IT Innovations for Better Life

以創新資訊技術來增進生命品質

林 仲彥

June 9, 2015

LAB OF System Biology & Network Biology
中央研究院資訊科學研究所 @iis, Academia Sinica, TAIWAN
系統生物學與網路生物學實驗室

中央研究院 資訊科學研究所 Institute of Information Science Academia Sinica

國家衛生研究院 National Health Research Institutes

The Pain-points for Current NGS Research

由於新世代分析技術的精進，使得整個基因體序列資料解析，得以實現，除癌症醫學外，也將應用於其他臨床檢驗與生農醫藥的研究上。然而，隨著定序技術的快速演進，資料的快速成長，與多維度資料的整合及可視化，都變成當今資訊技術的一大挑戰，對於傳統的生物醫學研究單位來說，這樣的數位障壁，已變成新一代生醫農學研究人員的首要面臨課題。這些因資料巨量畫所形成的**數位高壁**，可分述如下數點：

生物序列資料快速成長

- 以當前最新的機種illumine Hi-seq 2500定序儀來說，一次的定序便會有1.2 TB左右的序列檔案產出
- 如分析一般數量族群樣本，原始資料將以TB等級計算
- 原始檔案與分析過程及結果之大量中繼檔案龐大處理困難
- 隨高速定序儀的技術發展，同一樣本的產出量會遞年增加

巨量資料管理能力不足

- 大量資料遠端傳輸（如600GB序列資料）耗時且不穩定
- 無妥善資料擷取、再利用、管理與備份的機制

缺乏巨量資料運算平台與生物資訊人員支援

- 隨資料的增加，既有計算平台不敷使用
- 亦無專業IT人員的支援
- 分析工具軟硬體的需求偏高，安裝不易
- 如無優化程式或是引入高速平行化計算架構，計算資源耗用嚴重

需要不同研究領域專門知識

- 不同生物加值資料庫的引入與更新不易
- 專業生物資訊人員不足，無法引導分析方向
- 新的分析方法往往建置不易

多面向的分析結果難以判讀

- 沒有引入統計分析，或是統計方法的使用錯誤，造成結果無法判讀
- 文字型分析結果，缺乏多維度可視化圖像分析

Steps for NGS Data Analysis

數據前處理
- 定序品質校正

初步分析
- 進行序列重組或是比對參考基因體序列

進階分析
- 序列功能性解析
- 基因表現比對

Primary Analysis
- Base calling/ Sequence trimming

Secondary Analysis
- Assembly or Ref mapping

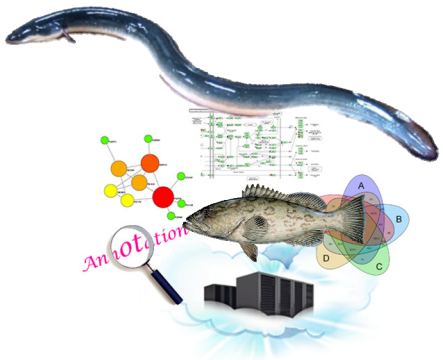
Tertiary Analysis
- Calculate Mapping data/ expression profile
- Functional inference

Penaeus Genome Database (PAGE)

LAVENDER: Litopenaeus vannamei Genome Database

研究範疇
Research Topics

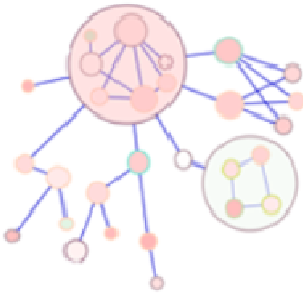
水產生物基因體與轉錄體解析及其調控機制



Chung-Yen Lin Ph.D. (林仲彥)

研究重點:
基因體研究與新世代高通量生物數據之整合分析

大規模生物網路拓樸分析:偵測重要關鍵因子與次網路



個人化醫學研究:
以次世代定序找尋致癌之融合基因與重要變異

生物資訊資料庫與線上分析平台之開發

PhyOgenetic WEb Repeater

Penaeus Genome Database (PAGE)

