence from each of the six reading-frames of a genomic DNA sequence against a non-redundant protein database, and accumulates all the found hits that we name protomotifs, including low-scored alignments [5]. These protomotifs are usually accumulated in coding sequences but rarely in non-coding sequences. Thus the graphic profile of accumulated AnABlast protomotifs yields peaks that may accurately mark putative protein-coding genes, pseudogenes, and fossil sequences, even in the presence of sequencing errors.

AnABlast runs a BLASTX search against the UniRef50 database and gathers hits with a default minimal bit-score of 30, though the user can choose to run AnABlast with a higher sensitivity (bit-score 28) or with a lower sensitivity but higher specificity (bit-score 35). The results appear in a genome browser that uses the JBrowse tool, including four tracks:

- AnABlast profile, diagram showing protomotifs accumulation; in green colors those from the forward strand, and in red colors those from the reverse strand.
- Peaks, regions with a protomotif accumulation above the threshold (70 is used by default, meaning that 70 different non-redundant proteins share this protomotif). The predicted functional annotation using Sma3s [8] is showed when an element is selected.
- ORFs, found open reading frames from a start codon to a stop codon.
- Predicted genes, gene structure predicted by a gene finder: AUGUSTUS for eukaryotic sequences [9], and Prodigal for prokaryotes sequences [4].

A representative landscape resulting from the AnABlast analysis of the fission yeast Schizosacchomyces pombe region coding for the annotated and well characterized cwf2 and atp-11 genes is shown in Fig. 2. Users can select one region from the last three tracks to obtain both the nucleotide and amino acid sequences. In addition, users can zoom in a region and gather the sequences from the AnABlast profile pull-down. At present, one execution of AnABlast takes approximately 1 minute for each Kb of input sequence.

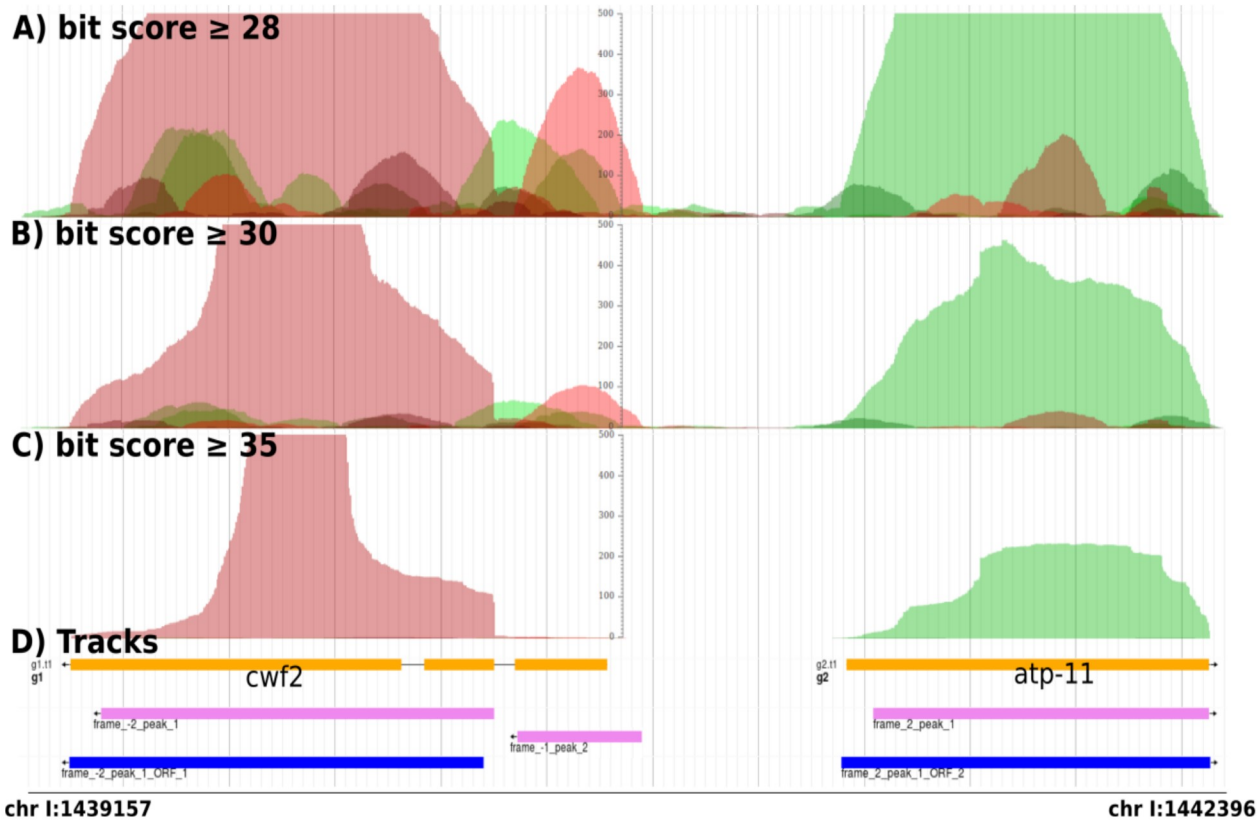


Fig. 2. Representative AnABlast profile highlighting exons in the annotated cwf2 and atp-11 protein-coding genes in the S. pombe genome (Pombase annotations). The six different colors represent profiles of accumulated alignments in protein sequences predicted from each of the six possible reading frames. A) Using a high sensitivity cut-off, several unspecific peaks appear in non-coding regions, B) using the default cut-off, peaks appear only matching with true protein-coding regions, C) using a low sensitivity cut-off, coding sequences are precisely delimited, but the first exon of cwf2 does not show any peak, D) Annotated gene exons (yellow), AnABlast peaks (pink), and ORF tracks (blue) are shown to give a reference on the genomic region where the peaks appear.

5.1.4 Examples of using AnABlast

AnABlast can be executed online at http://www.bioinfocabd.upo.es/anablast/. By default, it will use optimal parameters with a default sensitivity [7], making the use of this programme extremely simple. The most basic use of AnABlast is the identification of coding sequences in intergenic regions of annotated genomes.

5.1.4.1 Basic search of intergenic coding regions in annotated genomes

To illustrate the usage of the AnABlast web application as a new method for the search of new putative coding regions, we took the genome sequence and annotation of Salmonella

enterica subsp. enterica serovar Typhimurium str. LT2 (ASM694v2 assembly) from Ensembl Bacteria database [10], and extracted all the intergenic regions longer than 1Kb. We discarded the regions where the gene finder Prodigal predicted protein-coding genes, and used AnABlast to analyse the 8 remaining regions, which were potential candidates to harbour genes that escaped the gene finder. Two adjacent peaks were identified in a region flanked by the genes STM3083 (putative mannitol dehydrogenase) and STM3085 (putative gntR family regulatory protein) (Fig. 3). The two peaks in the forward strand were then BLASTed against UniProt to search for homologous protein sequences [11]. The first one (with the lowest signal) has an ORF with a predicted amino acid sequence similar to an Uroporphyrinogen decarboxylase (UniProt:A0A0V2D8Y5), while the second one (with the higher peak) is similar to a racemase (UniProt:A0A158N139), both of them from different species of Salmonella.

Fig. 3. AnABlast results for region AE006468:3242290-3248051 of S. enterica LT2. Protomotif accumulation for four previously annotated genes in the reverse strand are highlighted in red colors, and two novel signals appearing in the forward strand in green colors, one of which could represent a racemase protein not previously annotated. Complete ORF were found for two of the discovered regions that were missed in a conventional gene finder (Prodigal).

Overall, these results show how AnABlast is useful to discover new putative protein-coding genes where other methods have failed. In eukaryotic genomes, in addition to new putative genes as shown above in the prokaryotic Salmonella genome, AnABlast peaks also highlight new exons in annotated genes

5.1.4.2 Identification of small genes without significant homologs in the data base.

Genes encoding very small polypeptides and/or lacking significant homology to any others in the database are difficult to identify by conventional in silico methods. As an example, Fig. 4 shows the AnABlast profile of an intergenic genomic region discovering one of such putative small genes in the S. pombe genome [7].

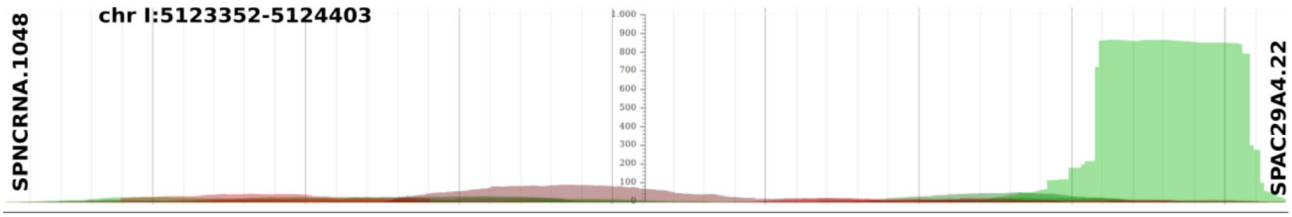


Fig. 4. AnABlast peak at Chr I: 2975642-2975772 (green color, forward strand) in the S. pombe genome encodes a small peptide with no significant similarity to known proteins in databases.

Therefore, as shown in this example, AnABlast is particularly useful in the identification of small ORFs and/or coding sequences lacking significant homology to others in databases, coding sequences that often escape to conventional searches.

5.1.4.3 Pseudogenes and rearranged DNA fragments.

The identification of ORFs within DNA regions underlined by AnABlast helps to identify coding sequences of putative new genes. However, AnABlast peaks often identify coding sequences lacking complete ORFs, accumulating both stop codons and frameshifts, which usually represent pseudogenes and fossil coding sequences. Therefore, it is also helpful in identifying genomic rearrangements (Fig. 5) or even sequences acquired by horizontal transfer, showing evolutionary remnants [7]. Furthermore, the simple visual inspection of the obtained profiles may identify near identical patterns in different chromosome locations which remark direct or inverse repeats of a genomic region (see example in Fig. 5).

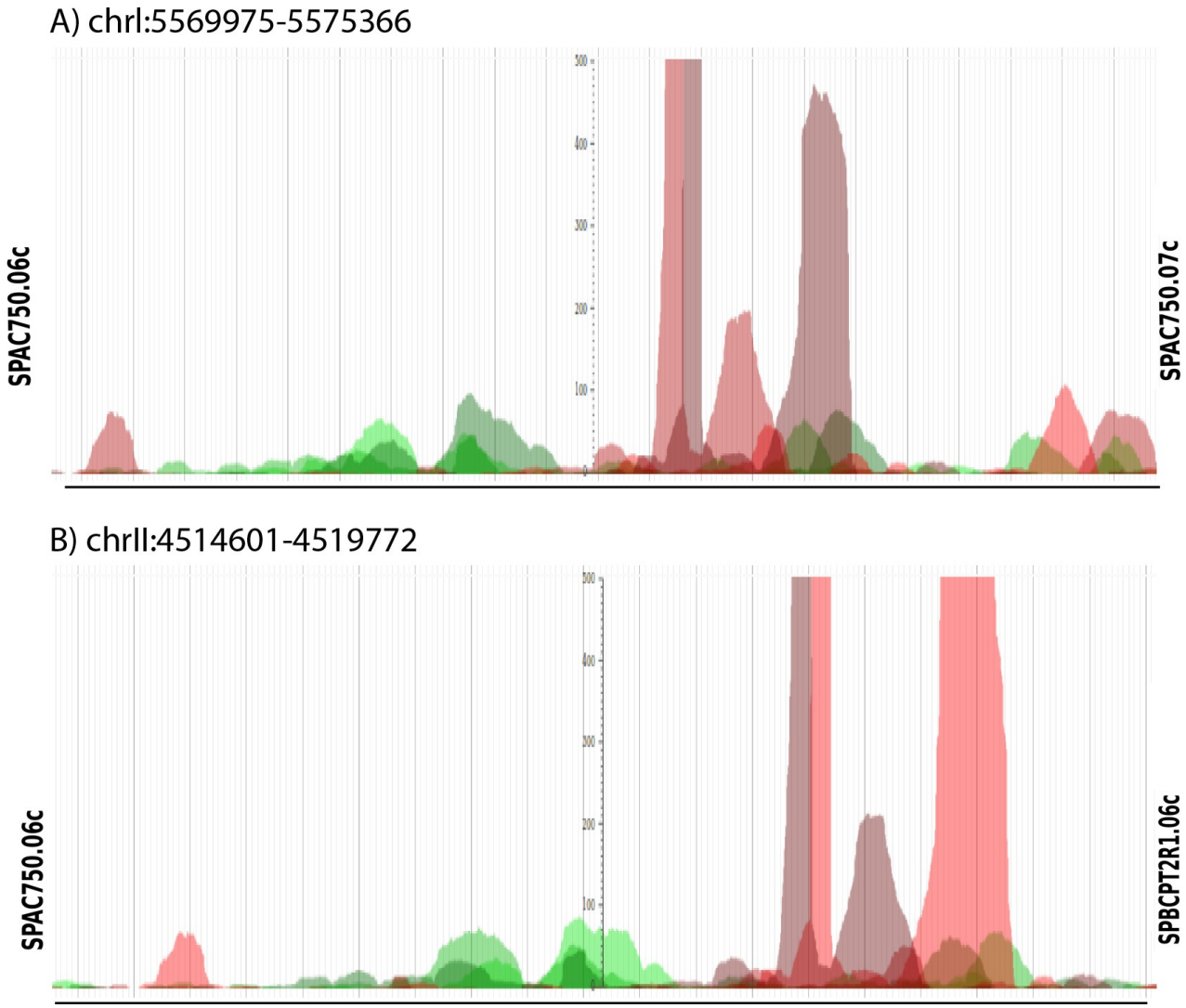


Fig. 5. Similar AnABlast profiles uncovering a rearranged region repeated in two different chromosomes (A and B) in the S. pombe genome. Annotated genes flanking the corresponding chromosomal intervals are indicated.

All the features described above make AnABlast a useful tool for the exhaustive analysis of the enormous amount of genomic data that is obtained in the present time, previous to the next post-genomic era.

5.1.5 References

1. Alioto,T. (2012) Gene prediction. Methods Mol. Biol. Clifton NJ, 855, 175–201.

2. Nesvizhskii, A.I. (2014) Proteogenomics: concepts, applications and computational strategies. Nat Methods 11: 1114-1125. doi: 10.1038/nmeth.3144

3. Guigó, R., Flicek, P., Abril, J.F., et al. (2006) EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biol 7 Suppl 1: S2.1-31

4. Hyatt, D., Chen, G.L., Locascio, P.F., et al. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11: 119 doi: 10.1186/1471-2105-11-119

5. Pérez, A.J., Thode, G., Trelles, O. (2004) AnaGram: protein function assignment, Bioinformatics 20: 291-2

6. Thode, G., García-Ranea, J.A., Jimenez, J. (1996) Search for ancient patterns in protein sequences, J Mol Evol 42: 224-33

7. Jimenez J., Duncan, C.D., Gallardo, M., et al. (2015) AnABlast: a new in silico strategy for the genome-wide search of novel genes and fossil regions. DNA Res 22: 439–449 doi: 10.1093/dnares/dsv025

8. Casimiro-Soriguer, C.S., Muñoz-Mérida, A., Pérez-Pulido, A.J. (2017) Sma3s: a universal tool for easy functional annotation of proteomes and transcriptomes. Proteomics 2017 17 doi: 10.1002/pmic.201700071

9. Stanke, M., Schöffmann, O., Morgenstern B., Waack, S. (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7: 62

10. Kersey, P.J., Allen, J.E., Armean, I., et al. (2016) Ensembl Genomes 2016: more genomes, more complexity. Nucleic Acids Res 44: D574-580 doi: 10.1093/nar/gkv1209

11. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. Nucleic Acids Res 45: D158–D169 doi: 10.1093/nar/gkw1099

6. Capítulo 3: *Drosophila melanogaster*

6.1 Ancient evolutionary signals of protein-coding sequences allow the discovery of new genes in the *Drosophila melanogaster* genome

Carlos S. Casimiro-Soriguer[1], Alejandro Rubio[1], Juan Jimenez[1], and Antonio J. Pérez-Pulido[1,*]

[1]Centro Andaluz de Biologia del Desarrollo (CABD, UPO-CSIC-JA). Facultad de Ciencias Experimentales (Área de Genética), Universidad Pablo de Olavide, 41013, Sevilla, Spain

6.1.1 Abstract

The current growth in DNA sequencing techniques makes of genome annotation a crucial task in the genomic era. Traditional gene finders focus on protein-coding sequences, but they are far from being exhaustive. The number of this kind of genes continuously increases due to new experimental data and development of improved bioinformatics algorithms. In this context, AnABlast represents a novel *in silico* strategy, based on the accumulation of short evolutionary signals identified by protein sequence alignments of low score. This strategy potentially highlights protein-coding regions in genomic sequences regardless of traditional homology or translation signatures. Here, we analyze the evolutionary information that the accumulation of these short signals encloses. Using the *Drosophila melanogaster* genome, we stablish optimal parameters for the accurate gene prediction with AnABlast and show that this new strategy significantly contributes to add genes, exons and pseudogenes regions, yet to be discovered in both already annotated and new genomes.

Keywords: *ancient sequences; gene finding; protein-coding genes; drosophila melanogaster; protomotifs*

6.1.2 Introduction

Research groups from all over the world are sequencing whole genomes as a common task, taking advantage of the current burst in the genomics era [1]. The analysis of the sequences from those genomes is essential for accurate annotation procedures. However, computational tools for gene discovery usually miss around 20% of protein-coding genes when annotating a whole genome, or even more in the case of eukaryotic organisms [2,3]. Thus, a significant number of protein-coding sequences and other functional genomic elements are missing when using currently available genomic annotation approaches.

One of the most intensively studied model organism is the fruit-fly *Drosophila melanogaster*. Its genome was sequenced in 2000, and 13,601 protein-coding genes were initially annotated, coming from the integration of the two used gene finders, which respectively predicted 13,189 and 17,464 genes [4]. From this milestone, the number of fruit-fly genes has changed, and numerous and significant discrepancies have arisen [5]. But nowadays the FlyBase database put this number at 14,133 [6], showing that the number of genes is constantly increasing over time, and a greater increase is expected to come from the discovery of new kinds of genes, such as those shorter than 100 amino acids, which in the fruit-fly genome could account for thousands of them [7].

Traditional gene finders are routinely based on both significant sequence similarity and sequence signatures such as those used to define open reading frames (ORF), signals involved in splicing [8], or combined protocols to get better results [9]. Among the new proposed methods, we have previously shown that accumulation of low-score alignments, which would represent footprints of ancient sequences, highlights present and ancient protein-coding regions which are hard to discover by conventional methods [10]. Briefly, this novel computational approach, that we named AnABlast, compares the putative amino acid sequences from the six reading frames of a genomic sequence against a non-redundant protein database, and collects the matches, including low-score alignments, which we call protomotifs. These are specifically accumulated in coding but rarely in non-coding sequences. Thus, the profile of AnABlast with peaks of accumulated protomotifs, accurately marks putative protein-coding genes, pseudogenes, and fossils of ancient

coding sequences, overcoming the effects of possible sequencing errors and reading frame shifts, since it does not search for reading frames but sequence coding signals.

Here, we use the well-studied *D. melanogaster* genome to determine optimal parameters for AnABlast and efficiently identify protein-coding sequences. By using AnABlast with early annotation versions of the database, we show that many new genes predicted by this algorithm are true genes that have been incorporated into the genome annotation of this organism. We also show that AnABlast is useful to discover small ORF and fossil sequences that are hidden to conventional gene finder algorithms, and show how this new strategy can contribute to discover the complete set of protein-coding regions of a whole genome.

6.1.3 Results and Discussion

Searching for protein-coding signals in the fruit-fly genome

AnABlast is a computational tool that searches for protein-coding regions in whole genomes by taking into account low-score alignments shared by multiple unrelated protein sequences. To this end, AnABlast uses the putative amino acid sequences translated from a genomic sequence to search for sequence similarity in a non-redundant protein database. Alignments obtained from this similarity search (called protomotifs), including those of a low score, are then piled up along the query sequence, and peaks accumulating protomotifs above a specific threshold will highlight potential protein-coding regions and will be considered coding-signals (Fig. 1). Finally, these coding-signals can be evaluated: those ones underlying exons from a protein-coding gene will be true positive predictions, exons without coding signals will be false negatives, and coding-signals underlying introns or intergenic regions will be putative false positives. But it should be noted that the false positives could potentially underlie new coding sequences which escaped to conventional annotation pipelines. Thus, false positives highlighted by AnABlast may represent genomic regions encoding putative new proteins, but also non-functional degenerated protein-coding regions, something of particular interest in current genome research.

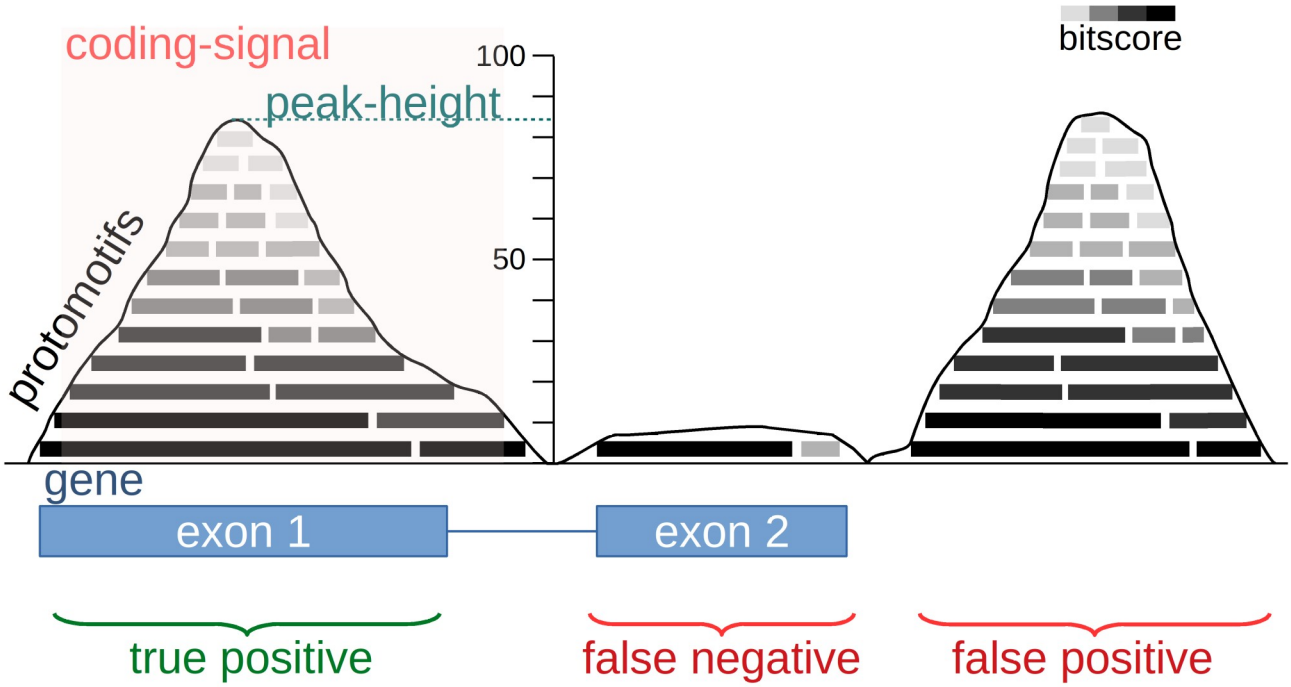


Fig. 1. Schematic diagram showing AnABlast profiles obtained in a theoretical genomic region with a two-exon gene. The peak-height is the maximum accumulation of protomotifs in a specific genomic position (BLAST alignments including low bit-scores). Peaks with a protomotif accumulation above a peak-height threshold are considered as putative protein-coding regions (coding-signal). Significant peaks matching a known exon represents true positive peaks, while those underlying a genomic region without known exons are considered false positive coding-signals. Well-known exons which do not significantly accumulate protomotifs (peak-height below the threshold) constitute false negatives.

