Desarrollo de herramientas bioinformáticas aplicadas a proteómica shotgun y proteómica dirigida

TESIS DOCTORAL

Vital Vialás Fernandez

Departamento de Microbiología II
Facultad de Farmacia
Universidad Complutense de Madrid

Desarrollo de herramientas bioinformáticas aplicadas a proteómica shotgun y proteómica dirigida

Memoria que presenta para optar al título de Doctor

Vital Vialás Fernandez

Dirigida por la Doctora

Concha Gil García

Departamento de Microbiología II
Facultad de Farmacia
Universidad Complutense de Madrid

Resumen

...

...

...

Índice

Resumen v

I Introducción 1

Introducción 3
Proteómica. Conceptos generales 3
Espectrometría de masas 4
Espectrometría de masas en Tandem. MS/MS 6
Digestión de proteínas en péptidos 7
Proteómica en gel 8
Huella Peptídica 9
Proteómica *shotgun* 9
Separación multidimensional de péptidos 9
Asignación Péptido-Espectro 10
Búsqueda basada en secuencia 10
Búsqueda basada en bibliotecas de espectros 10
Identificación por secuenciación *de novo* 10
Inferencia de proteínas a partir de péptidos 10
Evaluación estadística de los resultados 10
Proteomica dirigida. SRM/MRM 10
Candida albicans como organismo modelo 10
Repositorios publicos de proteómica shotgun y dirigida 10

II Desarrollo de una aplicacion web para datos de proteómica shotgun de *Candida albicans* 13

1. Proteopathogen, a protein database for studying *Candida albicans* - host interaction 15

2. Proteopathogen 2, adaptación al formato estándar de iden-

tificaciones .mzIdentML **23**
2.1. 23

III Creación de un PeptideAtlas de *Candida albicans* 25

3. A *Candida albicans* PeptideAtlas **27**
3.1. 35

IV Desarrollo de una base de datos para datos de Proteómica Dirigida (MRM) **37**

Bibliografía **39**

Índice de figuras

1. La complejidad del proteoma aumenta exponencialmente y bla 11

Índice de Tablas

Introducción

Introducción

Nada en Biología tiene sentido si no es bajo la luz de la Evolución

Theodosius Dobzhansky

Tradicionalmente, el gen se ha concebido como la unidad fundamental -el átomo- de la vida, sometida a la acción de la selección natural. Así definió Richard Dawkins en el Gen Egoísta al gen, la unidad indivisible auto-replicante, mientras que los individuos y sus conductas eran meras *máquinas de supervivencia*. Sin embargo, es el fenotipo y no el genotipo lo que interactúa con el ambiente y con otros organismos. Las proteínas, los *ladrillos* con que se construye la vida, sí son visibles, a diferencia de los genes, a la selección natural. Por otra parte, el clásico dogma central de la Biología Molecular caducó hace ya tiempo y hoy lo recordamos, más bien, como una sobresimplificación. Actualmente, el emergente campo de la Proteogenómica da cuenta de la intrincada red de procesos regulatorios de transferencia de información entre el gen y la proteína. La Proteómica por su parte, le debe a la Genómica el reconocimiento y agradecimiento de haber abierto camino en la Biotecnología moderna. La Bioinformática, en este panorama, tiene un papel integrador. Al igual que la Proteómica, se sirve de diferentes tecnologías que avanzan y se retroalimentan sinérgicamente. Así la Proteómica se beneficia de los avances en Espectrometría de Masas, y estos instrumentos progresan en función de la demanda en investigación. De la misma manera, la Proteómica Computacional, la parte de la Bioinformática mas cercana a la Proteómica, evoluciona para facilitar el análisis de los datos que los investigadores requieren, pero también se beneficia de la incesante, creciente capacidad de procesamiento en las computadoras actuales.

Proteómica. Conceptos generales

El concepto de Proteoma, fue acuñado originalmente por Marc Wilkins en 1994 en analogía al concepto de Genoma. Si el Genoma es la dotación génica de una célula u organismo, el Proteoma es entendido como *el conjunto*

total de proteínas expresadas por los genes de una célula, tejido u organismo. Sin embargo, mientras que el Genoma permanece constante en todas las células del organismo, el Proteoma es un concepto mas variable. Los genes se expresan en función de las condiciones en que se encuentra la célula, según el orgánulo, el tejido, y estadío del desarrollo entre otros factores. Además existen niveles de complejidad adicional en el curso de información desde el gen a la proteína como el *splicing* alternativo y las modificaciones post-traduccionales. Por eso el término Proteoma puede diversificarse, para ajustarse a definiciones mas específicas. Así, podemos hablar del proteoma (o fosfo-proteoma) de un orgánulo celular, como la mitocondria, en un tejido concreto, en unas condiciones ambientales definidas por los nutrientes disponibles, posiblemente sometida a condiciones de estrés, etc...

Proteómica es, por tanto, el estudio del proteoma, independientemente del conjunto o subconjunto de proteínas objeto de estudio. Pero además Proteómica se refiere a las tecnologías utilizadas para ello.

El establecimiento de la espectrometría de masas aplicada a moléculas biológicas a finales de los años 80 y el desarrollo de técnicas de separación de péptidos y proteínas como la electroforesis PAGE y la cromatografía líquida permitieron que la Proteómica se consolidara y extendiera como disciplina científica.

La Figura ilustra como el grado de complejidad biológica desde la unidad de información, es decir, el gen, hasta la unidad funcional, la proteína, aumenta exponencialmente.

Espectrometría de masas

El desarrollo de las técnicas de ionización *suave* de macromoléculas biológicas a finales de los años 80, además de valer el Nobel a John Fenn y Koichi Tanaka, permitió sentar las bases de la Espectrometría de Masas aplicada a la Proteómica. Las técnicas de ionización ESI (Ionización por ElectroSpray) (Fenn et al., 1989) y SLD (Desorción Suave por Láser) (Tanaka et al., 1988) permitieron que las grandes y frágiles moléculas biológicas como las proteínas pudieran ser ionizadas y volatilizadas para ser posteriormente introducidas en los espectrómetros de masas.

Como ocurre en muchas otras ocasiones en la ciencia, de forma paralela e independientemente habían surgido en distintas partes del mundo ideas muy similares. Así, el desarrollo de SLD que valió el Nobel a K. Tanaka, tuvo un precedente unos años antes. Franz Hillenkamp y Michael Karas en Frankfurt, Alemania (éstos discutiblemente no galardonados) habían ideado una técnica similar que, en este caso, denominaron MALDI (Desorción/Ionización Láser Asistida por Matriz) (Karas y Hillenkamp, 1988) Aunque MALDI no fue aplicado a la ionización de proteínas hasta la publicación del trabajo de Tanaka, actualmente éste es el acrónimo que se ha impuesto y es, de hecho,

una técnica muy extendida en laboratorios de espectrometría de masas.

Un espectrómetro de masas es, en esencia, una balanza de precisión molecular capaz de medir, hasta un determinado límite de sensiblidad, la masa (en relación a la carga) de moléculas (ionizadas). Consta básicamente de tres partes o secciones:

1. **Fuente de iones**: Las macromoléculas biológicas, como proteínas y péptidos, no son volátiles El desarrollo de las técnicas de ionización *suave* permitió que péptidos y proteínas ionizados relativamente intactos

    - En **ESI**, el analito se encuentra en fase líquida en un solvente orgánico volátil como metanol o acetonitrilo. Esta solución es conducida a través de un capilar sometido a un campo eléctrico de forma que las micro-gotas en el ápice del capilar, una vez que la carga supera un límite, adquieren una forma cónica y forman un aerosol. Se produce entonces la desolvatación por evaporación del solvente. Así, las micro-gotas del aerosol disminuyen su tamaño, reagrupándose en gotas más estables y pequeñas en un proceso reiterativo, hasta el punto en que las moléculas de analito se repelen con la fuerza suficiente para superar la tensión superficial y liberarse del solvente (explosión de Coulomb) quedando iones de analito en suspensión que son introducidos en un sistema vacío hacia el espectrómetro.
    - **MALDI** consiste en embeber la muestra en una matriz líquida con alta capacidad de absorber luz UV sobre la que inciden pulsos de luz láser, lo que permite que las moléculas sean ionizadas y volatilizadas.

2. **Analizador de masas** El analizador de masas es la parte del instrumento en la que los iones se separan. Esta separación se produce en base a la relación entre la masa y la carga (m/z) de los iones. Existen varios tipos de analizadores de masas, que pueden combinarse y es el elemento que se usa generalmente para definir el tipo de instrumento.

    - **Analizadores de sector**
    - **Analizadores por Tiempo de Vuelo**
    - **Cuadrupolos**
    - **Trampas Iónicas**

3. Detector

Dos parámetros importantes en un espectrómetro de masas son su *sensibilidad* y su *resolución* ya que determinan notablemente la cantidad y calidad de información del espectro generado, lo que a su vez, es esencial para identificar el péptido que origina el espectro.

La *sensibilidad* de un espectrómetro de masas es la capacidad para detectar masas muy pequeñas, y puede llegar a ser de hasta unas pocas partes por millon (ppm) en el caso de instrumentos de alta precisión como LTQ-Orbitrap, pero requiere un ajuste óptimo de múltiples parámetros como la calibración del instrumento, temperatura, etc...

La *resolución* es la capacidad para discernir señales que realmente corresponden a diferentes iones dentro de una ventana o margen de valores m/z Esto es esencial para evitar la co-fragmentación, es decir, obtener fragmentos de iones precursores diferentes con similares m/z.