UPS Unique Probe Selector

EVIDENCE Intervirus in Deep Conception

PDA Primer Design Assistant

PALM Phylogenetic reconstruction by Automatic Likelihood Model selector

MyBLAST

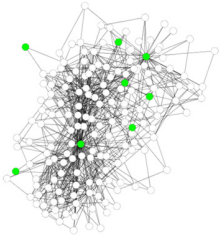
研究成果(2011-2015)

複雜生物網路拓樸解析 (與高明達老師合作)

① Spotlight：整合多種網路拓樸特性，偵測生物網路中潛藏的蛋白質複合體 (***Gene***, 2013)

② Hubba / Cytohubba: 以拓樸演算法鑑別出複雜生物網路中的重要樞紐與關鍵次網路 (***NAR***, 2008, ***BMC Bioinformatics***, 2011, ***BMC Systems Biology***, 2014)

- **Download >7500 times**
- **Scopus citations> 100 times**
- **Google Scholar> 120 times**



生物序列高通量基因表現與調控網路分析

① 日本鰻發育前期之轉錄體變化 (***PLoS ONE***, 2015)

② 不同光照波長下沈香轉錄體之調控(***BMC Plant Science***, 2015)

③ 溫度變化下珊瑚與共生藻之轉錄體調控 (***Molecular Ecology***, 2014)

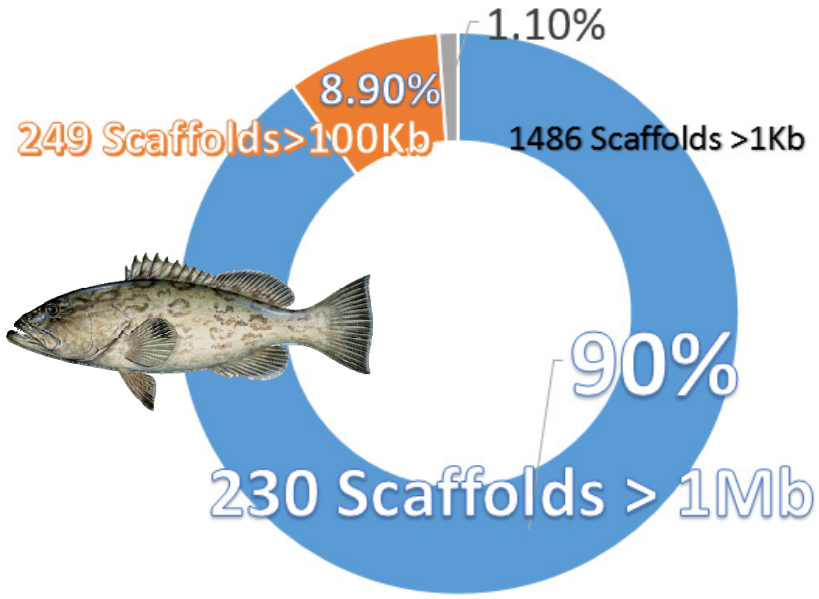
④ 脊索與頭索動物全轉錄體分析(***Marine Genomics***, 2014, ***Briefings in Functional Genomics***, 2012)

⑤ 台灣食葉飛鼠之腸道菌相基因群功能解析(***BMC Genomics***, 2012)

⑥ 對蝦類基因體與轉錄體之資料庫建置與功能註解分析(***Marine Biotechnology***, 2011)