To test the capability of AnABlast to discover protein-coding regions, we used this algorithm to analyze the whole genome of the fruit-fly *D. melanogaster* (2012 and 2017 releases). Protein sequences coming from the virtual translation of the complete fly genome (in all the six reading frames) were subjected to BLAST search against a non-redundant protein database (UniRef50) under a low restriction threshold, and the resulting alignments were accumulated along the query sequence to produce the AnABlast profile. Then, all well-known exons of this genome were compared with the set of putative coding regions identified by AnABlast. As expected, most of the AnABlast peaks with a high protomotif accumulation matched annotated exons (putative true positives), but a small fraction of them fell in both intronic and intergenic regions lacking of any annotated gene, exon or pseudogene (Suppl. file 1, genomic browser with AnABlast results). These false positive signals represent a particularly interesting set of genomic regions, since they could constitute new protein-coding regions.

Protomotifs underlie into the true reading frame

Protein-coding signals highlighted by AnABlast are mainly composed by protein sequence alignments of low score, but also occasionally high score. To test if such alignments are just random, or they actually match true protein-coding regions, we studied the distribution of protomotifs underlying protein-coding regions at different BLAST bit-scores, regarding to the different possible reading frames. Though millions of protomotifs were scattered throughout the fruit-fly genome within annotated exons, most of them were concentrated in the right reading frame, with a much lower number found in any of the other possible five reading frames (Fig. 2).

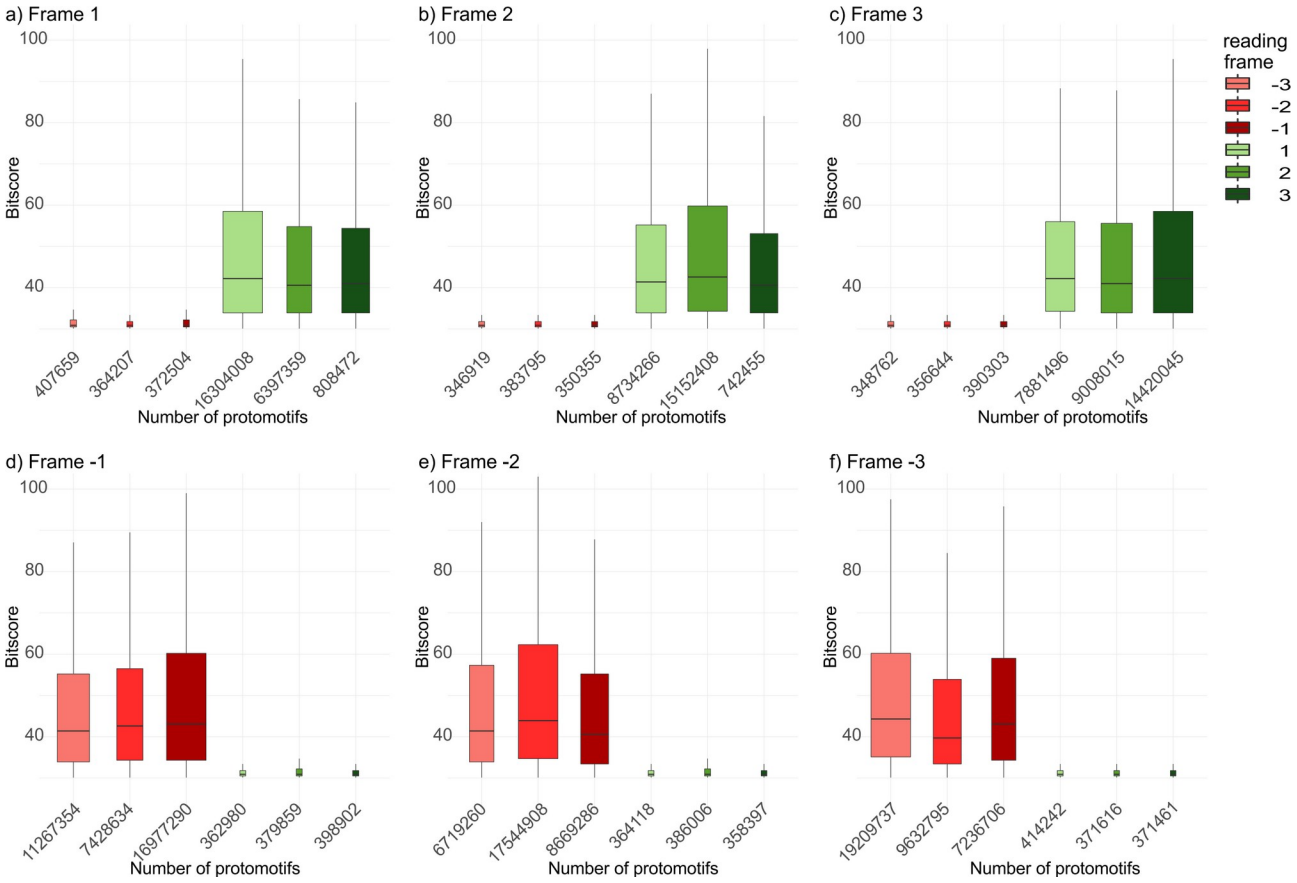


Fig. 2. Distribution of protomotifs coming from true positive coding-signals separated by the true reading frame of the protein-coding sequences where they accumulate. The different parts of the figure represent protomotifs accumulated in protein-coding sequences at different BLAST bit-score starting in a) frame +1, b) frame +2, c) frame +3, d) frame -1, e) frame -2, and f) frame -3. The box size is proportional to the number of protomotifs in that frame, and the exact number of protomotifs is also shown below the X-axis. The three reading frames coming from the forward strand are colored in green color, and the three coming from the reverse strand are colored in red color.

Thus, protomotifs are mainly accumulated in the true reading frame in spite of their low score. Interestingly, a significant number of them accumulate in the right strand but at a

different reading frame. Contrarily, the other strand shows an enormously reduced number of protomotifs, with a lower order of magnitude. In fact, the protomotifs accumulated in the contrary strand present bit-score values lower than 30, but those accumulated in the true strand present values near to 100. This result suggests, from an evolutionary point of view, that new protein-coding genes might putatively come just from shifting the reading frame in the same strand.

Optimization of AnABlast parameters for the efficient prediction of protein-coding signal

Until now we have seen how AnABlast coding-signals mainly match to protein-coding region in the genome, but we did not use any threshold to evaluate the results and measure the accuracy in the procedure of gene prediction. To optimize AnABlast parameters for the identification of new exons and genes, the distribution of true and false positive coding-signals were evaluated at different peak-height thresholds. AnABlast profiles depend on the bit-score value used to restrict alignment significance during the BLAST search, therefore, in addition to the value of bit-score 30, previously used by AnABlast [10], the evaluation was carried out also using the more and less restrictive bit-scores of 29 and 31 respectively. Regardless of the taken score, true positive coding-signals account for the highest peak-height (Fig. 3), though under more restricted score values (higher bit-score), AnABlast peaks were more selective and focused into the protein-coding regions of the genome (higher peak-heights). However, the absolute number of true protein-coding regions dropped down with such higher scores, decreasing the number of peaks underling protein-coding sequences (Fig. 3c). On average, predicted coding-signals falling in non-coding regions (false positives) have much lower peak-height values (Fig. 3b). However, the distribution of these false positives show outliers with peak-height values indistinguishable from the true positive set, which could be considered as new putative protein-coding regions.

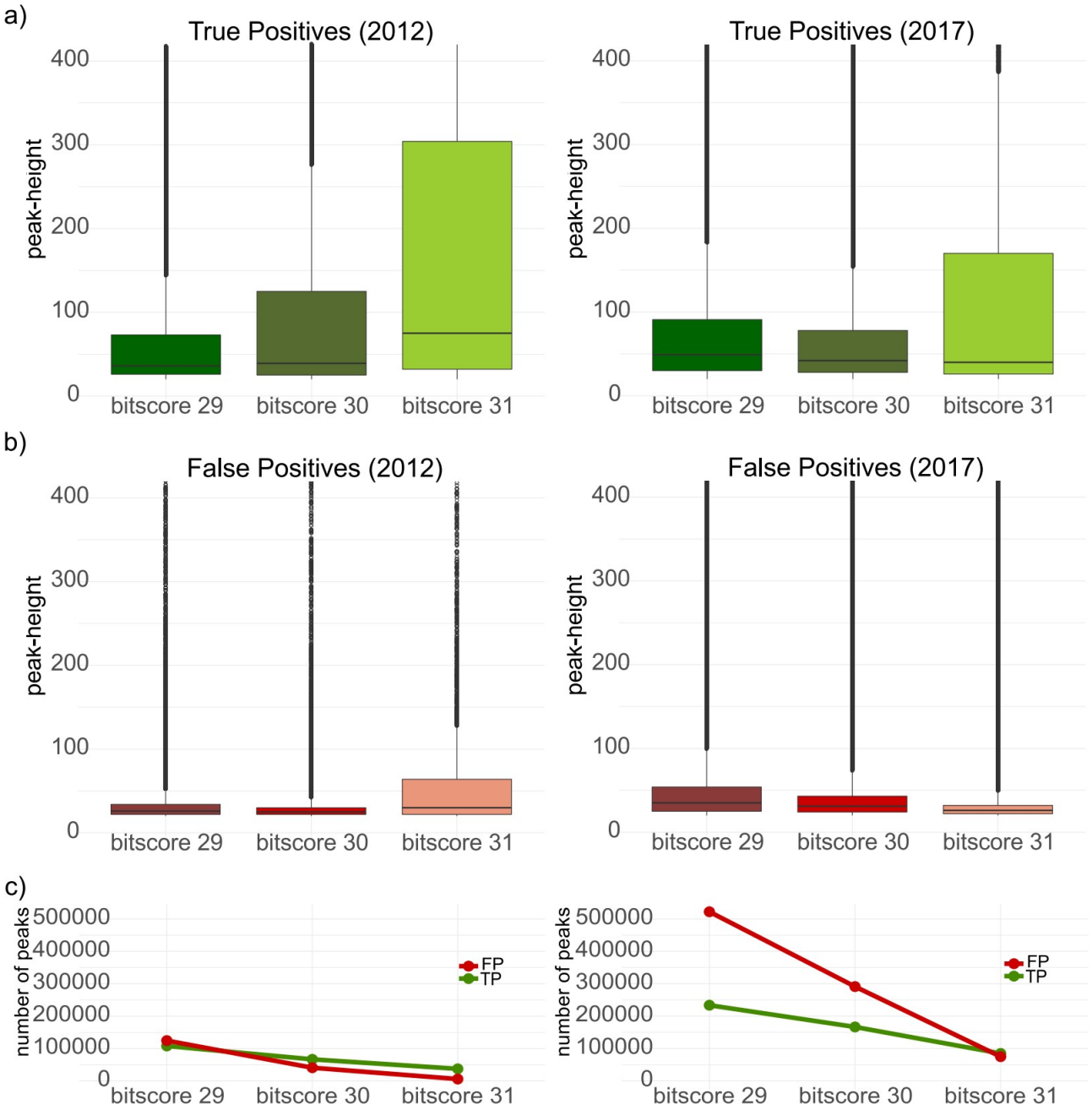


Fig. 3. Peak-height distribution and number of coding-signals found at different bit-score values. Peak-height distributions are separated by a) true positive and b) false positive, and they are shown by each database release (2012 and 2017) and three different bit-score thresholds. The outliers are shown as a chain of points above the boxes. c) The number of true and false positive coding-signals at any peak-height with the corresponding bit-score thresholds (note that it shows number of peaks with peak-heights as low as 20 and higher).

In a specific genome, the accuracy of the protein-coding sequence prediction by AnABlast not only depends on the bit-score value, but also on the peak-height threshold coming from the alignment accumulation. Under a bit-score value of 30, the optimal peak-height cutoff depends on the used database and the amount of sequences that it contains. In this way, when the most current database from 2017 was used, the false positive outliers

appear from a peak-height of 70, so proposing that peaks higher than this value could predict new protein-coding regions. However, when using the 2012 database, this peak-height value was around to 40.

To better test both the precision and recall of AnABlast in predicting protein-coding sequences, coding-signals were compared against the gene annotation of fruit-fly genome and the accuracy of AnABlast prediction in this set was analyzed. As expected, the recall is higher when using a more recent database (release 2017, with more than 21 million proteins) compared to an older one (release 2012, with around 4.5 million proteins) at the same peak-height. When using the database of 2017, the precision has an asymptote at around peak-height equal to 100 with a value of around 90% (only 1 in 10 predictions are not right), though the recall at this threshold is only of 65% (only 6.5 in 10 of the true protein-coding sequences are recovered) (Fig. 4a). However, this accuracy is reached with peak-height equal to 35, when the older release of the database is used (Fig. 4b).

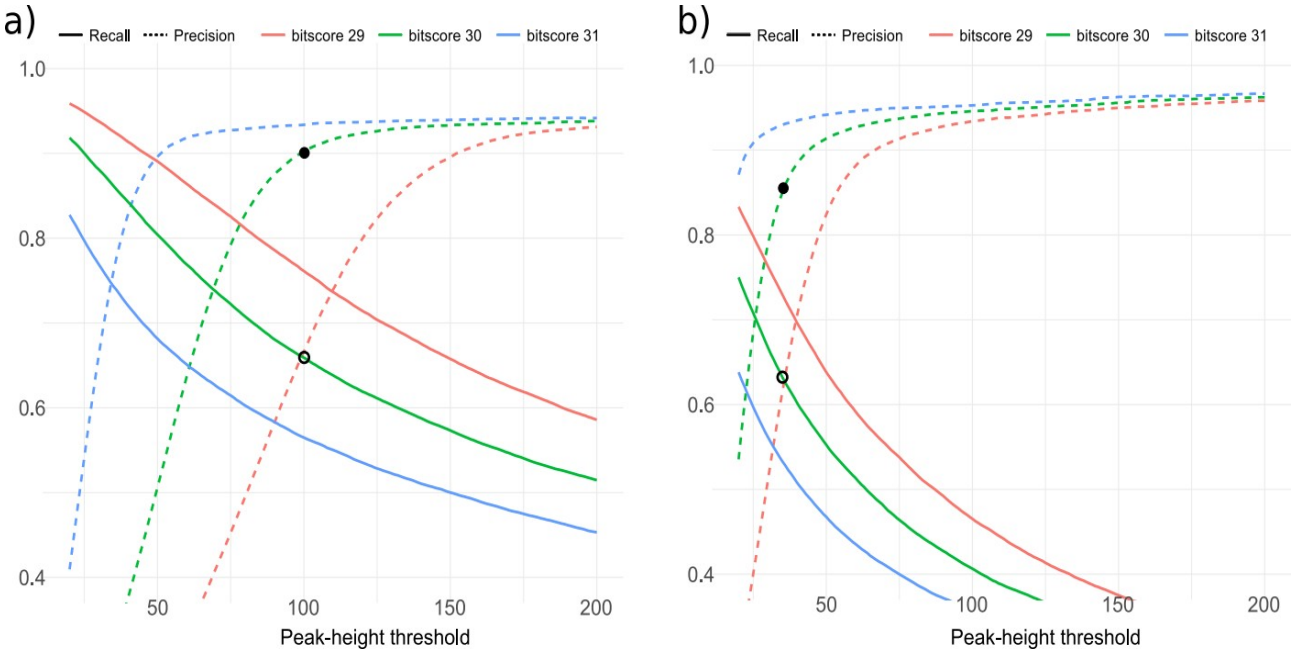


Fig. 4. Recall and precision of AnABlast at different bit-score thresholds. Values were calculated when using the databases: a) release 2017, and b) release 2012. The black dot marks the precision value at bit-score 30, and the unfilled dot marks the recall. The complete results and values for all the used parameters can be found in Suppl. file 2.

As described above, the precision varies regarding to the bit-score threshold used, and a higher precision and lower recall are reached when more restrictive values are used. So, to ensure a high accuracy we chose bit-score 30 and a peak-height threshold of 100. By

using these parameters, we expect that AnABlast could discover new unknown protein-coding genes and exons inside the 10% putative false positives. However, with the older database, with a number of sequences almost five times lower, we should take a peak-height threshold of 35. All of this gives a great number of AnABlast coding-signals matching with protein-coding region spread over the fruit-fly chromosomes, and between 4500-7000 (depending of the used database) candidates to be new protein-coding sequences (Table 1; Suppl. file 3). These restricted parameters allow finding more than 30,000 exons from the current fruit-fly annotation.

Table 1. Number of true and false positives predicted by AnABlast using the 2012 and 2017 databases, and separated by chromosomes. The peak-height threshold used was 40 (2012) and 100 (2017).

		chromosomes							
Coding-signals		**2L**	**2R**	**3L**	**3R**	**4**	**X**	**Y**	**Total**
True positives (2012)		6,691	7,515	6,954	9,085	387	6211	81	36,924
False positives (2012)		1,035	1,390	1,384	1,517	40	1,143	483	6,992
True positives (2017)		6,139	6,798	6,164	8,325	379	5,390	72	33,267
False positives (2017)		535	1,068	947	913	41	443	518	4,465

AnABlast is able to discover current genes using an old database

The number of annotated protein-coding sequences is continuously revisited in annotated genomes, and new genes, exons and pseudogenes continuously appear as a consequence of experimental results and new *in silico* approaches. For instance, when the FlyBase database released in 2012 is compared to the current 2017 release, it can be found that 38 protein-coding genes, 91 exons from well-known genes, and 74 pseudogenes entered into the database later than 2012. This dataset of true protein-coding sequences absent in the 2012 allows us to carry out a simulation to estimate the efficiency of AnABlast in discovering new protein-coding sequences. Remarkably, when using the 2012 FlyBase database, with the parameters previously suggested (bit-score 30 and peak-height 35), we found that AnABlast highlights the majority of the protein-coding sequences from this dataset (Table 2). More than 60% of the protein-coding genes are found, and also the 80% of the pseudogenes were predicted by AnABlast. These results

improve when using a less restricted peak-height value. In the case of new exons, their small length (some of them are coding for only a few amino acids) makes extremely difficult the *in silico* identification. However, up to 11% of them were also discovered by AnABlast, increasing up to 60% when changing the default peak-height to 26, which present a precision of 70% (Table 2; Suppl. file 4). Overall, it is important to highlight that the most of these new protein-coding sequences predicted by AnABlast were not found by the widely used gene finder AUGUSTUS [11].

Table 2. Genomic elements from the database release 2017 discovered using the database release 2012, separated by peak-height threshold. Note that '<5'show the false negatives (when less than 5 protomotifs were found), and the last column show the most significant true positives.

Sequence type	annotated (2017)	found by AUGUSTUS gene finder	peak-height (<5)	peak-height (<26)	peak-height (26-35)	peak-height (>35)
Protein-coding gene	38	9	0 (0%)	14 (37%)	1 (2%)	23 (61%)
Exon	91	9	12 (13%)	55 (60%)	14 (16%)	10 (11%)
Pseudogene	74	15	0 (0%)	7 (9%)	8 (11%)	59 (80%)

The identification of very small genes is still challenging for *in silico* strategies, including AnABlast. One of the new genes that AnABlast failed to identify in the 2012 database (CG45546) is coding for a short protein of 93 amino acids (Fig. 5a). Interestingly, AnABlast efficiently identified it when using the 2017 database, due to the fact that this sequence and its putative homologs were now included in the database, increasing the peak-height to a significant level. This gene is still lost by AUGUSTUS, even when using the current database release. To discard that these coding-signals underlined by AnABlast occur by chance, the reverse sequence of this gene was used as a negative control. When this control is analyzed, AnABlast profiles present no accumulation of protomotifs (Fig. 5a, below). Furthermore, we shuffled the sequence of the gene, and the 85% of the simulations did not present any protomotif, and the remaining 15% gave peak-height values lower than 18 (Suppl. File 5).