Espectrometría de masas en Tandem. MS/MS

Los péptidos, separados en el espectrómetro de masas en base a su relación m/z generan señales cuyas intensidades son medidas en el detector produciendo un espectro. Por otra parte, para organismos que estén secuenciados, existen bases de datos de las secuencias de sus proteinas. La elección de la mejor proteína candidata (o péptido candidato) que ha generado ese espectro, como se explicará en detalle más adelante, consiste, en esencia, en medir el grado de similitud entre los valores de m/z empíricos obtenidos en el espectro y los valores de m/z que teóricamente se producen a partir de una digestion *in silico* de las secuencias presentes en la base de datos

En ocasiones, cuando la proteína original se encuentra relativamente aislada, el espectro que generan los péptidos que se detectan en el instrumento, es suficientemente específico de la proteína original y ésta puede ser identificada. Este es el principio de la técnica conocida como Huella Peptídica, descrita en detalle en una sección posterior. Sin embargo, esta técnica requiere que la proteína se encuentre aislada y el rendimiento que ofrece, por tanto, es limitado.

En la espectrometría de masas en tandem (MS/MS), los péptidos, una vez ionizados y dentro del espectrómetro, vuelven a ser sometido a una fragmentación adicional (en la cámara de colisión). Los péptidos se fragmentan, generando iones lo que hace que el patrón de fragmentación sea más específico de la secuencia original. Esto permite aumentar el poder de resolución, permitiendo distinguir péptidos con masas parecidas, y por tanto permite que en la muestra puedan co-existir mezclas proteicas mas complejas aumentando así el rendimiento del experimento.

Digestión de proteínas en péptidos

Tras la obtención de una muestra de proteinas, ya sea una mezcla compleja de éstas o una proteína mas o menos aislada y purificada, el primer paso consiste en someter a las proteinas a la acción de una enzima proteasa que corta en puntos específicos de la secuencia y que las digiere en un conjunto de péptidos. Pero, sabiendo que los espectrómetros de masas sí pueden medir masas de proteínas intactas,

¿por qué hacer una digestión que aumenta el grado de complejidad de la muestra y que supone el problema añadido de la inferencia de la proteína original a partir de sus péptidos constituyentes?

o dicho de otra manera

¿es necesario el paso intermedio de digestión en péptidos para luego inferir las proteínas originales?

La respuesta a estas preguntas tiene que ver, sobre todo, con limitaciones técnicas. Las proteínas intactas pueden ser difíciles de manipular, algunas, como las proteínas de membrana son insolubles en condiciones en que otras sí lo son. Muchos detergentes comunmente usados interfieren en MS ya que son fácilmente ionizables y se encuentran en gran cantidad en proporción a las proteinas. Además la sensibilidad de los espectrómetros es menor para proteinas intactas que para péptidos. La cantidad de posibles formas en que una proteína es procesada, incluyendo modificaciones post-traduccionales en sus péptidos, y variaciones conformacionales entre otras hace que la combinación de isoformas posibles y sus masas sean imposibles de discernir por MS. Por otra parte, para identificar a la proteína originaria se requiere información de la secuencia y para esto los espectrómetros son mas eficientes si se analizan secuencias de un tamaño de unos 20 aminoácidos aproximadamente.

A pesar de esto, los espectrómetros sí permiten inferir, al menos parcialmente, secuencias a partir de proteinas intactas y con ello identificarlas. Este es el objetivo de la llamada Proteómica *top-down*

La digestión consiste en la rotura de proteínas en péptidos por acción de una enzima proteolítica. Tradicionalmente se ha utilizado para esto tripsina, que rompe la secuencia aminoacídica a continuación, en el lado carboxilo-, de Arginina(R) o Lisina (K) a menos que exista una Prolina (P) adyacente. Los péptidos generados por acción de la tripsina, llamados péptidos *trípticos*, tienen un tamaño adecuado, dada la frecuencia media de R y K, para el análisis por espectrometría de masas lo que explica la popularidad de esta proteasa.

También es posible la utilización de otras proteasas siempre que se conozca su patrón de corte. Es de hecho una aproximación inevitable para aquellos casos en que la tripsina no sea útil, por ejemplo, debido a una baja frecuencia de R y K que no generen peptidos del tamaño adecuado.

Proteómica en gel

La separación de proteínas en geles de poli-acrilamida (PAGE, *Polyacrilamide Gel Electrophoresis*), es una técnica, o serie de técnicas con variantes, que consiste en separar proteinas presentes en una muestra inicial en base a sus propiedades fisico-químicas diferenciadoras como su carga, tamaño y/o su punto isoeléctrico. En función del número de estas propiedades que se aprovechan para separar, en mayor o menor grado, las proteinas de una muestra se distinguen básicamente dos tipos de PAGE

- Geles monodimensionales, 1D-PAGE. En este tipo de geles las proteinas se separan en función de su peso molecular. La electroforesis hace que las proteínas mas pequeñas, de menor peso molecular, se desplacen mas lejos en el gel, sometido a una diferencia de potencial.

- Geles bidimensionales. 2D-PAGE. En este caso, las proteínas se separan en una primera dimensión en función de su punto isoeléctrico. Las proteinas se desplazan sobre una tira con un gradiente de pH hasta situarse en un punto donde su carga neta se equilibra con la de su entorno. A continuación la tira se coloca en la cabecera de un gel y se aplica la segunda dimensión, de modo que se las proteínas se separan más, en este caso por peso molecular, al igual que en un gel 1D.

Otra clasificación posible de las técnicas PAGE puede establecerse en función de si se usan condiciones desnaturalizantes o no

- Geles desnaturalizantes. SDS-PAGE

- Condiciones nativas o no desnaturalizantes Blue Native

La proteómica en gel ha sido (y continúa siendo) una técnica muy empleada en laboratorios de todo el mundo. Tiene algunas limitaciones, como el hecho de que proteínas de bajo peso molecular no son fácilmente observables, o que el número de proteínas identificables a partir de un gel difícilmente pueda superar el millar. Sin embargo, este tipo de estudios sigue teniendo un nicho en la Proteómica actual (Rogowska-Wrzesinska et al., 2013) como en el caso de que el organismo de estudio no tenga su genoma secuenciado,

Most importantly, it allows the visualisation, identification and quantitation of intact proteins. The particular ability to separate proteins with small changes in their pI allows for efficient separation and subsequent identification of protein isoforms and proteins modified by for example glycosylation or proteolytic cleavage. Combined with dedicated visualisation methods, 2D gels support detection of modified protein for which no efficient enrichment method currently exists.

Huella Peptídica

La Huella Peptídica de una proteína se refiere al hecho de que el patrón de fragmentación de una proteína en los péptidos que la constituyen utilizando una enzima proteolítica determinada, es muy específico de la proteína original (siempre y cuando se conozca el patrón de corte de la enzima, como es el caso de la tripsina) y el espectro que generan puede ser utilizado para identificarla. Sin embargo, a pesar de esta especificidad, la altísima variedad de proteinas implica una mayor aun variedad de posibles péptidos generados a partir de ellas que pueden tener masas muy similares. Por ese motivo, esta técnica requiere que la proteína se encuentre previamente aislada, generalmente a partir de una mancha o *spot* proteico de 2D-PAGE

Generalmente la técnica de la Huella Peptídica se lleva a cabo por espectrometría MALDI-TOF(TOF). Esto significa que, una vez obtenidos los péptidos correspondientes a la proteina del *spot*, éstos se situan en una matriz MALDI, donde son ionizados e introducidos en un analizador de Tiempo de Vuelo (TOF).

Una vez obtenido el espectro patrón de masas peptídicas, la identificación del péptido se realiza

Proteómica *shotgun*

La proteómica *shotgun* es la técnica de elección para la mayoría de estudios proteómicos a gran escala. El nombre *shotgun* proviene de una analogía con las técnicas clásicas de secuenciación genómica donde el ADN es fragmentado en secuencias más pequeñas que posteriormente son ensambladas. En la proteómica *shotgun* las proteínas son fragmentadas en péptidos a partir de los cuales se infiere finalmente la proteína original. Implica varios pasos descritos a continuación.

Separación multidimensional de péptidos

A diferencia de la técnica de la Huella Peptídica donde cada proteína se encuentra (en mayor o menor grado) aislada, en la Proteómica de alto rendimiento o *shotgun*, dado que el objetivo es detectar el máximo número de proteínas en un solo experimento, se parte de una mezcla más compleja. Este hecho es importante porque, si contamos con que a partir de cada proteína presente en la muestra, se generan múltiples péptidos trípticos, el grado de complejidad aumenta enormemente tras la digestión. Por este motivo, para evitar que la mezcla de péptidos sea demasiado compleja (para la resolución en el análisis MS), opcionalmente se puede realizar una separación previa de las proteinas que reduzca el grado de complejidad. Esta separación puede realizarse de varias maneras, por ejemplo la separación puede compenzar a

nivel de proteína por electroforesis en un gel 1D-PAGE, por fraccionamientos sub-celulares correspondientes a distintos orgánulos, etc...

Pero además de esta separación a nivel de proteína se hace también un fraccionamiento a nivel de péptido en este caso, generalmente, por cromatografía líquida. Al igual que en las técnicas basadas en gel, en cromatografía se habla de dimensiones de separación en función de las propiedades fisico-químicas que se aprovechan para el fraccionamiento. Además frecuentemente son combinables y pueden emplearse secuencialmente. Algunos de los tipos de LC más utilizados se describen a continuación.

