



中国科学院水生生物研究所

INSTITUTE OF HYDROBIOLOGY, CHINESE ACADEMY OF SCIENCES

CNCP2021

Proteogenomic analyses for Genome Annotation and Global Profiling of Post-Translational Modifications

Mingkun Yang

Lab of Functional Proteomics

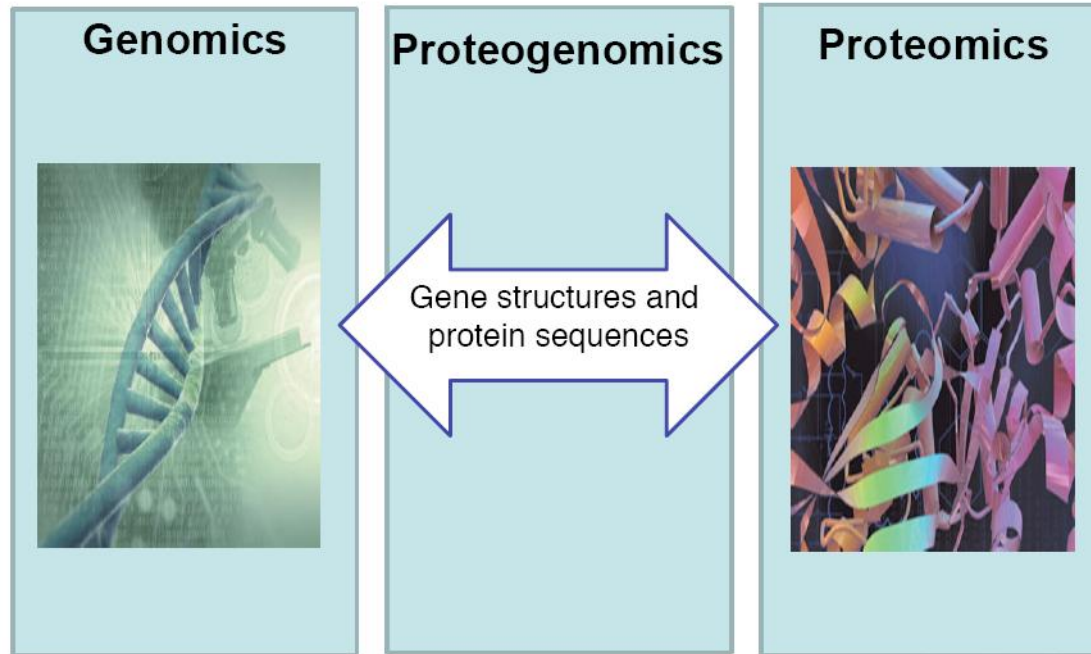
Institute of Hydrobiology, Chinese Academy of Sciences

August 26, 2021, Shanghai

什么是蛋白基因组学？

Proteomics + Genomics = Proteogenomics

[Wikipedia](#): use proteomic information, often derived from mass spectrometry, to improve gene annotations. (利用蛋白质组的信息，通常是来自于质谱的数据，来提高基因的注释)



基因注释方法

传统的基因注释方法：

- 基于算法的基因预测：从头计算法、同源建模法
- 表达序列标签(expression sequence tag)测序
- 转录组测序
- 设计引物进行目标序列PCR测序

2004年Jaffe，首次提出了蛋白质基因组学proteogenomics概念，将蛋白质组研究方法应用于支原体基因组的注释过程中

然而，基于质谱技术的蛋白质组学方法注释基因组，一直被看作是少数派的做法

人类蛋白质组草图

ARTICLE

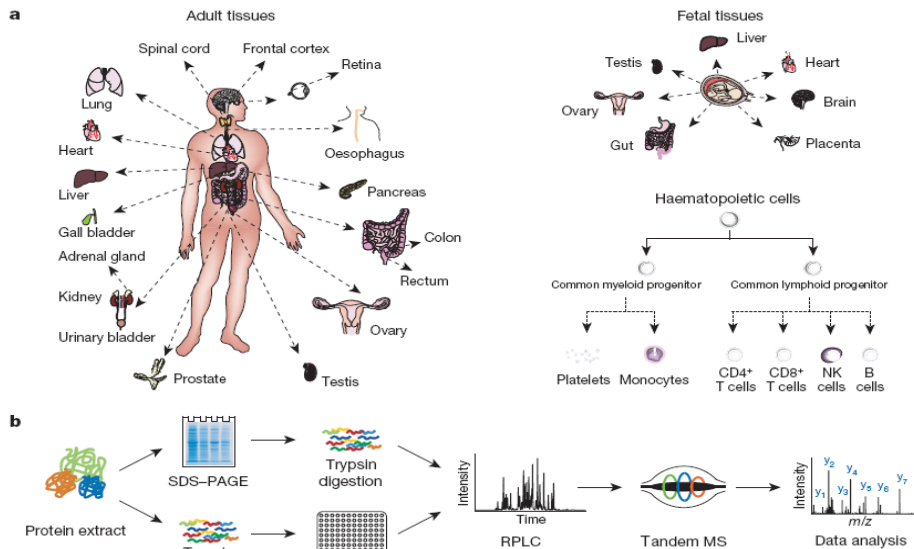
doi:10.10

A draft map of the human proteome

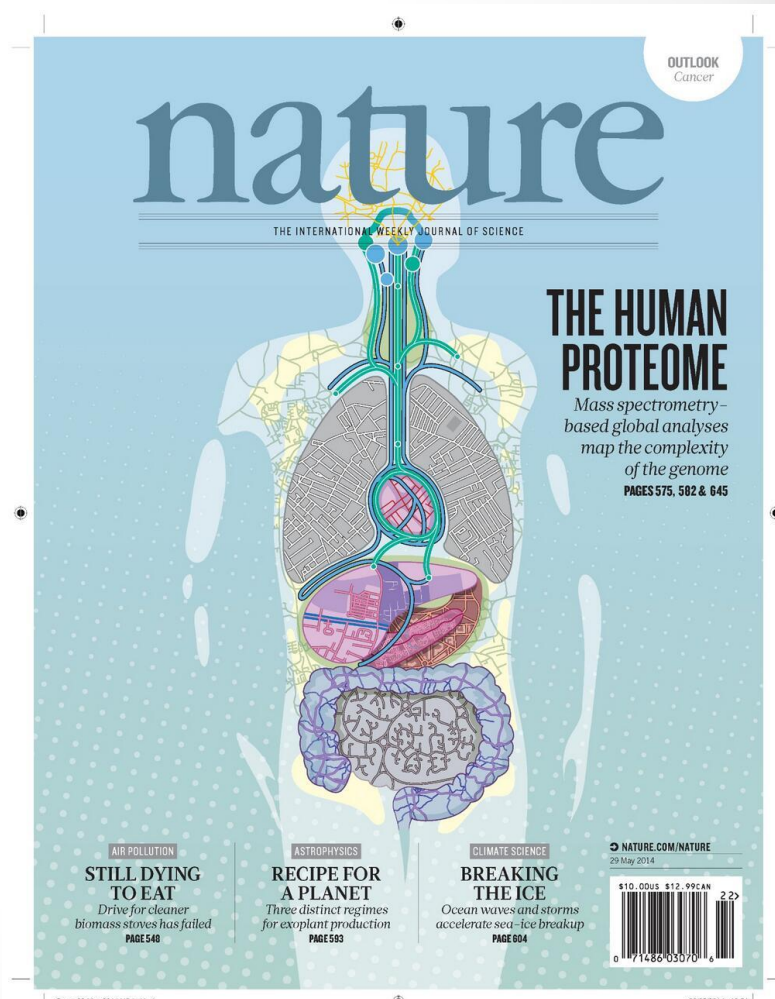
ARTICLE

doi:10.1038/nat

Mass-spectrometry-based draft of the human proteome



2014



The quantitative and condition-dependent *Escherichia coli* proteome

Alexander Schmidt¹, Karl Kochanowski², Silke Vedelaar³, Erik Ahrné¹, Benjamin Volkmer², Luciano Callipo², Kévin Knoops⁴, Manuel Bauer¹, Ruedi Aebersold^{2,5} & Matthias Heinemann^{2,3}

Journal of
proteome
research

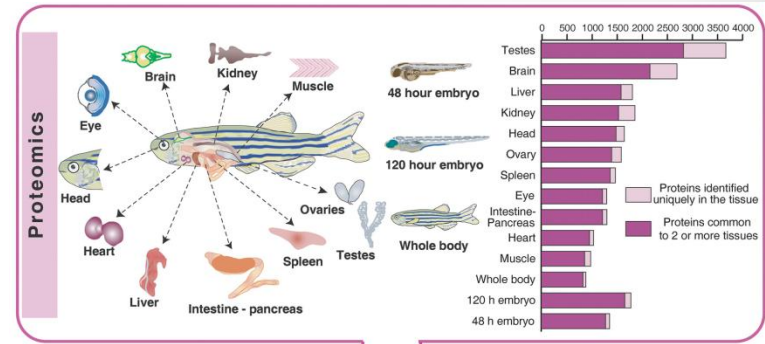
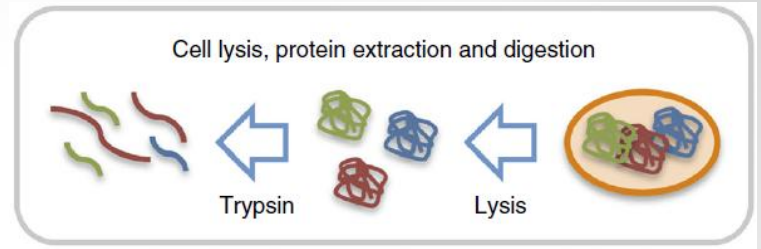
Subscriber access provided by Kaohsiung Medical University

Article

Enrichment-based proteogenomics identifies microproteins, missing proteins, and novel smORFs in *Saccharomyces cerevisiae*

Cuitong He, Chenxi Jia, Yao Zhang, and Ping Xu

J. Proteome Res., Just Accepted Manuscript • DOI: 10.1021/acs.jproteome.8b00032 • Publication Date (Web): 13 Jun 2018



MCP MOLECULAR & CELLULAR PROTEOMICS

Published by the American Society for Biochemistry and Molecular Biology

Technological Innovation and Resources

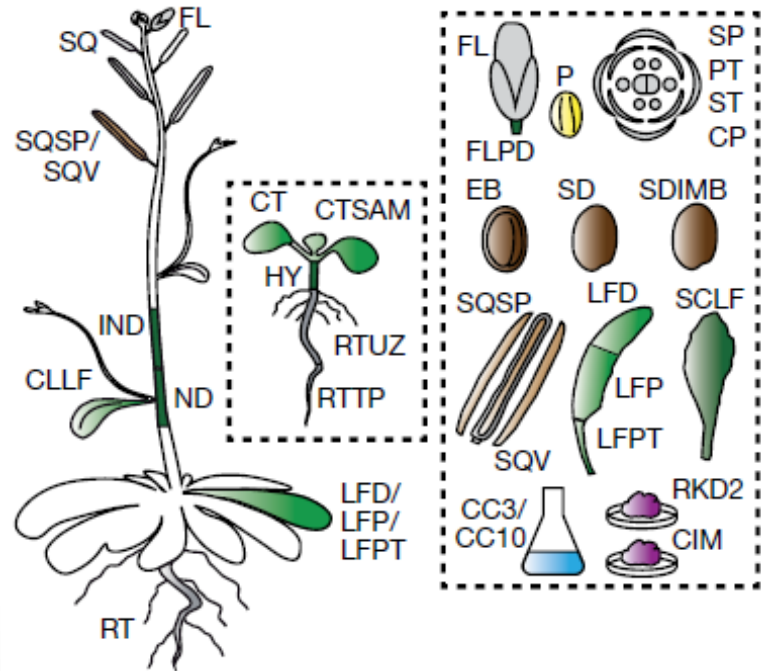
© 2014 by The American Society for Biochemistry and Molecular Biology, Inc. This paper is available on line at <http://www.mcponline.org>

Annotation of the Zebrafish Genome through an Integrated Transcriptomic and Proteomic Analysis

[nature](#) > [articles](#) > [article](#)

Article | Published: 11 March 2020

Mass-spectrometry-based draft of the *Arabidopsis* proteome



为什么需要蛋白质基因组学？


BMC Genomics. 2007 Jul 27;8:255.