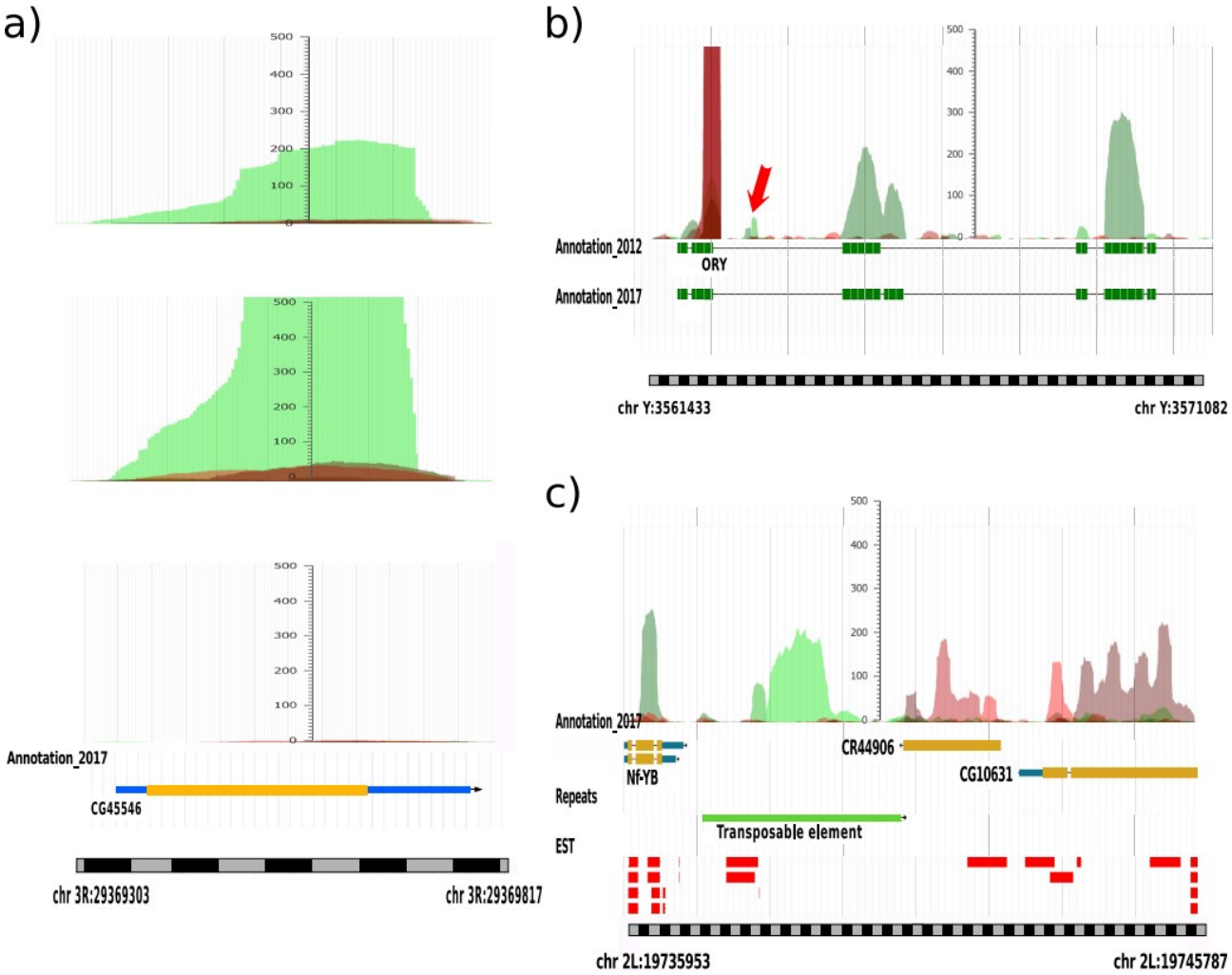


Fig. 5. AnABlast profile for three regions of the fruit-fly genome. Green color represents protomotif accumulation in the forward strand, and red color in the reverse strand. a) Different analysis for the gene CG45546 region, from top to bottom: using database release 2012, 2017 and using the reverse sequence as random query; b) region including part of the ory gene (CG40446), together with the exons annotated in both database releases (the red arrow marks two peaks corresponding to an ancient mobile element); c) region of the pseudogene CR44906, including surrounding genes and a transposable element in the 5' end. An additional track with EST signals (Expressed Sequence Tags) is shown, which suggests expression for the transposable sequence.

Some new exons are also found and highlighted within well-annotated genes. One of this exon was found in the *Ory* gene (CG40446). The exon appearing in the 2017 database is found by AnABlast using the 2012 release and a peak-height higher than 35 (Fig. 5b). Interestingly, AnABlast produced two weak peaks within an intronic sequence of this gene, with a peak-height around 50, very similar to others in the 3' end. Conventional search for homologous sequences to these AnABlast coding-signals revealed similarity with retrotransposons from invertebrate organisms, suggesting that this genomic region is coding for ancient proteins of a mobile element. In addition, a high peak overlapping with an exon in the 5' end is also emerging in the reverse strand, which represents a tri-

nucleotide region coding for amino acid repeats in all the reading frames. This artefact is characteristic of nucleotide repeats, and it can be avoided by enabling the low-complexity filter in the similarity search step with BLAST.

In addition to new genes and exons, AnABlast was able to discover 59 pseudogenes which did not appear in the used 2012 database (Table 2). This shows the ability of AnABlast for discovering protein-coding regions regardless of the presence of a complete open reading frame. One of these pseudogenes (CR44906), included in FlyBase in 2013, is clearly highlighted by AnABlast in the reverse strand of the 2012 database (Fig. 5c). Remarkably, another coding-signal is found in the forward strand, upstream of this pseudogene. In deep analysis of this sequence revealed that it encodes the transposase of an annotated transposable element. The presence of numerous expression sequence tags (EST) support the expression of this sequence. However, this transposase is not yet annotated in FlyBase database.

Putative new protein-coding sequences in the present database

Finally, according to the efficient identification of protein-coding sequences highlighted by AnABlast, it is expected that after a future further characterization, a considerable fraction of the false positive sequences predicted when the 2017 database is used become true positives. One of these candidates is found 3' upstream to the genes CG44014 and CG44013, coding for uncharacterized proteins bearing a calycin domain related to extracellular proteins and involved on lipid transport. AnABlast suggests a significant coding-signal in this region (Fig. 6). The putative protein sequence encoded by this AnABlast region has no homologues in other organisms, but it is located in an evolutionary conserved region, again matching with EST signals which also support the putative expression of this genomic region. A list of false positives which could likely propose new putative protein-coding sequences is available in Suppl. file 3, and in tracks FP (False Positives) in Suppl. File 1.

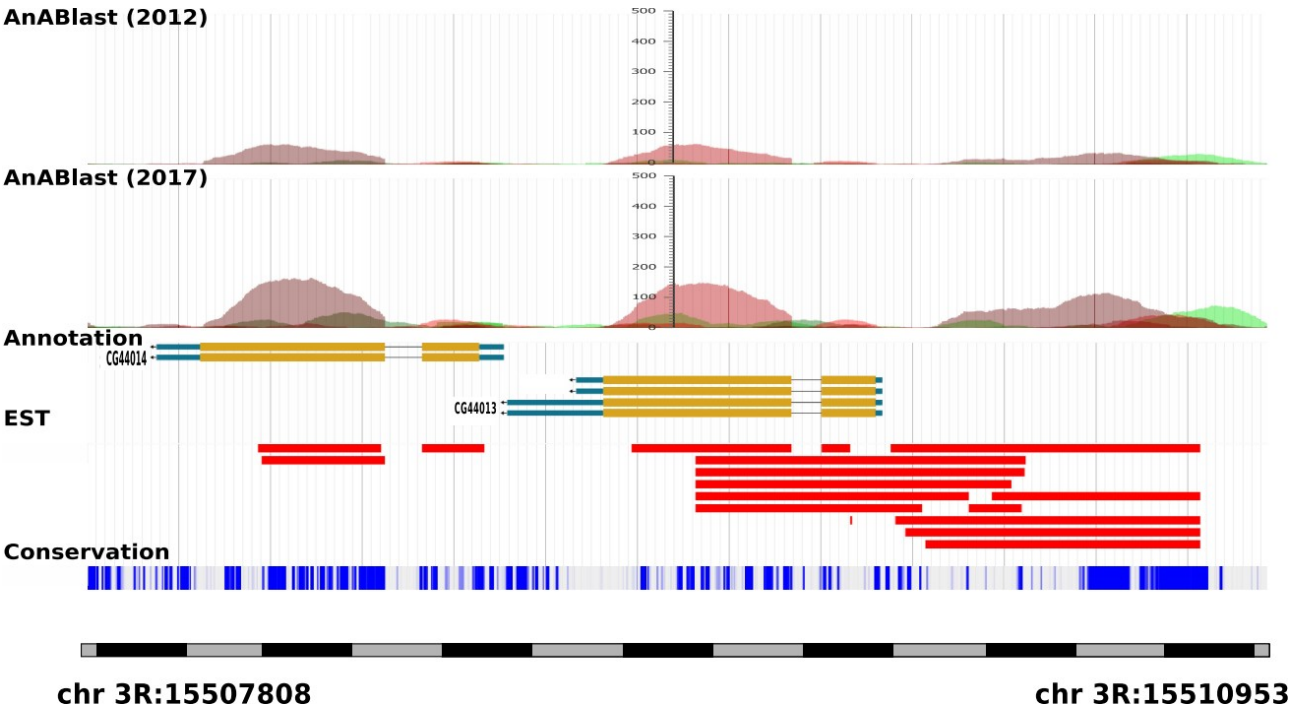


Fig. 6. AnABlast profile for a region with a putative new gene. The profiles were created with both 2012 and 2017 release databases, and they are shown together with the gene, the EST track and one additional track taken from the UCSC browser representing the evolutionary conservation of the sequence versus 27 different insect genomes, which shows a high conservation for the proposed new gene.

Discussion

Currently, gene finder algorithms have a limited recall and usually lose the 10-20% of the true coding regions [3], especially those lacking homologs and/or having non-conventional characteristics such as small ORFs or pseudogenes. It leads to the necessary development of new algorithms based on different ideas [12,13]. In this context, we proposed a new *in silico* strategy, named AnABlast that uses low-score alignments coming from multiple non-redundant proteins [10]. As shown in this study, in agreement with previous reports [14–16], these alignments (protomotifs) do not accumulate randomly but in true genomic protein-coding regions (Supplementary file 1; Fig. 3). By using the *D. melanogaster* genome as a model system, we set optimal parameters for using AnABlast as a new protein-coding finder in whole sequenced genomes.

AnABlast has not the aim of annotating an entire genome, since it allows the identification of only 60-85% of the actual genes annotated in a genome (Table 2). However, the accumulation of protein sequences using a non-redundant database and low bit-scores in the BLAST search enables AnABlast to discover new genes that scape to conventional

strategies, producing a precision up to 90% with exons, genes and pseudogenes among the identified protein-coding signals (Fig. 4). Thus, AnABlast is particularly useful to research for new genes in already annotated genomes. Another advantage of the AnABlast strategy is the fact that protomotifs are accumulated within the true reading frame, and not scattered throughout the genome (Fig. 2). Importantly, coding signals are also identified by AnABlast in the coding strand, but at different reading frames. It rarely occurs in any of the three reading frames coming from the reverse strand, suggesting that new protein-coding exons or genes may emerge by frameshift mutations in preexisting ORFs [17]. In fact, the peak-height distribution matching ORFs in true genes has the highest value, followed by peaks identified in the next frame, which suggests that new protein-coding regions may emerge by point deletions in the original frame. This observation agree with previous evidences in mammals suggesting a higher frequency of evolutionary fixation for deletion than for insertion mutations [18,19], a trend that have also been found in the *D. melanogaster* genome [20]. Remarkably, AnABlast coding-signal are sometimes found in the ends of well-known genes, overlapping with the right reading frame and suggesting than C-terminus and N-terminus of genes are subjected to evolutionary contractions and expansions that are efficiently identified by AnABlast [10].

The discovery of protein-coding genes by AnABlast is independent of the appearance of an open reading frame, a feature that allows the discovery of sequences without canonical structures, such as pseudogenes and transposable sequences. Disabled or unitary pseudogenes originated from inactivated genes are particularly difficult to identify due to its high sequence divergence after long-term evolution [21]. Since AnABlast searches for the accumulation of footprints of common ancient protein sequences (low-score patterns), this strategy is particularly useful in underlying fossil sequences in which significant homology is lost (Fig. 5c).

Another important challenge to the whole annotation of genomes is the discovery of short ORFs [7]. These short protein-coding sequences were missed in the past, since it is difficult to distinguish between functional open reading frames and non-functional ones arisen by chance [22]. Albeit less efficiently, AnABlast is also useful for assisting in this task (Fig. 5a). Altogether, we encourage the use of AnABlast as a good *in silico* method that complement current gene finder algorithms and conventional genome annotation tools.

6.1.4 Conclusions

The present study shows how AnABlast, which uses a strategy based of the accumulation of low bit-score protein homologs along a query protein sequence, is able to discover new putative protein-coding genes/exons where other methods fail. AnABlast is also able to locate pseudogenes showing evolutionary remnants or even small ORFs that escape the conventional searches. All these features makes of AnABlast a meaningful tool for the exhaustive analysis of genomic data, currently produced at an increasingly rapid rate. To allow the analysis of genomic regions and searching for new protein-coding genes, we have built a web application which is available at http://www.bioinfocabd.upo.es/anablast/ [23]. Our results aim to analyze new genomes as well as to revisit annotated ones in order to discover new hidden genes.

6.1.5 Materials and Methods

Search for protein-coding signals

AnaBlast was used to search for protomotifs using the release 6 from *D. melanogaster* genome versus the UniRef50 database from January 2012 (with 4,606,913 sequences) and January 2017 (with 21,859.863 sequences), independently. UniRef50 is a protein database with non-redundant sequences in 50% identity threshold [24]. Blastx was used to get hits (that we call protomotif) with a threshold e-value of 10 and a bit-score between 29-31, which gave significant results in other projects [10], though the e-value has been decreased in order to optimize the analysis of a complete genome.

Protomotifs were classified by reading frame, when they matched to well-known exons in the genome. The distribution of protomotifs was made using the ggplot library of R programming language.

The genome analysis was performed in a HPC cluster, using 100 threads and it lasted around 1 week. The remaining analysis with sequences up to 10kb were performed in the web application of AnABlast, which allows to analyze genomic sequences up to 25Kb, or

longer if the user provides the precalculated BLAST report: http://www.bioinfocabd.upo.es/anablast/

Testing protocol

The *D. melanogaster* genome annotation release dmel-all-r5.43 from January 2012 was converted to the release dmel-all-r6.19 from January 2017 using the conversion tool from the FlyBase database. Genes were compared with bedtools intersect, obtaining all the new exons, complete genes, and pseudogenes appearing in release 6 but not in release 5. To a higher constraint, the sequences were searched in UniProt database to discard previously described protein-coding genes, and only genes not appearing in any database release before 2017 were maintained. The remaining sequences were taken and used as the testing dataset. For the testing protocol, Blastx was run with the genome release 6 and the Uniref50 database release from January 2012. The sequences from the testing dataset were taken with 100 nucleotides both in the 5' and 3' ends, previous to analyze by AnABlast.

Both tracks for EST sequences and conservation (27 insects conservation by PhastCons) were obtained from the UCSC browser [25].

Accuracy measurement

A coding-signal is considered to match with an annotated exon when at least the 20% of the exon is covered, or the 20% of the peak underlies the exon. To check this, AnaBlast results were converted to bed format and compared to the GFF file with the annotated genes from the *D. melanogaster* release 6. Accuracy was measured by comparing exons and protein-coding signal from AnABlast, considering true positives (TP, AnABlast coding-signals matching to exons, or pseudogenes), false positives (FP, AnABlast coding-signals matching to introns or intergenic regions), and false negatives (FN, exons or pseudogenes without AnABlast coding-signals). Then, precision (specificity of the analysis: percentage of right predictions in the results) and recall (sensitivity of the analysis: percentage of right elements which are predicted) was calculated:

Precision = (TP/TP+FP) x 100

Recall = (TP/TP+FN) x 100

Acknowledgements

This research was supported by the Ministry of Economy and Competitiveness of the Spanish Government grant BFU2016-77297-P. We would like to thank C3UPO for the HPC support, and Manuel S. Casimiro-Soriguer for designing the logo.

Authors' contributions

C.S.C. performed the algorithm construction and training, A.R. performed the predictions into the different databases, and both of them contributed into the web application and genomic browser design. J.J. contributed to the algorithm and test design, A.J.P. conceived and coordinated the study together with J.J., carried out the design, and wrote the manuscript.

6.1.6 References

[1] N.J. Loman, M.J. Pallen, Twenty years of bacterial genome sequencing, Nat. Rev. Microbiol. 13 (2015) 787–794. doi:10.1038/nrmicro3565.

[2] R. Guigó, P. Flicek, J.F. Abril, A. Reymond, J. Lagarde, F. Denoeud, S. Antonarakis, M. Ashburner, V.B. Bajic, E. Birney, R. Castelo, E. Eyras, C. Ucla, T.R. Gingeras, J. Harrow, T. Hubbard, S.E. Lewis, M.G. Reese, EGASP: the human ENCODE Genome Annotation Assessment Project, Genome Biol. 7 Suppl 1 (2006) S2.1-31. doi:10.1186/gb-2006-7-s1-s2.

[3] S.J. Goodswen, P.J. Kennedy, J.T. Ellis, Evaluating high-throughput ab initio gene finders to discover proteins encoded in eukaryotic pathogen genomes missed by laboratory techniques, PloS One. 7 (2012) e50609. doi:10.1371/journal.pone.0050609.

[4] M.D. Adams, S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S.E. Scherer, P.W. Li, R.A. Hoskins, R.F. Galle, R.A. George, S.E. Lewis, S. Richards, M. Ashburner, S.N. Henderson, G.G. Sutton, J.R. Wortman, M.D. Yandell, Q. Zhang, L.X. Chen, R.C. Brandon, Y.H. Rogers, R.G. Blazej, M. Champe, B.D. Pfeiffer, K.H. Wan, C. Doyle, E.G. Baxter, G. Helt, C.R. Nelson, G.L. Gabor, J.F. Abril, A. Agbayani, H.J. An, C. Andrews-Pfannkoch, D. Baldwin, R.M. Ballew, A. Basu, J. Baxendale, L. Bayraktaroglu, E.M. Beasley, K.Y. Beeson, P.V. Benos, B.P. Berman, D. Bhandari, S. Bolshakov, D. Borkova, M.R. Botchan, J. Bouck, P. Brokstein, P. Brottier, K.C. Burtis, D.A. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, J.M. Cherry, S. Cawley, C. Dahlke, L.B. Davenport, P. Davies, B. de Pablos, A. Delcher, Z. Deng, A.D. Mays, I. Dew, S.M. Dietz, K. Dodson, L.E. Doup, M. Downes, S. Dugan-Rocha, B.C. Dunkov, P. Dunn, K.J. Durbin, C.C. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Fosler, A.E. Gabrielian, N.S. Garg, W.M. Gelbart, K. Glasser, A. Glodek, F. Gong, J.H. Gorrell, Z. Gu, P. Guan, M. Harris, N.L. Harris, D. Harvey, T.J. Heiman, J.R. Hernandez, J. Houck, D. Hostin, K.A. Houston, T.J. Howland, M.H. Wei, C. Ibegwam, M. Jalali, F. Kalush, G.H. Karpen, Z. Ke, J.A. Kennison, K.A. Ketchum, B.E. Kimmel, C.D. Kodira, C. Kraft, S. Kravitz, D. Kulp, Z. Lai, P. Lasko, Y. Lei, A.A. Levitsky, J. Li, Z. Li, Y. Liang, X. Lin, X. Liu, B. Mattei, T.C. McIntosh, M.P. McLeod, D. McPherson, G. Merkulov, N.V. Milshina, C. Mobarry, J. Morris, A. Moshrefi, S.M. Mount, M. Moy, B. Murphy, L. Murphy, D.M. Muzny, D.L. Nelson, D.R. Nelson, K.A. Nelson, K. Nixon,

D.R. Nusskern, J.M. Pacleb, M. Palazzolo, G.S. Pittman, S. Pan, J. Pollard, V. Puri, M.G. Reese, K. Reinert, K. Remington, R.D. Saunders, F. Scheeler, H. Shen, B.C. Shue, I. Sidén-Kiamos, M. Simpson, M.P. Skupski, T. Smith, E. Spier, A.C. Spradling, M. Stapleton, R. Strong, E. Sun, R. Svirskas, C. Tector, R. Turner, E. Venter, A.H. Wang, X. Wang, Z.Y. Wang, D.A. Wassarman, G.M. Weinstock, J. Weissenbach, S.M. Williams, null WoodageT, K.C. Worley, D. Wu, S. Yang, Q.A. Yao, J. Ye, R.F. Yeh, J.S. Zaveri, M. Zhan, G. Zhang, Q. Zhao, L. Zheng, X.H. Zheng, F.N. Zhong, W. Zhong, X. Zhou, S. Zhu, X. Zhu, H.O. Smith, R.A. Gibbs, E.W. Myers, G.M. Rubin, J.C. Venter, The genome sequence of Drosophila melanogaster, Science. 287 (2000) 2185–2195.

[5] S. Karlin, A. Bergman, A.J. Gentles, Genomics: Annotation of the *Drosophila* genome, Nature. 411 (2001) 259–260. doi:10.1038/35077152.

[6] J. Thurmond, J.L. Goodman, V.B. Strelets, H. Attrill, L.S. Gramates, S.J. Marygold, B.B. Matthews, G. Millburn, G. Antonazzo, V. Trovisco, T.C. Kaufman, B.R. Calvi, N. Perrimon, S.R. Gelbart, J. Agapite, K. Broll, L. Crosby, G. dos Santos, D. Emmert, L.S. Gramates, K. Falls, V. Jenkins, B. Matthews, C. Sutherland, C. Tabone, P. Zhou, M. Zytkovicz, N. Brown, G. Antonazzo, H. Attrill, P. Garapati, A. Holmes, A. Larkin, S. Marygold, G. Millburn, C. Pilgrim, V. Trovisco, P. Urbano, T. Kaufman, B. Calvi, B. Czoch, J. Goodman, V. Strelets, J. Thurmond, R. Cripps, P. Baker, FlyBase 2.0: the next generation, Nucleic Acids Res. 47 (2019) D759–D765. doi:10.1093/nar/gky1003.

[7] J.-P. Couso, P. Patraquim, Classification and function of small open reading frames, Nat. Rev. Mol. Cell Biol. 18 (2017) 575–589. doi:10.1038/nrm.2017.58.

[8] T. Alioto, Gene prediction, Methods Mol. Biol. Clifton NJ. 855 (2012) 175–201. doi:10.1007/978-1-61779-582-4_6.

[9] F. Zickmann, B.Y. Renard, IPred - integrating ab initio and evidence based gene predictions to improve prediction accuracy, BMC Genomics. 16 (2015) 134. doi:10.1186/s12864-015-1315-9.

[10] J. Jimenez, C.D.S. Duncan, M. Gallardo, J. Mata, A.J. Perez-Pulido, AnABlast: a new in silico strategy for the genome-wide search of novel genes and fossil regions, DNA Res. 22 (2015) 439–449. doi:10.1093/dnares/dsv025.

[11] M. Stanke, O. Schöffmann, B. Morgenstern, S. Waack, Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources, BMC Bioinformatics. 7 (2006) 62. doi:10.1186/1471-2105-7-62.