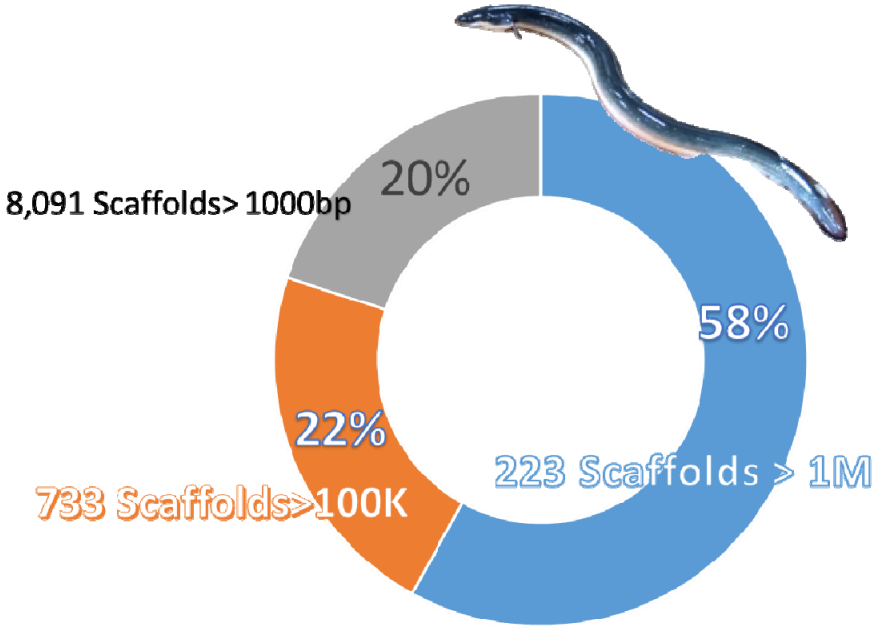
研究成果 (2013-2015)

① Grouper Genome Project



Genome size: 1.06G
N50: 5.1M base
Coverage: 140x

② Japanese Eel Genome Project

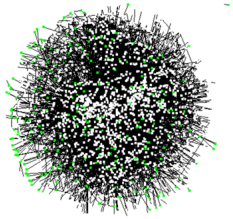


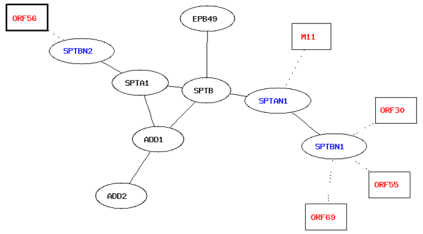
Genome size: 1.20G
N50: 1.6 M base
Coverage: 150x

研究成果(2011-2015)

① **以系統生物學策略分析病毒感染機制** (Virus-host interactomics/ transcriptomics: from network biology to bench research)

- 與UCLA腫瘤病毒研究室合作，分析流感、B/C型肝炎與乳突病毒等用藥前後病毒序列變異與抗藥性之關係，及與人類細胞之交互動態網路，還有鑑別出可供新藥物設計使用之高保守區域 (***PLoS Pathogens***, 2014, ***Scientific Reports***, 2014, ***mBio***, 2014)

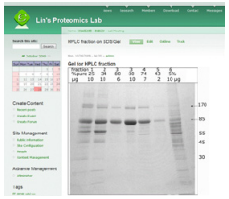




UCLA

② **發展電子實驗記錄本**(Developing Electronic Laboratory Notebook (ELN) for Research Community for Knowledge Management)

- **微軟公司贊助提供MS AZURE雲端平台兩年無償使用權 (2011-2013)**
- 已研發各式平台之電子實驗室記錄本自動安裝程式，並有多國語系版本，live-DVD等，雲端平台之多實驗室管理介面研發中。



Windows Azure

實驗室電子記錄本

ELECTRONIC LAB NOTEBOOK

Elegance
Electronic Lab Notebook

實驗室電子記錄本

Electronic Lab Notebook

雲端連線研究合作
Access and collaborate via Cloud

實驗室研究智慧數位化
Experimental Records and Wisdom

不包含
Not included

Elegance Electronic Lab Notebook

重要功能
Essential Functions for ELN

- Friendly Installation/ non installation
- Content generator
- PDF printout
- User management
- Search
- Web Access worldwide
- Calendar/ Event
- Webpage clip
- Web share
- Succinct control panel
- Image gallery
- Backup /restore
- Digital signature
- Security
- Personalization
- Print
- Data Exchange

BioLinux Windows Azure

World Cloud Research Collaboration Project hosted by Microsoft Research, 2011~ 2013，微軟公司贊助

ELN on QNAP NAS (Coming soon)