Plasmodium falciparum (恶性疟原虫)

- 2002年就完成了全基因组测序
- **约25%的基因注释是错误的** (393个基因)
- 这对其它物种基因组注释意味着什么？

Likely that high percentage of gene predictions are incorrect!

对于所有已测序的基因组，可能都有这么高的基因组注释错误率！

BMC Genomics 

Research article **Open Access**

cDNA sequences reveal considerable gene prediction inaccuracy in the *Plasmodium falciparum* genome

Fangli Lu^{1,2}, Hongying Jiang¹, Jinhui Ding³, Jianbing Mu¹, Jesus G Valenzuela¹, José MC Ribeiro¹ and Xin-zhuan Su*¹

Address: ¹Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, USA. ²Department of Parasitology, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, Guangdong 510080, P.R.C and ³Bioinformatics Unit, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, Maryland, USA

Email: Fangli Lu - fangliu@yahoo.com; Hongying Jiang - hojiang@niaid.nih.gov; Jinhui Ding - ding@mail.nih.gov; Jianbing Mu - jmu@niaid.nih.gov; Jesus G Valenzuela - jvalenzuela@niaid.nih.gov; José MC Ribeiro - jrbeiro@niaid.nih.gov; Xin-zhuan Su* - xsu@niaid.nih.gov

* Corresponding author

Published: 27 July 2007 Received: 14 February 2007
BMC Genomics 2007, 8:255 doi:10.1186/1471-2164-8-255 Accepted: 27 July 2007

This article is available from: <http://www.biomedcentral.com/1471-2164/8/255>

© 2007 Lu et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: The completion of the *Plasmodium falciparum* genome represents a milestone in malaria research. The genome sequence allows for the development of genome-wide approaches such as microarray and proteomics that will greatly facilitate our understanding of the parasite biology and accelerate new drug and vaccine development. Designing and application of these genome-wide assays, however, requires accurate information on gene prediction and genome annotation. Unfortunately, the genes in the parasite genome databases were mostly identified using computer software that could make some erroneous predictions.

Results: We aimed to obtain cDNA sequences to examine the accuracy of gene prediction *in silico*. We constructed cDNA libraries from mixed blood stages of *P. falciparum* parasite using the SMART cDNA library construction technique and generated 17332 high-quality expressed sequence tags (EST), including

为什么需要蛋白质基因组学？



The Human Proteome Project: Current State and Future Direction

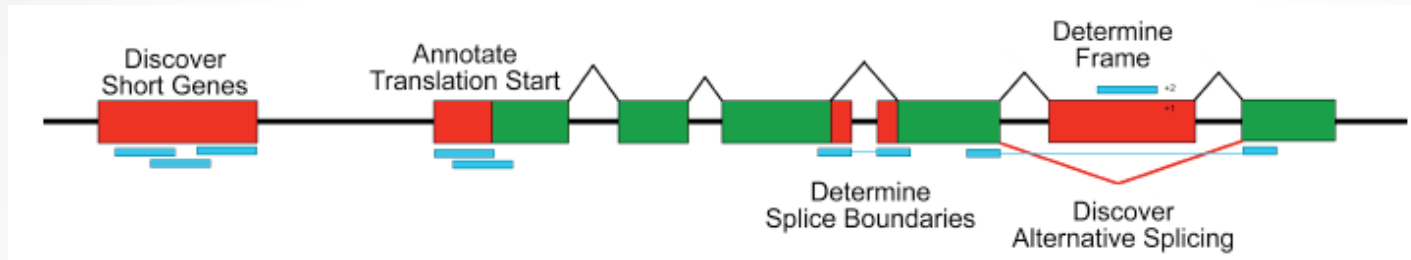
- 1) 人类基因组计划中预测的20300个蛋白质编码基因，约2/3的蛋白质的功能不太清楚；
- 2) 约6000（30%）个基因缺乏蛋白质水平上的证据支持；
- 3) 还有许多其他蛋白质的丰度、分布、细胞亚定位、相互作用和翻译后修饰也是不清楚的

Legrain *et al.*, MCP, 2011

 Human Proteome Project



蛋白质基因组学可以做什么？



确认预测的基因

校正错误的基因

发现新的基因

确定信号肽

确定起始位点

发现蛋白的翻译后修饰

非常规起始密码子

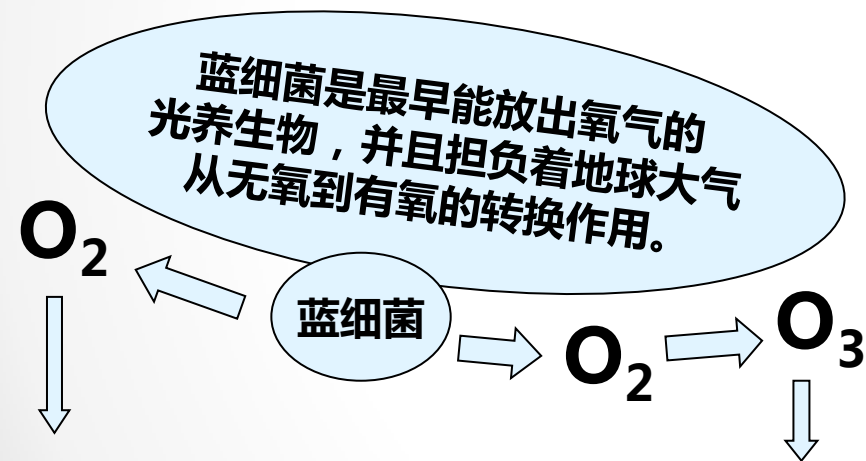
终止密码子通读

蛋白表达及翻译后修饰的动态变化

蓝细菌(Cyanobacteria)

蓝细菌（又名蓝藻）是一大类群分布极广的、可进行产氧光合作用的、古老的原核微生物。

(25-35亿年前, 蓝细菌出现, 光合作用开始)



地球变成了高度氧化性的, 使生物进化速度产生了一个巨大的爆发。

有效阻挡紫外辐射。使高密度种群得以发展。

蓝藻水华



蓝细菌的蛋白基因组学 (Proteogenomics) 分析

PNAS PNAS PNAS

Proteogenomic analysis and global discovery of posttranslational modifications in prokaryotes

Ming-kun Yang^{a,1}, Yao-hua Yang^{a,1}, Zhuo Chen^a, Jia Zhang^a, Yan Lin^a, Yan Wang^a, Qian Xiong^a, Tao Li^{a,2}, Feng Ge^{a,2}, Donald A. Bryant^b, and Jin-dong Zhao^a

^aKey Laboratory of Algal Biology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China; and ^bDepartment of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA 16802

Edited* by Jiayang Li, Chinese Academy of Sciences, Beijing, China, and approved November 13, 2014 (received for review July 6, 2014)

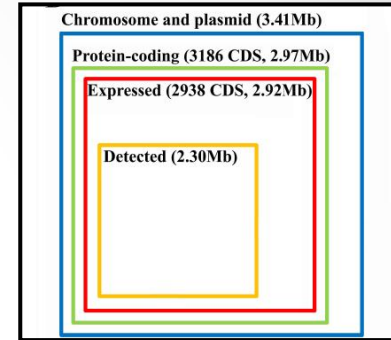
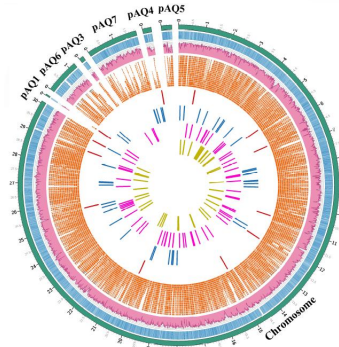
We describe an integrated workflow for proteogenomic analysis and global profiling of posttranslational modifications (PTMs) in prokaryotes and use the model cyanobacterium *Synechococcus* sp. PCC 7002 (hereafter *Synechococcus* 7002) as a test case. We found more than 20 different kinds of PTMs, and a holistic view of PTM events in this organism grown under different conditions was obtained without specific enrichment strategies. Among 3,186 predicted protein-coding genes, 2,938 gene products (>92%) were identified. We also identified 118 previously unidentified proteins and corrected 38 predicted gene-coding regions in the *Synechococcus* 7002 genome. This systematic analysis not only provides comprehensive information on protein profiles and the diversity of PTMs in *Synechococcus* 7002 but also provides some insights into photosynthetic pathways in cyanobacteria. The entire proteogenomics pipeline is applicable to any sequenced prokaryotic organism, and we suggest that it should become a standard part of genome annotation projects.

and cyanobacterial cells adjust their cellular activities in response to a wide range of environmental cues and stimuli. Recently, cyanobacteria have attracted great interest due to their crucial roles in global carbon and nitrogen cycles and their ability to produce clean and renewable biofuels such as hydrogen (14–16). *Synechococcus* 7002 is a unicellular, marine cyanobacterium and a model organism for studying photosynthetic carbon fixation and the development of biofuels (17, 18). However, whereas the genome of *Synechococcus* 7002 is fully sequenced, it is annotated only by in silico methods (www.ncbi.nlm.nih.gov/), with a large portion (1,210 out of 3,186) of protein-coding genes annotated as hypothetical proteins (17). Therefore, a comprehensive analysis is needed to provide experimental support for the genome annotation so as to facilitate systems-level analysis. Using our method, we performed the validation of the predicted protein-coding genes, identified previously unidentified genes, and corrected gene initiation and stop-codon positions in *Synechococcus* 7002, and directional RNA-Seq was used to determine the existence of a

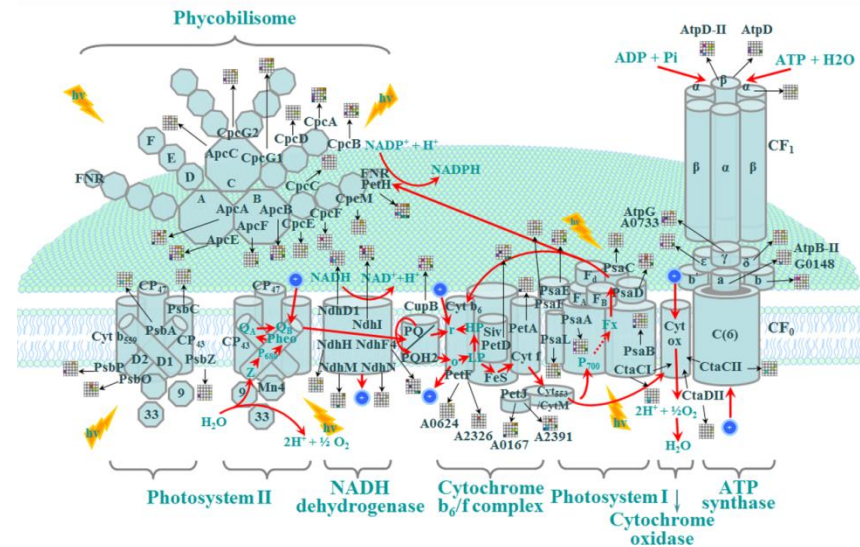


PNAS PLUS

APPLIED BIOLOGICAL SCIENCES



鉴定到蓝细菌92.2%的全部预测的编码基因；发现118个新基因；发现了23种蛋白质的翻译后修饰...