- Cromatografía líquida en fase reversa (RP-LC)
- Cromatografía de intercambio catiónico

Asignación Péptido-Espectro

Búsqueda basada en secuencia

Búsqueda basada en bibliotecas de espectros

Identificación por secuenciación *de novo*

Inferencia de proteínas a partir de péptidos

Evaluación estadística de los resultados

Proteomica dirigida. SRM/MRM

Candida albicans como organismo modelo

Repositorios publicos de proteómica shotgun y dirigida

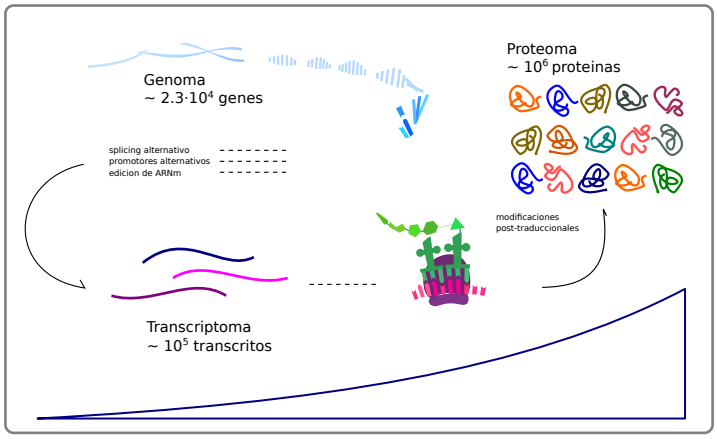


Figura 1: La complejidad del proteoma aumenta exponencialmente y bla

Desarrollo de una aplicacion web para datos de proteómica shotgun de *Candida albicans*

Capítulo 1

Proteopathogen, a protein database for studying *Candida albicans* - host interaction

4664 DOI 10.1002/pmic.200900023 *Proteomics* 2009, *9*, 4664–4668

TECHNICAL BRIEF

Proteopathogen, a protein database for studying *Candida albicans* – host interaction

Vital Vialás[1], Rubén Nogales-Cadenas[2], César Nombela[1], Alberto Pascual-Montano[2] and Concha Gil[1,3]

[1] Departamento de Microbiología II, Facultad de Farmacia, Universidad Complutense de Madrid, Madrid, Spain
[2] Departamento de Arquitectura de Computadores y Automática, Facultad de Ciencias Físicas, Universidad Complutense de Madrid, Madrid, Spain
[3] Unidad de Proteómica UCM-Parque Científico de Madrid, Facultad de Farmacia, Universidad, Complutense de Madrid, Madrid, Spain

Received: January 13, 2009
Revised: June 25, 2009
Accepted: July 2, 2009

There exist, at present, public web repositories for management and storage of proteomic data and also fungi-specific databases. None of them, however, is focused to the specific research area of fungal pathogens and their interactions with the host, and contains proteomics experimental data. In this context, we present Proteopathogen, a database intended to compile proteomics experimental data and to facilitate storage and access to a range of data which spans proteomics workflows from description of the experimental approaches leading to sample preparation to MS settings and peptides supporting protein identification. Proteopathogen is currently focused on *Candida albicans* and its interaction with macrophages; however, data from experiments concerning different pathogenic fungi species and other mammalian cells may also be found suitable for inclusion into the database. Proteopathogen is publicly available at http://proteopathogen.dacya.ucm.es

Keywords:
Candida albicans / Database / Host / MS / Microbiology / Pathogen

Candida albicans is an opportunistic pathogenic fungus, which can be found as a component of the usual flora in human mucoses. Although it does not normally cause disease in immunocompetent colonized hosts, in the case of immunosuppressed patients *Candida* cells can overproliferate and become pathogenic. Cells in yeast form (oval cells) may produce hyphae, penetrate tissues and eventually cause invasive candidiasis. At present, the frequency of this fatal opportunistic mycosis continues to be distressing and, unfortunately, solution is hindered by the reduced effectiveness and serious side effects of the few available drugs, the appearance of antifungal-drug resistance, and the lack of accurate and prompt diagnostic procedures [1].

Addressing proteomic studies involving the way *Candida* interacts with immune cells is thus essential in order to improve our comprehension of the process of infection and represents the primary step of investigation that could lead to future development of diagnosis methods, vaccines and antifungal drugs [2–5].

Experimental techniques in proteomics have quickly evolved in such a way that nowadays we have to deal with vast amounts of complex data originated by the combination of multi-dimensional separation techniques and MS analysis together with the bioinformatics software reports [6]. Existing public repositories for management and storage of proteomic data such as World 2-D PAGE [7], the Proteome Database System for Microbial Research 2-D PAGE [8], or PRIDE [9]; and fungi-specific databases such as BioBase MycoPathPD [10], Candida Genome Database (CGD) [11] or Candida DB [12] are very popular and useful tools. However, none of them deals with proteomic experimental

Correspondence: Dr. Concha Gil, Departamento de Microbiología II, Facultad, de Farmacia, Universidad Complutense, Plaza de Ramón y Cajal s/n, 28040 Madrid, Spain
E-mail: conchagil@farm.ucm.es
Fax:+34-913941745

Abbreviations: CGD, Candida Genome Database; **GO,** *Gene Ontology*; **KEGG,** Kyoto Encyclopedia of Genes and Genomes; **PDB,** Protein Data Bank

data related to the specific research area of fungal pathogens and their interaction with the host. In this context, we present Proteopathogen, a protein database, currently focused on the *C. albicans* – macrophage interaction model – which enables a framework for the access and submission of proteomic workflow data, from description of the experimental approaches leading to sample preparation to MS settings and identification-supporting peptides. Through its interface web site, the database can easily be queried to allow an efficient browsing through all the stored data, improving the quality of eventual analysis of MS results.

Regarding the compilation of information used to populate the database, data from three different studies were considered suitable to be present in Proteopathogen. The first two correspond to publish works relating to proteomics of the *Candida* – macrophage interaction [2, 3], where the former reports 66 different *C. albicans* identified proteins and the latter, 38 murine macrophage proteins. The third study represents an analysis of the *C. albicans* plasma membrane proteome [13]. It compiles a set of experiments aimed at extraction and identification of membrane proteins and a set of experiments intended to obtain enrichment in glycosylphosphatidylinositol-anchored surface proteins, which have been reported to be involved in cell wall biogenesis, cell–cell adhesion and interaction with the host [14].

In all cases, protein identifications lists are collected together with the pertinent experimental context specified by descriptions of the experimental approaches, MS settings and peptides supporting identification for each of the proteins (Table 1).

Along with the experimental information, and in order to provide a deeper view of the data, complementary information is retrieved from public web repositories. In the case of *C. albicans* proteins, identifiers, synonyms, aminoacid sequence of the translated open reading frame, *Saccharomyces cerevisiae* orthologs, *Gene Ontology* (GO) annotation, pathway annotations and scientific literature references were obtained from CGD [11], whereas in the case of murine macrophage proteins, the equivalent information was obtained from UniProt KnowledgeBase [15] and the Mouse Genome Database [16]. Additionally, pathways annotations were retrieved from Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Database [17] and structure information from the Protein Data Bank (PDB) [18].

Concerning the architecture of the software, the back-end layer consists of a MySQL database managed by the web application development framework Ruby on Rails that sets up structure and relations of data, handles queries to the database and displays the user web-based interface.

The experimental context is addressed in Proteopathogen in a hierarchical manner, where a main general approach, which may correspond to a published article, is characterized by a description or title, authors, target species and Pubmed identifier when available; and experiments within it, are in they turn, characterized by the description of the particular experiment, the date when it was performed and number of identified proteins.

Information on one particular protein is split into several sections in Proteopathogen. *Protein Basic Information* displays the UniProt accession number, description, species, evidence for the existence, standard gene name, organism-specific database identifiers, yeast orthologs for *Candida* proteins and human orthologs for mouse proteins and sequence. The Section 2 lists experiments in which the particular protein has been identified. Where available, one or more of the following sections will be displayed as well: the table entitled GO showing GO annotations along with the pertinent scientific references, the *KEGG Pathways* and *CGD Pathways* tables rendering annotations from KEGG and CGD respectively, and *PDB*, a table specifying structural information. Where no PDB identifiers are found for *C. albicans* proteins, *S. cerevisiae* orthologs are used instead, and similarly, when a PDB identifier cannot be found for mouse proteins, the human ortholog is used.

In all cases, proteins are unambiguously related to their corresponding experiment, thus enabling a relation to the data concerning experimental parameters of identification and identification-supporting peptides. This data comprise, on the one hand, common MS settings for all proteins identified in the particular experiment, including search database, MS type, analysis software, digestion enzyme, fixed aminoacid modifications, variable modifications and maximum allowed number of miscleavages; and on the other hand, particular parameters and peptides list for each protein, including number of matched peptides, score,

Table 1. Overview of the stored data in Proteopathogen as well as their published evidences

References	Description of experimental approach	Species	♯Protein identifications
[2]	*C. albicans* differentially expressed proteins after 3 h interaction with RAW 264.7 murine macrophages. 2-D silver-stained gel. MS/MS (MALDI/TOF-TOF)	*C. albicans*	66
[3]	Proteins identified from cytoplasmic extracts of RAW 264.7 cells after 45 min interaction with *C. albicans*	*Mus. musculus*	38
[13]	Identification of Glycosyl phosphatidil inositol (GPI)-anchored membrane proteins Identification of membrane proteins	*C. albicans*	292 1273

observed peptide mass, calculated peptide mass, start and end coordinates, number of missed cleavages and the sequence of the peptide.

The web interface to Proteopathogen offers multiple ways to query the database. Through the *Browse Experiments* search option, a list containing all sets of experimental approaches is displayed. In its turn, one particular experiment can be browsed through all the proteins identified in it.

The *Search* form may be used in different manners. Queries for one particular protein can be performed by supplying one of the multiple supported identifiers, namely standard gene names, Candida feature name, Candida DB identifiers, CGD identifiers, MGI identifiers and UniProt accession numbers. Free text queries can be performed as well, which will retrieve a list of proteins showing coincidences in the description field of the Proteopathogen protein entry. As an additional feature, peptide sequences can also be searched for retrieving in this case, proteins in any experiment having the searched sequence in any of the identification-supporting peptides. Wild characters (''*'') and Boolean operators are supported for free text queries and for peptide sequence queries.