Storage space: 6-8 Tb

Elegance

Electronic Lab Notebook

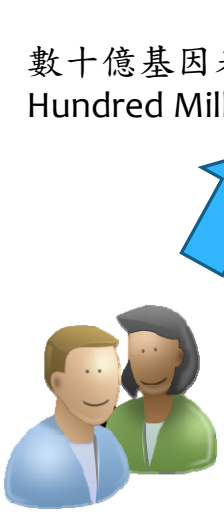
Develop Platforms for Omics Studies

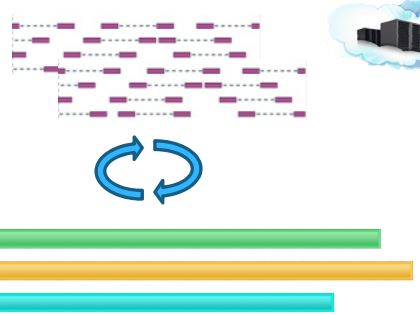
發展基因體與轉錄體等多重體學之線上分析平台

基因體與轉錄體等多重體學之線上分析平台

Multi-Omics onLine Annotation System (MOLAS)

數十億基因表現定序短片段
Hundred Millions Reads in RNA-seq





序列組裝與註解
Contigs Assembly/ Annotation

MOLAS

Transcriptome
Methylome
Interactome
Genome
Network Biology
Metabolome

多重體學資料整合

全文檢索與序列比對/
Full Text search and BLAST

BLAST Result

Program	Database	Job Note	Query	Blast Result (text)	Blast Result (csv)
blastn	Pen contig				

Advance Option

Max target	E-value	Match/Mismatch penalty	Gap creation	Word size	Complexity filter
10	10	2, -3	Linear	11	Turn on

Show Hits per Sequences by Rank Top 3 (Strand = Both / Both)

Seq	Rank	Hits	Length	Score	Bits	E-value	Identities	Strand
your sequence name	1	6687_TUS_4385	469	26.5	28	7.6e-1	14 / 14 (100%)	Plus / Plus
your sequence name	2	6687_TUS_3060	586	26.5	28	7.6e-1	14 / 14 (100%)	Plus / Plus
your sequence name	3	6687_TUS_3024	700	26.5	28	7.6e-1	16 / 17 (94%)	Plus / Plus

以多重資料庫對序列進行註解/
Contigs annotated by SignalP, NR, GO, Interpro and KEGG

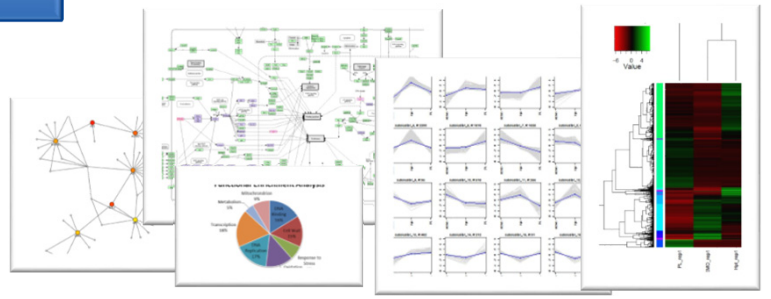
Contig information

GO the Gene Ontology

Pfam

HMMER

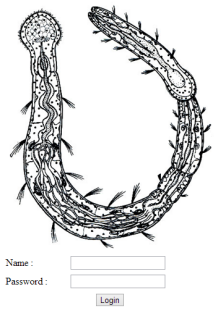
功能性解析與動態網路分析/
Functional Inference, Dynamic Network Analysis



以MOLAS作為轉錄體解析與線上資料庫的快速建置平台

MOLAS as Rapid and Handy Platform for Transcriptome Analysis and Database Construction