建立了完整的蛋白质基因组学研究和分析流程, 可适用于各种原核生物; 实现了全蛋白质组水平的蛋白质翻译后修饰的系统发现

针对蓝细菌中蛋白质翻译后修饰的功能及其分子调控机制这一科学问题,建立了蛋白质翻译后修饰的全局发现技术和分析方法,系统研究了蓝细菌中磷酸化、乙酰化、琥珀酰化等重要翻译后修饰的功能和分子作用机制

翻译后修饰的全局发现及其分析软件

(*PNAS*, 2014,111 (52) :E5633-E5642;

Molecular & Cellular Proteomics, 2016;15(11):3529-3539;

Journal of Proteomics, 2016,16;134:57-64)

乙酰化

(*Molecular & Cellular Proteomics*,
2017,16(7):1297-1311;

Journal of Proteome Research,
2015,14 (2) :1275-1286;

Molecular & Cellular Proteomics,
2014,13(12):3352-66)

琥珀酰化

(*Plant and Cell Physiology*,
2018,59(7):1466-1482;

Molecular & Cellular Proteomics,
2018,17(3):457-471;

Molecular & Cellular Proteomics,
2015,14(4):796-811)

蛋白质翻译后修饰 组学及功能研究

丙二酰化

(*Journal of Proteome Research*,
2017; 16(5):2030-2043)

甲基化

(*Genomics Proteomics & Bioinformatics* ,
2020:S1672-0229(18)30224-9)

丙酰化

(*Plant Physiology*, 2020;184(2):762-776;
International Journal of Molecular Sciences ,
2019;20(19):4792)

磷酸化

(*Plant and Cell Physiology*, 2015, 56 (10):1997-2013;
Molecular & Cellular Proteomics, 2014;13(2):503-19;
Journal of Proteome Research, 2013;12(4):1909-23)

蓝细菌的蛋白基因组学 (Proteogenomics) 分析

PNAS PNAS PNAS

Proteogenomic analysis and global discovery of posttranslational modifications in prokaryotes

Ming-kun Yang^{a,1}, Yao-hua Yang^{a,1}, Zhuo Chen^a, Jia Zhang^a, Yan Lin^a, Yan Wang^a, Qian Xiong^a, Tao Li^{a,2}, Feng Ge^{a,2}, Donald A. Bryant^b, and Jin-dong Zhao^a

^aKey Laboratory of Algal Biology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China; and ^bDepartment of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA 16802

Edited* by Jiayang Li, Chinese Academy of Sciences, Beijing, China, and approved November 13, 2014 (received for review July 6, 2014)

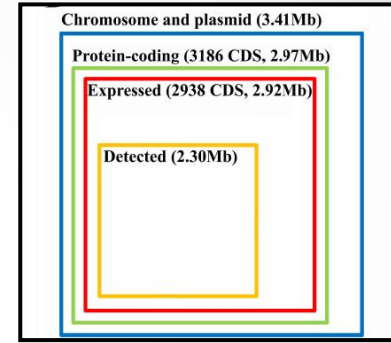
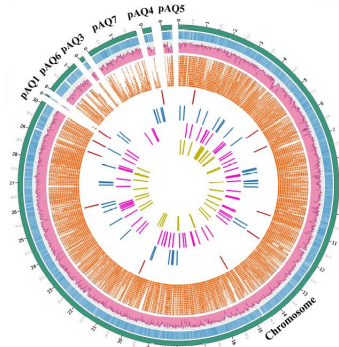
We describe an integrated workflow for proteogenomic analysis and global profiling of posttranslational modifications (PTMs) in prokaryotes and use the model cyanobacterium *Synechococcus* sp. PCC 7002 (hereafter *Synechococcus* 7002) as a test case. We found more than 20 different kinds of PTMs, and a holistic view of PTM events in this organism grown under different conditions was obtained without specific enrichment strategies. Among 3,186 predicted protein-coding genes, 2,938 gene products (>92%) were identified. We also identified 118 previously unidentified proteins and corrected 38 predicted gene-coding regions in the *Synechococcus* 7002 genome. This systematic analysis not only provides comprehensive information on protein profiles and the diversity of PTMs in *Synechococcus* 7002 but also provides some insights into photosynthetic pathways in cyanobacteria. The entire proteogenomics pipeline is applicable to any sequenced prokaryotic organism, and we suggest that it should become a standard part of genome annotation projects.

and cyanobacterial cells adjust their cellular activities in response to a wide range of environmental cues and stimuli. Recently, cyanobacteria have attracted great interest due to their crucial roles in global carbon and nitrogen cycles and their ability to produce clean and renewable biofuels such as hydrogen (14–16). *Synechococcus* 7002 is a unicellular, marine cyanobacterium and a model organism for studying photosynthetic carbon fixation and the development of biofuels (17, 18). However, whereas the genome of *Synechococcus* 7002 is fully sequenced, it is annotated only by in silico methods (www.ncbi.nlm.nih.gov/), with a large portion (1,210 out of 3,186) of protein-coding genes annotated as hypothetical proteins (17). Therefore, a comprehensive analysis is needed to provide experimental support for the genome annotation so as to facilitate systems-level analysis. Using our method, we performed the validation of the predicted protein-coding genes, identified previously unidentified genes, and corrected gene initiation and stop-codon positions in *Synechococcus* 7002, and directional RNA-Seq was used to determine the existence of a

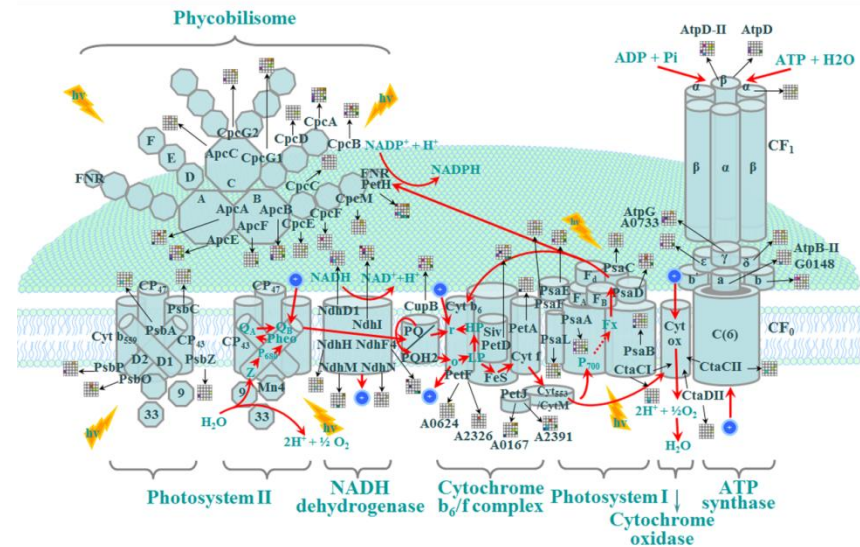


PNAS PLUS

APPLIED BIOLOGICAL SCIENCES



鉴定到蓝细菌92.2%的全部预测的编码基因；发现118个新基因；发现了23种蛋白质的翻译后修饰...



建立了完整的蛋白质基因组学研究和分析流程, 可适用于各种原核生物; 实现了全蛋白质组水平的蛋白质翻译后修饰的系统发现

GAPP: 一种应用于原核生物的基因组注释和翻译后修饰全局发现的蛋白基因组分析软件

Proteogenomic analysis and global discovery of posttranslational modifications in prokaryotes

Ming-kun Yang^{1,†}, Yao-hua Yang^{1,†}, Zhuo Chen², Jia Zhang³, Yan Lin³, Yan Wang³, Qian Xiong³, Tao Li^{1,2}, Feng Ge^{1,2}, Donald A. Bryant⁴, and Jin-dong Zhao³

¹Key Laboratory of Algal Biology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China; and ²Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA 16802

Edited* by Jiayang Li, Chinese Academy of Sciences, Beijing, China, and approved November 13, 2014 (received for review July 6, 2014)

We describe an integrated workflow for proteogenomic analysis and global profiling of posttranslational modifications (PTMs) in prokaryotes and use the model cyanobacterium *Synechococcus* sp. PCC 7002 (hereafter *Synechococcus* 7002) as a test case. We found more than 20 different kinds of PTMs, and a holistic view of PTM events in this organism grown under different conditions was obtained without specific enrichment strategies. Among 3,186 predicted protein-coding genes, 2,938 gene products (>92%) were identified. We also identified 118 previously unidentified proteins and corrected 38 predicted gene-coding regions in the *Synechococcus* 7002 genome. This systematic analysis not only provides comprehensive information on protein profiles and the diversity of PTMs in *Synechococcus* 7002 but also provides some insights into photosynthetic pathways in cyanobacteria. The entire proteogenomics pipeline is applicable to any sequenced prokaryotic organism, and we suggest that it should become a standard part of genome annotation projects.

and cyanobacterial cells adjust their cellular activities in response to a wide range of environmental cues and stimuli. Recently, cyanobacteria have attracted great interest due to their crucial roles in global carbon and nitrogen cycles and their ability to produce clean and renewable biofuels such as hydrogen (14–16). *Synechococcus* 7002 is a unicellular, marine cyanobacterium and a model organism for studying photosynthetic carbon fixation and the development of biofuels (17, 18). However, whereas the genome of *Synechococcus* 7002 is fully sequenced, it is annotated only by *in silico* methods (www.ncbi.nlm.nih.gov/), with a large portion (1,210 out of 3,186) of protein-coding genes annotated as hypothetical proteins (17). Therefore, a comprehensive analysis is needed to provide experimental support for the genome annotation so as to facilitate systems-level analysis. Using our method, we performed the validation of the predicted protein-coding genes, identified previously unidentified genes, and corrected gene initiation and stop-codon positions in *Synechococcus* 7002, and di-



PNAS PLUS

APPLIED BIOLOGICAL SCIENCES



整个蛋白质基因组学研究和分析流程适用于各种已经测序的原核生物, 我们建议应成为基因组注释的一项标准流程

MCP MOLECULAR & CELLULAR PROTEOMICS

A message from A.L. Burlingame, Editor-in-Chief

Research

© 2016 by The American Society for Biochemistry and Molecular Biology, Inc.
This paper is available on line at <http://www.mcponline.org>

GAPP: A Proteogenomic Software for Genome Annotation and Global Profiling of Post-translational Modifications in Prokaryotes*

Jia Zhang[‡]||, Ming-kun Yang[‡]||, Honghui Zeng[§], and Feng Ge[‡]§||

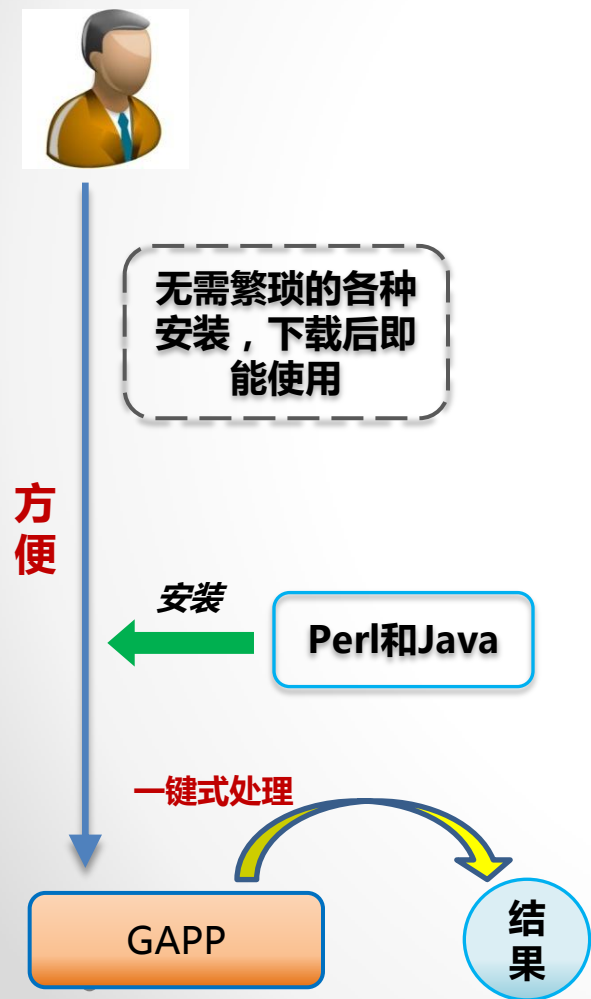
Although the number of sequenced prokaryotic genomes is growing rapidly, experimentally verified annotation of prokaryotic genome remains patchy and challenging. To facilitate genome annotation efforts for prokaryotes, we developed an open source software called GAPP for genome annotation and global profiling of post-translational modifications (PTMs) in prokaryotes. With a single command, it provides a standard workflow to validate and refine predicted genetic models and discover diverse PTM events. We demonstrated the utility of GAPP using proteomic data from *Helicobacter pylori*, one of the major human pathogens that is responsible for many gastric diseases. Our results confirmed 84.9% of the existing predicted *H. pylori* proteins, identified 20 novel protein coding genes, and corrected four existing gene models with regard to translation initiation sites. In particular, GAPP revealed a large repertoire of PTMs using the same proteomic data and provided a rich resource that can be used to examine the functions of reversible modifications in this human pathogen. This software is a powerful tool for genome annotation and global discovery of PTMs and is applicable to any sequenced prokaryotic organism; we expect that it will become an integral part of ongoing genome annotation efforts for prokaryotes. GAPP is freely available at <https://sourceforge.net/projects/gappproteogenomic/>. *Molecular & Cellular Proteomics* 15: 10.1074/mcp.M116.060046, 1–11, 2016.

regions and gene models from raw sequencing data (1, 2). Computational *ab initio* algorithms, which are hard to avoid errors such as missing genes and incorrect gene boundaries, are limited in their prediction accuracy. Recently, the use of information from RNA-seq or expressed sequence tag libraries has dramatically improved the genome annotation confidence (3). However, abundance of transcripts does not necessarily reflect cellular protein levels, and the correlation between protein and mRNA levels is generally slight (4–6). The existence and function of many gene products remain to be confirmed. Thus, MS has emerged as an invaluable high-throughput method for accurate genome annotation, a concept called “proteogenomics” (7–9).