In order to enhance interactivity and collaboration with users, a submission form is included in the web interface to allow the upload of more proteomic experimental approaches as long as they concern the topics addressed in Proteopathogen. Sequential steps request from the user the following information: a description of the experimental context, a related protein list, MS parameters

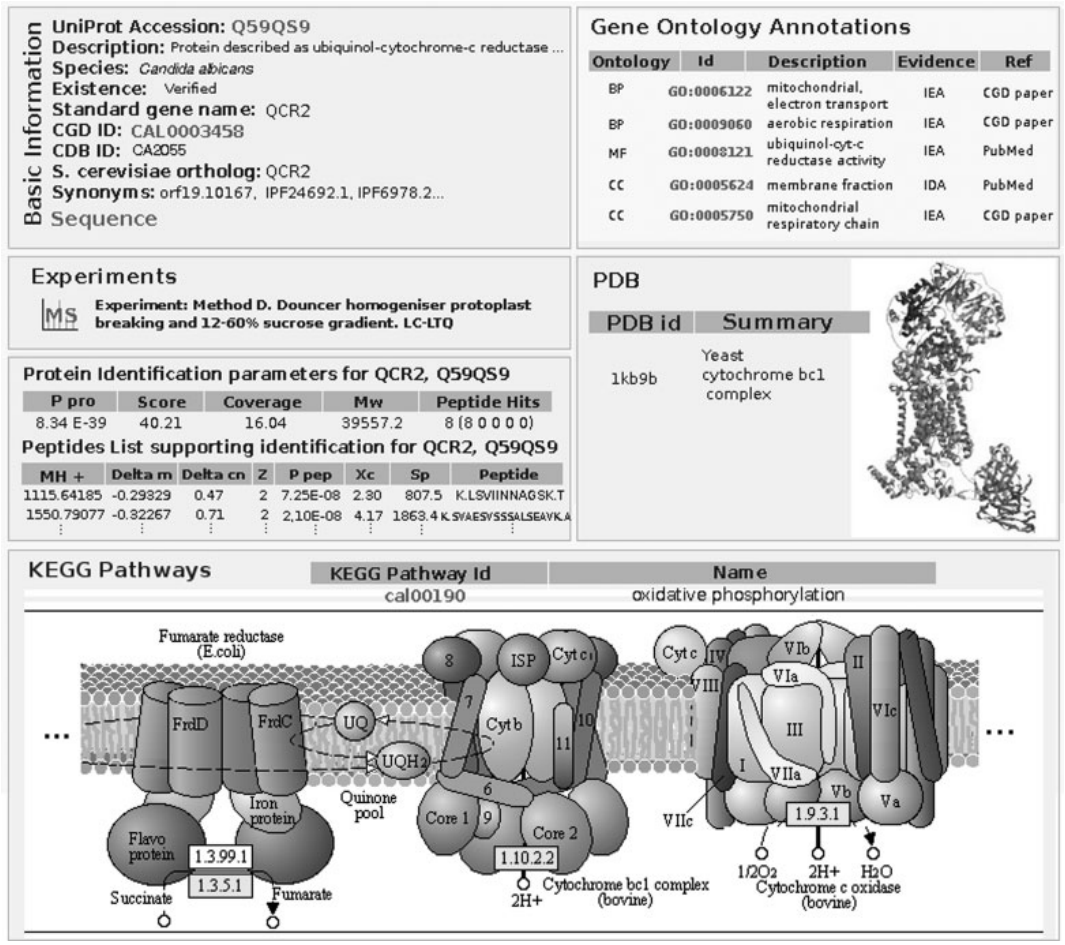


Figure 1. Use case: Search for *C. albicans* ubiquinol-cytochrome-*c* reductase QCR2. The different sections in the result comprise information on protein description and identifiers, experiments in which it has been identified, GO annotation, KEGG and CGD pathway annotation and structural information from PDB.

and identification-supporting peptides lists. These data are subject to revision prior to eventual insertion into Proteopathogen by the database curators. Besides, the whole relational database and the MS data reports are available for download at the web site.

All the information that is retrievable from Proteopathogen when queried for one particular protein is shown in Fig. 1 for the specific case of ubiquinol-cytochrome-*c* reductase QCR2 of *C. albicans* which has been reported to show antigenic properties in human [19].

The *Protein Basic Information* section displays the Uniprot accession number, a brief description of the protein as stated at CGD, evidence for its existence, standard gene name, feature name, CGD and Candida Database identifiers, yeast ortholog gene name, synonyms and sequence.

The Section 2 lists all the experiments in which QCR2 has been identified. All of them belong to the same general approach aimed at purification of membrane proteins. In every case, the corresponding links to the MS identification parameters and supporting peptides are displayed as well. This experimental data are shown in Fig. 1 for identification of QCR2 in the experiment described as ''Method D. Douncer homogenizer protoplast breaking and 12–60% sucrose gradient. LC-LTQ''.

The section entitled *GO annotations* shows terms related to the electron transport chain, but more interestingly, it also shows an *inferred from direct assay* (IDA) annotation to the term *membrane fraction* [20], which fits to the fact that the protein is identified in five of the methods aimed at purification of membrane proteins.

KEGG Pathways table provides a link to the KEGG Pathway entry for *Oxidative phosphorylation*, and provides the feature to show in place the image corresponding to the map from KEGG. *CGD Pathways* displays an analogous link to the pathway entry at CGD that, in this case, is named *aerobic respiration (cyanide sensitive)–electron donors*.

Finally, in the *PDB* section, there are four structure images available along with links to the PDB entries, corresponding to a cytochrome bc1 complex from *S. cerevisiae*. Orthologs were used since no structure could be found for the *Candida* protein.

In conclusion, Proteopathogen represents, up to date, the first public web-based repository for proteomics data related to studies involving *C. albicans* pathogenicity and its interaction with immune system cells in the host. Moreover, it enables a framework for public access and submission of this type of data and it is intended to be more actively populated in the near future, including data from different pathogenic fungi and mammalian cells, becoming a reference database in its field. Unlike other protein identification databases, Proteopathogen is focused to a specific topic but, at the same time, includes a wide range of data including descriptions of the experimental contexts, information on proteins such as GO and pathway annotations, structural information and detailed MS parameters. Therefore, Proteopathogen will contribute to save time and facilitate analysis of proteomic workflow reports for researchers interested in this area.

The authors are grateful to César Vicente from the Computer Architecture Department, Complutense University of Madrid for his excellent technical assistance. This work was supported by BIO 01989-2006 from the Comision Interministerial de Ciencia y Tecnología (CYCIT, Spain), DEREMICROBIANA – CM from Comunidad Autónoma de Madrid, and REIPI, Spanish Network for the Research in Infectious Diseases, RD06/0008/1027 from the Instituto de Salud Carlos III. The Proteomics work was carried out in the Proteomics Unit UCM-Parque Científico, a member of the National Institute for Proteomics PROTEORED, funded by Genoma España. APM and RNC are partially supported by Spanish grants BIO2007-67150-C03-02, S-Gen-0166/2006 and PS-010000-2008-1.

The authors have declared no conflict of interest.

References

[1] Calderone, R. A. (Ed.), *Candida and Candidiasis*, ASM Press, Washington D.C 2002.

[2] Fernández-Arenas, E., Cabezón, V., Bermejo, C., Arroyo, J., *et al.*, Integrated genomic and proteomic strategies bring new insight into *Candida albicans* response upon macrophage interaction. *Mol. Cell. Proteomics* 2007, *6*, 460–478.

[3] Martínez-Solano, L., Nombela, C., Molero, G., Gil, C., Differential protein expression of murine macrophages upon interaction with *Candida albicans*. *Proteomics* 2006, *6*, 133–144.

[4] Pitarch, A., Nombela, C., Gil, C., *Candida albicans* biology and pathogenesis: insights from proteomics. *Methods Biochem. Anal.* 2006a, *49*, 285–330.

[5] Pitarch, A., Nombela, C., Gil, C., Contributions of proteomics to diagnosis, treatment, and prevention of candidiasis. *Methods Biochem. Anal.* 2006b, *49*, 331–361.

[6] Monteoliva, L., Albar, J. P., Differential proteomics: an overview of gel and non-gel based approaches. *Brief Funct. Genomic Proteomics* 2004, *3*, 220–239.

[7] Hoogland, C., Mostaguir, K., Appel, R. D., Lisacek, F., The World-2DPAGE Constellation to promote and publish gel-based proteomics data through the ExPASy server. *J. Proteomics* 2008, *71*, 245–248.

[8] Pleissner, K. P., Eifert, T., Buettner, S., Schmidt, F. *et al.*, Web-accessible proteome databases for microbial research. *Proteomics* 2004, *4*, 1305–1313.

[9] Martens, L., Hermjakob, H., Jones, P., Adamski, M. *et al.*, PRIDE: the proteomics identifications database. *Proteomics* 2005, *5*, 3537–3545.

[10] Csank, C., Costanzo, M. C., Hirschman, J., Hodges, P. *et al.*, Three yeast proteome databases: YPD, PombePD, and CalPD (MycoPathPD). *Methods Enzymol,* 2002, *350*, 347–373.

[11] Arnaud, M. B, Costanzo, M. C., Skrzypek, M. S., Binkley, G. *et al.*, The Candida Genome Database (CGD), a community resource for *Candida albicans* gene and protein information. *Nucleic Acids Res.* 2005, *33*, 358–363.