[12] S.S. Gross, C.B. Do, M. Sirota, S. Batzoglou, CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction, Genome Biol. 8 (2007) R269. doi:10.1186/gb-2007-8-12-r269.

[13] D.R. Kelley, B. Liu, A.L. Delcher, M. Pop, S.L. Salzberg, Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering, Nucleic Acids Res. 40 (2012) e9. doi:10.1093/nar/gkr1067.

[14] G. Thode, J.A. García-Ranea, J. Jimenez, Search for ancient patterns in protein sequences, J. Mol. Evol. 42 (1996) 224–233.

[15] M.A. Andrade, Position-specific annotation of protein function based on multiple homologs, Proc. Int. Conf. Intell. Syst. Mol. Biol. (1999) 28–33.

[16] A.J. Pérez, G. Thode, O. Trelles, AnaGram: protein function assignment, Bioinforma. Oxf. Engl. 20 (2004) 291–292.

[17] J. Raes, Y. Van de Peer, Functional divergence of proteins through frameshift mutations, Trends Genet. TIG. 21 (2005) 428–431. doi:10.1016/j.tig.2005.05.013.

[18] Z. Zhang, M. Gerstein, Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes, Nucleic Acids Res. 31 (2003) 5338–5348.

[19] M.S. Taylor, C.P. Ponting, R.R. Copley, Occurrence and consequences of coding sequence insertions and deletions in Mammalian genomes, Genome Res. 14 (2004) 555–566. doi:10.1101/gr.1977804.

[20] A. Massouras, S.M. Waszak, M. Albarca-Aguilera, K. Hens, W. Holcombe, J.F. Ayroles, E.T. Dermitzakis, E.A. Stone, J.D. Jensen, T.F.C. Mackay, B. Deplancke, Genomic variation and its impact on gene expression in Drosophila melanogaster, PLoS Genet. 8 (2012) e1003055. doi:10.1371/journal.pgen.1003055.

[21] L. Salmena, Pseudogene redux with new biological significance, Methods Mol. Biol. Clifton NJ. 1167 (2014) 3–13. doi:10.1007/978-1-4939-0835-6_1.

[22] F. Hubé, C. Francastel, Coding and Non-coding RNAs, the Frontier Has Never Been So Blurred, Front. Genet. 9 (2018) 140. doi:10.3389/fgene.2018.00140.

[23] A. Rubio, C.S. Casimiro-Soriguer, P. Mier, M.A. Andrade-Navarro, A. Garzón, J. Jimenez, A.J. Pérez-Pulido, AnABlast: Re-searching for Protein-Coding Sequences in Genomic Regions, Methods Mol. Biol. Clifton NJ. 1962 (2019) 207–214. doi:10.1007/978-1-4939-9173-0_12.

[24] B.E. Suzek, Y. Wang, H. Huang, P.B. McGarvey, C.H. Wu, UniProt Consortium, UniRef clusters: a comprehensive and scalable alternative for improving sequence

similarity searches, Bioinforma. Oxf. Engl. 31 (2015) 926–932. doi:10.1093/bioinformatics/btu739.

[25] D. Karolchik, A.S. Hinrichs, T.S. Furey, K.M. Roskin, C.W. Sugnet, D. Haussler, W.J. Kent, The UCSC Table Browser data retrieval tool, Nucleic Acids Res. 32 (2004) D493-496. doi:10.1093/nar/gkh103.

7. Capítulo 4:

Caenorhabditis elegans

7.1 Identification of small protein-coding regions in the *Caenorhabditis elegans* genome: an application of AnABlast

CS Casimiro-Soriguer [1,2],MM Rigual[1,2],AM Brokate-Llanos[1], MJ Muñoz[1], A Garzon[1*], A Perez-Pulido[1*], J Jimenez[1*]

[1]Centro Andaluz de Biología del Desarrollo (CABD, UPO-CSIC). Universidad Pablo de Olavide, Ctra. de Utrera, km.1, 41013, Sevilla, Spain

*Co-corresponding authors

7.1.1 Summary

The sequencing of higher eukaryotic genomes has become routine, but the accurate identification of genes and their protein-coding regions is less straightforward. AnABlast is a new approach that we recently described for the *in silico* recognition of protein-coding regions in DNA sequences. By using AnABlast, here we report the identification of 82 putative new protein-coding sequences that were identified only by this algorithm in the well annotated *Caenorhabditis elegans* genome. Conventional homology/motif searches, available RNA expression data and RNA interference experiments supported that AnABlast efficiently predicts functional protein-coding sequences, including proteins encoded in small ORFs. Therefore, AnABlast provides a novel tool for the accurate identification of protein-coding regions that escape to other *in silico* strategies, as well as a new approach towards the detection of small bioactive proteins in annotated eukaryotic genomes.

Keywords: AnABlast, *C. elegans*, new genes, *in silico* gene-prediction tool, protein-coding sequences, sORFs

7.1.2 Introduction

Obtaining the complete inventory of genes in sequenced genomes is a main goal in the current genomic age [1]. Thanks to advances in bioinformatics and computing power, it is now possible to scan the genome in unprecedented scrutiny [2,3]. Different *in silico* methods has been devised as useful predictors of protein-coding regions [4-7], but such methods are generally not sufficiently robust for finding small exons, small protein-coding genes and highly divergent genes, sharing very short stretches of sequence identity in databases, which remain reluctant to identification even in model organisms [8].

Bioinformatics tool AnABlast has been recently developed as a reliable new computational approach for locating protein-coding regions in genomic DNA sequences [9,10]. This novel gene predictor is based on the fact that protein-coding genes mostly arise from previous ones during evolution. Thus, even small and/or highly divergent coding sequences may harbour ancient footprints to be found among the millions of proteins available in databases. AnABlast accumulates these footprints searched which are low-stringency BLAST alignments common between the query sequence and known protein sequences. Since protein-coding footprints are not present in non-coding DNA sequences [11,12], the significant accumulation of insignificant BLAST score alignments allows AnABlast to highlight DNA protein-coding regions that, at present, can only be uncovered *in silico* by this algorithm [9,10].

In a pilot study, the system was trained on the fission yeast genome, and its performance was evaluated by examining RNA expression on the predicted AnABlast coding regions of the genome of this lower eukaryotic model [9]. However, identifying genes in higher eukaryotes is more complex. In higher eukaryotes, genes may span hundreds or thousands of Kbs with the protein-coding sequences accounting for only a few percent of the total sequence [2,3]. Thus, the accurate identification of such sequences in complex eukaryotic genomes is a difficult undertaking. The nematode *Caenorhabditis elegans* has been established as a multicellular eukaryote model for the study of genetics and developmental biology. Initial analysis of the complete genome sequence of *C. elegans* by the WormBase consortium revealed over 19000 coding genes, but this number has been continuously increasing as a consequence of both, new experimental data and improved

protein-coding gene prediction algorithms [13]. The latest version of the *C. elegans* genome sequence (WS228) predicts 24610 coding genes (Genome and biological are available in the WormBase database [14]. The *in silico* identification of novel protein-coding regions in this model organism *in silico* is challenging. By analysing the *C. elegans* genome, here we show that AnABlast is highly efficient in locating yet unknown protein-coding sequences in this complex genome, including small protein-coding genes and new exons of known genes.

7.1.3 Results and Discussion

Discovery of new protein-coding regions in the *C. elegans* genome

Sequence-sequence and sequence-profile alignment algorithms have been widely adopted for the identification of related genes [15]. Methods that can identify remote homologues sharing insignificant BLAST scores between each other have also been implemented by detecting intermediate homologue sequences to connect them [16, 17]. AnABlast efficiently uncovers protein-coding sequences not by significant homology or intermediate homologues, but by the significant abundancy of short-stretches of amino acid sequences (protomotifs) in a protein coming from the virtual translation of a query DNA sequence [10]. In order to perform a search for unidentified protein-coding sequences in a multicellular eukaryote, AnABlast profiles were generated for the entire genome of the *C. elegans* nematode. The complete set of AnABlast results of the *C. elegans* genome analysis, including signals for known genes, can be accessed at the genomic browser http://www.bioinfocabd.upo.es/jbrowse/JBrowse-1.12.1/?data=datasets/celegans. As expected, the vast majority of AnABlast signals were related to already annotated genes, and only those non-annotated AnABlast sequences were then selected.

AnABlast is designed to highlight protein-coding sequences from unknown genes and exons, but also from highly degenerated pseudogenes and relic sequences [9,10]. To focus on the search of new actual genes, or new exons of annotated genes, AnABlast signals predicting peptides with less than 30 amino acids were discarded. To evaluate the singularity of AnABlast, the new protein-coding sequences were subjected to other available gene finders [6,15], and those predicted also by other bioinformatics tools (about

60%) were discarded as well. Overall 92 AnABlast regions were selected as new candidate to be actual coding-protein sequences uncovered only by AnABlast. Among them, 82 were located at least at 500 nucleotides to any annotated gene, suggesting that the new signals could identify new genes (S1 Table), while the other 10 signals were adjacent (less than 500 nucleotides) to annotated genes, which therefore could rather be new exons of known genes (S2 Table).

It was long assumed that proteins are at least 100 amino acids long. Interestingly, predicted AnaBlast protein-coding sequences account for relatively small ORFs (see ORF length in S1 and S2 Tables). While long ORFs can be efficiently identified (6,15), the *in silico* detection of small ORFs is very difficult as the short length makes it hard to distinguish true coding ORFs from ORFs occurring by chance. With the advances in technology, notably ribosome profiling assays and proteogenomics, the identification of functional small peptides has drastically increased, rising the number of potentially functional ORFs within the eukaryotic genomes [19-22]. Based on the remarkable property of AnABlast in highlighting small coding regions (less than 100 residues), we believe that this computer approach may provide a powerful tool for the identification of these elusive small ORFs in sequenced genomes, complementing proteogenomic methods.

Characterization of novel putative *C. elegans* genes

Different approaches can be used to validate predicted AnABlast sequences as functional genes. Often, putative protein-coding regions detected by AnABlast escape to conventional homology searches. However, AnABlast signals occasionally uncover short sequences with significant homology to others in databases that had been missed or ignored from previous bioinformatics analysis. Thus, in a first approach, open reading frames (ORFs) in the corresponding DNA sequence (chain and frame) highlighted by AnABlast were identified, and a conventional search of motifs and homologs were performed to uncover putative functions of the inferred protein sequences. Since small ORFs may initiate with start codons other than AUG [20], predicted protein sequences lacking an initiation methionine were equally considered as possible exons from other gene. As shown in S1 Table, some of the identified sequences share stretches of significant similarity to proteins in reference data bases, and/or match recognized motif signatures [23, 24]. AnABlast signals G71 and G75 are representative examples of

predicted genes with both, motifs and significant homolog proteins found in related *Caenorhabditis* species (Fig 1).

Fig 1. Putative new protein-coding sequences identified by AnABlast in signals G71 (X:11701412-11701730) and G75 (V:4970301-4970493). A) Signals G31 (left) andG75 (right) (squares). Annotated exons of their respective adjacent genesC44C10.3 and F26G5.10 are shown (yellow boxes). B) F-box associatedFBA2-Motif signature and Serpentine type 7TM GPCR chemoreceptor motif underlined in signals G71 (left) and G 75 (right) respectively. C) Top 10 Blast hits of G71 (left) and G75 (right) protein sequences. D)G71 (left) and G75 (right) protein sequence alignments to proteins found in the indicated Caenorhabditis specie. Sequence ID and alignment significance (E-value) are indicated.

One important question regarding this study is why the annotation of these genes had been missed from the WormBase database (www.WormBase.org). The fact that many of these candidate genes encode relatively small peptides with short stretches of similarity to other proteins in databases explain why the annotation of these putative protein-coding sequences have been missed from previous *in silico* analysis. The WormBase database provides genome-wide data of gene expression from Ribo-seq, RNA-seq and proteomic experimental results. Therefore, we also analysed the existence of reported RNA and peptide expression in the selected sequences to support the accuracy of AnABlast in searching for actual functional genes. Significant expression levels were observed in 18 of the proposed new genes (see in S1 Table). Among them, some showed either significant

homology to other proteins (G30, G45 and G54), harboured significant motifs (G26, G58), or both (G31) (Fig 2).

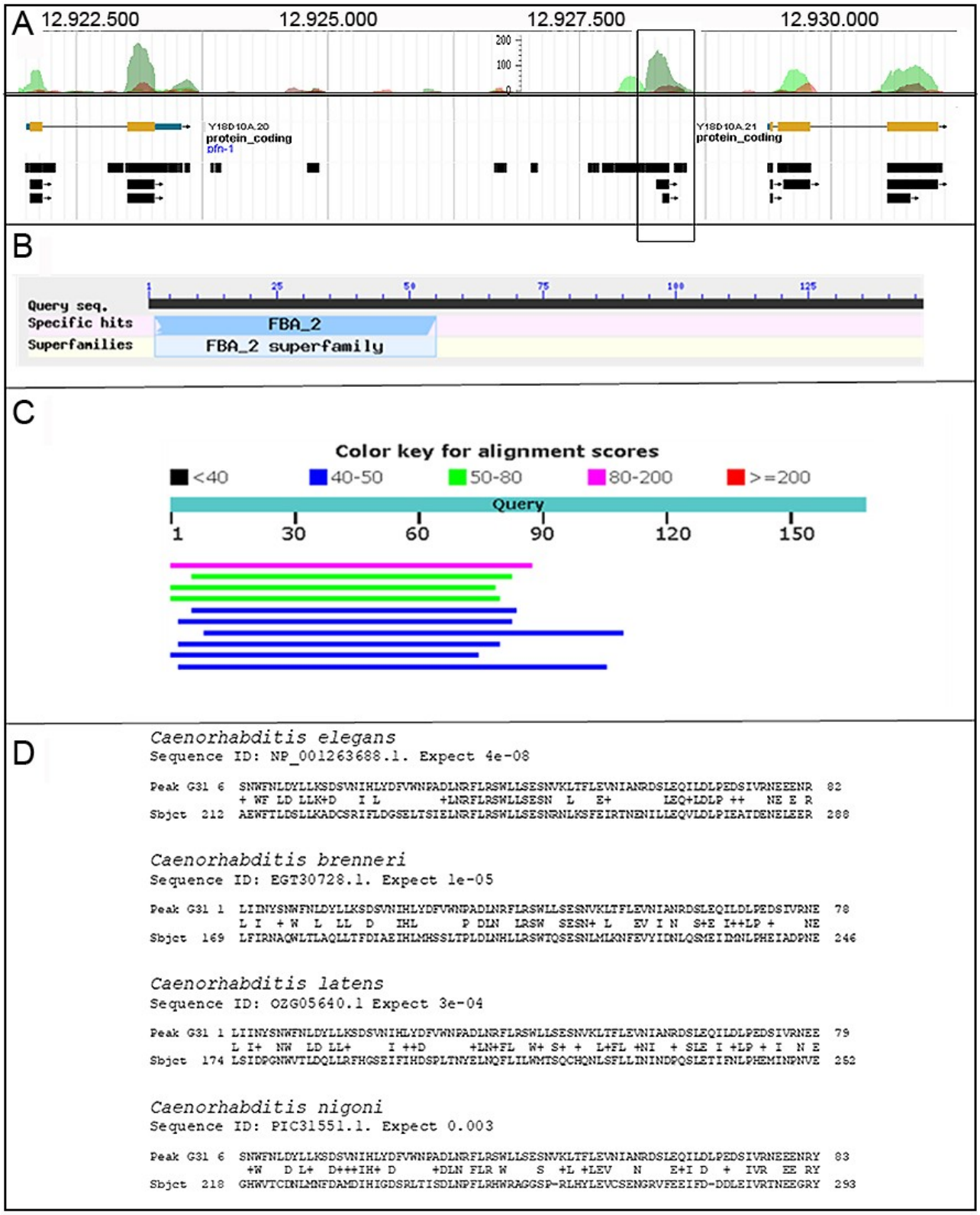


Fig 2. Putative new protein-coding sequence identified by AnABlast in peak G31 (I:12928143-12928563). A) AnABlast profile of peak G31 (square) and annotated exons (yellow boxes) of the flanking genes Y18D10A.20 andY18D10A. RNA expression profiles (RNA-Seq, black boxes) available in WormBase are shown. B) FBA2-Motif signature identified in predicted G31 protein sequence. C) Top 10 Blast hits of G31

protein sequence. D) G31 Protein sequence alignments to proteins in the indicated Caenorhabditis specie. Sequence ID and alignment significance (E-value) are indicated.

Expression itself is a useful clue for the identification of functional genes. Remarkably, 12 out of 18 AnABlast-predicted coding-regions showing significant expression levels lack any detectable homology or motif signature (S1 Table), reinforcing the potential use of AnABlast in re-searching for putative genes that escape to conventional *in silico* strategies.

Characterization of novel putative exons

In the identified putative protein-coding sequences, 30 AnABlast signals were located at less than 500 bp to the 5' or to the 3' end of known genes, suggesting that these coding regions could well be exons of the adjacent genes. Among them, 10 were exclusively underlined by AnABlast as putative exons (S2 Table) and were further studied (the remaining 20, also predictable by other algorithms, were discarded). To explore the possibility that these10 AnABlast signals could identify new exons of known genes, amino acid sequences predicted by AnABlast and by the adjacent gene were concatenated, and the resulting sequence was used to search for homolog proteins in data bases. When the new predicted exon belongs to the annotated gene, homolog proteins expanding significant similarity along the added exon may be identified in data bases. This approach suggested that at least E6, E13, E26 and E28 encode putative new exons belonging to their adjacent genes.

AnABlast E6 DNA region is located at the 5' end of the gene encoding the hypothetical protein M03A8.3, a protein containing a PH-like domain. Remarkably, Blast search of concatenated E6-M03A8.3 protein sequences indicates that a stretch of E6 expands the amino-terminal end of M03A8.3 protein homologs identified in *C. briggsae* and *C. nigoni* (Fig 3).

Based on microarray and RNA-seq data, expression of M03A8.3 is affected by several genes including daf-16, daf-12, and sir-2.1 [WormBase data], suggesting a role of this protein in different developmental processes of the nematode. According to RNA expression data available in WormBase, DNA regions coding for both E6 and M03A8.3 are expressed at similar levels in *C. elegans* (see in Fig. 3).Thus, we suggest that genomic DNA predicted by AnABlast to encode E6 is likely a new non-annotated exon of the gene encoding M03A8.3in the *C. elegans* genome.

AnABlast-coding sequence E13 locates at the 5’ end of the gene encoding two protein-isoforms, F20G2.3a and F20G2.3b. WhileF20G2.3a homologs containing E13 sequences were not found, the concatenated E13-F20G2.3b sequence was similar to protein F28H7.8 found in the parasitic nematode *Toxocara canis*, suggesting that E13 could be a putative exon of the F20G2.3b isoform (see in Fig. 4). Some weak intronic signals can be predicted in the DNA sequence encoding E13. However, while RNA-seq data indicate that both F20G2.3a and F20G2.3b DNA regions are expressed, expression is not observed in the E13genomic interval. Thus, the E13 DNA sequence may represent a pseudo exon, rather than an actual exon of this gene.

Signals E23 and E28 can also be identified in homolog proteins when concatenated to predicted protein sequences coded by their respective adjacent genes (S1 Fig), suggesting that both are putative new exons of these hypothetical genes.

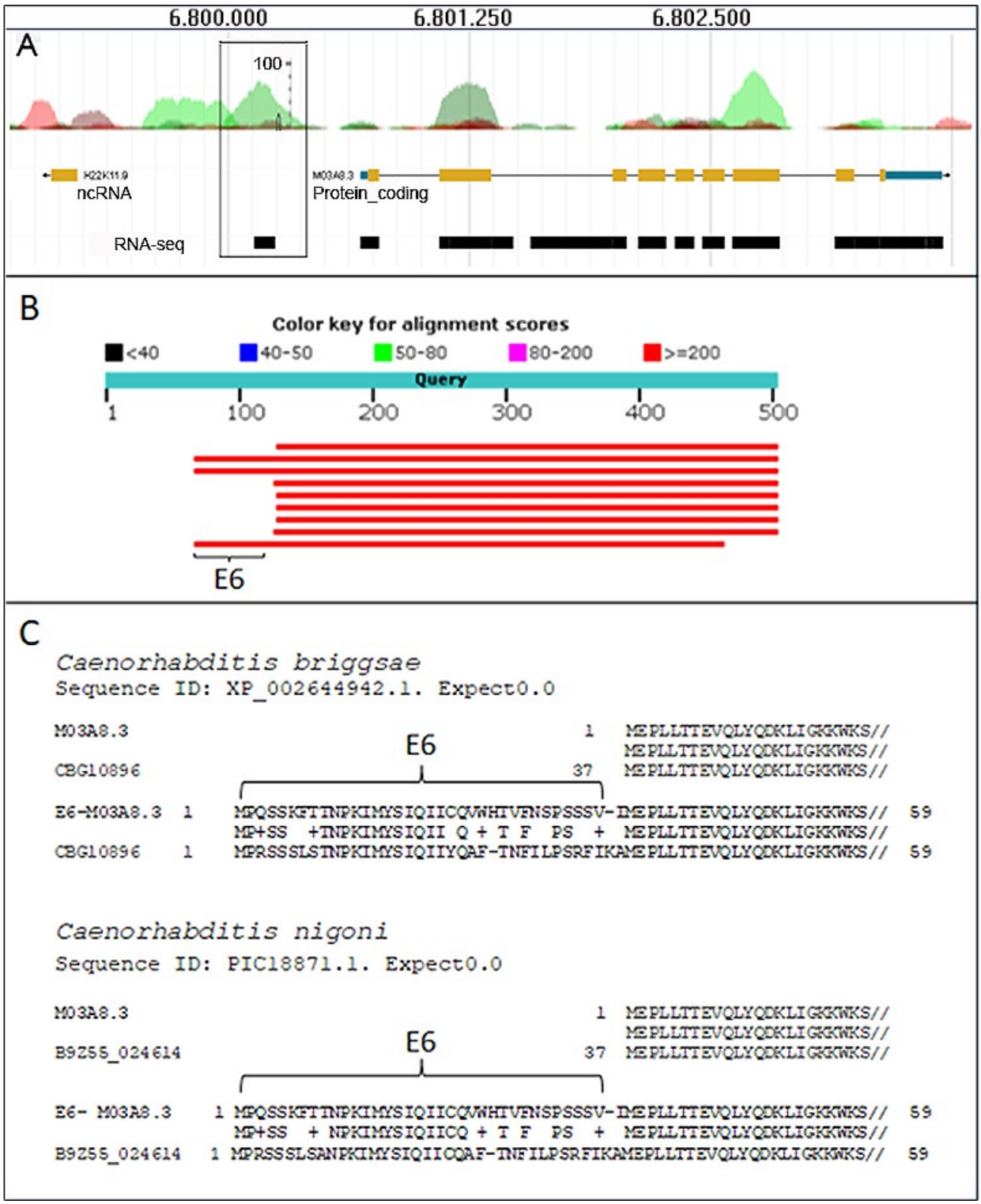


Fig 3. Putative new exon identified by AnABlast peak E6 (X:6799982-6800363). A) AnABlast profiles showing peak E6 (square) and the adjacent signals matching exons encoding the M03A8.3 protein. RNA expression data (RNA-seq) are shown. B) Top 10 hits in Blast search of concatenated E6-M03A8.3 sequences. E6 sequence is indicated. C) Protein sequence alignment of the annotated M03A8.3 protein sequence (upper) and the concatenated E6-M03A8.3protein sequence (lower) to protein CBG10896 (Sequence ID: XP_002644942.1) of C. briggsae and protein B9Z55_024614 (Sequence ID: PIC18871.1) of C. nigonias indicated. Alignment significance (E-value), and E6 Amino acid sequences are indicated.