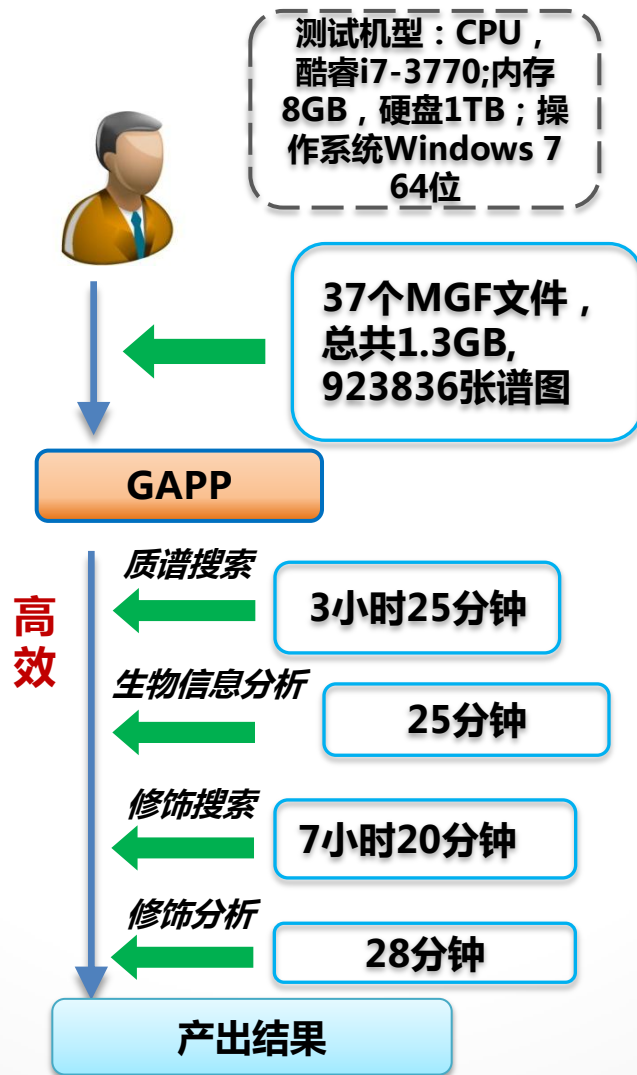
Proteogenomics is defined as the use of proteomic data, often derived from MS, to improve and refine genome annotation (10, 11). Pioneering work by Yates *et al.* (12) and Jaffe *et al.* (13) opened new avenues to high-throughput gene annotation. In recent years, proteogenomics has emerged as a promising and indispensable approach to genome annotation (10, 11). It has been applied for genome annotation including identification of novel genes, correction and validation of predicted genes in various organisms (14–21). Proteogenomics has become particularly important for prokaryotes where genomes are sequenced on a daily basis and *in silico* gene

GAPP: 一种应用于原核生物的基因组注释和翻译后修饰全局发现的蛋白基因组分析软件

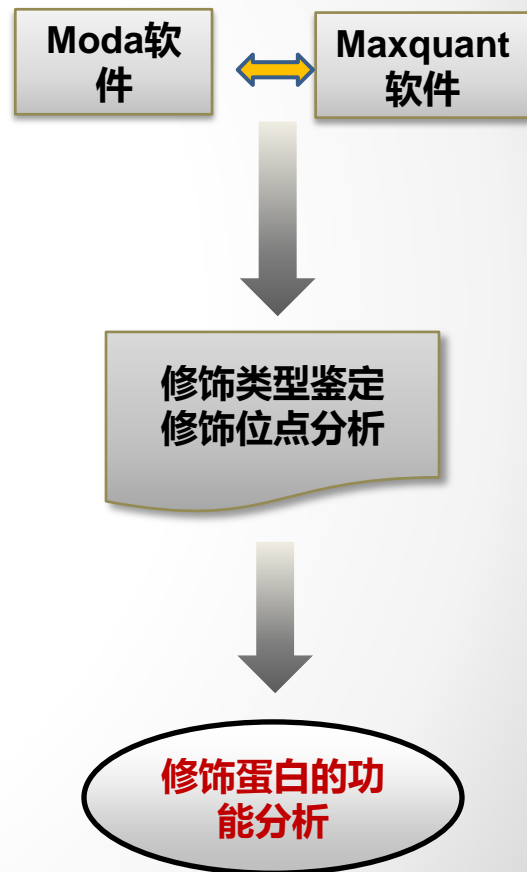
1. “一键式”，简单方便



2. 快速高效。



3. 大规模翻译后修饰鉴定



国家重点研发计划 课题任务书

课题名称: 精准鉴定技术体系在蛋白质基因组学中的应用研究

所属项目: 蛋白质组精准鉴定搜索引擎及技术体系

所属专项: 蛋白质机器与生命过程调控

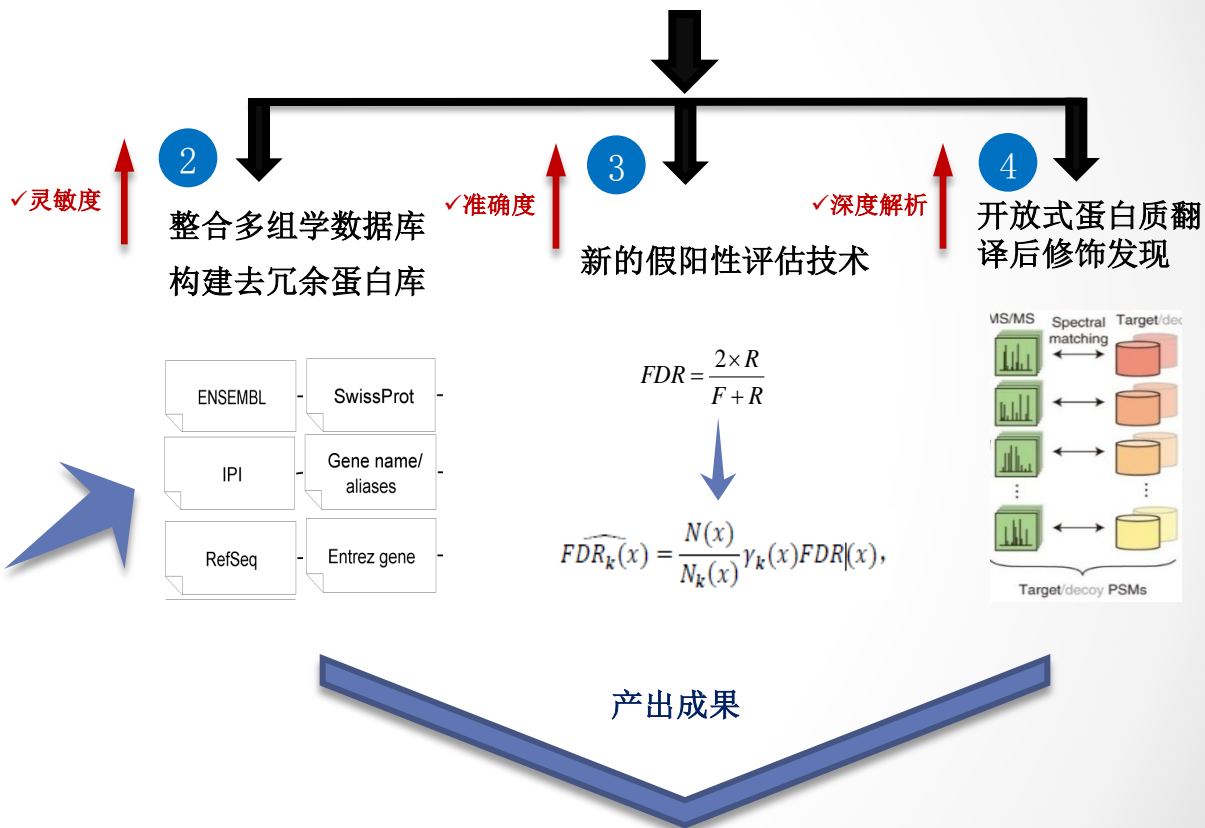
项目牵头承担单位: 复旦大学

课题承担单位: 中国科学院水生生物研究所

课题负责人: 葛峰

执行期限: 2016年07月至2021年06月

5 开发自主知识产权的专业的蛋白基因组分析软件



建立标准化样本预处理和质谱鉴定方法

✓标准化

1

重要模式真核生物蛋白基因组的精准鉴定和深度分析

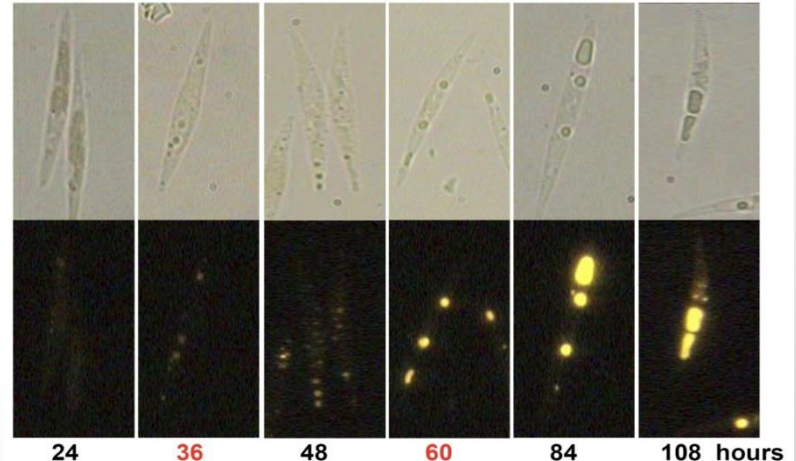
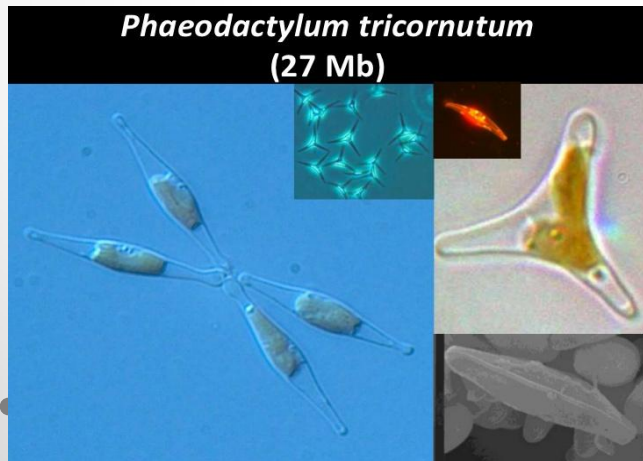
硅藻(Diatom)

硅藻是一类重要的单细胞光合真核浮游生物，遍布于海洋、湖沼、河流等水域。硅藻对自然界物质循环（如硅、碳）起着重要作用，地球上约20%的初级生产力由硅藻承担，构造奇特，呈现许多美妙的不可思议的花纹和图案。



三角褐指藻(*Phaeodactylum tricornutum*)