[12] Rossignol, T., Lechat, P., Cuomo, C., Zeng, Q. *et al.*, CandidaDB: a multi-genome database for Candida species and related Saccharomycotina. *Nucleic Acids Res.* 2008, *36*, 557–561.

[13] Cabezón, V., Llama-Palacios, A., Nombela, C., Monteoliva, L., Gil, C., Analysis of *Candida albicans* plasma membrane proteome. *Proteomics* 2009, *9*, in press, DOI: 10.1002/pmic.200800988.

[14] Plaine, A., Walker, L., Da Costa, G., Mora-Montes, M. *et al.*, Functional analysis of *Candida albicans* GPI-anchored proteins: roles in cell wall integrity and caspofungin sensitivity. *Fungal Genet. Biol.* 2008, *45*, 1404–1414.

[15] UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* 2008, *36*, 190–195.

[16] Bult, C. J., Eppig, J. T., Kadin, J. A., Richardson, J. E., *et al.*, The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res.* 2008, *36*, 724–728.

[17] Kanehisa, M., Araki, M., Goto, S., Hattori, M. *et al.*, KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 2008, *36*, 480–484.

[18] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G. *et al.*, The Protein Data Bank. *Nucleic Acids Res.* 2000, *28*, 235–242.

[19] Pitarch, A., Abian, J., Carrascal, M., Sanchez, M. *et al.*, Proteomics-based identification of novel *Candida albicans* antigens for diagnosis of systemic candidiasis in patients with underlying hematological malignancies. *Proteomics* 2004, *4*, 550–559.

[20] Insenser, M., Nombela, C., Molero, G., Gil, C., Proteomic analysis of detergent-resistant membranes from *Candida albicans*. *Proteomics* 2006, *6*, S74–S81.

...
...

Capítulo 2

Proteopathogen 2, adaptación al formato estándar de identificaciones .mzIdentML

2.1.

...
...

Creación de un PeptideAtlas de *Candida albicans*

Capítulo 3

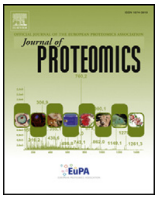
A *Candida albicans* PeptideAtlas

JOURNAL OF PROTEOMICS 97 (2014) 62–68

Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot



A *Candida albicans* PeptideAtlas☆

Vital Vialas[a,b,*], *Zhi Sun*[c], *Carla Verónica Loureiro y Penha*[a,b], *Montserrat Carrascal*[d], *Joaquín Abián*[d], *Lucía Monteoliva*[a,b], *Eric W. Deutsch*[c], *Ruedi Aebersold*[e,f], *Robert L. Moritz*[c], *Concha Gil*[a,b,*]

[a]*Dept. Microbiología II, Universidad Complutense de Madrid, Madrid, Spain*
[b]*IRYCIS: Instituto Ramón y Cajal de Investigación Sanitaria, Madrid, Spain*
[c]*Institute for Systems Biology, Seattle, WA, USA*
[d]*CSIC/UAB Proteomics Laboratory, Instituto de Investigaciones Biomédicas de Barcelona—Consejo Superior de Investigaciones Científicas, Spain*
[e]*Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland*
[f]*Faculty of Science, University of Zurich, Zurich, Switzerland*

ARTICLE INFO

Available online 26 June 2013

Keywords:
Candida albicans
PeptideAtlas
Proteotypic peptides

ABSTRACT

Candida albicans public proteomic datasets, though growing steadily in the last few years, still have a very limited presence in online repositories. We report here the creation of a *C. albicans* PeptideAtlas comprising near 22,000 distinct peptides at a 0.24% False Discovery Rate (FDR) that account for over 2500 canonical proteins at a 1.2% FDR. Based on data from 16 experiments, we attained coverage of 41% of the *C. albicans* open reading frame sequences (ORFs) in the database used for the searches. This PeptideAtlas provides several useful features, including comprehensive protein and peptide-centered search capabilities and visualization tools that establish a solid basis for the study of basic biological mechanisms key to virulence and pathogenesis such as dimorphism, adherence, and apoptosis. Further, it is a valuable resource for the selection of candidate proteotypic peptides for targeted proteomic experiments *via* Selected Reaction Monitoring (SRM) or SWATH-MS.

Biological significance
This *C. albicans* PeptideAtlas resolves the previous absence of fungal pathogens in the PeptideAtlas project. It represents the most extensive characterization of the proteome of this fungus that exists up to the current date, including evidence for *uncharacterized* ORFs. Through its web interface, PeptideAtlas supports the study of interesting proteins related to basic biological mechanisms key to virulence such as apoptosis, dimorphism and adherence. It also provides a valuable resource to select candidate proteotypic peptides for future (SRM) targeted proteomic experiments.
This article is part of a Special Issue entitled: Trends in Microbial Proteomics.

Abbreviations: SRM, Selected Reaction Monitoring; CGD, *Candida* Genome Database; FDR, False Discovery Rate; PSM, Peptide–Spectrum Match; PRIDE, Protein Identifications Database; PSS, Predicted Suitability Score; ESS, Empirical Suitability Score

☆ This article is part of a Special Issue entitled: Trends in Microbial Proteomics.

* *Corresponding authors at:* Departamento de Microbiología II, Facultad de Farmacia, Universidad Complutense de Madrid, Plaza Ramón y Cajal s/n. 28040 Madrid, Spain. Tel.: +34 91 394 17 55; fax: +34 91 394 17 45.
E-mail addresses: vvialasf@ucm.es (V. Vialas), conchagil@ucm.es (C. Gil).

1874-3919/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jprot.2013.06.020

1. Introduction

Candida albicans is a fungus of great clinical importance. In addition to asymptomatically colonizing mucous membranes as a commensal in a large percentage of the population, it may cause severe opportunistic infections in specific cases such as patients with weakened immune defenses, a common circumstance in cancer and AIDS patients. *C. albicans* infections are also a threat to patients in post-surgical situations and intensive care unit stays. In this respect, invasive candidiasis remains nowadays one of the major types of nosocomial infections and a challenge in terms of economical and health costs [1–3]. From the perspective of proteomics, recent studies have provided new insights into the *C. albicans* biology and suggested new clinical biomarker candidates for diagnosis and prognosis of invasive candidiasis [4–7].

However, the clinical relevance of this organism is not reflected in the number of large-scale publicly available proteomics resources. Up to the current date, the PRIDE [8] database includes only 15 experiments accounting for 1786 identified proteins. The more *C. albicans*-focused Proteopathogen database [9] comprises several hundred protein identifications including data from gel based proteomics, and other major proteomic online resources such as the Global Proteome Machine Database (GPMDB [10]) or Tranche [11] contain no *C. albicans* data whatsoever.

As for the genomic data, according to *Candida* Genome Database (CGD), currently the most comprehensively annotated *C. albicans* sequence repository [12], the *C. albicans* genome contains 6215 ORFs (as of May 28, 2013), out of which 1497 are annotated as *verified*, *i.e.* representing genes for which there is empirical evidence that the ORF actually encodes a functionally characterized protein. In contrast, 4566 ORFs are termed *uncharacterized*, indicating that there exists no conclusive evidence for the existence of a protein product. This data implies that most part of the predicted proteome, over 70% of the ORFs, is still unknown or has not been properly annotated yet. An extensive characterization of the *C. albicans* proteome will therefore be of great value to increase our knowledge in proteins involved in mechanisms of virulence and infection and, thus serves as a basis to design strategies for diagnosis, vaccination and treatment of invasive candidiasis.

Since its inception, the PeptideAtlas project [13] has encouraged mass spectrometry data submission by the community and has thus grown to a large compilation of atlases of different species including human tissue and body fluid specific builds (brain, plasma [14] and urine), microbial builds (*Halobacterium* [15], *Mycobacterium tuberculosis* [16], *Streptococcus* [17], *Leptospira*, *Plasmodium* [18], *Saccharomyces* [19] and *Schizosaccharomyces* [20]); invertebrate builds (*Caenorhabditis elegans*, *Drosophila* [21] and *Apis mellifera* [22]); and a pig and a bovine milk [23] builds. The PeptideAtlas project, as a multi-species compendium of proteomes, is continuously increasing its biological diversity. The recent *Schizosaccharomyces pombe* atlas [23] attains a large coverage of its proteome by *ad hoc* extensive fractionation and high-resolution LC–MS/MS, and contributes in the sense that some of the fission yeast biological processes have a high degree of conservation with the corresponding pathways in mammalian cells. The incorporation of *C. albicans* resolves the previous absence of fungal pathogens in the PeptideAtlas and their under representation in any public proteomic data repository.

Furthermore, the proven utility of PeptideAtlas as a resource for selecting proteotypic peptides for Selected Reaction Monitoring (SRM) [24] or SWATH-MS [25] will enable a starting point for future targeted proteomics workflows in *C. albicans*.

2. Material and methods

2.1. *Empirical data compilation*

Large amounts of mass spectrometry data corresponding to many and diverse measurements of the *C. albicans* proteome initially intended for different purposes were assembled in order to build the PeptideAtlas. A range of proteomic methods, protocols and different biological conditions were used to generate the data as shown in Table 1. These include membrane protein extractions [26], morphological yeast to hypha transition experiments [27] and phosphoprotein enrichment treatments. The combination of these diverse datasets resulted in an

Table 1 – List of experiments collected to construct the *C albicans* PeptideAtlas.