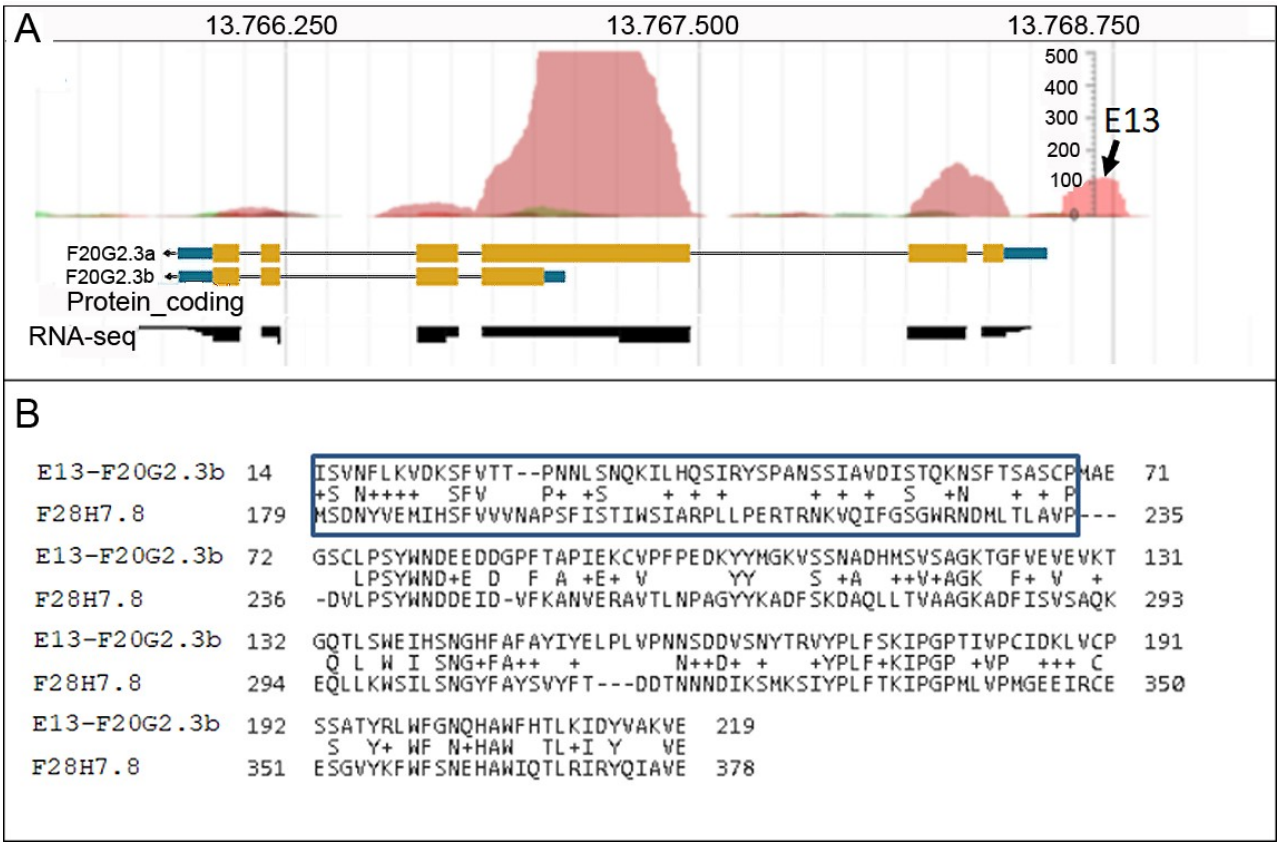


Fig 4. Putative new exon underlined by AnABlast peak E13 (V: 13768591-13768795). A) AnABlast Profiles indicating the peak E13 and adjacent signals matching Exons encoding F20G2.3a and F20G2.3b isoforms of the adjacent gene. RNA expression (RNA-seq) is shown. B) Sequence alignment of concatenated E13-F20G2.3b protein sequence to protein F28H7.8 (ID: KHN82554.1) (expected 1e-31) of the nematode Toxocana canis. E13 amino acid sequence is indicated (square box)

Functional analysis of the putative new protein-coding sequences by RNA interference

As described above, detailed analysis of directed homology searches, or RNA-expression data, support AnABlast as a useful algorithm in the re-search of protein-coding sequences that are hidden to other established methods. However, the great challenge for AnABlast is the possibility to uncover new genes that escape both, to conventional *in silico* analysis and to available expression data. The nematode *C. elegans* is a genetically tractable model system that has been widely used to investigate the molecular mechanisms of aging and longevity, and the development of RNA interference (RNAi) technology has provided a powerful tool for performing large-scale genetic screens in this organism. RNAi is an endogenous cellular mechanism triggered by double-stranded RNA (dsRNA), which leads to the degradation of homologous RNAs in the cytoplasm. RNAi degradation of mRNA is typically not complete, often giving rise to hypomorphic phenotypes [25].

However, RNAi is relatively easy to use since feeding worms with bacteria expressing dsRNA is enough to knockdown gene expression [25]. Thus, to gain insight into the function of the candidate genes discovered only by AnABlast, all the selected DNA sequences were knocked down with RNAi, and developmental-associated phenotypes were examined. Only a fraction of the functional genes should be expected to develop a distinguishable phenotype on these characteristics by using RNAi strategies [26]. Thus, it is remarkable that three of the selected sequences (signals G71, G98 andG107) yielded RNAi-dependent phenotypic defects, suggesting that at least these three DNA sequences uncovered only by AnABlast likely identify new functional genes.

Peak G71 encodes a predicted 82 amino acids peptide, with neither significant homologs in databases nor reported RNA expression (S1 Table). When this DNA sequence was knocked down with RNAi, a significant delay in the L4 to adult transition was observed (Fig 5). Detailed *in silico* analysis indicated that the encoded 82 amino acid sequence contains an F-box domain. About 326 *C. elegans* annotated genes contain known F-box sequences, indicating that it is a frequent domain in the worm genome. F-box containing proteins usually bind SCF complexes, which in turn function in the ubiquitination of cell cycle regulatory proteins [27]. In *C. elegans*, however, the F-box domain is also found in FOG-2 proteins triggering spermatogenesis during development [28]. According to the observed phenotype in peak 71 knocked down worms (Fig 5), the proposed F-box containing peptide could play a role in exit of the L4 state during the nematode development.

Peak G98 encodes a putative 35 amino acids peptide without reported RNA expression, as well as without significant homologs (S1 Table). RNAi knockdown experiments of this sequence yielded a slight delay in development that can be visualized in the transition of the last developmental stage (L4 larval stage to adult) (S2A Fig). This small peptide harbours a BEACH domain, a highly conserved motif which functions in lysosomal protein trafficking, but also endomembrane signalling during development [29]. Thus, the identified BEACH-containing peptide encoded by AnABlast peak 98 could also affect different developmental steps.

Finally, peak 107 encodes a 38 amino acids peptide, which according to RNAi results could also play a role during L4 state progression, with a percentage of animals that never

reach adulthood (S2B Fig). A BAH domain is found in this peptide (S1 Table). Proteins containing this domain are usually involved in chromatin remodelling, histone recognitions [30], and PCNA ubiquitination [31]. Therefore, there is a large repertoire of different functions in which this peptide could act during worm development.

The analysis of the knockdown efficiency of a publicly available RNAi resource revealed that >90% of *in vivo* lines exhibited residual gene expression of 25% or more RNAi, suggesting that RNAi is likely insufficient to functionally identify a large number of genes [25]. In fact, only about 19% of *C. elegans* genes yield observable phenotypes when subjected to RNAi knockdown [26]. In our characterization, this frequency would significantly drop down since our phenotypic assay only covers a limited set of developmental defects. Remarkably, RNAi-induced phenotypes were observed in 3 out of 94 sequences assessed, suggesting that sequences highlighted only by AnABlast efficiently uncover genomic regions encoding functional protein.

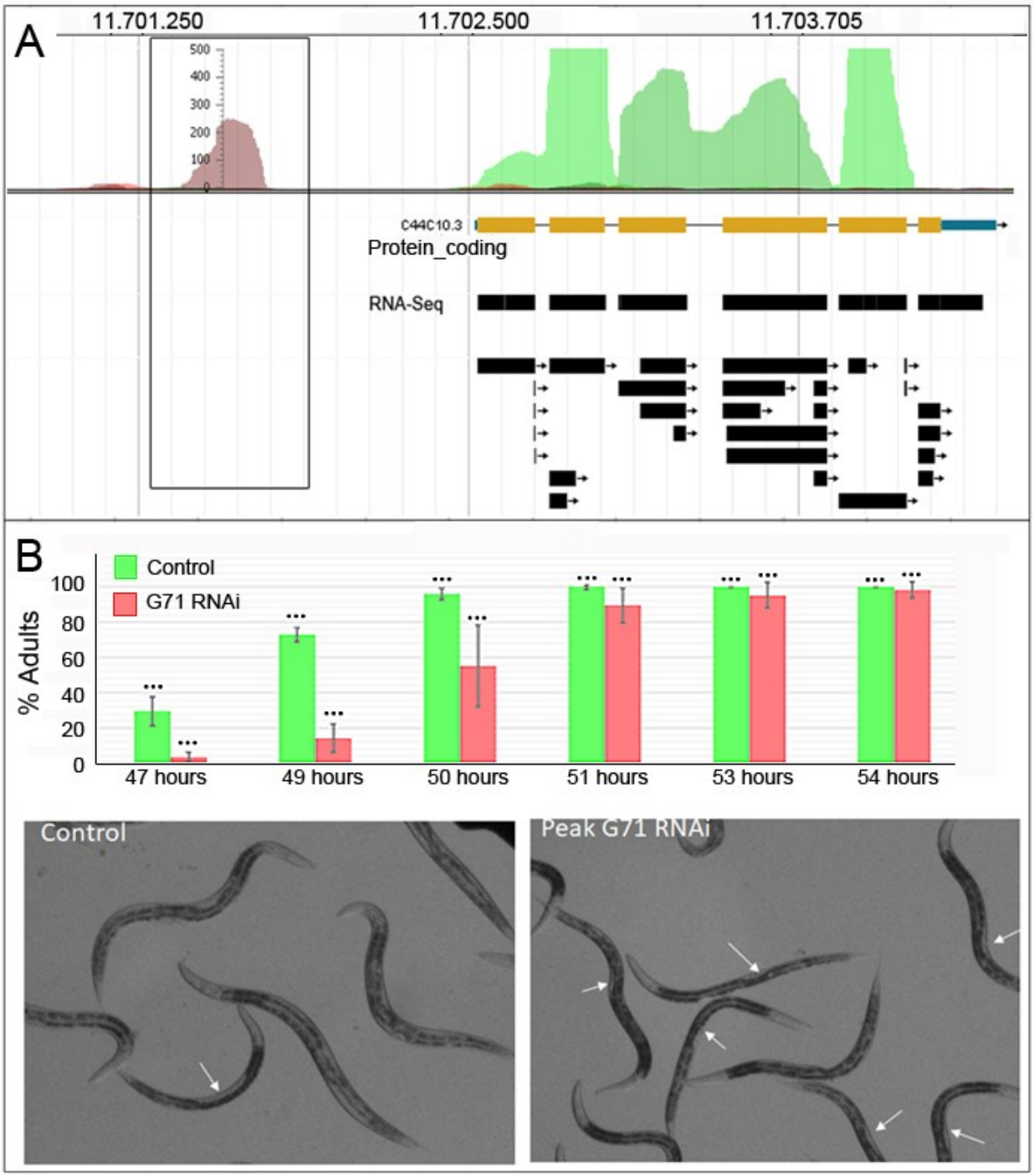


Fig 5. Analysis of AnABlast Peak G71(X: 11701412-11701730). A) AnABlast profiles showing peak G71 (square box) and the adjacent signals matching exons encoding protein C44C10.3. RNA expression data (RNA-seq) are shown. B) Average worms (%) that reaches adulthood from L1 (time 0 hours) in worms subject to E71 RNAi (pink) with respect to the control (green). Standard deviation bars are indicated. Photographs of the phenotype caused by G71 RNAi with respect to control at 49 hours are shown (lower panels). Arrows indicate the typical non-mature vulva of L4 animals.

Perhaps the most fundamental question that can be asked about a DNA sequence is whether or not it encodes protein. Large ORFs are easily uncovered *in silico* by ORF-Finder algorithms, but finding small ORFs represents an extremely difficult task [19,22].

Our approach leads to the prediction of a set of putatively coding sequences including small ORFs in the *C. elegans* genome. We encourage the use of AnABlast for the *in silico* re-search of protein-coding sequences in other sequenced genomes, as well as a new tool for the *in silico* identification of putative bioactive peptides.

Acknowledgments

We thank Genetics Group members at the Pablo de Olavide University for their useful comments on the manuscript, and Victor Carranco for technical assistance.

Funding

This research was supported by the Ministry of Economy and Competitiveness of the Spanish Government grant BFU2016-77297-P. Authors declare no competing interest.

Authors' contribution

In this study, JJ coordinated the research, designed experiments, analysed data and wrote the paper. APP designed, supervised and analysed *in silico* experiments. AG designed, supervised and analysed DNA cloning for RNAi experiments. MJM designed and supervised RNAi experiments. CSCS employed AnABlast analysis on the entire *C. elegans* genome, sequence selection and *in silico* characterization. MMR carried out cloning and RNAi expression of all AnABlast sequences in bacteria. AMBL carried out RNAi assays and phenotype analysis in *C. elegans*.

7.1.4 Materials and Methods

AnABlast Search strategy

AnABlast search for putative protein-coding sequences was employed following described methods [10] but analysing the complete genome. Due to the long length of the *C. elegans* chromosomes, they were used as the reference database in a similarity search using BLASTX and the millions of protein sequences of non-redundant UniRef50 database (2014_02 version) as query sequences [15]. To get non-restricted alignments, a threshold

bit-score of 30 was used. Then, AnABlast takes the positions from the acquired hits and counts the number of alignments (belonging to different non-redundant proteins) that matches each genomic position. As a result, AnABlast yields a profile along the *C. elegans* genome where discrete peaks highlight specific regions. The found hits including low-scored alignments (protomotifs) are usually accumulated in coding sequences but rarely in non-coding sequences [11,12]. Thus, profile of accumulated AnABlast protomotifs yields peaks that accurately marks putative protein-coding regions even in the presence of sequencing errors, or evolutionary degenerated sequences such as pseudogenes and relic sequences. Peaks with less than 70 accumulated alignments were discarded.

In silico analysis of the selected sequences

For *in silico* analysis, *C. elegans* annotations in the genomic regions of the candidates were downloaded from WormBase database at 1 February 2017. Annotations were gathered from the tracks of WormBase browser, including gene coordinates, RNA expression, proteomics, and similarity sequences. An expression evidence is associated to an AnABlast candidate when the evidence overlaps at least 20% with the candidate signal. The amino acid sequence delimited by each AnABlast peak was further studied by using BLAST [15], Pfam for domain sequences [27], and Sma3s for functional annotation [3].

Knockdown RNAi assay by feeding

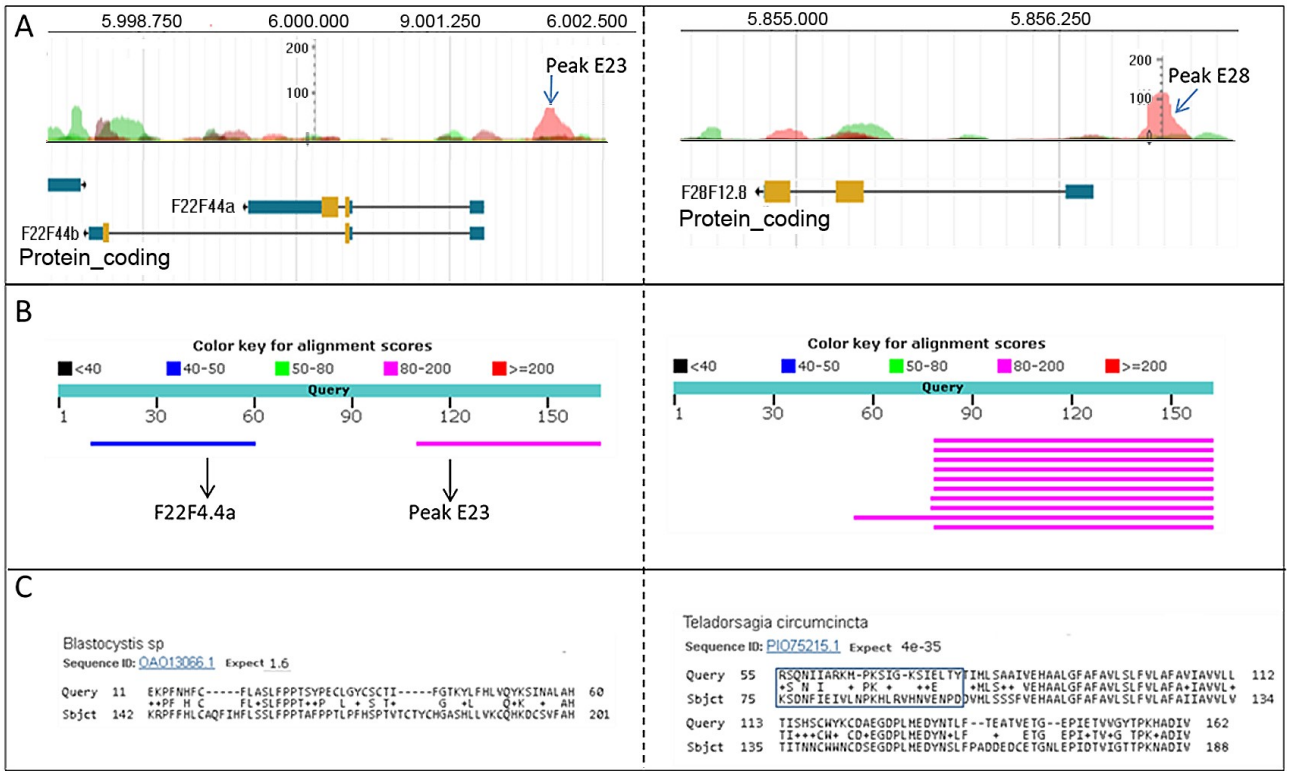
The RNAi clones used in this study were obtained from the selected DNA sequenced underlined by AnABlast as putative protein-coding sequences. DNA fragments ranging between 0.2 to 0.4 kb were PCR amplified and cloned into a pL4440 vector by ligation after digestion with restriction enzymes. Oligonucleotides designed for each clone are listed in S3 Table. All RNAi clones were verified by DNA sequencing. *C. elegans* strains were cultured and maintained using standard procedures [32). L1 of N2 were synchronized in M9 for 16 hours at 20°C and seeded in plates with *E. coli* strains that carry either the empty vector pL4440 (control) or the AnABlast DNA sequence-targeting as previously described in [25]. Plates were incubated at 20°C and young adults were counted at 47-48 hours each hour for 6-7 hours. The images were taken at 50 hours in Olympus SZX16 stereoscope equipped with a PLAPO 1x lens and an Olympus DP73 camera.