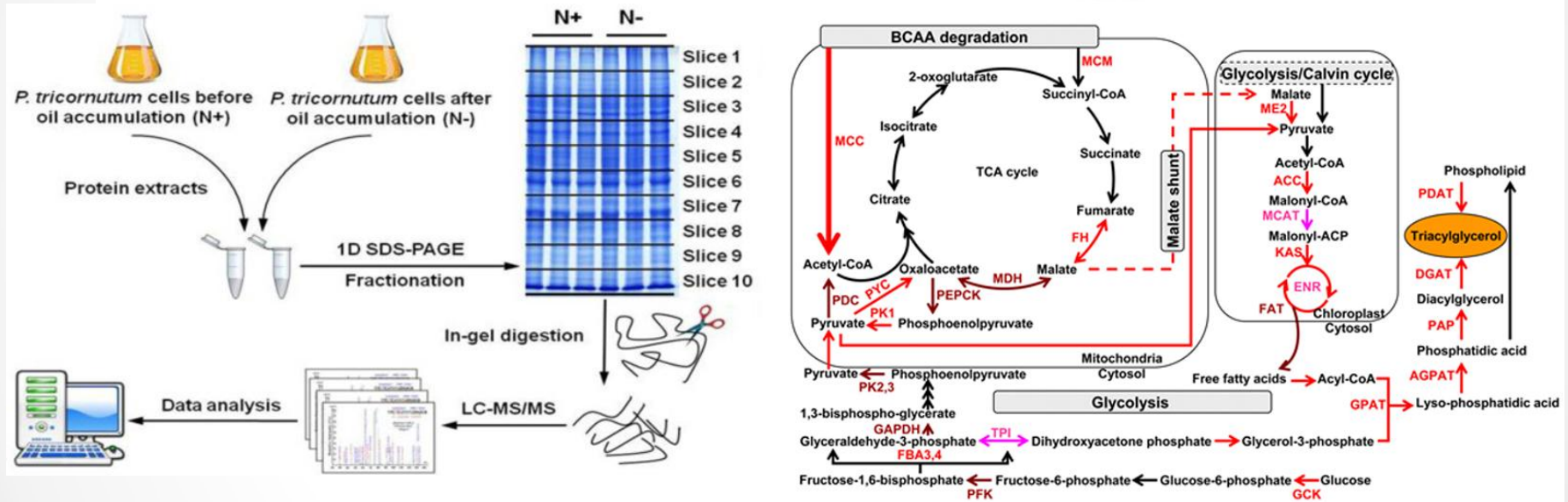
三角褐指藻，海洋硅藻的模式生物，生长迅速，在胁迫条件下积累油脂(达到细胞干重的20~30%)，是目前最受关注的能源藻种之一。



三角褐指藻—基于定量蛋白质组的研究

The Plant Cell, Vol. 26: 1681–1697, April 2014, www.plantcell.org © 2014 American Society of Plant Biologists. All rights reserved.

Methylcrotonyl-CoA Carboxylase Regulates Triacylglycerol Accumulation in the Model Diatom *Phaeodactylum tricornutum* OPEN



Plant Cell 2014; 26(4):1681-1697

- 非标记定量蛋白质组学技术研究三角褐指藻甘油三酯积累的蛋白调控网络。
- 非标定量方法，鉴定到**1193**个蛋白，然而基因组注释中包含**10402**个 **predicted genes** ?

三角褐指藻基因组注释的问题

Vol 456 | 13 November 2008 | doi:10.1038/nature07410

nature

LETTERS

The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes

Table 1 | Major features of the *P. tricornutum* and *T. pseudonana* genomes

	<i>P. tricornutum</i>	<i>T. pseudonana</i>
Genome size	27.4 Mb	32.4 Mb
Predicted genes	10,402	11,776
Core genes*	3,523	4,332
Diatom-specific genes*	1,328	1,407
Unique genes*	4,366	3,912
Introns	8,169	17,880
Introns per gene	0.79	1.52
Long-terminal-repeat retrotransposon content	5.8%	1.1%

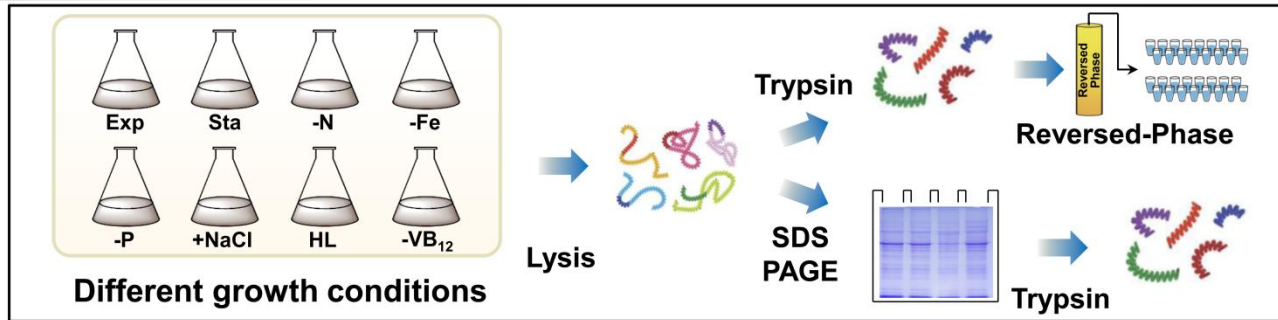
在如下两个数据库85%–90% 的三角褐指藻的基因被注释为 “predicted protein”

NCBI (https://www.ncbi.nlm.nih.gov/protein?LinkName=genome_protein&from_uid=418)

Uniprot (<http://www.uniprot.org/uniprot/?query=Phaeodactylum+tricornutum+%28strain+CCAP+1055%2F1%29&sort=score>)

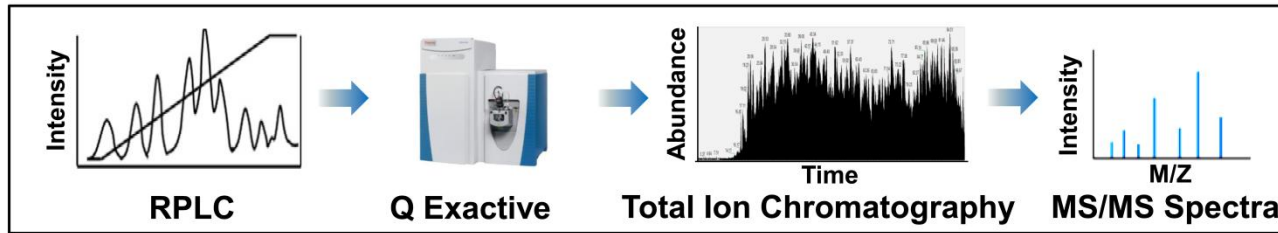
三角褐指藻：样本预处理、质谱鉴定和数据解析等流程

Samples



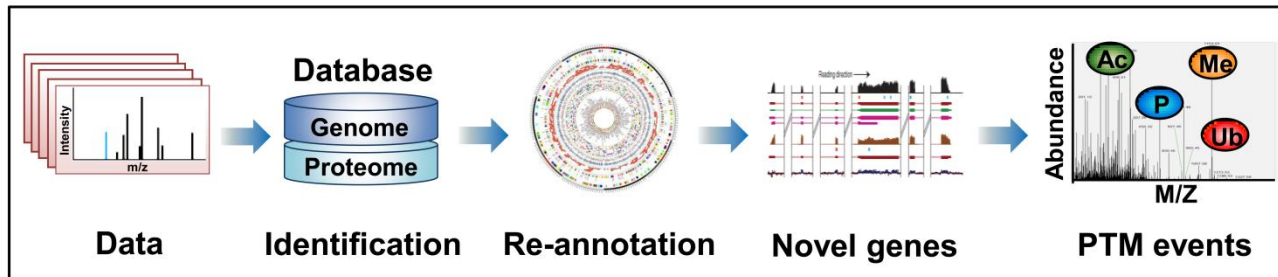
- 不同胁迫条件下的三角褐指藻；基于蛋白和肽段的样品预分离技术；多酶切技术体系；