# experiment	Sample (as named in the web interface)	Labeling/treatment	Instrument type	# raw files
1	Calb_acidic_subproteome	–	LTQ	3
2	Calb_memb	–	LTQ	8
3	SILAC_phos_OrbitrapVelos_1	SILAC. IMAC + TiO2	Orbitrap Velos	3
4	SILAC_phos_OrbitrapVelos_2	SILAC. IMAC + TiO2	Orbitrap Velos	3
5	SILAC_phos_OrbitrapVelos_3	SILAC. IMAC + TiO2	Orbitrap Velos	3
6	SILAC_phos_OrbitrapVelos_4	SILAC. IMAC + TiO2	Orbitrap Velos	3
7	SILAC_phos_OrbitrapXL_1A	SILAC. IMAC	Orbitrap XL	11
8	SILAC_phos_OrbitrapXL_1A_TiO2	SILAC. IMAC + TiO2	Orbitrap XL	5
9	SILAC_phos_OrbitrapXL_1B	SILAC. IMAC	Orbitrap XL	6
10	SILAC_phos_OrbitrapXL_1B_TiO2	SILAC. IMAC + TiO2	Orbitrap XL	6
11	SILAC_phos_OrbitrapXL_2	SILAC. IMAC	Orbitrap XL	6
12	SILAC_phos_OrbitrapXL_3	SILAC. IMAC	Orbitrap XL	6
13	SILAC_phos_OrbitrapXL_4	SILAC. IMAC	Orbitrap XL	5
14	Calb_extract_3TOF	–	Triple TOF	2
15	Hyphal_extract_OrbitrapVelos	–	Orbitrap Velos	4
16	Yeast_extract_OrbitrapVelos	–	Orbitrap Velos	4

unprecedented overall coverage of the *C. albicans* proteome. Protein samples were obtained as previously described in [27]. Briefly, cells of the clinical isolate SC5314 were grown in YPD medium for standard growth, whereas hyphal form growth was induced using either Lee medium pH 6.7 or heat-inactivated fetal bovine serum. Protein extracts were then obtained by mechanical cell disruption using either glass beads in the MSK cell homogenizer or the Fast-Prep cell breaker. Protein digests were obtained by trypsinization and separated *via* HPLC. All spectra acquisition runs were performed by LC–MS/MS in a data-dependent manner in different instruments and setups. Table 1 provides an overview of the experiments along with the instruments used for the mass spectrometry and the corresponding number of raw spectra data files that were acquired.

In addition, raw MS data from unpublished, SILAC labeled and phosphoprotein enriched samples generated from studies focused on *Candida* interaction with host immune cells and from experiments studying the hyphal and yeast-form proteomes, were added to the collection.

2.2. Peptide and protein identification

PeptideAtlas ensures consistency and quality of the stored data by processing the raw spectra sets by the Trans-Proteomic Pipeline (TPP) [28], a suite of software tools for processing shotgun proteomic datasets. The TPP tools are run in a well-established sequential pipeline spanning steps from creating appropriate standard files to be used as input by the search engine to statistical validation of protein inference and calculation of the False Discovery Rate (FDR).

The collected raw spectra files in different proprietary file formats were converted to the standard format for mass spectrometry output data mzML [29], searched using X!Tandem [30] with the K-score algorithm plug-in [31] and the output search results were converted to the search engine-independent pepXML format [32].

The target fasta sequence file used for the search was obtained from the *Candida* Genome Database (CGD) [12] at: http://www.candidagenome.org/download/sequence/C_albicans_SC5314/Assembly21/.

Common contaminants from the common Repository of Adventitious Proteins (cRAP) were appended. Then for each of these sequences, counterpart reversed decoy sequences were appended.

PeptideProphet [33] was then run on the search results to model the distributions of correctly and incorrectly assigned Peptide-to-Spectrum Matches (PSMs). It then assigns probabilities of being correct for each PSM, yielding a sensitive and flexible approach to report results in a comparable manner. Next, iProphet [34] was used to combine additional sources of evidence including multiple identifications of the same peptide across spectra, experiments, and charge and modification states, allowing a more precise integration of evidence supporting the identification of each unique peptide sequence. ProteinProphet [35] was then run to refine iProphet probabilities by adding the information at the protein level, like the number of sibling peptides within a protein and to compute final protein level probabilities. The prophet tools together combine multiple layers of evidence and refine the model iteratively to achieve an optimal analysis of the data. Finally MAYU [36] estimated FDR at different levels for each contributing experiment and for the entire dataset based on the PSMs to decoy proteins.

This process followed the pipeline first implemented in the construction of the human plasma PeptideAtlas described in [14] and successfully applied to other builds such as the bovine milk and mammary gland PeptideAtlas [23].

2.3. Construction of the PeptideAtlas

The PeptideAtlas building process calculates the cumulative number of identified peptide and proteins across the experiments, gathers information on protein to genome location mappings and estimates the peptides' Empirical Suitability Score and Predicted Suitability Score (ESS and PSS). The genomic mappings, since *C. albicans* is not present in the Ensembl database, which is the default PeptideAtlas uses to that purpose, were extracted from a generic feature file located at the following url: http://www.candidagenome.org/download/gff/C_albicans_SC5314/C_albicans_SC5314_version_A21-s02-m05-r10_features.gff.

An overview of how the different experiments contribute, in terms of the number of identified spectra and peptides, to the atlas build is depicted in Fig. 1.

Besides, and due to the particularly rich number of identifications in experiments aimed at the detection of phosphorylated proteins (experiments #3 to #13), a similarly processed version of the PeptideAtlas was created including in this case PTMProphet results which provide, alongside each modified residue, the probability that the post-translational modification is truly detected at that site.

3. Results and discussion

3.1. Assessment of proteome coverage and functional enrichment analysis

The assembled proteomic datasets (Table 1) were subject to uniform data processing in order to build the *C. albicans*

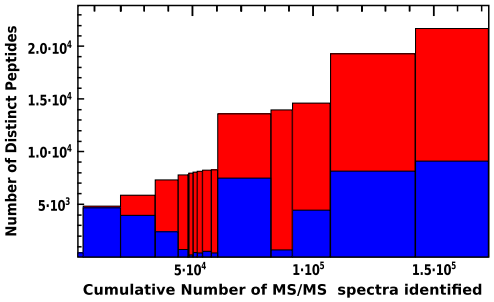


Fig. 1 – Histogram showing the cumulative number of distinct peptides in the *C. albicans* PeptideAtlas. Each bar represents a different experiment that has contributed to the build. Bar width is proportional to the number of high confidence PSMs. Height of the blue section of the bar represents the number of distinct peptides in each experiment and total height of the bar (red plus blue sections) indicates the cumulative number of peptides. The order of experiments is the same as in Table 1.

PeptideAtlas. The PSM assignment and protein inference processes were conducted by means of the consistent and robust pipeline TPP. The prophet tools integrate various levels of information and report identification results in statistical terms so that spectrum assignments, peptide to protein mappings and protein groups are statistically validated, leading to an overall improved sensitivity for a defined FDR level. As a result the generated *C. albicans* PeptideAtlas comprises 21,938 peptides identified at a 0.24% FDR allocated to 2562 proteins at a 1.2% FDR, that is, a coverage of 41.3% of the 6209 *C. albicans* translated ORF sequences from the fasta database used for searches. While the presented instance of the *C. albicans* PeptideAtlas has reached unprecedented coverage, it does not represent a final representation of the respective proteome. Like other PeptideAtlas instances for other species, the *C. albicans* atlas will be expanded upon submission and processing of new MS data generated in ongoing projects.

To determine the biological functions encompassed by the covered part of the proteome in this PeptideAtlas a Gene Ontology (GO) annotation enrichment analysis was carried out for the list of all detected *C. albicans* canonical proteins, excluding decoy hits, using the biological process ontology and Genecodis software [37]. Predictably, it generated a diverse array of clusters heterogeneously annotated, among which the largest in number of proteins are associated with the GO terms *oxidation–reduction process*, *cellular response to drug*, *pathogenesis* and *hyphal growth* respectively (Fig. 2). The enrichment in some very generic GO terms such as *oxidation–reduction process*, *cellular response to drug* and *translation* supports the hypothesis that the diversity of experiments assembled to build the atlas provides a representative, unbiased subset of the *C. albicans* proteome. In contrast, the more precise groups resulting from the analysis related to *pathogenesis*, *hyphal growth* and *fungal-type cell wall organization* are consistent with the large contribution to the atlas by the experiment aimed at identifying proteins from cells in hyphal form and by the profusion of these sort of annotations in the source database.

As for the set of proteins present in the fasta database used for the searches that are not covered in the PeptideAtlas, they were subject to a similar analysis and were found to be enriched in annotations related to the *transmembrane transport* GO term (Fig. 2). These proteins are not easily observed by LC–MS/MS techniques as previously reported [20]. Also, we observed enrichment in *regulation of transcription, DNA-dependent* in the undetected part of the proteome. Given the short life span and low abundance of many transcription factors it is plausible that they were not detected in the collected datasets and their under representation in proteomic data has also been reported in other proteomic studies and in PeptideAtlas instances from other species [20,38,39]. The low number of protein groups significantly associated with GO annotations in the undiscovered set is understandably due to the fact that 2460 out of 3665 of the undetected protein sequences, roughly two thirds, correspond to unnamed ORFs, meaning, that little is known about their biological function.

In addition to the groups of functionally characterized proteins, this PeptideAtlas offers solid empirical evidence for the existence of 1564 proteins, showing a ProteinProphet probability score greater than 0.9, corresponding to *uncharacterized* ORFs in the CGD database (*i.e.*, one-third of all 4566 *uncharacterized* ORFs).