7.1.5 References

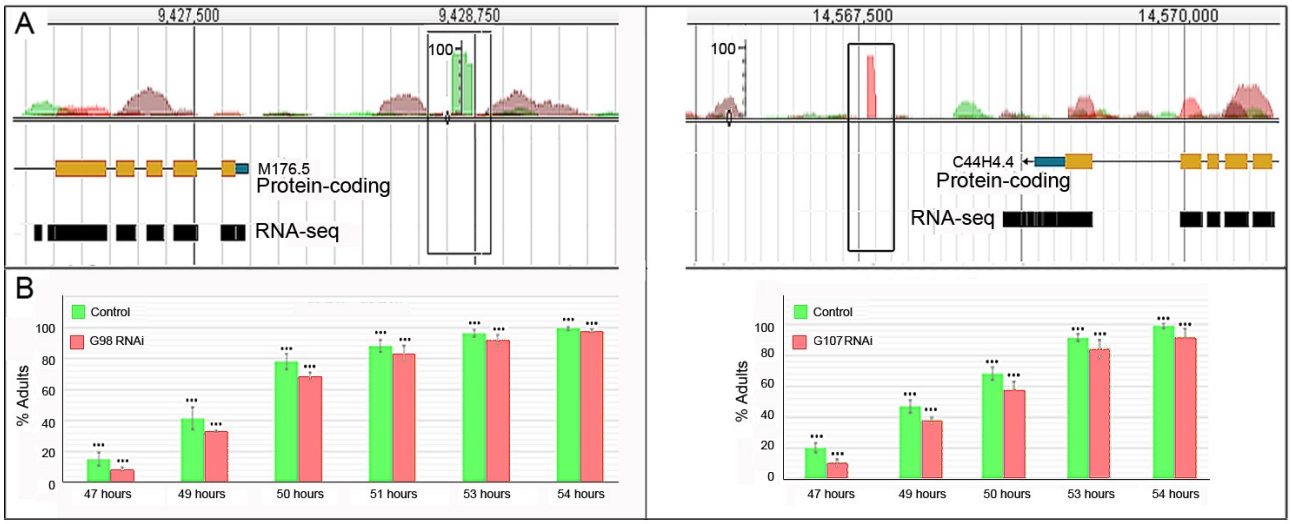
1. Kersey, P.J., Allen, J.E., Armean, I., et al.(2016) Ensembl Genomes 2016: more genomes, more complexity. Nucleic Acids Res44: D574-580 doi: 10.1093/nar/gkv1209
2. Brent MR1.Steady progress and recent breakthroughs in the accuracy of automated genome annotation.Nat Rev Genet. 2008 Jan;9(1):62-73:
3. Casimiro-Soriguer CS, Muñoz-Mérida A, Pérez-Pulido AJ. Sma3s: A universal tool for easy functional annotation of proteomes and transcriptomes. Proteomics. 2017 Jun;17(12). doi: 10.1002/pmic.201700071. PubMed PMID: 28544705.
4. Uberbacher EC, Mural RJ. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. Proc Natl Acad Sci U S A. 1991 Dec 15;88(24):11261-5
5. Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. Nucleic Acids Res. 2012;40:e9
6. Alioto,T. (2012) Gene prediction. Methods Mol. Biol. Clifton NJ, 855, 175–201.
7. Nesvizhskii, A.I. (2014) Proteogenomics: concepts, applications and computational strategies. Nat Methods 11: 1114-1125. doi: 10.1038/nmeth.3144
8. Stanke, M., Schöffmann, O., Morgenstern B., Waack, S.(2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics7: 62
9. Jimenez J., Duncan, C.D., Gallardo, M., et al.(2015) AnABlast: a new in silico strategy for the genome-wide search of novel genes and fossil regions. DNA Res22: 439–449 doi: 10.1093/dnares/dsv025
10. Rubio A, Casimiro-Soriguer CS, Mier P, Andrade-Navarro MA, Garzón A, Jiménez J, and Pérez-Pulido AJ. (2019) AnABlast: Re-searching for Protein-Coding Sequences in Genomic Regions. Methods in Molecular Biology. 1962:207-214. doi: 10.1007/978-1-4939-9173-0_12
11. Thode, G., García-Ranea, J.A., Jimenez, J. (1996) Search for ancient patterns in protein sequences, J MolEvol 42: 224-33
12. Pérez, A.J., Thode, G., Trelles, O. (2004) AnaGram: protein function assignment. Bioinformatics 20: 291-2
13. Yoshimura J et al., Recompleting the *Caenorhabditis elegans* genoma Genome Res. 2019; 29: 1009-1022.
14. Dubaj Price M, Hurd DD. WormBase: A Model Organism Database. Med Ref Serv Q. 2019 Jan-Mar;38(1):70-80. doi: 10.1080/02763869.2019.1548896. PubMed PMID: 30942676.
15. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25(17): 3389–3402.
16. Li W, Pio F, Pawlowski K, Godzik A (2000) Saturated BLAST: An automated multiple intermediate sequence search used to detect distant homology. Bioinformatics 16(12): 1105–1110.
17. Dargahi D, Baillie D, Pio F (2013) Bioinformatics Analysis Identify Novel OB Fold Protein Coding Genes in C. elegans. PLoS ONE 8(4): e62204. doi:10.1371/ journal.pone.0062204

18. Goodswen, S. J., Kennedy, P. J. & Ellis, J. T. Evaluating high-throughput ab initio gene finders to discover proteins encoded in eukaryotic pathogen genomes missed by laboratory techniques. PloS One 7, e50609 (2012).
19. Crappé J1, Van Criekinge W, Trooskens G, Hayakawa E, Luyten W, Baggerman G, Menschaert G. Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. BMC Genomics. 2013 Sep 23;14:648. doi: 10.1186/1471-2164-14-648
20. Hellens RP, Brown CM, Chisnall MAW, Waterhouse PM, MacknightRC. The Emerging World of Small ORFs.Trends Plant Sci. 2016 Apr;21(4):317-328. doi: 10.1016/j.tplants.2015.11.005.
21. Couso J-P, Patraquim P. Classification and function of small open reading frames. Nat Rev Mol Cell Biol. 2017;18:575–89.
22. Kroll JE1,2,3, da Silva VL2,3, de Souza SJ2,3, de Souza GA2,3,4. A tool for integrating genetic and mass spectrometry-based peptide data: Proteogenomics Viewer: PV: A genome browser-like tool, which includes MS data visualization and peptide identification parameters. Bioessays. 2017 Jul;39(7). doi: 10.1002/bies.201700015. Epub 2017 Jun 5.
23. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. Nucleic Acids Res45: D158–D169 doi: 10.1093/nar/gkw1099
24. El-Gebali S, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019 Jan 8;47(D1):D427-D432. doi: 10.1093/nar/gky995.
25. Kamath, R. S., A. G. Fraser, Y. Dong, G. Poulin, R. Durbin et al., 2003 Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature 421: 231–237.
26. Lizabeth A. at al., The Transgenic RNAi Project at Harvard Medical School: Resources and Validation. GENETICS November 1, 2015 vol. 201 no. 3 843-852; https://doi.org/10.1534/genetics.115.180208
27. Kipreos ET, Pagano M. The F-box protein family. Genome Biol. 2000;1(5):REVIEWS3002. Epub 2000 Nov 10. Review. PubMed PMID: 11178263; PubMed Central PMCID: PMC138887.
28. Hu S, Skelly LE, Kaymak E, Freeberg L, Lo TW, Kuersten S, Ryder SP, Haag ES. Multi-modal regulation of C. elegans hermaphrodite spermatogenesis by the GLD-1-FOG-2 complex. Dev Biol. 2019 Feb 15;446(2):193-205. doi: 10.1016/j.ydbio.2018.11.024. Epub 2018 Dec 30.
29. Khodosh R1, Augsburger A, Schwarz TL, Garrity PA Bchs, a BEACH domain protein, antagonizes Rab11 in synapse morphogenesis and other developmental events. Development. 2006 Dec;133(23):4655-65. Epub 2006 Nov 1
29. Yang N, Xu RM. Structure and function of the BAH domain in chromatin biology. Crit Rev Biochem Mol Biol. 2013 May-Jun;48(3):211-21. doi: 10.3109/10409238.2012.742035. Epub 2012 Nov 27.
30. Niimi A, Hopkins SR, Downs JA, Masutani C. The BAH domain of BAF180 is required for PCNA ubiquitination. Mutat Res. 2015 Sep;779:16-23. doi: 10.1016/j.mrfmmm.2015.06.006. Epub 2015 Jun 17.
31. Stiernagle T1. Maintenance of C. elegans. WormBook. 2006 Feb 11:1-11.

7.1.6 Supplemental Figures



S1 Fig. Putative new exons identified by AnABlast peaks E23 (X:6001927-6002254) and E28 (V:5856625-5856859). A) Signals E23 (left) and E28 (right) (squared). Annotated exons of their respective adjacent genes F22F4.4a and F28F12.8 are shown. B) Top Blast hits of E23 (left) and E28 (right) predicted protein sequences. C) E23 (left) and E28 (right) protein sequence alignments to proteins found in the indicated specie. Exon E28 sequences is highlighted in box. Sequence ID and alignment significance (expected value) are indicated.



S2 Fig. RNAi analysis of AnABlast signals G98 (II: 9428645-9428735) and G107 (X: 14567563-14567626). A) AnABlast profiles showing peak G98 (left) and peak G107 (right) and signals matching exons encoding their respective adjacent proteinM175.5 and C44H4. Signals G98 and G107are indicated (squares). RNA expression data (RNA-seq) are shown. B) Average worms (%) that reaches adulthood from L1 (time 0 hours) in worms subject to E98 (left) or to E107 (right) RNAi (pink) with respect to their controls (green). Standard deviation bars are shown.

8. Discusión General

8.1 Discusión

La secuenciación de genomas completos permite obtener el catálogo completo de las instrucciones de fabricación de herramientas moleculares que posee un organismo. Pero el hecho de obtener la secuencia de nucleótidos de un organismo, no se traduce de forma instantánea en el conjunto de funciones que puede realizar dicho organismo. Para obtener estas funciones se necesita de un análisis en profundidad de esta secuencia para localizar de esta forma todos y cada uno de los elementos funcionales que la componen. Con este objetivo se creo en 2003 el Proyecto de la Enciclopedia de los Elementos del DNA (ENCODE en sus siglas en ingles) (The ENCODE Project Consortium 2004). Este proyecto se centraba en el genoma humano pero pronto se amplio a organismos modelo mediante modENCODE (modENCODE Consortium et al. 2009).

Desde entonces, se han empleado diversas técnicas de laboratorio para localizar estos elementos funcionales como son RNAseq (Gerstein et al. 2010), espectrometría de masas (Krijgsveld et al. 2003) o perfiles ribosómicos (Ribo-seq), especialmente en el caso de los sORFs (Olexiouk, Van Criekinge, and Menschaert 2018). Todas estas técnicas, al basarse en la detección de proteínas, o sus precursores, como es el caso del RNAm, tienen la desventaja de que solo pueden detectar aquellos elementos funcionales que se están empleando por el organismo en un momento dado. Siendo necesarios muestreos a lo largo del desarrollo del organismo y en diferentes ambientes, para obtener el conjunto completo de las funciones codificadas en un genoma. Pero incluso realizando este tipo de muestreos, no se llegan a detectar la totalidad de los genes codificantes de proteínas, como evidencia el hecho de que el número de estos genes haya variado a lo largo de los años en diferentes organismos modelo. Este tipo de técnicas se enfrentan a su vez a problemas de escalabilidad, especialmente en el momento actual donde la secuenciación de genomas completos es algo habitual

Por tanto, la localización *in silico* de genes codificantes de proteínas en genomas completos, sigue siendo uno de los grandes retos en la actual era genómica. Por un lado, en genomas de organismos procariotas o a la hora de localizar genes canónicos, se obtienen buenos resultados con los algoritmos actuales, aunque sigue existiendo margen

de mejora como ha evidenciado AnABlast en el desarrollo de esta tesis. Por otro lado, el hecho de la existencia de genes no canónicos como pseudogenes, secuencias fósiles, pequeños ORF, codones de inicio alternativos y cambios en el marco de lectura (*frameshifts*) hace que este reto siga vigente. En especial en los sORF ya que su búsqueda presenta una gran dificultad debido a su pequeño tamaño (menos de 100 aa) (Crappé et al. 2013; Kroll et al. 2017).

A la dificultad de encontrar estructuras génicas no canónicas, hay que sumar la dificultad que genera la presencia de errores en la secuenciación y en el ensamblado. Estos errores pueden ocultar la presencia de genes a los algoritmos y programas tradicionales, los cuales pueden dejar sin localizar hasta un 20% de los genes presentes en un genoma eucariota (Goodswen, Kennedy, and Ellis 2012). Esto ha puesto en evidencia la necesidad de implementar nuevas ideas para la localización de genes codificantes de proteínas (Gross et al. 2007; Kelley et al. 2012).

La aparición de los secuenciadores de tercera generación, que producen lecturas de gran tamaño, junto con las lecturas cortas, pero de gran precisión, de las tecnologías de secuenciación de segunda generación, ha puesto en evidencia la necesidad de mejorar los ensamblados de los genomas actuales. Esto es necesario incluso en genomas de organismos modelo como *C. elegans* (Yoshimura et al. 2019) donde se han localizado nuevos genes mediante algoritmos tradicionales como AUGUSTUS (Stanke et al. 2004).

Esta falta de precisión en los algoritmos tradicionales puede ser debida a la rigidez de los mismos, ya que tradicionalmente intentan buscar estructuras o secuencias génicas utilizando parámetros estrictos y bien definidos. Pero la línea que separa las regiones codificantes de las no codificantes no es clara. Ejemplos de este límite difuso los tenemos en los lncRNA, pseudogenes, 5'UTR y sORF, que se traducen a proteínas que pueden tener funciones clave para el organismo o pueden servir como reservorio para la creación de nuevos genes (Couso and Patraquim 2017; Ji et al. 2015; Pueyo, Magny, and Couso 2016; Ruiz-Orera et al. 2014).

Es en este campos, incluyendo la localización de genes no canónicos, crípticos o fósiles, es donde encaja AnABlast y su búsqueda de acumulaciones de pequeñas secuencias fósiles de aminoácidos denominados *protomotivos*. Se ha descrito en otros trabajos y con

mayor soporte en este trabajo, como los protomotivos no se acumulan de forma aleatoria (Andrade 2009; Perez, Thode, and Trelles 2004; Thode, García-Renea, and Jimenez 1996), sino en gran medida en la misma hebra y marco de lectura donde se encuentra la secuencia codificante de proteínas. Esto hace que las señales de AnABlast permitan localizar genes codificantes de proteínas incluso si estos presentan varios codones de STOP y/o cambios de fase de lectura. En este trabajo además se ha mostrado como esta estrategia puede ser aplicable sobre genomas completos permitiendo un análisis masivo de los mismos, lo cual es fundamental en la época actual donde se secuencian toda clase de organismos diariamente.

Durante el desarrollo de esta tesis se ha trabajado en todo momento con genomas de organismos modelo, bien conocidos y anotados y en todos ellos se han encontrado indicios de la existencia de posibles nuevos genes codificantes de proteínas. Este hecho no es algo fuera de lo común, ya que a lo largo de los años el número de genes codificantes de proteínas, tanto en *D. melanogaster* como en *C. elegans*, ha ido aumentando. Esto no implica que se hallan localizado todos los genes presentes en dichos genomas, como se ha mostrado recientemente en un ensamblado mejorado del genoma de *C. elegans* por Yoshimura et al. (Yoshimura et al. 2019), el cual ha permitiendo la localización de nuevos genes codificantes de proteínas.

A pesar de las evidencias, estos posibles nuevos genes no dejan de ser una predicción, por lo tanto una confirmación de su existencia en laboratorio permitiría validar los resultados predichos por AnABlast. Por este motivo se realizó la búsqueda de fenotipos mediante la técnica de RNAi. Se detectaron 3 fenotipos al silenciar 96 regiones en el genoma de *C. elegans.* Este número, no puede considerarse bajo, ya que hay que tener en cuenta que en la gran mayoría de las líneas de *C. elegans* (>90%) solo se silencia aproximadamente el 75% de la expresión del gen, existiendo un 25 % de la proteína que se sigue expresando (Kamath et al. 2003). Además el muestreo de fenotipos que se realizó, solo cubría una limitada cantidad de defectos en el desarrollo. Por lo tanto obtener 3 fenotipos de un total de 96 ensayos, supone un número que pone de relieve y confirma la alta especificidad que tienen AnABlast a la hora de localizar genes codificantes de proteínas, y a su vez su utilidad en futuros flujos de trabajo de localización de genes sobre nuevos genomas.

La localización de genes dentro de genomas completos es un paso fundamental para descifrar la información contenida en estas moléculas biológicas. Pero para desentrañar en profundidad la información contenida en los genomas, es necesario saber qué funciones realizan las proteínas que en ellos se codifican. Por lo tanto, la anotación funcional de estas proteínas o de sus precursores, los RNAm, es imprescindible.

Actualmente existen poco más de una decena de aplicaciones capaces de predecir la función que va a tener una secuencia biológica (nucleótidos o aminoácidos), y la gran mayoría tienen algún tipo de limitación: sólo es posible su utilización mediante una aplicación web, el número de secuencias que se pueden anotar es limitado, la anotación esta limitada a transcriptomas, son complejas de usar o son aplicaciones de pago. Además de estas limitaciones, muchas de estas aplicaciones utilizan bases de datos de referencia que han crecido exponencialmente en los últimos años. Por lo tanto los tiempo de anotación de grandes conjuntos de secuencias problema también han aumentado, o en el peor de los casos siguen empleando bases de datos pequeñas pero desactualizadas.

Por estos motivos se desarrolló la segunda versión de la aplicación Sma3s. Esta versión permite la anotación masiva de secuencias mediante una base de datos actualizada y de un tamaño y crecimiento contenido, UniRef90. A su vez se eliminaron prácticamente las dependencias y se mejoró el formato y salida de los resultados. Todo esto junto a un aumento de la precisión y una reducción de los coste de computación. Todas estas características permitieron integrar, de forma sencilla, la anotación funcional dentro del algoritmo de AnABlast y por consiguiente de la aplicación web, debido a la mejora en la facilidad de uso e instalación de la aplicación. Además Sma3s v2 mediante su versión de línea de comandos es empleada para proyectos genómicos de gran envergadura como pueden ser: La reconstrucción de rutas metabólicas y transcriptomas de *Quercus ilex (L*ópez-Hidalgo et al. 2018; Guerrero-Sanchez et al. 2019)*,* la reanotación de los genes presentes en el organismo modelo *Nicotiana benthamiana (Kourelis 2018)* o la anotación del proteoma completo de organismos no modelo como la paloma mensajera (Gazda et al. 2018).

Precisamente, la facilidad de uso, es una de las características a la que menos importancia se da durante el desarrollo de una nueva aplicación en el mundo académico.

Esto puede estar motivado por que comúnmente este tipo de aplicaciones se presentan por primera vez al público en revistas científicas especializadas donde la mayoría de sus lectores tienen los conocimientos necesarios para hacer uso de complicadas aplicaciones. Pero si se quiere que una aplicación sea realmente útil, es necesario ampliar el espectro de usuarios a aquellos científicos que carecen de habilidades de programación o computación.

Con este objetivo se desarrolló también la aplicación web de AnABlast, la cual ha puesto a disposición de la comunidad científica, de una forma sencilla y gráfica, todas las posibilidades de las que hace gala AnABlast. Pero, Además, esta aplicación web ha sido realmente útil para el avance de nuestra investigación, al permitirnos realizar pruebas y experimentos de forma rápida y sencilla.

8.1.1 Planes futuros

Durante el desarrollo de esta tesis se han planteado diversos planes de futuro y mejoras tanto en la aplicación Sma3s como en AnABlast.

Uno de los cuellos de botella que presentan tanto la aplicación Sma3s como AnABlast es la utilización del algoritmo BLAST para el alineamiento de secuencias biológicas. Entre los diferentes algoritmos de alineamiento que existen uno de los más prometedores el DIAMOND, por lo que posiblemente se desarrollen versiones de prueba tanto de AnABlast y Sma3s que empleen este algoritmo. Además de probar nuevos algoritmos de alineamiento quizás sea necesario implementar nuevos algoritmos para estas aplicaciones que permitan aumentar tanto el rendimiento como la especificidad y sensibilidad de las aplicaciones.

Otra mejora que se ha planteado en la aplicación Sma3s, es ampliar la cantidad y variedad de anotaciones que se obtienen como resultado mediante la utilización de las bases de datos relacionadas que presenta UniProtKB. Además, se podrían usar perfiles HMMER de proteínas bien anotadas, para así acelerar la anotación.

Volviendo a AnABlast. Actualmente solo se ha aplicado este algoritmo sobre organismos modelos. Por lo tanto se plantea su utilización y estudio en organismos tan diferentes como procariotas y humano. Existiendo pruebas iniciales realizadas sobre la aplicación web que han permitido detectar nuevos posibles sORF en procariotas y abundantes secuencias de baja complejidad LINE y SINE en el cromosoma Y de humano.

Otro aspecto importante a la hora de analizar los perfiles de AnABlast, en el futuro, son la forma de los picos, ya que se ha visto que tiene una relación directa con la secuencia, pudiéndose detectar de forma visual repeticiones en tandem y en espejo de regiones genómicas (Figura 17) . Además se ha visto como cierto tipo de secuencias mantienen unas formas de pico muy características. Por ejemplo las regiones CRISPR son fácilmente detectables de forma visual al presentar los picos de AnABlast una característica forma de sierra (Figura 18).

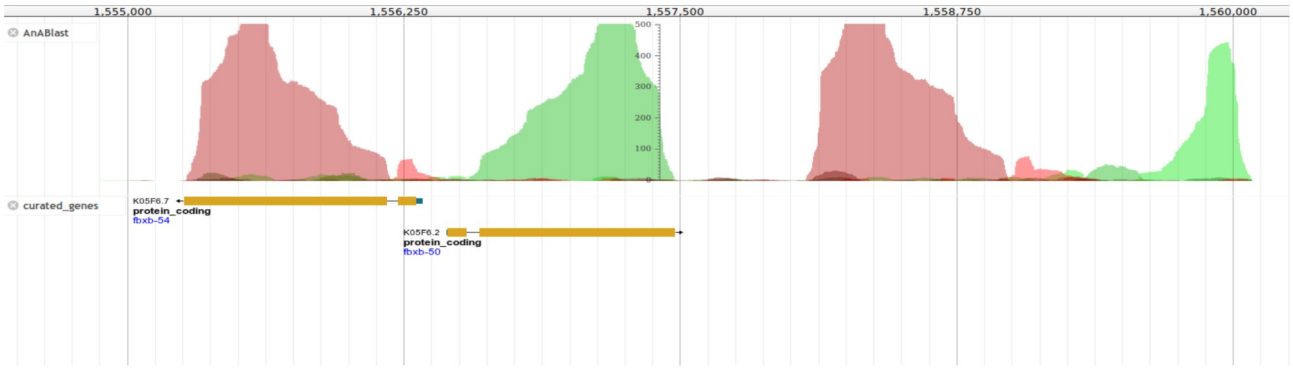


Figura 17: Posible duplicación en espejo invertida y duplicación en tandem. Imagen capturada del explorador de genes (JBrowse) de *C. elegans*. Se muestra dos proteínas F-box B (lado izquierdo) procedentes de los genes curados de Wormbase. Además, se muestra también los picos de AnABlast correspondientes a estas proteínas y dos picos más, aguas arriba, con formas similares (los picos en rojo se corresponden con señales de la cadena reversa y en verde de la cadena adelantada). La forma de los picos de AnABlast sugiere que existió una primera duplicación en espejo e inversión que dio lugar a las dos proteínas F-box B seguida de una duplicación en tandem que dio lugar a los dos picos, aguas arriba, de los cuales el que está en la cadena reversa (rojo) posiblemente se traduzca a aminoácidos a pesar de no encontrarse descrito en WormBase.