LC-MS/MS



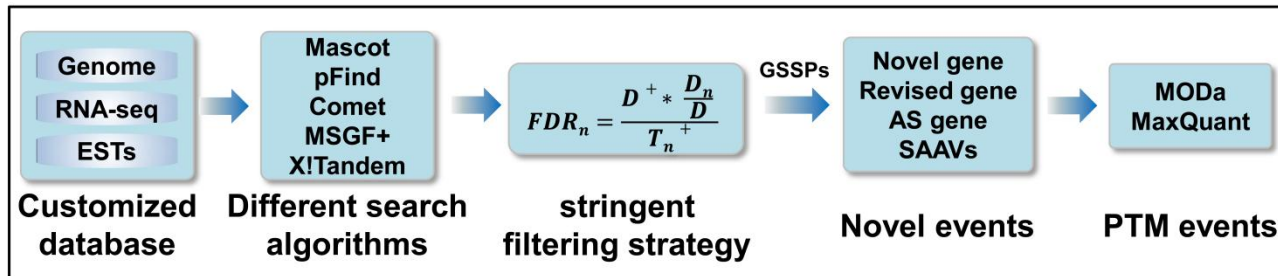
- 高分辨质谱分析；

Identification



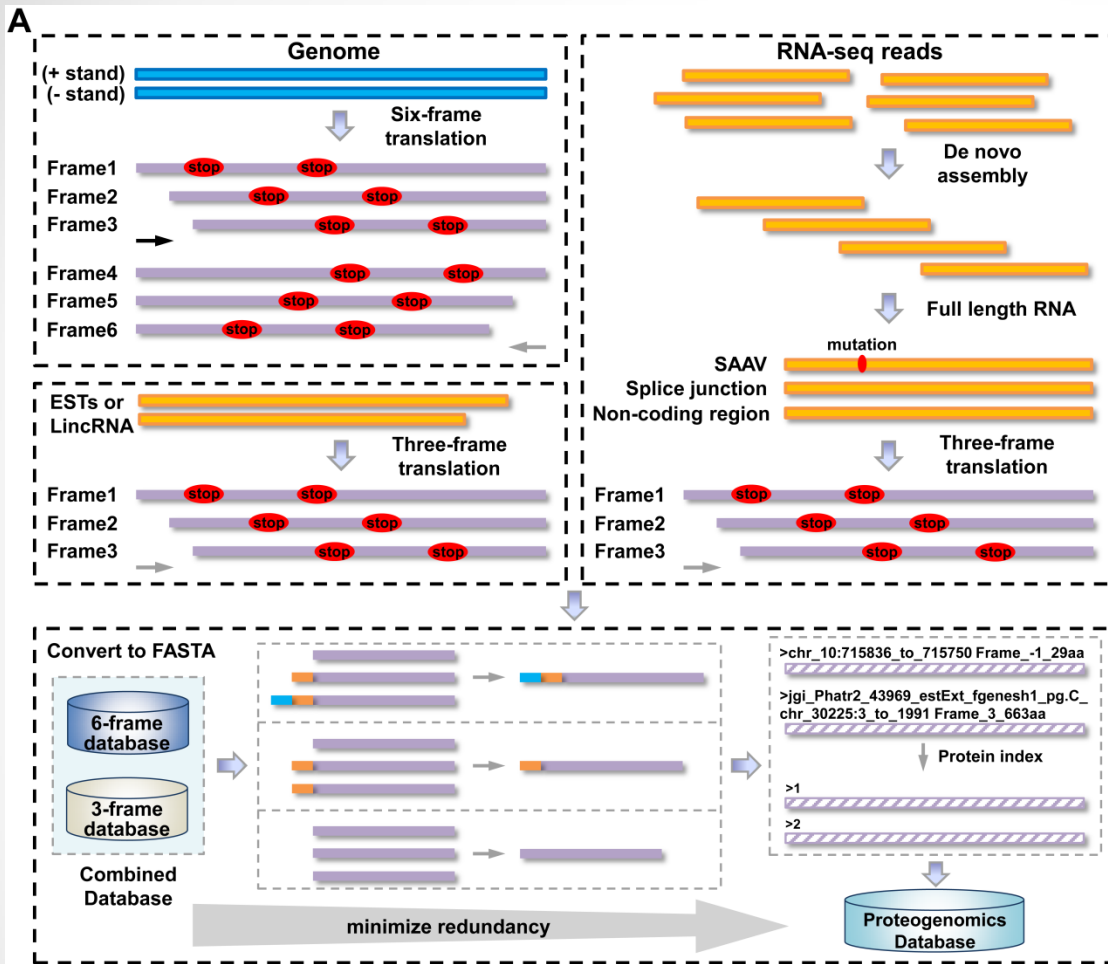
- 数据处理；

Pipeline



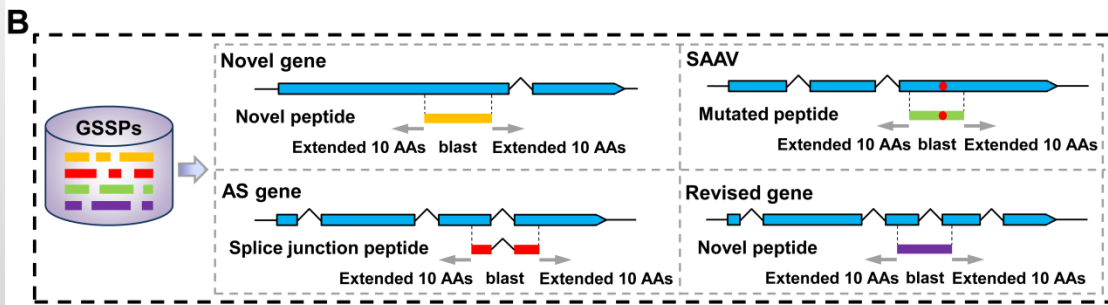
- 建立了一套适用于真核生物蛋白质基因组学研究和分析的流程。

生物信息学分析策略

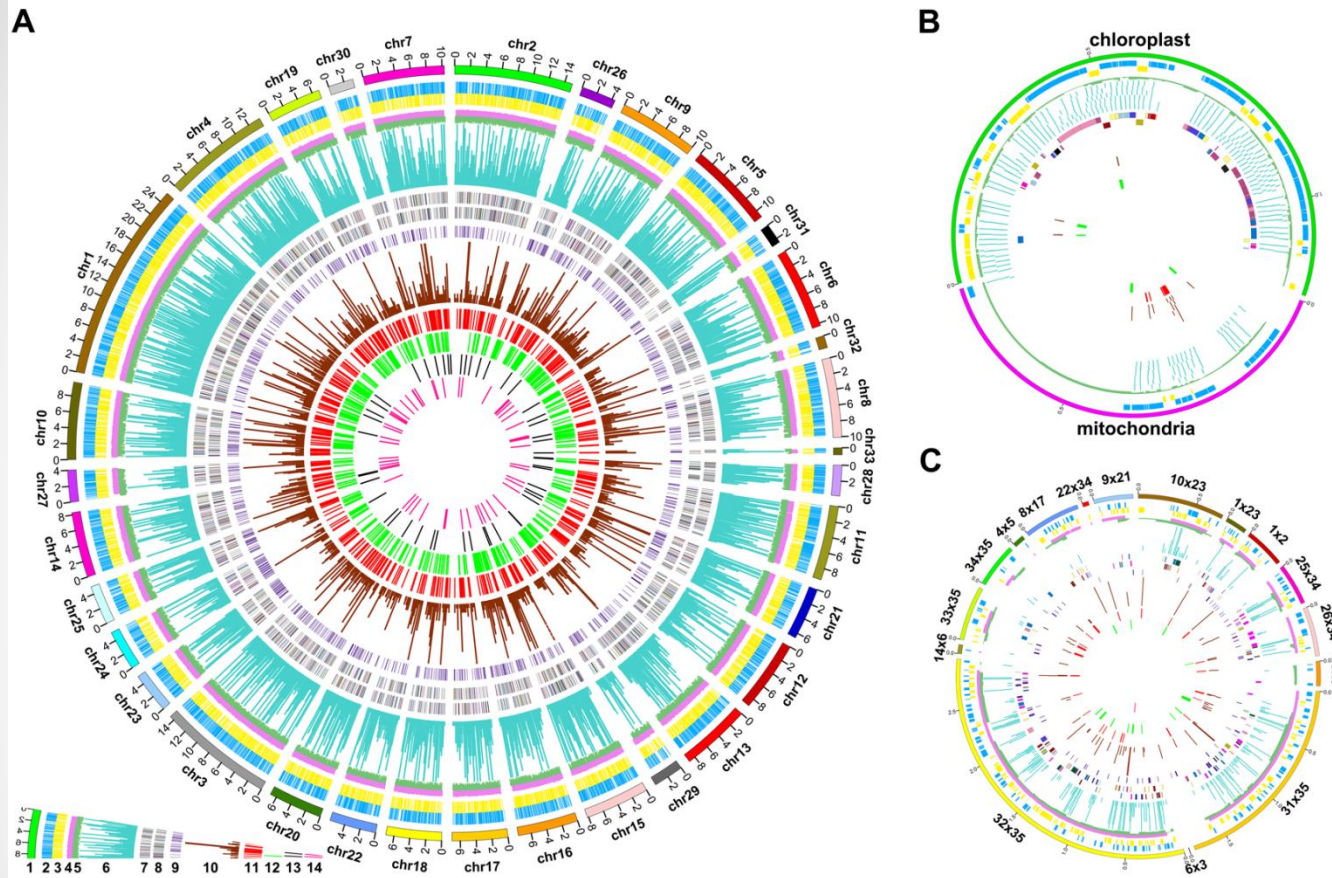


➤ 建库与缩减数据库：
整合基因组、转录组、ESTs序列等多组学数据，并对数据库进行了缩减，最终获得非冗余的蛋白质基因组学数据库；

➤ 基因注释：
严格的新基因发现方法，可变剪切体与点突变从头发现算法。



三角褐指藻蛋白质组精细图谱



1: chromosomes; 2: genes on plus strand; 3: genes on minus strand; 4: GC (>50%); 5: GC (<50%); 6: identified peptides mapping to existing proteins; 7: KOG annotation of existing proteins (plus strand); 8: KOG annotation of existing proteins (minus strand); 9: LncRNA; 10: novel peptides; 11: novel genes; 12: Revised genes; 13: alternative splicing genes; 14: single amino acid variants (SAAVs).

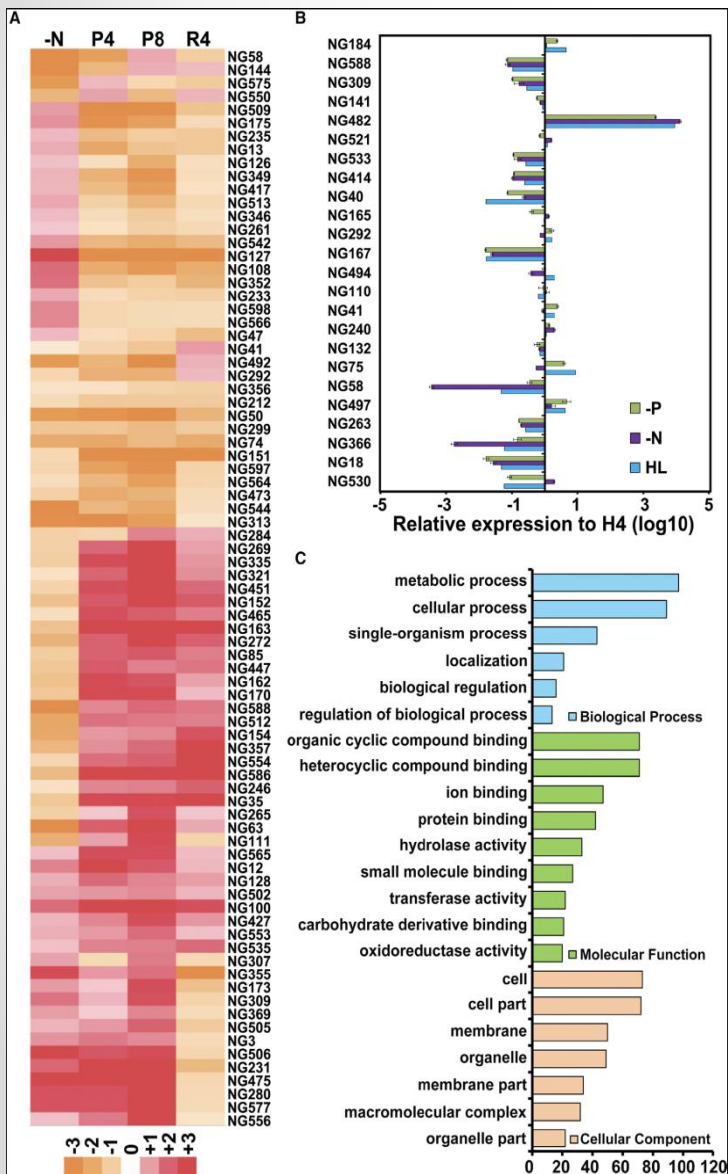
- 鉴定到**6628**个已注释的编码基因；
- 发现了**606**个新的蛋白编码基因，**56**个新发现的蛋白编码基因，在之前的研究中被预测为长链非编码RNA；校正了**506**已注释的编码基因结构；
- 鉴定到**21**个新的可变剪切体，修正了**73**个已注释基因的可变剪切位点；发现了**58**个发生氨基酸突变的基因；
- 鉴定到 **268**个微小短肽（micropeptides）。