3.2. *Proteins of interest. Case of use*

From the clinical angle, the characterization of the *C albicans* proteome is focused on particular subproteomes, including cell surface constituents, and the set of proteins involved in the yeast-to-hypha transition. The cell wall, as the outermost cell structure represents the contact surface with host cells and therefore gathers many antigens, virulence factors and Pathogen Associated Molecular Patterns (PAMPs) [40]. Proteins

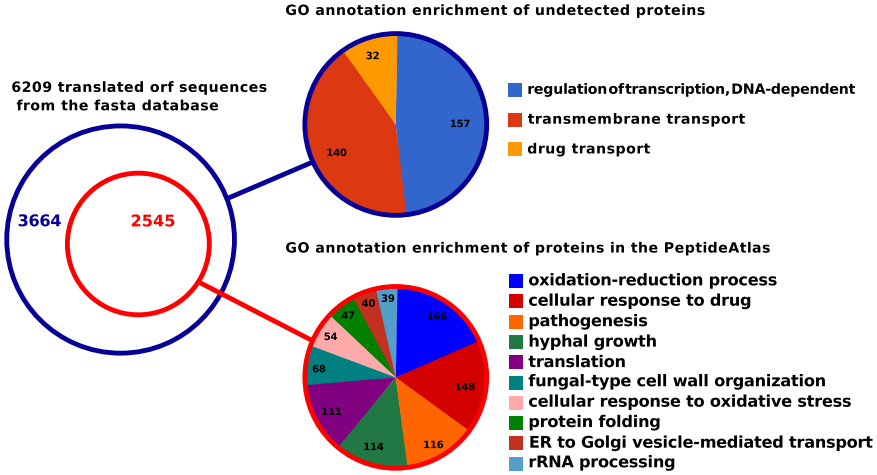


Fig. 2 – Gene Ontology annotation enrichment analysis for both the covered and undetected proteome subsets. All shown GO annotations correspond to the biological process ontology and were found significant for a p-value cut-off below 0.01.

involved in hyphal growth are also relevant in pathogenesis, in the sense that hyphae have been proven as key for invasiveness whereas the switch back to yeast form plays a role in dissemination [41].

Within these groups, a selected set of proteins of interest present in the atlas, are the adhesins from the ALS family with a role in invasiveness Als2p and Als3p; those required for cell wall biogenesis and organization glycosidases Phr1p, Phr2p and Utr2p; mannosyltransferases Pmt1p, Pmt4 and Pmt6; those involved in the cell-wall glucan metabolism Mp65p and Ecm33p, and the hyphal cell wall constituents Hwp1, Csp37p and Rbt1p.

Other relevant proteins in the atlas are the ones related to apoptosis, since those would make an ideal target for the treatment of invasive candidiasis. Among those, the atlas contains Mca1p, Bcy1p, Ras1p and three unnamed ORFs with orthologous in other species showing roles in the apoptotic process (orf19.713, orf19.967 and orf19.7365).

For any particular proteins of interest, the PeptideAtlas web interface provides tools to explore the data. A user can

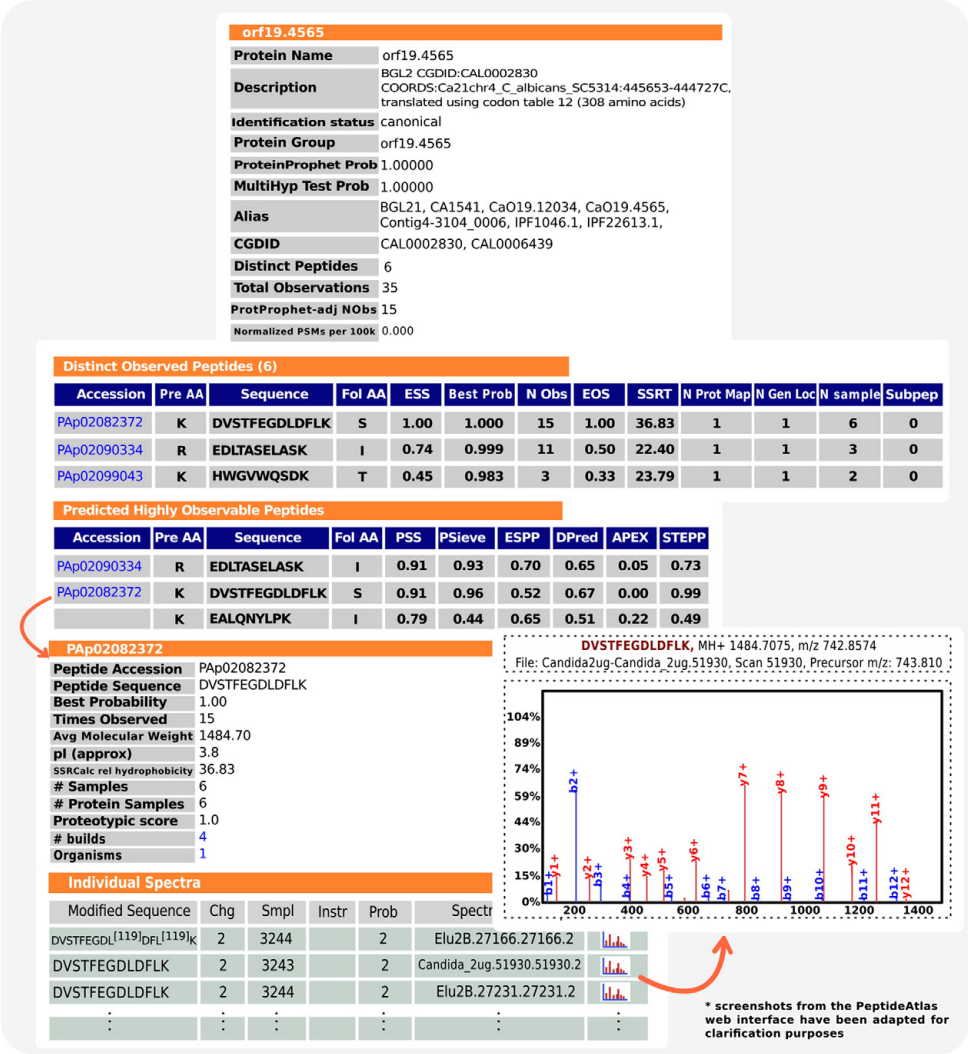


Fig. 3 – Protein- and peptide-centric views for Bgl2p are depicted. *Distinct observed peptides* are ranked by the *BestProb* parameter (representing the PeptideProphet probability). Of those, most probably, some will also be present in the following *Predicted Highly Observable Peptides* table were peptides are ranked by PSS, a combination of different prediction algorithms. For all observed peptides, spectra from the different experiments are also available.

browse through a set of protein and peptide-centric views as illustrated in Fig. 3 for the specific case of Bgl2p, a cell wall glucosyltransferase. Its corresponding observed peptides are highlighted in the protein sequence and sorted by the Empirical Suitability Score (ESS), which represents the proportion of the number of samples in which the peptide is observed with regard to the number of samples in which the original protein is observed. This parameter, in combination with others, such as a number of protein mappings, genome location and amino acid composition will help the user to select candidate proteotypic peptides for a targeted proteomics (SRM, Selected Reaction Monitoring) experiment.

Concerning those cases where a selected protein of interest is not observed in the selected build, the PeptideAtlas also provides the Predicted Suitability Score (PSS), a value resulting from the combination of different observability prediction algorithms based upon physico-chemical properties derived from the amino acid composition and previous training datasets as described in [42].

The build that assembles the phosphoprotein enrichment experiments may be of great potential interest to study biological processes such as signal transduction, since it encompasses a number of kinases and phosphatases. A total of 421 different phosphopeptides were detected and allocated to 210 phosphoproteins. The largest number of phosphorylation sites occurs in S, 410 phosphopeptides contain, at least, one phosphorylation in S; 79 phosphopeptides contain, at least, one phosphorylation in T; and 10 phosphopeptides contain one phosphorylation in Y.

4. Conclusions

This *C albicans* PeptideAtlas build provides empirical identification evidence for 21,938 unique peptides including 421 phosphopeptides at a 0.24% peptide-level FDR that account for a high-confidence set (as defined in [14]) of 2562 canonical proteins at a 1.2% protein-level FDR representing thus a significant advance in the proteomic characterization of *C. albicans*.

Through the web interface, an important set of tools are made available to the scientific community, enabling a solid foundation to study different basic biological processes like dimorphism, signal transduction, apoptosis and the interaction with the human host. Furthermore, its value as a resource for proteotypic peptide selection is of great potential interest for future SRM experiments.

The current version of the PeptideAtlas can be found at:https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/buildDetails?atlas_build_id=323and the version including PTM results at:https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/buildDetails?atlas_build_id=324.

Acknowledgments

The Proteomics Unit UCM-Parque Científico de Madrid is a member of the ProteoRed-Spanish National Institute for Proteomics.

We are thankful to María Luisa Hernáez and Jose Antonio Reales for helping in sample obtention from the hyphal and yeast form protein extracts and to Antonio Serna for providing the tandem mass spectra from the triple-TOF instrument. Also Aida Pitarch helped in the preparation of the manuscript.

This work was supported by BIO 2009-07654 and BIO 2012-31767 from the Ministerio de Economía y Competitividad, PROMPT (S2010/BMD-2414) from the Comunidad de Madrid, and Instituto de Salud Carlos III, Subdirección General de Redes y Centros de Investigación Cooperativa, Ministerio de Economía y Competitividad, Spanish Network for Research in Infectious Diseases (REIPI RD12/0015) -co-financed by the European Development Regional Fund "A way to achieve Europe" ERDF.

EWD, ZS, and RLM are supported in part by the National Institute of General Medical Sciences, under Grant No. R01 GM087221, 2P50 GM076547/Center for Systems Biology, the National Science Foundation MRI [Grant No. 0923536], the EU FP7 grant 'ProteomeXchange' [Grant No. 260558], and by the Luxembourg Centre for Systems Biomedicine and the University of Luxembourg.

RA is supported in part by ERC advanced grant 'Proteomics v3.0' (Grant No. 233226) of the European Union.