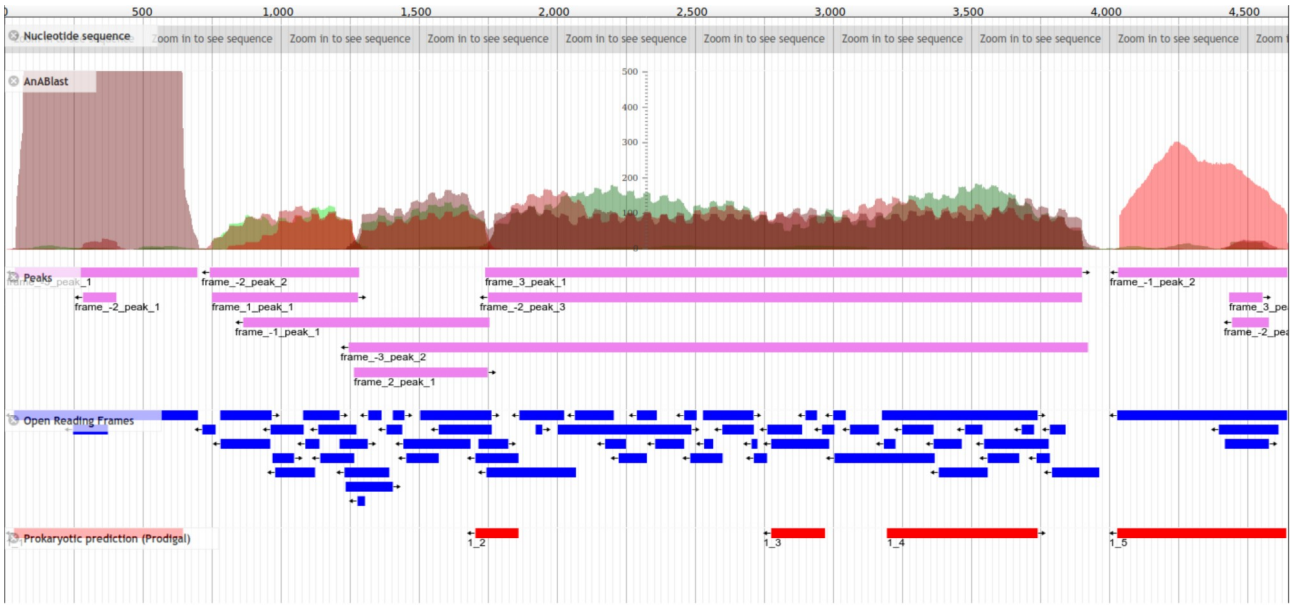


Figura 18: Resultados de la aplicación web de AnABlast de una secuencia genómica bacteriana. A la derecha se puede ver la característica forma de sierra que presentan las regiones CRISPR. El primer track se corresponde con las señales de AnABlast (picos). El segundo track (violeta) son las regiones genómicas dentro de cada pico. El tercero (azul) los ORF solapantes con cada pico. El último track (rojo) es la predicción de la aplicación Prodigal para esta secuencia genómica.

9. Conclusiones

9.1 Conclusiones

1. Sma3s v2 es una herramienta bioinformática que permite la anotación funcional tanto de proteomas como de transcriptomas completos.

2. Sma3s v2 reduce los tiempos de ejecución y los requerimientos computacionales con respecto a la versión 1 manteniendo la precisión en las anotaciones y aumentando la sensibilidad.

3. Sma3s v2 es una aplicación que puede ser empleada por toda la comunidad científica, sin requerir apenas conocimientos computacionales.

4. Sma3s v2 permite comparar anotaciones de genomas completos, por medio de las clases funcionales de su fichero de resultados summary.

5. Las aplicaciones diseñadas para localizar genes en genomas completos, actualmente no son capaces de localizar todos lo genes, existiendo un número significativo de genes que se escapan a sus predicciones.

6. El acúmulo de alineamientos de secuencia no significativos de proteínas sobre genomas completos no se produce al azar, sino en regiones codificantes de proteínas y principalmente en su mismo marco de lectura

7. El acúmulo de alineamientos no significativos de proteínas sobre genomas completos indica ,en un alto porcentaje, la presencia de regiones codificantes de proteínas, siendo capaces de discriminar entre zonas codificantes y no codificantes. Además, aumenta la probabilidad de ser una región codificante de proteínas cuanto mayor es el acúmulo de alineamientos no significativos de BLAST

8. Los falsos positivos de AnABlast pueden indicar la presencia de posibles nuevas regiones codificantes de proteínas que se han escapado a la predicción mediante otros métodos.

9. Utilizar un bitscore de 30 en BLAST como límite inferior para los alineamientos no significativos, muestra los mejores resultados para detectar regiones codificantes de proteínas, y no acumular así secuencias al azar.

10. Para un determinado bitscore la altura de pico más precisa depende del tamaño de la base de datos que se emplea como referencia. Para una referencia actual (Uniref50 2017), una altura de pico de 70 es la más adecuada para detectar posibles nuevas regiones codificantes de proteínas.

11. En AnABlast, el empleo de bases de datos más recientes produce una mayor sensibilidad con respecto a utilizar bases de datos antiguas. Siendo los picos de AnABlast significativamente más bajos cuando se emplean bases de datos menos recientes.

12. AnABlast es capaz de detectar secuencias codificantes de proteínas y pseudogenes actuales de *Drosophila melanogaster*, no incluidas en versiones antiguas de la base de datos FlyBase ni detectables por otros métodos como AUGUSTUS, empleando como referencia versiones antiguas de UniRef50.

13. AnABlast es capaz de detectar nuevas regiones codificantes de proteínas que presentan evidencias claras de expresión pero que aún no están incluidas como codificantes en versiones actuales de las bases de datos de *Drosophila melanogaster*.

14. AnABlast es capaz de predecir nuevas regiones codificantes de proteínas en el genoma de *Caenorhabditis elegans*, incluso aquellas consideradas smallORF (menos de 100 aa).

15. Muchas de las regiones codificantes de proteínas predichas por AnABlast en el genoma de *Caenorhabditis elegans* presentan evidencias de transcripción y traducción y algunas de ellas pueden servalidadas mediante experimentos de RNA interferente en laboratorio.

16. La aplicación web facilita el uso de AnABlast en regiones de hasta 25 Kb permitiendo analizar el acúmulo de protomotivos en regiones menores de 25 kb.

10. Bibliografía General

A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, et al. 1996. “Life with 6000 Genes.” *Science* 274.

Abebrese, Emmanuel L., Syed H. Ali, Zachary R. Arnold, Victoria M. Andrews, Katharine Armstrong, Lindsay Burns, Hannah R. Crowder, et al. 2017. “Identification of Human Short Introns.” Edited by Emanuele Buratti. *PLOS ONE* 12 (5): e0175393. https://doi.org/10.1371/journal.pone.0175393.

Adams, M. D. 2000. “The Genome Sequence of Drosophila Melanogaster.” *Science* 287 (5461): 2185–95. https://doi.org/10.1126/science.287.5461.2185.

Altschul, S. 1997. “Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs.” *Nucleic Acids Research* 25 (17): 3389–3402. https://doi.org/10.1093/nar/25.17.3389.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Andrade, Miguel A. 2009. “Position-Specific Annotation of Protein Function Based on Multiple Homologs,” 6.

Andrews, Shea J., and Joseph A. Rothnagel. 2014. “Emerging Evidence for Functional Peptides Encoded by Short Open Reading Frames.” *Nature Reviews Genetics* 15 (3): 193–204. https://doi.org/10.1038/nrg3520.

Antipov, Dmitry, Anton Korobeynikov, Jeffrey S. McLean, and Pavel A. Pevzner. 2016. “HybridSPAdes: An Algorithm for Hybrid Assembly of Short and Long Reads.” *Bioinformatics* 32 (7): 1009–15. https://doi.org/10.1093/bioinformatics/btv688.

Attwood, T. K. 2000. “PRINTS-S: The Database Formerly Known as PRINTS.” *Nucleic Acids Research* 28 (1): 225–27. https://doi.org/10.1093/nar/28.1.225.

Au, Kin Fai, Jason G. Underwood, Lawrence Lee, and Wing Hung Wong. 2012. “Improving PacBio Long Read Accuracy by Short Read Alignment.” Edited by Yi Xing. *PLoS ONE* 7 (10): e46679. https://doi.org/10.1371/journal.pone.0046679.

Aziz, Ramy K, Daniela Bartels, Aaron A Best, Matthew DeJongh, Terrence Disz, Robert A Edwards, Kevin Formsma, et al. 2008. “The RAST Server: Rapid Annotations Using Subsystems Technology.” *BMC Genomics* 9 (1): 75. https://doi.org/10.1186/1471-2164-9-75.

Badger, J. H., and G. J. Olsen. 1999. “CRITICA: Coding Region Identification Tool Invoking Comparative Analysis.” *Molecular Biology and Evolution* 16 (4): 512–24. https://doi.org/10.1093/oxfordjournals.molbev.a026133.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al. 2012. “SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing.” *Journal of Computational Biology* 19 (5): 455–77. https://doi.org/10.1089/cmb.2012.0021.

Benson, Dennis A., Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. 2012. “GenBank.” *Nucleic Acids Research* 41 (D1): D36–42. https://doi.org/10.1093/nar/gks1195.

Borodovsky, Mark, and James McIninch. 1993. “GENMARK: Parallel Gene Recognition for Both DNA Strands.” *Computers & Chemistry* 17 (2): 123–33. https://doi.org/10.1016/0097-8485(93)85004-V.

Bru, C. 2004. “The ProDom Database of Protein Domain Families: More Emphasis on 3D.” *Nucleic Acids Research* 33 (Database issue): D212–15. https://doi.org/10.1093/nar/gki034.

Bryant, Donald M., Kimberly Johnson, Tia DiTommaso, Timothy Tickle, Matthew Brian Couger, Duygu Payzin-Dogru, Tae J. Lee, et al. 2017. “A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors.” *Cell Reports* 18 (3): 762–76. https://doi.org/10.1016/j.celrep.2016.12.063.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson. 2015. “Fast and Sensitive Protein Alignment Using DIAMOND.” *Nature Methods* 12 (1): 59–60. https://doi.org/10.1038/nmeth.3176.

Buels, Robert, Eric Yao, Colin M. Diesh, Richard D. Hayes, Monica Munoz-Torres, Gregg Helt, David M. Goodstein, et al. 2016. “JBrowse: A Dynamic Web Platform for Genome Visualization and Analysis.” *Genome Biology* 17 (1): 66. https://doi.org/10.1186/s13059-016-0924-1.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. 2009. “BLAST+: Architecture and Applications.” *BMC Bioinformatics* 10 (1): 421. https://doi.org/10.1186/1471-2105-10-421.

“CGI.” 2019. Página web. 2019. https://metacpan.org/pod/CGI.

Chen, Ting-Wen, Ruei-Chi Richie Gan, Timothy H Wu, Po-Jung Huang, Cheng-Yang Lee, Yi-Ywan M Chen, Che-Chun Chen, and Petrus Tang. 2012. “FastAnnotator- an Efficient Transcript Annotation Web Tool,” 8.

Chin, Chen-Shan, Paul Peluso, Fritz J Sedlazeck, Maria Nattestad, Gregory T Concepcion, Alicia Clum, Christopher Dunn, et al. 2016. “Phased Diploid Genome Assembly with Single-Molecule Real-Time Sequencing.” *Nature Methods* 13 (12): 1050–54. https://doi.org/10.1038/nmeth.4035.

Collins, F. S. 2003. “The Human Genome Project: Lessons from Large-Scale Biology.” *Science* 300 (5617): 286–90. https://doi.org/10.1126/science.1084564.

Conesa, A., S. Gotz, J. M. Garcia-Gomez, J. Terol, M. Talon, and M. Robles. 2005. “Blast2GO: A Universal Tool for Annotation, Visualization and Analysis in Functional Genomics Research.” *Bioinformatics* 21 (18): 3674–76. https://doi.org/10.1093/bioinformatics/bti610.

Couso, Juan-Pablo, and Pedro Patraquim. 2017. “Classification and Function of Small Open Reading Frames.” *Nature Reviews Molecular Cell Biology* 18 (9): 575–89. https://doi.org/10.1038/nrm.2017.58.

Crappé, Jeroen, Wim Van Criekinge, Geert Trooskens, Eisuke Hayakawa, Walter Luyten, Geert Baggerman, and Gerben Menschaert. 2013. “Combining in Silico Prediction and Ribosome Profiling in a Genome-Wide Search for Novel Putatively Coding SORFs.” *BMC Genomics* 14 (1): 648. https://doi.org/10.1186/1471-2164-14-648.

Dawson, Natalie L., Tony E. Lewis, Sayoni Das, Jonathan G. Lees, David Lee, Paul Ashford, Christine A. Orengo, and Ian Sillitoe. 2017. “CATH: An Expanded Resource to Predict Protein Function through Structure and Sequence.” *Nucleic Acids Research* 45 (D1): D289–95. https://doi.org/10.1093/nar/gkw1098.

Dayhoff, M. O, Schwartz, R. M. 1979. *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation.

Deamer, David, Mark Akeson, and Daniel Branton. 2016. “Three Decades of Nanopore Sequencing.” *Nature Biotechnology* 34 (5): 518–24. https://doi.org/10.1038/nbt.3423.

Delcher, A. 1999. “Improved Microbial Gene Identification with GLIMMER.” *Nucleic Acids Research* 27 (23): 4636–41. https://doi.org/10.1093/nar/27.23.4636.

Delcher, Arthur L., Kirsten A. Bratke, Edwin C. Powers, and Steven L. Salzberg. 2007. “Identifying Bacterial Genes and Endosymbiont DNA with Glimmer.” *Bioinformatics* 23 (6): 673–79. https://doi.org/10.1093/bioinformatics/btm009.

Dubaj Price, Michelle, and Daryl D. Hurd. 2019. “WormBase: A Model Organism Database.” *Medical Reference Services Quarterly* 38 (1): 70–80. https://doi.org/10.1080/02763869.2019.1548896.

El-Gebali, Sara, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, et al. 2019. “The Pfam Protein Families Database in 2019.” *Nucleic Acids Research* 47 (D1): D427–32. https://doi.org/10.1093/nar/gky995.

“Elsiklab, Github.” 2019. 2019. https://github.com/elsiklab/multibigwig.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. “CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data.” *Bioinformatics* 28 (23): 3150–52. https://doi.org/10.1093/bioinformatics/bts565.

Gazda, Małgorzata A, Pedro Andrade, Sandra Afonso, Jolita Dilytė, John P Archer, Ricardo J Lopes, Rui Faria, and Miguel Carneiro. 2018. “Signatures of Selection on Standing Genetic Variation Underlie Athletic and Navigational Performance in Racing Pigeons.” Edited by Emma Teeling. *Molecular Biology and Evolution* 35 (5): 1176–89. https://doi.org/10.1093/molbev/msy030.

Gene Ontology Consortium. 2004. “The Gene Ontology (GO) Database and Informatics Resource.” *Nucleic Acids Research* 32 (90001): 258D – 261. https://doi.org/10.1093/nar/gkh036.

Gerstein, M. B., C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, J. O. Korbel, O. Emanuelsson, Z. D. Zhang, S. Weissman, and M. Snyder. 2007. “What Is a Gene, Post-ENCODE? History and Updated Definition.” *Genome Research* 17 (6): 669–81. https://doi.org/10.1101/gr.6339607.

Gerstein, M. B., Z. J. Lu, E. L. Van Nostrand, C. Cheng, B. I. Arshinoff, T. Liu, K. Y. Yip, et al. 2010. “Integrative Analysis of the Caenorhabditis Elegans Genome by the ModENCODE Project.” *Science* 330 (6012): 1775–87. https://doi.org/10.1126/science.1196914.

Goodswen, Stephen J., Paul J. Kennedy, and John T. Ellis. 2012. “Evaluating High-Throughput Ab Initio Gene Finders to Discover Proteins Encoded in Eukaryotic Pathogen Genomes Missed by Laboratory Techniques.” Edited by Anna Tramontano. *PLoS ONE* 7 (11): e50609. https://doi.org/10.1371/journal.pone.0050609.

Gross, Samuel S, Chuong B Do, Marina Sirota, and Serafim Batzoglou. 2007. “CONTRAST: A Discriminative, Phylogeny-Free Approach to Multiple Informant de Novo Gene Prediction.” *Genome Biology* 8 (12): R269. https://doi.org/10.1186/gb-2007-8-12-r269.

Guerrero-Sanchez, Victor M., Ana M. Maldonado-Alconada, Francisco Amil-Ruiz, Andrea Verardi, Jesús V. Jorrín-Novo, and María-Dolores Rey. 2019. “Ion Torrent and Lllumina, Two Complementary RNA-Seq Platforms for Constructing the Holm Oak (Quercus Ilex) Transcriptome.” Edited by Mukesh Jain. *PLOS ONE* 14 (1): e0210356. https://doi.org/10.1371/journal.pone.0210356.

Guigó, Roderic, Steen Knudsen, Neil Drake, and Temple Smith. 1992. “Prediction of Gene Structure,” January, 17.

Haft, Daniel H., Jeremy D. Selengut, Roland A. Richter, Derek Harkins, Malay K. Basu, and Erin Beck. 2012. “TIGRFAMs and Genome Properties in 2013.” *Nucleic Acids Research* 41 (D1): D387–95. https://doi.org/10.1093/nar/gks1234.

Hamidian, Mohammad, Ryan R. Wick, Rebecca M. Hartstein, Louise M. Judd, Kathryn E. Holt, and Ruth M. Hall. 2019. “Insights from the Revised Complete Genome Sequences of Acinetobacter Baumannii Strains AB307-0294 and ACICU Belonging to Global Clones 1 and 2.” *Microbial Genomics*, September. https://doi.org/10.1099/mgen.0.000298.

Hanada, Kousuke, Kenji Akiyama, Tetsuya Sakurai, Tetsuro Toyoda, Kazuo Shinozaki, and Shin-Han Shiu. 2010. “SORF Finder: A Program Package to Identify Small Open Reading Frames with High Coding Potential.” *Bioinformatics* 26 (3): 399–400. https://doi.org/10.1093/bioinformatics/btp688.

Harris, Todd W., Igor Antoshechkin, Tamberlyn Bieri, Darin Blasiar, Juancarlos Chan, Wen J. Chen, Norie De La Cruz, et al. 2010. “WormBase: A Comprehensive Resource for Nematode Research.” *Nucleic Acids Research* 38 (suppl_1): D463–67. https://doi.org/10.1093/nar/gkp952.

Heath, Simon C, Ivo G Gut, Paul Brennan, James D McKay, Vladimir Bencko, Eleonora Fabianova, Lenka Foretova, et al. 2008. “Investigation of the Fine Structure of European Populations with Applications to Disease Association Studies.” *European Journal of Human Genetics* 16 (12): 1413–29. https://doi.org/10.1038/ejhg.2008.210.

Hellens, Roger P., Chris M. Brown, Matthew A.W. Chisnall, Peter M. Waterhouse, and Richard C. Macknight. 2016. “The Emerging World of Small ORFs.” *Trends in Plant Science* 21 (4): 317–28. https://doi.org/10.1016/j.tplants.2015.11.005.

Henikoff, S., and J. G. Henikoff. 1992. “Amino Acid Substitution Matrices from Protein Blocks.” *Proceedings of the National Academy of Sciences* 89 (22): 10915–19. https://doi.org/10.1073/pnas.89.22.10915.

Huerta-Cepas, Jaime, Kristoffer Forslund, Luis Pedro Coelho, Damian Szklarczyk, Lars Juhl Jensen, Christian von Mering, and Peer Bork. 2017. “Fast Genome-Wide Functional

Annotation through Orthology Assignment by EggNOG-Mapper.” *Molecular Biology and Evolution* 34 (8): 2115–22. https://doi.org/10.1093/molbev/msx148.

Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, et al. 2019. “EggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses.” *Nucleic Acids Research* 47 (D1): D309–14. https://doi.org/10.1093/nar/gky1085.

Hyatt, Doug, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. 2010. “Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification.” *BMC Bioinformatics* 11 (1): 119. https://doi.org/10.1186/1471-2105-11-119.

Jackson, Ruaidhrí, Lina Kroehling, Alexandra Khitun, Will Bailis, Abigail Jarret, Autumn G. York, Omair M. Khan, et al. 2018. “The Translation of Non-Canonical Open Reading Frames Controls Mucosal Immunity.” *Nature* 564 (7736): 434–38. https://doi.org/10.1038/s41586-018-0794-7.

“JavaScript.” 2019. 2019. https://www.javascript.com/.

Ji, Zhe, Ruisheng Song, Aviv Regev, and Kevin Struhl. 2015. “Many LncRNAs, 5’UTRs, and Pseudogenes Are Translated and Some Are Likely to Express Functional Proteins.” *ELife* 4 (December): e08890. https://doi.org/10.7554/eLife.08890.

Jimenez, Juan, Caia D. S. Duncan, María Gallardo, Juan Mata, and Antonio J. Perez-Pulido. 2015. “AnABlast: A New in Silico Strategy for the Genome-Wide Search of Novel Genes and Fossil Regions.” *DNA Research* 22 (6): 439–49. https://doi.org/10.1093/dnares/dsv025.

Jones, P., D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, et al. 2014. “InterProScan 5: Genome-Scale Protein Function Classification.” *Bioinformatics* 30 (9): 1236–40. https://doi.org/10.1093/bioinformatics/btu031.

Kamath, Ravi S., Andrew G. Fraser, Yan Dong, Gino Poulin, Richard Durbin, Monica Gotta, Alexander Kanapin, et al. 2003. “Systematic Functional Analysis of the Caenorhabditis Elegans Genome Using RNAi.” *Nature* 421 (6920): 231–37. https://doi.org/10.1038/nature01278.

Kelley, David R., Bo Liu, Arthur L. Delcher, Mihai Pop, and Steven L. Salzberg. 2012. “Gene Prediction with Glimmer for Metagenomic Sequences Augmented by Classification and Clustering.” *Nucleic Acids Research* 40 (1): e9–e9. https://doi.org/10.1093/nar/gkr1067.

Kent, W. J., A. S. Zweig, G. Barber, A. S. Hinrichs, and D. Karolchik. 2010. “BigWig and BigBed: Enabling Browsing of Large Distributed Datasets.” *Bioinformatics* 26 (17): 2204–7. https://doi.org/10.1093/bioinformatics/btq351.

Kersey, Paul Julian, James E. Allen, Irina Armean, Sanjay Boddu, Bruce J. Bolt, Denise Carvalho-Silva, Mikkel Christensen, et al. 2016. “Ensembl Genomes 2016: More Genomes, More Complexity.” *Nucleic Acids Research* 44 (D1): D574–80. https://doi.org/10.1093/nar/gkv1209.