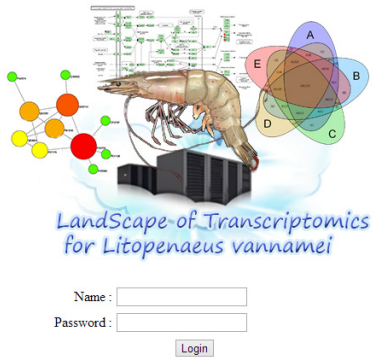
蠕蟲頭部再生機制研究



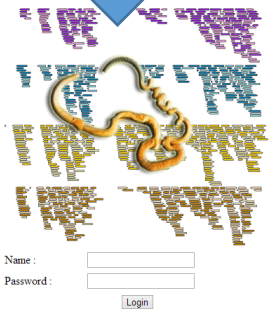
重組序列
Assembly

重組序列與其基因表現
Assembly + Expression

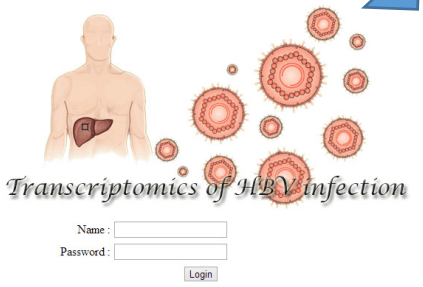
白蝦轉錄體與抗病機制之探討



日本鰻幼體營養需求與優良種鰻之鑑別



半索動物發育機制研究



肝炎病毒致病機制之研究

早發型乳癌基因表現分析

完整系統架構/ Our Goal of Whole System

External Databases

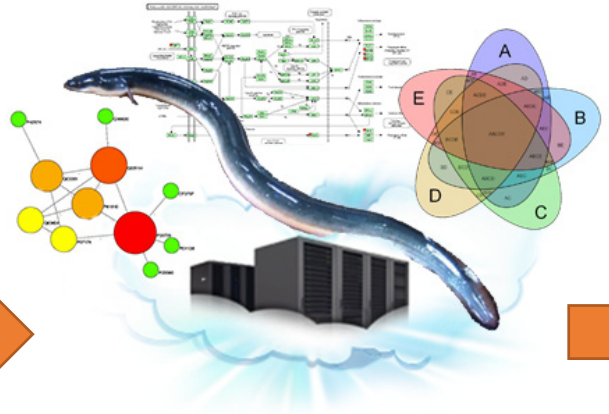
UniProt

NCBI

GO the Gene Ontology

KEGG

Pfam



Omics Database for Specific Organism

Approach

- Transcriptome
- Genome
- Methylome
- Network Biology
- Proteome
- Eel biology/ physiology

Infrastructure

MOLAS

Galaxy

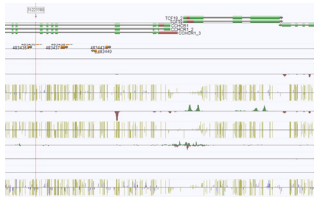
Windows Azure

amazon web services™

Our own Platforms/Cloud Computing/ Data Storage

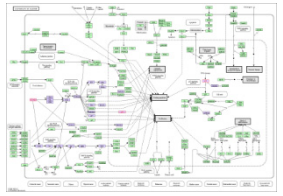
Basic functions of Omic Database

Browser for Genome/ Methylome/ Transcriptome/ etc

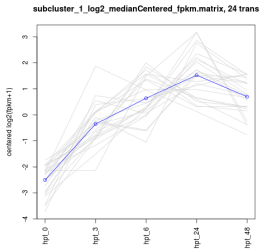


Sequence Annotation

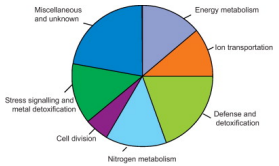
Pathway Analysis



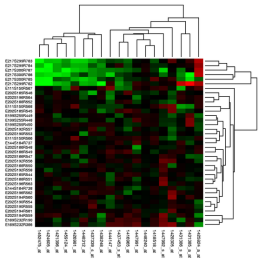
Expression Pattern Clustering



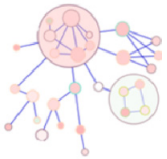
Functional Enrichment



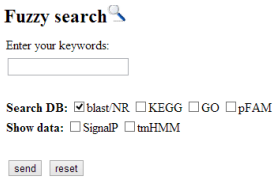
Heatmap



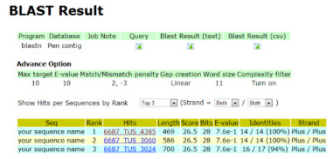
Protein Network Analysis



Full text Search



Similarity Search



全基因體甲基化雲端分析平台

Genome-wide DNA Methylation Analysis on Cloud

與植微所陳柏仰老師合作

Case A/ treatment

M

- - C G A T G A -
- - G C T A C T -

M M

VS.