REFERENCES

[1] Wisplinghoff H, Bischoff T, Tallent SM, Seifert H, Wenzel RP, Edmond MB. Nosocomial bloodstream infections in US hospitals: analysis of 24,179 cases from a prospective nationwide surveillance study. Clin Infect Dis 2004;39:309–17.

[2] Moran C, Grussemeyer CA, Spalding JR, Benjamin DK, Reed SD. Comparison of costs, length of stay, and mortality associated with *Candida glabrata* and *Candida albicans* bloodstream infections. Am J Infect Control 2010;38:78–80.

[3] Tong KB, Murtagh KN, Lau C, Seifeldin R. The impact of esophageal candidiasis on hospital charges and costs across patient subgroups. Curr Med Res Opin 2008;24:167–74.

[4] Fernández-Arenas E, Cabezón V. Integrated proteomics and genomics strategies bring new insight into *Candida albicans* response upon macrophage interaction. Mol Cell Proteomics 2007;6:460–78.

[5] Pitarch A, Nombela C, Gil C. Prediction of the clinical outcome in invasive candidiasis patients based on molecular fingerprints of five anti-*Candida* antibodies in serum. Mol Cell Proteomics 2011;10, doi:10.1074/mcp.M110.004010 [M110.004010].

[6] Pitarch A, Nombela C, Gil C. *Candida albicans* biology and pathogenicity: insights from proteomics. Methods Biochem Anal 2006;49:285–330.

[7] Pitarch A, Nombela C, Gil C. Contributions of proteomics to diagnosis, treatment, and prevention of candidiasis. Methods Biochem Anal 2006;49:331–61.

[8] Vizcaíno JA, Côté RG, Csordas A, Dianes Ja, Fabregat A, Foster JM, et al. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res 2013;41:D1063–9.

[9] Vialás V, Nogales-Cadenas R, Nombela C, Pascual-Montano A, Gil C. Proteopathogen, a protein database for studying *Candida albicans*—host interaction. Proteomics 2009;9:4664–8.

[10] Craig R, Cortens JP, Beavis RC. Open source system for analyzing, validating, and storing protein identification data. J Proteome Res 2004;3:1234–42.

[11] Smith BE, Hill JA, Gjukich MA, Andrews PC. Tranche distributed repository and ProteomeCommons.org. Methods Mol Biol 2011;696:123–45.

[12] Costanzo MC, Arnaud MB, Skrzypek MS, Binkley G, Lane C, Miyasato SR, et al. The *Candida* Genome Database: facilitating

research on *Candida albicans* molecular biology. FEMS Yeast Res 2006;6:671–84.
[13] Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, et al. The PeptideAtlas project. Nucleic Acids Res 2006;34:D655–8.
[14] Farrah T, Deutsch EW, Omenn GS, Campbell DS, Sun Z, Bletz J a, et al. A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. Mol Cell Proteomics 2011;10, doi:10.1074/mcp.M110.006353 [M110.006353].
[15] Van PT, Schmid AK, King NL, Kaur A, Pan M, Whitehead K, et al. *Halobacterium salinarum* NRC-1 PeptideAtlas: toward strategies for targeted proteomics and improved proteome coverage. J Proteome Res 2008;7:3755–64.
[16] Schubert OT, Mouritsen J, Ludwig C, Röst HL, Rosenberger G, Arthur PK, et al. The Mtb Proteome Library: a resource of assays to quantify the complete proteome of *Mycobacterium tuberculosis*. Cell Host Microbe 2013;13:602–12.
[17] Lange V, Malmström Ja, Didion J, King NL, Johansson BP, Schäfer J, et al. Targeted quantitative analysis of *Streptococcus pyogenes* virulence factors by multiple reaction monitoring. Mol Cell Proteomics 2008;7:1489–500.
[18] Lindner SE, Swearingen KE, Harupa A, Vaughan AM, Sinnis P, Moritz RL, et al. Total and putative surface proteomics of malaria parasite salivary gland sporozoites. Mol Cell Proteomics 2013;12, doi:10.1074/mcp.M112.024505 [M110.024505].
[19] King NL, Deutsch EW, Ranish JA, Nesvizhskii AI, Eddes JS, Mallick P, et al. Analysis of the *Saccharomyces cerevisiae* proteome with PeptideAtlas. Genome Biol 2006;7:R106.
[20] Gunaratne J, Schmidt A, Quandt A, Neo SP, Sarac OS, Gracia T, et al. Extensive mass spectrometry-based analysis of the fission yeast proteome: the *S. pombe* PeptideAtlas. Mol Cell Proteomics 2013;12, doi:10.1074/mcp.M112.023754 [M112.023754].
[21] Loevenich SN, Brunner E, King NL, Deutsch EW, Stein SE, Aebersold R, et al. The *Drosophila melanogaster* PeptideAtlas facilitates the use of peptide data for improved fly proteomics and genome annotation. BMC Bioinformatics 2009;10:59.
[22] Chan QWT, Parker R, Sun Z, Deutsch EW, Foster LJ. A honey bee (*Apis mellifera* L.) PeptideAtlas crossing castes and tissues. BMC Genomics 2011;12:290.
[23] Bislev S, Deutsch E, Sun Z. A bovine PeptideAtlas of milk and mammary gland proteomes. Proteomics 2012;12:2895–9.
[24] Deutsch E, Lam H, Aebersold R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. EMBO Rep 2008;9:429–34.
[25] Gillet LC, Navarro P, Tate S, Röst H, Selevsek N, Reiter L, et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics 2012;11, doi:10.1074/mcp.O111.016717 [O111.016717].
[26] Cabezón V, Llama-Palacios A, Nombela C, Monteoliva L, Gil C. Analysis of *Candida albicans* plasma membrane proteome. Proteomics 2009;9:4770–86.
[27] Monteoliva L, Martinez-Lopez R. Quantitative proteome and acidic subproteome profiling of *Candida albicans* yeast-to-hypha transition. J Proteome Res 2010;10:502–17.
[28] Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, Tasman N, et al. A guided tour of the Trans-Proteomic Pipeline. Proteomics 2010;10:1150–9.
[29] Martens L, Chambers M, Sturm M. mzML—a community standard for mass spectrometry data. Mol Cell Proteomics 2011;10, doi:10.1074/mcp.R110.000133 [R110.000133].
[30] Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics 2004;20:1466–7.
[31] MacLean B, Eng J, Beavis R, McIntosh M. General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. Bioinformatics 2006;22:2830–2.
[32] Keller A, Eng J, Zhang N. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol Syst Biol 2005;1 [2005.0017].
[33] Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 2002;74:5383–92.
[34] Shteynberg D, Deutsch EW, Lam H, Eng JK, Sun Z, Tasman N, et al. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Mol Cell Proteomics 2011;10, doi:10.1074/mcp.M111.007690 [M111.007690].
[35] Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 2003;75:4646–58.
[36] Reiter L, Claassen M, Schrimpf S. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. Mol Cell Proteomics 2009;8:2405–17.
[37] Tabas-Madrid D, Nogales-Cadenas R, Pascual-Montano A. GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. Nucleic Acids Res 2012;40:W478–83.
[38] Ding C, Chan DW, Liu W, Liu M, Li D, Song L, et al. Proteome-wide profiling of activated transcription factors with a concatenated tandem array of transcription factor response elements. Proc Natl Acad Sci U S A 2013;110:6771–6.
[39] Simicevic J, Schmid AW, Gilardoni PA, Zoller B, Raghav SK, Krier I, et al. Absolute quantification of transcription factors during cellular differentiation using multiplexed targeted proteomics. Nat Methods 2013;10:570–6.
[40] Vialás V, Perumal P, Gutierrez D, Ximénez-Embún P, Nombela C, Gil C, et al. Cell surface shaving of *Candida albicans* biofilms, hyphae and yeast form cells. Proteomics 2012;8:2331–9.
[41] Saville SP, Lazzell AL, Monteagudo C, Lopez-Ribot JL. Engineered control of cell morphology *in vivo* reveals distinct roles for yeast and filamentous forms of *Candida albicans* during infection. Eukaryot Cell 2003;2:1053–60.
[42] Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, et al. Computational prediction of proteotypic peptides for quantitative proteomics. Nat Biotechnol 2007;25:125–31.

3.1.

...
...

Desarrollo de una base de datos para datos de Proteómica Dirigida (MRM)

Bibliografía

Y así, del mucho leer y del poco dormir,
se le secó el celebro de manera que vino
a perder el juicio.

Miguel de Cervantes Saavedra

Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. y Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science (New York, N.Y.)*, vol. 246(4926), páginas 64–71, 1989. ISSN 0036-8075.

Karas, M. y Hillenkamp, F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical chemistry*, vol. 60(20), páginas 2299–301, 1988. ISSN 0003-2700.

Rogowska-Wrzesinska, A., Le Bihan, M.-C., Thaysen-Andersen, M. y Roepstorff, P. 2D gels still have a niche in proteomics. *Journal of proteomics*, vol. 88, páginas 4–13, 2013. ISSN 1876-7737.

Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., Yoshida, T. y Matsuo, T. Protein and polymer analyses up tom/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, vol. 2(8), páginas 151–153, 1988. ISSN 0951-4198.

–¿Qué te parece desto, Sancho? – Dijo Don Quijote –
Bien podrán los encantadores quitarme la ventura,
pero el esfuerzo y el ánimo, será imposible.

Segunda parte del Ingenioso Caballero
Don Quijote de la Mancha
Miguel de Cervantes

–Buena está – dijo Sancho –; fírmela vuestra merced.
–No es menester firmarla – dijo Don Quijote–,
sino solamente poner mi rúbrica.

Primera parte del Ingenioso Caballero
Don Quijote de la Mancha
Miguel de Cervantes