Khan, Ishita K., Qing Wei, Meghana Chitale, and Daisuke Kihara. 2015. “PFP/ESG: Automated Protein Function Prediction Servers Enhanced with Gene Ontology Visualization Tool.” *Bioinformatics* 31 (2): 271–72. https://doi.org/10.1093/bioinformatics/btu646.

Kong, Lei, Yong Zhang, Zhi-Qiang Ye, Xiao-Qiao Liu, Shu-Qi Zhao, Liping Wei, and Ge Gao. 2007. “CPC: Assess the Protein-Coding Potential of Transcripts Using Sequence Features and Support Vector Machine.” *Nucleic Acids Research* 35 (suppl_2): W345–49. https://doi.org/10.1093/nar/gkm391.

Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. “Canu: Scalable and Accurate Long-Read Assembly via Adaptive *k* -Mer Weighting and Repeat Separation.” *Genome Research* 27 (5): 722–36. https://doi.org/10.1101/gr.215087.116.

Kourelis, Jiorgos. 2018. “Re-Annotated Nicotiana Benthamiana Gene Models for Enhanced Proteomics and Reverse Genetics,” 16.

Krijgsveld, Jeroen, René F Ketting, Tokameh Mahmoudi, Janik Johansen, Marta Artal-Sanz, C Peter Verrijzer, Ronald H A Plasterk, and Albert J R Heck. 2003. “Metabolic Labeling of C. Elegans and D. Melanogaster for Quantitative Proteomics.” *Nature Biotechnology* 21 (8): 927–31. https://doi.org/10.1038/nbt848.

Kroll, José Eduardo, Vandeclécio Lira da Silva, Sandro José de Souza, and Gustavo Antonio de Souza. 2017. “A Tool for Integrating Genetic and Mass Spectrometry-Based Peptide Data: Proteogenomics Viewer: PV: A Genome Browser-like Tool, Which Includes MS Data Visualization and Peptide Identification Parameters.” *BioEssays* 39 (7): 1700015. https://doi.org/10.1002/bies.201700015.

Lavezzo, Enrico, Marco Falda, Paolo Fontana, Luca Bianco, and Stefano Toppo. 2016. “Enhancing Protein Function Prediction with Taxonomic Constraints – The Argot2.5 Web Server.” *Methods* 93 (January): 15–23. https://doi.org/10.1016/j.ymeth.2015.08.021.

Letunic, Ivica, and Peer Bork. 2018. “20 Years of the SMART Protein Domain Annotation Resource.” *Nucleic Acids Research* 46 (D1): D493–96. https://doi.org/10.1093/nar/gkx922.

Li, Heng. 2016. “Minimap and Miniasm: Fast Mapping and de Novo Assembly for Noisy Long Sequences.” *Bioinformatics* 32 (14): 2103–10. https://doi.org/10.1093/bioinformatics/btw152.

Lin, M. F., I. Jungreis, and M. Kellis. 2011. “PhyloCSF: A Comparative Genomics Method to Distinguish Protein Coding and Non-Coding Regions.” *Bioinformatics* 27 (13): i275–82. https://doi.org/10.1093/bioinformatics/btr209.

Loman, Nicholas J., and Mark J. Pallen. 2015. “Twenty Years of Bacterial Genome Sequencing.” *Nature Reviews Microbiology* 13 (12): 787–94. https://doi.org/10.1038/nrmicro3565.

López-Hidalgo, Cristina, Victor M. Guerrero-Sánchez, Isabel Gómez-Gálvez, Rosa Sánchez-Lucas, María A. Castillejo-Sánchez, Ana M. Maldonado-Alconada, Luis Valledor, and Jesus V. Jorrín-Novo. 2018. “A Multi-Omics Analysis Pipeline for the Metabolic Pathway Reconstruction in the Orphan Species Quercus Ilex.” *Frontiers in Plant Science* 9 (July): 935. https://doi.org/10.3389/fpls.2018.00935.

Luo, Ruibang, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, et al. 2012. “SOAPdenovo2: An Empirically Improved Memory-Efficient Short-Read de Novo Assembler.” *GigaScience* 1 (1): 18. https://doi.org/10.1186/2047-217X-1-18.

Magny, E. G., J. I. Pueyo, F. M. G. Pearl, M. A. Cespedes, J. E. Niven, S. A. Bishop, and J. P. Couso. 2013. “Conserved Regulation of Cardiac Calcium Uptake by Peptides Encoded in Small Open Reading Frames.” *Science* 341 (6150): 1116–20. https://doi.org/10.1126/science.1238802.

Majoros, W. H., M. Pertea, and S. L. Salzberg. 2004. “TigrScan and GlimmerHMM: Two Open Source Ab Initio Eukaryotic Gene-Finders.” *Bioinformatics* 20 (16): 2878–79. https://doi.org/10.1093/bioinformatics/bth315.

Maxam, A. M., and W. Gilbert. 1977. “A New Method for Sequencing DNA.” *Proceedings of the National Academy of Sciences* 74 (2): 560–64. https://doi.org/10.1073/pnas.74.2.560.

Mi, Huaiyu, Anushya Muruganujan, Dustin Ebert, Xiaosong Huang, and Paul D Thomas. 2019. “PANTHER Version 14: More Genomes, a New PANTHER GO-Slim and Improvements in Enrichment Analysis Tools.” *Nucleic Acids Research* 47 (D1): D419–26. https://doi.org/10.1093/nar/gky1038.

Mitchell, Alex, Hsin-Yu Chang, Louise Daugherty, Matthew Fraser, Sarah Hunter, Rodrigo Lopez, Craig McAnulla, et al. 2015. “The InterPro Protein Families Database: The Classification Resource after 15 Years.” *Nucleic Acids Research* 43 (D1): D213–21. https://doi.org/10.1093/nar/gku1243.

modENCODE Consortium, Susan E. Celniker, Laura A. L. Dillon, Mark B. Gerstein, Kristin C. Gunsalus, Steven Henikoff, Gary H. Karpen, et al. 2009. “Unlocking the Secrets of the Genome.” *Nature* 459 (7249): 927–30. https://doi.org/10.1038/459927a.

Mount, D.W. 2001. *Genome Analysis. In Bioinformatics: Sequence and Genome Analysis.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Mukherjee, Supratim, Dimitri Stamatis, Jon Bertsch, Galina Ovchinnikova, Hema Y Katta, Alejandro Mojica, I-Min A Chen, Nikos C Kyrpides, and Tbk Reddy. 2019. “Genomes OnLine Database (GOLD) v.7: Updates and New Features.” *Nucleic Acids Research* 47 (D1): D649–59. https://doi.org/10.1093/nar/gky977.

Munoz-Merida, A., E. Viguera, M. G. Claros, O. Trelles, and A. J. Perez-Pulido. 2014. “Sma3s: A Three-Step Modular Annotator for Large Sequence Datasets.” *DNA Research* 21 (4): 341–53. https://doi.org/10.1093/dnares/dsu001.

Needleman, Saul B., and Christian D. Wunsch. 1970. “A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins.” *Journal of Molecular Biology* 48 (3): 443–53. https://doi.org/10.1016/0022-2836(70)90057-4.

Nirenberg, M. W., and J. H. Matthaei. 1961. “The Dependence of Cell-Free Protein Synthesis in E. Coli upon Naturally Occurring or Synthetic Polyribonucleotides.” *Proceedings of the National Academy of Sciences* 47 (10): 1588–1602. https://doi.org/10.1073/pnas.47.10.1588.

Olexiouk, Volodimir, Wim Van Criekinge, and Gerben Menschaert. 2018. “An Update on SORFs.Org: A Repository of Small ORFs Identified by Ribosome Profiling.” *Nucleic Acids Research* 46 (D1): D497–502. https://doi.org/10.1093/nar/gkx1130.

Pandurangan, Arun Prasad, Jonathan Stahlhacke, Matt E Oates, Ben Smithers, and Julian Gough. 2019. “The SUPERFAMILY 2.0 Database: A Significant Proteome Update and a New Webserver.” *Nucleic Acids Research* 47 (D1): D490–94. https://doi.org/10.1093/nar/gky1130.

Pearson, W. R., and D. J. Lipman. 1988. “Improved Tools for Biological Sequence Comparison.” *Proceedings of the National Academy of Sciences* 85 (8): 2444–48. https://doi.org/10.1073/pnas.85.8.2444.

Pedruzzi, Ivo, Catherine Rivoire, Andrea H. Auchincloss, Elisabeth Coudert, Guillaume Keller, Edouard de Castro, Delphine Baratin, et al. 2015. “HAMAP in 2015: Updates to the Protein Family Classification and Annotation System.” *Nucleic Acids Research* 43 (D1): D1064–70. https://doi.org/10.1093/nar/gku1002.

Perez, A. J., G. Thode, and O. Trelles. 2004. “AnaGram: Protein Function Assignment.” *Bioinformatics* 20 (2): 291–92. https://doi.org/10.1093/bioinformatics/btg414.

Piovesan, D., P. Luigi Martelli, P. Fariselli, A. Zauli, I. Rossi, and R. Casadio. 2011. “BAR-PLUS: The Bologna Annotation Resource Plus for Functional and Structural Annotation of Protein Sequences.” *Nucleic Acids Research* 39 (suppl): W197–202. https://doi.org/10.1093/nar/gkr292.

Pueyo, Jose I., Emile G. Magny, and Juan P. Couso. 2016. “New Peptides Under the s(ORF)Ace of the Genome.” *Trends in Biochemical Sciences* 41 (8): 665–78. https://doi.org/10.1016/j.tibs.2016.05.003.

Radivojac, Predrag, Wyatt T Clark, Tal Ronnen Oron, Alexandra M Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, et al. 2013. “A Large-Scale Evaluation of Computational Protein Function Prediction.” *Nature Methods* 10 (3): 221–27. https://doi.org/10.1038/nmeth.2340.

Rhoads, Anthony, and Kin Fai Au. 2015. “PacBio Sequencing and Its Applications.” *Genomics, Proteomics & Bioinformatics* 13 (5): 278–89. https://doi.org/10.1016/j.gpb.2015.08.002.

Ruiz-Orera, Jorge, Xavier Messeguer, Juan Antonio Subirana, and M Mar Alba. 2014. “Long Non-Coding RNAs as a Source of New Peptides.” *ELife* 3 (September): e03523. https://doi.org/10.7554/eLife.03523.

Salzberg, S. L., A. L. Delcher, S. Kasif, and O. White. 1998. “Microbial Gene Identification Using Interpolated Markov Models.” *Nucleic Acids Research* 26 (2): 544–48. https://doi.org/10.1093/nar/26.2.544.

Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. 1977. “Nucleotide Sequence of Bacteriophage ΦX174 DNA.” *Nature* 265 (5596): 687–95. https://doi.org/10.1038/265687a0.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. “DNA Sequencing with Chain-Terminating Inhibitors.” *Proceedings of the National Academy of Sciences* 74 (12): 5463–67. https://doi.org/10.1073/pnas.74.12.5463.

Schadt, E. E., S. Turner, and A. Kasarskis. 2010. “A Window into Third-Generation Sequencing.” *Human Molecular Genetics* 19 (R2): R227–40. https://doi.org/10.1093/hmg/ddq416.

Seemann, T. 2014. “Prokka: Rapid Prokaryotic Genome Annotation.” *Bioinformatics* 30 (14): 2068–69. https://doi.org/10.1093/bioinformatics/btu153.

Siepel, A. 2005. “Evolutionarily Conserved Elements in Vertebrate, Insect, Worm, and Yeast Genomes.” *Genome Research* 15 (8): 1034–50. https://doi.org/10.1101/gr.3715005.

Sigrist, Christian J. A., Edouard de Castro, Lorenzo Cerutti, Béatrice A. Cuche, Nicolas Hulo, Alan Bridge, Lydie Bougueleret, and Ioannis Xenarios. 2012. “New and Continuing Developments at PROSITE.” *Nucleic Acids Research* 41 (D1): D344–47. https://doi.org/10.1093/nar/gks1067.

Smith, T.F., and M.S. Waterman. 1981. “Identification of Common Molecular Subsequences.” *Journal of Molecular Biology* 147 (1): 195–97. https://doi.org/10.1016/0022-2836(81)90087-5.

Stanke, M., R. Steinkamp, S. Waack, and B. Morgenstern. 2004. “AUGUSTUS: A Web Server for Gene Finding in Eukaryotes.” *Nucleic Acids Research* 32 (Web Server): W309–12. https://doi.org/10.1093/nar/gkh379.

The 1000 Genomes Project Consortium, Laura Clarke, Xiangqun Zheng-Bradley, Richard Smith, Eugene Kulesha, Chunlin Xiao, Iliana Toneva, et al. 2012. “The 1000 Genomes Project: Data Management and Community Access.” *Nature Methods* 9 (5): 459–62. https://doi.org/10.1038/nmeth.1974.

The 1000 Genomes Project Consortium, Gil A. McVean, David M. Altshuler (Co-Chair), Richard M. Durbin (Co-Chair), Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, et al. 2012. “An Integrated Map of Genetic Variation from 1,092 Human Genomes.” *Nature* 491 (October): 56.

The ENCODE Project Consortium. 2004. “The ENCODE (ENCyclopedia Of DNA Elements) Project.” *Science* 306 (5696): 636–40. https://doi.org/10.1126/science.1105136.

The UniProt Consortium. 2017. “UniProt: The Universal Protein Knowledgebase.” *Nucleic Acids Research* 45 (D1): D158–69. https://doi.org/10.1093/nar/gkw1099.

Thode, Guillermo, Juan Antonio García-Renea, and Juan Jimenez. 1996. “Search for Ancient Patterns in Protein Sequences.” *Journal of Molecular Evolution* 42 (2): 224–33. https://doi.org/10.1007/BF02198848.

Tweedie, S., M. Ashburner, K. Falls, P. Leyland, P. McQuilton, S. Marygold, G. Millburn, et al. 2009. “FlyBase: Enhancing Drosophila Gene Ontology Annotations.” *Nucleic Acids Research* 37 (Database): D555–59. https://doi.org/10.1093/nar/gkn788.

White, Owen, Carol J Bult, Jean-Francois Tomb, Granger Sutton, Chris Fields, Li-lng Liu, and Tracy Spriggs. 1995. “Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd.” *SCIENCE* 216: 15.

Wilson, Richard K. 1999. “How the Worm Was Won: The C. Elegans Genome Sequencing Project.” *Trends in Genetics* 15 (2): 51–58. https://doi.org/10.1016/S0168-9525(98)01666-7.

Wood, V., M. A. Harris, M. D. McDowall, K. Rutherford, B. W. Vaughan, D. M. Staines, M. Aslett, et al. 2012. “PomBase: A Comprehensive Online Resource for Fission Yeast.” *Nucleic Acids Research* 40 (D1): D695–99. https://doi.org/10.1093/nar/gkr853.

Wu, C. H. 2004. “PIRSF: Family Classification System at the Protein Information Resource.” *Nucleic Acids Research* 32 (90001): 112D – 114. https://doi.org/10.1093/nar/gkh097.

Yoshimura, Jun, Kazuki Ichikawa, Massa J. Shoura, Karen L. Artiles, Idan Gabdank, Lamia Wahba, Cheryl L. Smith, et al. 2019. “Recompleting the *Caenorhabditis Elegans* Genome.” *Genome Research* 29 (6): 1009–22. https://doi.org/10.1101/gr.244830.118.

Zerbino, D. R., and E. Birney. 2008. “Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs.” *Genome Research* 18 (5): 821–29. https://doi.org/10.1101/gr.074492.107.

11. Apéndice

11.1 Códigos fuentes

11.1.1 Sma3s v2

https://github.com/UPOBioinfo/sma3s

11.1.2 AnABlast

https://github.com/UPOBioinfo/AnABlast

11.1.3 Aplicación web AnABlast

https://github.com/UPOBioinfo/AnABlast_webapp

11.2 Tablas de las posibles regiones codificantes encontrados en *C. elegans*

11.2.1 Posibles genes codificantes de proteínas

id	position	Expression	Homologs	motifs	ORF length	Protomotifs
G5	IV:1213206..1213443	+ (Polysome)	-		80	88
G14	II:11736898..11737084	+ (RNAseq)	-		39	175
G22	III:670901..671255	+ (RNAseq)	-		162	74
G23	III:5838386..5838494	+ (RNAseq)	-		41	71
G25	X:17515475..17515919	+ (RNAseq)	-		175	70
G26	II:11717583..11718477	+ (RNAseq)	-	PF07735 (F-box associated)	269	297
G30	II:13644506..13644671	+ (RNAseq)	+ (O02112_CAEEL)		75	196
G31	I:12928143..12928563	+ (RNAseq)	+ (G8JLX0_CAEEL)	PF07735 (F-box associated)	167	164
G32	V:13396252..13396714	+ (RNAseq)	-		176	84
G43	X:15015384..15015639	+ (RNAseq)	-		49	233
G45	III:7833431..7833716	+ (RNAseq)	+ (Q9XWD9_CAEEL)		97	102
G46	I:7316251..7316590	+ (RNAseq)	-		122	70
G49	V:14367725..14368037	+ (RNAseq)	-		119	74
G50	X:16728266..16728758	+ (RNAseq)	-		126	93
G51	V:8646870..8647347	+ (RNAseq)	-		165	98
G52	IV:379001..379340	+ (RNAseq)	-		140	71
G54	III:4101011..4101389	+ (RNAseq)	+ (A0A0U5FV96_9EURO)		138	139
G57	X:541358..541586	-	+ (E3LU80_CAERE)		59	1820

G58	V:19477929..19478157	+ (RNAseq)	-	PF00097 (Zinc finger, C3HC4 type (RING finger))	72	243
G61	V:10709701..10709983	-	-		102	424
G63	IV:1278944..1279091	-	-	PS50883 (EAL domain profile.)	37	363
G64	I:14308479..14308887	-	-		59	354
G65	V:5768239..5768632	-	-		61	343
G67	V:3576900..3577317	-	+ (G0PEE4_CAEBE)	PF10318 (Serpentine type 7TM GPCR chemoreceptor Srh)	67	288
G68	II:6978237..6978540	-	-	PF07735 (F-box associated)	113	271
G69	X:14916816..14916993	-	+ (Q84LU4_GINBI)		22	271
G70	V:13060163..13060358	-	+ (G0P3Q5_CAEBE)		46	251
G71	X:11701412..11701730	-	+ (Q4R178_CAEEL)	PF07735 (F-box associated)	86	251
G75	V:4970301..4970493	-	+ (A8WQ28_CAEBR)	PF10319 (Serpentine type 7TM GPCR chemoreceptor Srj)	47	147
G76	II:11805687..11805840	-	+ (Q9XX13_CAEEL)		52	136
G78	V:6699013..6699454	-	-		108	134
G83	II:3733728..3734019	-	-		154	116
G85	III:669337..669397	-	+ (A0A0P5LGG1_9CRUS)	mobidb-lite (consensus disorder prediction)	91	112
G86	II:4548339..4548606	-	-		95	111
G89	X:7713027..7713081	-	-		51	108
G93	I:10067869..10067959	-	-		24	104
G95	II:3857180..3857426	-	-		92	101
G97	I:6983892..6983949	-	-		41	97
G98	II:9428645..9428735	-	-		34	97
G99	X:13784500..13784599	-	-		42	96

G100	X:7369398..7369593	-	-		76	95
G103	III:7619994..7620480	-	-		180	92
G106	III:7546511..7546649	-	-		70	91
G107	X:14567563..14567626	-	+ (H3D5Q5_TETNG)	BAH domain	24	91
G108	II:3997148..3997499	-	-		121	90
G109	X:16024200..16024737	-	-		237	90
G110	IV:2212713..2213040	-	-		106	89
G111	IV:1212666..1213077	-	-		133	88
G112	X:13380046..13380358	-	-		104	88
G114	X:14344402..14344645	-	-		84	86
G115	III:214502..214871	-	-		133	85
G116	IV:1287156..1287513	-	-		126	84
G117	IV:9866734..9867040	-	-		79	84
G121	IV:3654868..3655246	-	-		182	82
G122	X:743277..743580	-	-		118	82
G123	III:12582510..12582588	-	+ (A0A0L8G6E5_OCTBM)		30	81
G126	I:14315133..14315517	-	-		59	78
G127	III:9992028..9992310	-	-		90	78
G129	X:15255109..15255442	-	-		127	78
G130	X:9603752..9604172	-	-		143	78
G131	II:12979623..12979845	-	-		39	77
G133	II:6310391..6310769	-	-		128	76
G134	III:11729720..11729903	-	-		76	76
G135	III:1446243..1446597	-	+ (A8Y423_CAEBR)	mobidb-lite (consensus disorder	66	76

				prediction) / mobidb-lite (consensus disorder prediction)		
G136	I:13395200..13395593	-	-		88	75
G137	V:10032341..10032656	-	-		79	75
G138	V:8678462..8678762	-	-		69	75
G139	X:9646607..9646910	-	-		121	75
G140	I:11500352..11500649	-	-		113	74
G141	I:12612423..12612753	-	-		135	74
G142	I:4819997..4820081	-	-		17	74
G144	II:1696730..1696958	-	-		89	74
G145	V:12890577..12890904	-	-		110	73
G146	X:1120485..1120578	-	-		88	73
G147	V:6564739..6564919	-	-		51	72
G148	I:12541606..12541939	-	-		87	71
G149	V:3831077..3831377	-	-		113	71
G152	X:10433971..10434304	-	-	mobidb-lite (consensus disorder prediction)	119	71
G154	X:854404..854797	-	-	mobidb-lite (consensus disorder prediction) / mobidb-lite (consensus disorder prediction)	179	71
G155	V:11699793..11700117	-	-		97	70
G156	X:14348630..14348903	-	-		137	70
G157	X:5994407..5994674	-	-		87	70

11.2.2 Posibles exones

id	position	Expression	Homologs	motifs	ORF_length	Knockdown phenotype of the adjacent gene	protomotifs
E5	X:3495047..3495548	+ (RNAseq)			127	Ninguno	83
E6	X:6799982..6800363	+ (RNAseq)	+ (E3LDB8_CAERE)		176	Ninguno	79
E7	II:13888312..13888573	+ (RNAseq)			92	Retención larvaria y letalidad	73
E13	V:13768591..13768795				64	Ninguno	121
E14	X:10674619..10674925				98	Retención larvaria y letalidad	80
E18	V:7882127..7882430				114	Ninguno	79
E23	X:6001927..6002254				149	Ninguno	74
E28	V:5856625..5856859				84	Ninguno	120
E29	V:3445206..3445602				155	Anormal desarrollo del linaje neural	80
E30	X:3170746..3171163				186	Ninguno	70