基于转录组与合成肽段的方法验证新基因

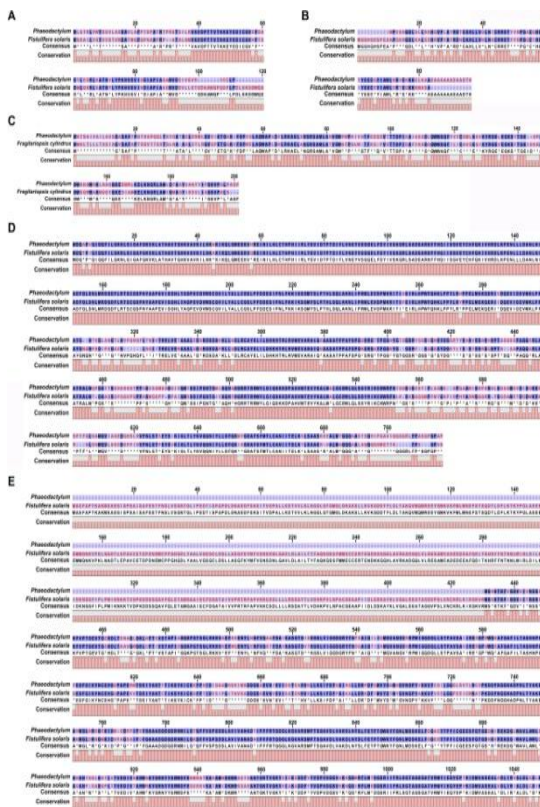
转录组分析

qPCR分析

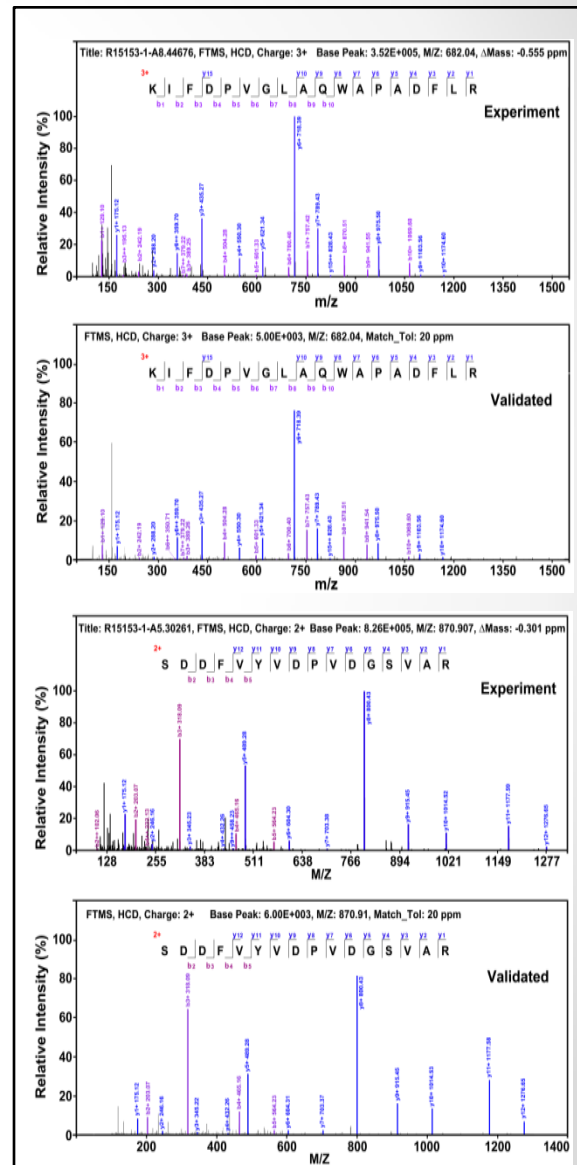
合成肽段验证新基因peptide



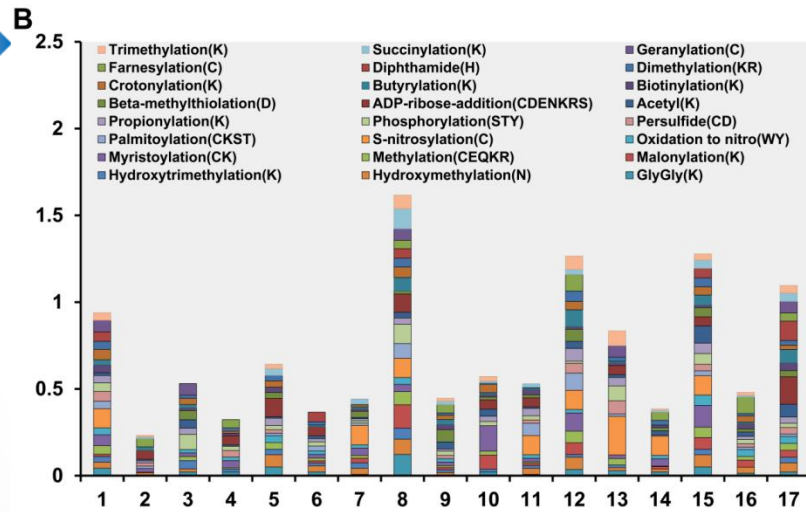
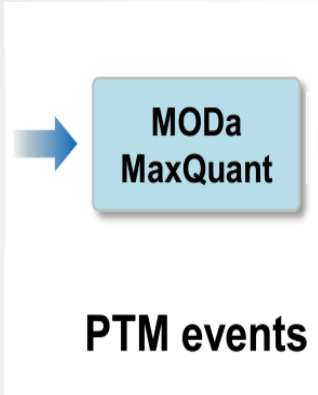
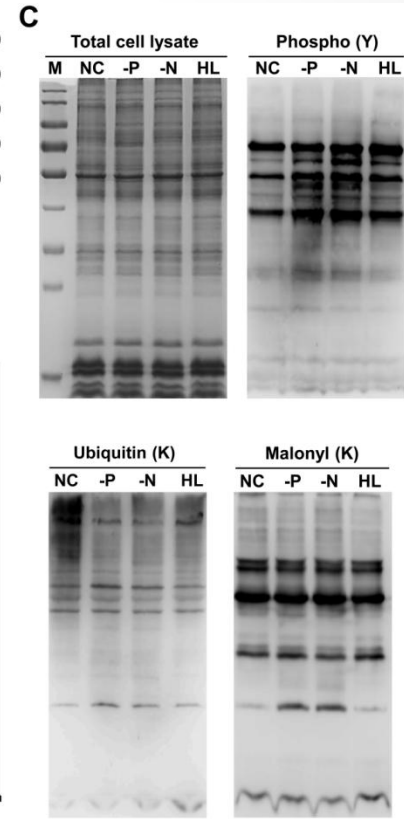
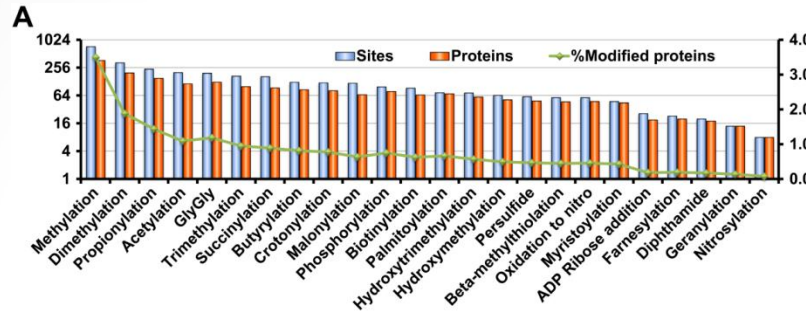
同源比对



GO 功能注释



全蛋白质组水平发现蛋白翻译后修饰

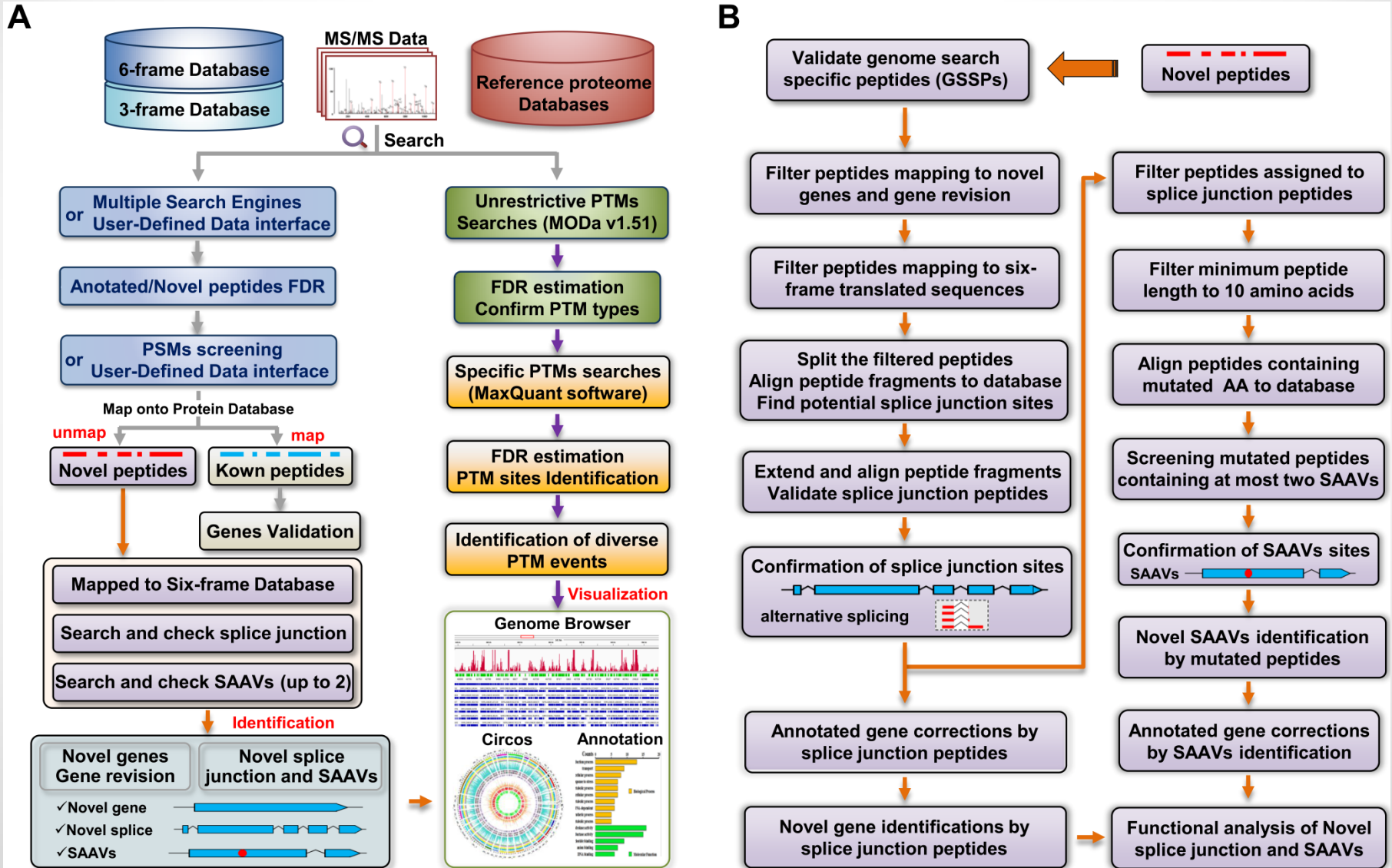


1: Energy production and conversion; 2: Cell cycle, cell division, chromosome partitioning; 3: Amino acid transport and metabolism; 4: Nucleotide transport and metabolism; 5: Carbohydrate transport and metabolism; 6: Coenzyme transport and metabolism; 7: Lipid transport and metabolism; 8: Translation, ribosomal structure and biogenesis; 9: Transcription; 10: Replication, recombination and repair; 11: Cell wall/membrane/envelope biogenesis; 12: PTMs, protein turnover, chaperones; 13: Inorganic ion transport and metabolism; 14: Secondary metabolites biosynthesis, transport and catabolism; 15: General function prediction only; 16: Function unknown; 17: Signal transduction mechanisms.

建立了完整的蛋白质基因组学研究和分析流程,可适用于各种真核生物

The pipeline and its source code are freely available at:

<https://sourceforge.net/projects/gapeproteogenomic>



模式硅藻的蛋白质组精细图谱

实现了三角褐指藻基因组的深度注释，并构建了蛋白质组精细图谱，建立了完整的适用于各种已经测序的真核生物蛋白基因组学分析流程。研究成果以封面论文的形式发表于Molecular Plant杂志。



- 精准鉴定到6628个已注释的编码基因；
- 发现606个新的蛋白编码基因；
- 校正了506个已注释的编码基因，其中有56个新发现的蛋白编码基因，在之前的研究中被错误预测为长链非编码RNA（LncRNA）；
- 鉴定到 268个可能具有重要功能的微小短肽；
- 21个新的可变剪切体；
- 修正了73个已注释基因的可变剪切位点；
- 58个发生氨基酸突变的基因；
- 全蛋白质组水平发现蛋白的翻译后修饰；
- 这些修饰参与众多的生物学过程；
- 建立了适用于真核生物的蛋白基因组学分析流程.....**

小结

CellPress
PARTNER JOURNAL

Molecular Plant
Resource Article

Genome Annotation of a Model Diatom *Phaeodactylum tricornutum* Using an Integrated Proteogenomic Pipeline

Mingkun Yang¹, Xiaohuang Lin^{1,2}, Xin Liu^{1,2}, Jia Zhang¹ and Feng Ge^{1,2,*}

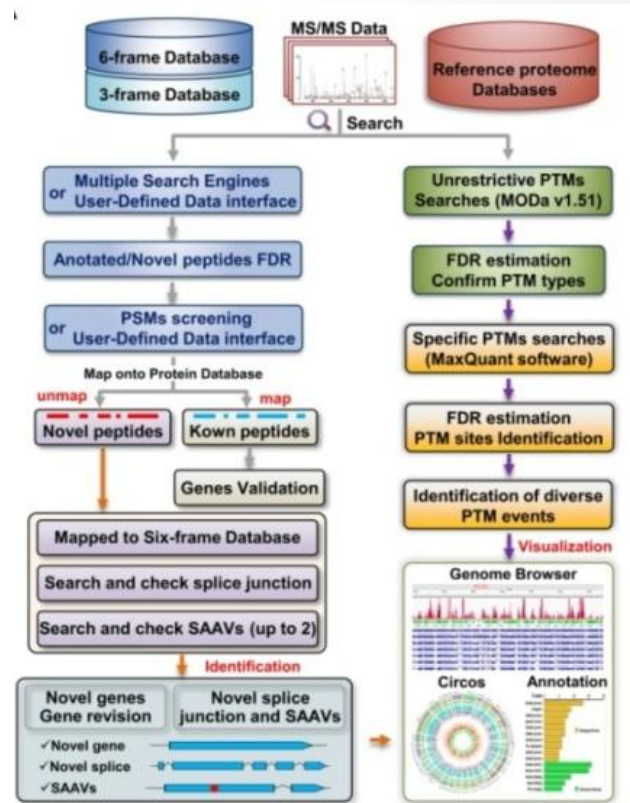
¹Key Laboratory of Algal Biology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China