Case B/ Control

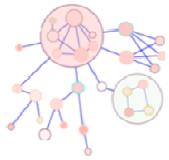
- - C G A T G A -
- - G C T A C T -

Bisulfite Converted Reads in Giga Byte Level

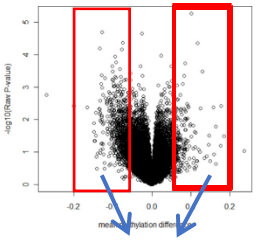
Our Approach on Cloud

① Align million reads on the reference genome
② Locate methylated sites
③ Identify differentially methylated regions
④ Functional enrichment analysis
⑤ Data Visualization

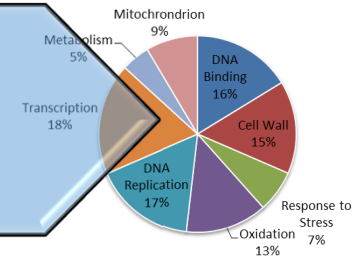
Possible Involved Pathways



Identify Differential Methylated Region (DMR) combined with Transcriptome



Functional Enrichment Analysis



Experiment Design, Data generation in NGS

Data Deluge Analysis by Cloud Computing

Identify the regulatory pathways affected by methylation

World Cloud Research Collaboration Project hosted by Microsoft Research, 2011~ 2013，微軟公司贊助

Windows Azure

Licensing to Taiwan Biotech Company
技轉授權國內生物科技公司使用

已於2014五月簽訂合作意向書
2015六月開始進行技術平台移轉

研究成果提供全球相關研究社群使用
Our Works for Research Community Worldwide

2003- 2015

➢線上資料庫 Web databases: 10

- ➢Database of interactome in *Helicobacter* (*hp*-DPI, Bioinformatics, 2005)
- ➢Protein Interactome of Fruit fly (flydpi, BMC Bioinformatics, 2006)
- ➢Shrimp Genome Database (PAGE, Marine Biotechnology, 2011)
- ➢AfterGenbank (http://interactome.nhri.org.tw/AfterGenbank, submitted)
- ➢Hemichordate Transcriptome Database (Marine Genomics, 2014)
- ➢Evidence (EnteroVirus in Deep Conception, submitted, 2015)
- ➢…….

➢線上分析工具平台 Web Applications: 10

- ➢Primer Design Assistant (PDA) (NAR, 2003)
- ➢Phylogenetic Web Repeater (POWER, NAR, 2005)
- ➢Unique Probe Selector (UPS, BMC Bioinformatics, 2008, 2010)
- ➢Hub Object Analyzer (Hubba, NAR, 2008)
- ➢Phylogenetic reconstruction by Automatic Likelihood Model selector (PALM, PLoS ONE, 2009)
- ➢Assembly protein complex from Interactome (Spotlight, Gene, 2013)
- ➢Cytohubba (BMC Systems Biology, 2014)
- ➢myBLAST (Submitted, 2014)
- ➢MOLAS (Preparing)
- ➢ELN (Preparing)

Usage:

Submit Jobs: >230,000 times

Processed data: over 1,600,000 sequences

Agenda Today

Time	Topic/ Speaker
9:40 - 10:20	**次世代高通量基因組測序 - NGS平台原理和實驗設計的考量**/ Next-Generation Sequencing – principles of NGS platforms and experimental considerations Mei-Yeh Lu Ph.D., 呂美曄博士，中央研究院定序中心
10:30 - 11:10	**次世代定序之線上基因概況分析平台**/ Multi-Omics onLine Analysis System (MOLAS) Shu-Hwa Chen Ph.D., 陳淑華 博士，中央研究院資訊科學研究所
11:20 - 12:00	**電子實驗室記錄本**/Elegance: Electronic Lab Notebook-- Digitize your experimental designs and results into wisdom from Discovery to Publication 黃智偉 先生 (Mr. Chi-Wei Huang)，中央研究院資訊科學研究所

Thanks for your Attention

LAB OF System Biology & Network Biology
中央研究院資訊科學研究所 @iis, Academia Sinica, TAIWAN
系統生物學與網路生物學實驗室