²University of Chinese Academy of Sciences, Beijing 100039, China

*Correspondence: Feng Ge (gefeng@ihb.ac.cn)

<https://doi.org/10.1016/j.molp.2018.08.005>

pAnno: A proteogenomic software for eukaryotes



Molecular & Cellular Proteomics
MCP
About MCP | This Article | Submissions | Subscriptions | Contact

Research

© 2016 by The American Society for Biochemistry and Molecular Biology, Inc.
This paper is available on line at <http://www.mcponline.org>

GAPP: A Proteogenomic Software for Genome Annotation and Global Profiling of Post-translational Modifications in Prokaryotes*

Jia Zhang[†], Ming-kun Yang[†], Honghui Zeng[§], and Feng Ge^{‡§¶}

Proteogenomic analysis and global discovery of posttranslational modifications in prokaryotes

Ming-kun Yang^{†,1}, Yao-hua Yang^{†,1}, Zhuo Chen[¶], Jia Zhang[¶], Yan Lin[¶], Yan Wang[¶], Qian Xiong[¶], Tao Li^{‡,2}, Feng Ge^{¶,2}, Donald A. Bryant[¶], and Jin-dong Zhao[¶]

¹Key Laboratory of Algal Biology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China; and ²Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA 16802

Edited* by Jiayang Li, Chinese Academy of Sciences, Beijing, China, and approved November 13, 2014 (received for review July 6, 2014)

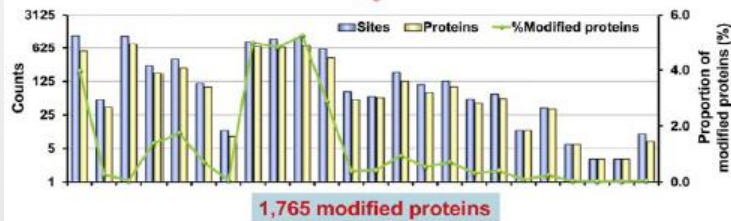
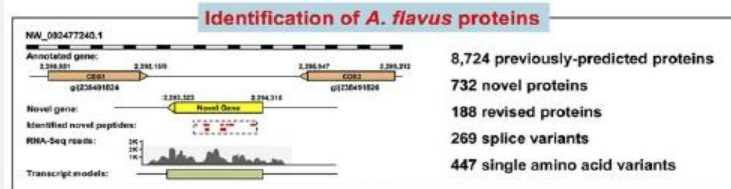
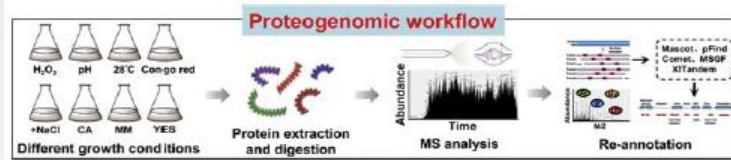
We describe an integrated workflow for proteogenomic analysis and cyanobacterial cells adjust their cellular activities in response to a wide range of environmental cues and stimuli. Recently, and global profiling of posttranslational modifications (PTMs) in

我们建立了完整的蛋白质基因组学研究和分析流程,适用于各种已经测序的物种,应当成为基因组测序和注释相关项目的一项标准流程。

黄曲霉菌的蛋白组精细图谱

MCP RESEARCH

Proteogenomic Characterization of the Pathogenic Fungus *Aspergillus flavus* Reveals Novel Genes Involved in Aflatoxin Production



VIRULENCE
2020, VOL. 12, NO. 1, 96–113
<https://doi.org/10.1080/21505594.2020.1859820>

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

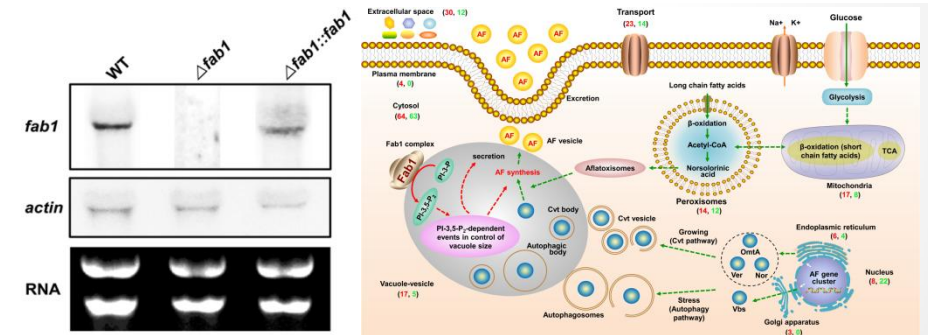
OPEN ACCESS Check for updates

A novel phosphoinositide kinase Fab1 regulates biosynthesis of pathogenic aflatoxin in *Aspergillus flavus*

Mingkun Yang^{a,b}, Zhuo Zhu^a, Youhuang Bai^a, Zhenhong Zhuang^a, Feng Ge^b, Mingzhu Li^a, and Shihua Wang^a



Antioxidant-related catalase CTA1 regulates development, aflatoxin biosynthesis, and virulence in pathogenic fungus *Aspergillus flavus*

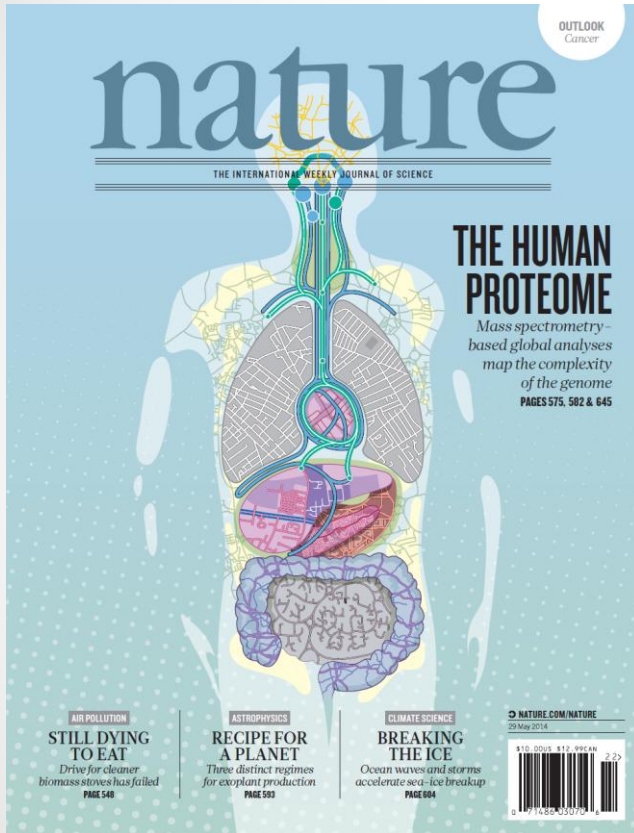


➤ 新基因验证与功能研究

- 鉴定到8724个已注释的蛋白；发现了732个新的蛋白编码基因；校正了188个已注释的编码基因结构；鉴定到88个新的可变剪切体和447个发生点突变的基因。

蛋白质基因组学分析面临的挑战

人类蛋白质组草图



These results suggest that the two studies are substantially overestimating the number of protein coding genes they identify. We conclude that the experimental data from these two studies should be used with caution.

J Proteome Res. 2014 Aug 1;13(8):3854-5.

假阳性率高！

Thus, our results suggest that the current drafts of the human proteome are far from complete, and that the data have to be used with caution especially in terms of personalized medicine.

J Proteome Res. 2015 Feb 6;14(2):1330-2.

假阴性率高！

The findings include an abundance of poor spectra, low-scoring peptide-spectrum matches and incorrectly identified proteins in both these studies, highlighting clear issues with the application of false discovery rates.

Expert Review of Proteomics. 2015, 12(6), 579-593.

质谱数据质量差！控制假阳性标准低！

Total citations: 769

如何快速精准解析

pAnno 2.0 的设计

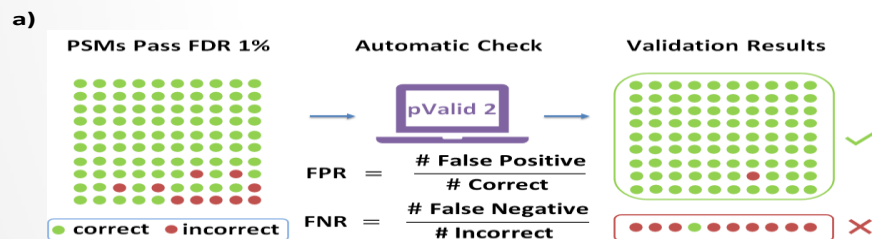
✓更快

	文件大小	运行时间
DNA序列文件	33.6M	3.7s
RNA序列文件	100.7M	5.0 s
已注释蛋白文件	6.4M	0.2s
总库（去冗余）	107.4M	65.2s
写入文件	61.3M	0.8s

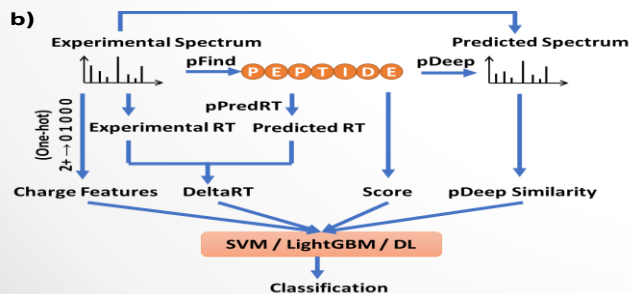


- GAPE建库需要数小时，完成基因注释需要1-2天；改进后建库只需不到2min，注释过程不超过10min；

✓更准



- 建立新肽段评估体系：pValid 2



致谢

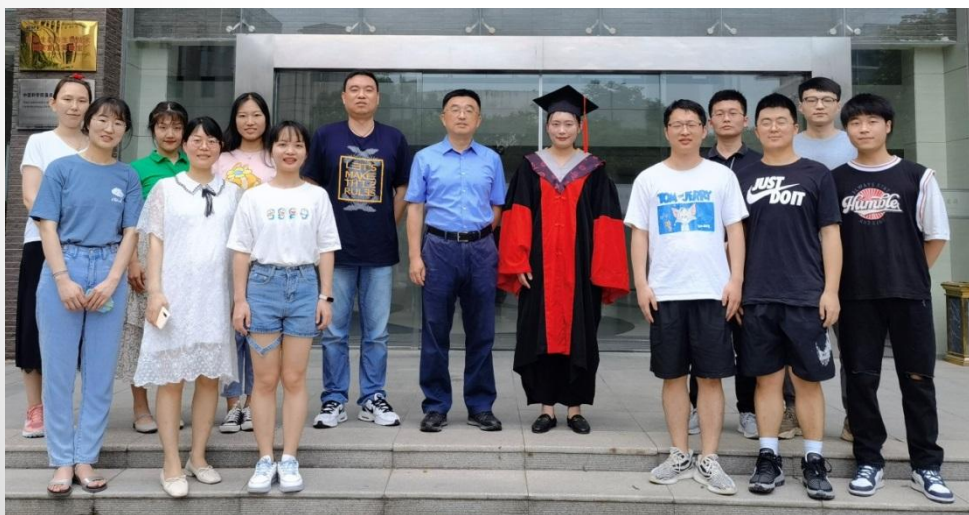
➤ 感谢贺老师团队的支持

贺思敏老师 迟浩老师

卫卓虹 硕士
周文婧 博士生
王凯菲 硕士生



➤ 水生生物功能蛋白质学学科组



研究方向：

- (1) 重要模式生物的蛋白基因组学研究；
- (2) 蛋白质翻译后修饰组学及其分子调控机制。

联系方式：中国科学院水生生物研究所
2号实验楼1009房间

电话：027-68780730

邮箱：yangmingkun@ihb.ac.cn

谢谢！
请指正！

中科院水生所

中科院水生所