CNCP2010

从新一代测序技术的基因组学到基于质谱仪的蛋白质组学

华大基因的生物信息学

From NGS Genomics to MS-based Proteomics

BGI's bioinformatics activities

ZHANG Yong (zhangy@genomics.org.cn)

BGI

November, 2010, Beijing

BGI - one of the Biggest Genome Centers in the world.

Beijing
1999-now

Hangzhou
2001-now

Shenzhen
2007-now

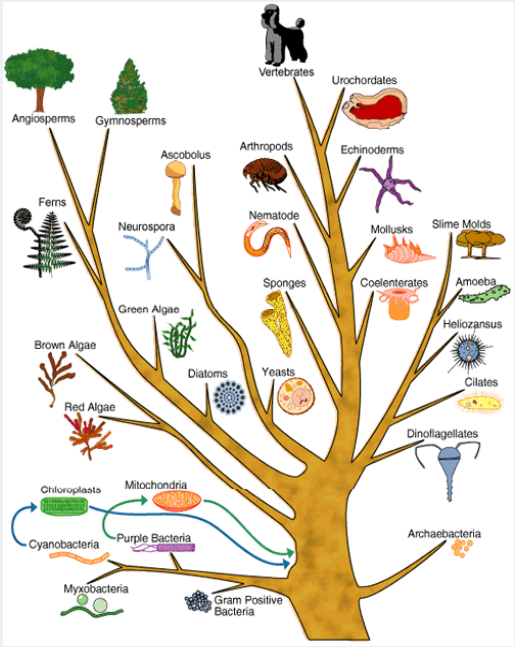
Wuhan
2010`-now

Hong Kong
2009-now

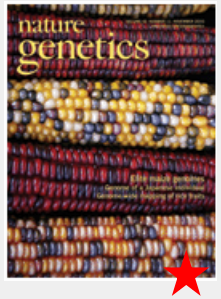
Our two major research directions

- Human health
  - Complex diseases
    - Metabolic disorders (type 2 diabetes, obesity, etc)
    - Cancer
    - Neurodegenerative disease
  - Personal genome sequencing

- Plant & animal genomes
  - Sequencing new genomes
    - Plant, animal
    - Bacteria, metagenomics (industrial, energy, pathogen, environmental, etc)
  - Molecular breeding
    - Crop
    - Livestock



Selected publications



Since 2009,
3 Science Papers
5 Nature Papers
3 Nature Biotech. Papers
3 Nature Genetics Paper
…

★ Papers in Shenzhen

Sequencing

- 1st generation sequencer
  - Sanger sequencer (capillary sequencer)
- 2nd generation sequencer
  - Illumina Solexa/GAII/HiSeq2000
  - AB SOLiD4 and Roche 454
- 3rd generation sequencer
  - Pacific Bioscience, … (Single molecular sequencer)

BGI current sequencing capacity

- 128 Illumina/HiSeq 2000
- ~~28 Illumina/Solexa~~
- 10 AB/SOLiD
- 6 AB/3730xl
- ~100 MegaBACE

- 200TB raw data per day;
- 900G base pairs per day;
- Data produced in 2008 is 8 times all data accumulated in Genbank till 2007.

Supercomputing Cluster

Beijing Site

Hong Kong Site

Shenzhen Site

102 TFlops
10,000 CPUs
20Tb Memory
10Pb Storage

- 800 Bioinformatians
  - Employee 职工
  - PhD students 学生
  - Young undergraduated students (a few) 本科
- Background
  - Physics, mathematics (30%) 物理数学
  - Biology, medical, biochemistry (30%) 生物、医学
  - Computer science, informatics (30%) 计算机、信息
  - Others (10%)
- Average age 平均年龄
  - 23

Bioinfo. Team

生物信息团队

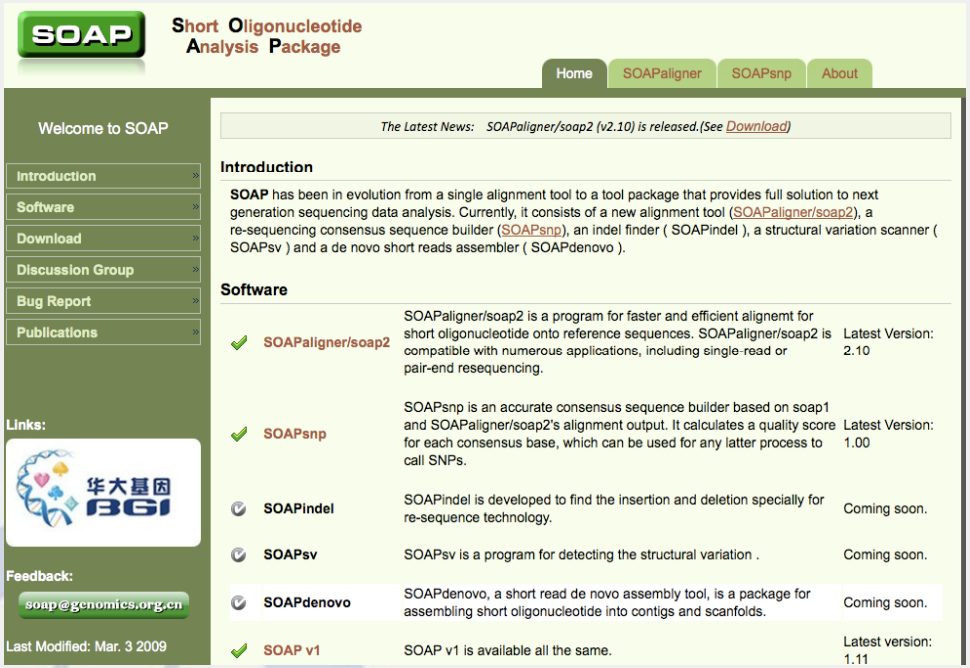
SOAP

- Short Oligonucleotide Alignment Program

- Website:

  http://soap.genomics.org.cn

- >10,000 users



Related publications

- SOAP:

Ruiqiang Li, Yingrui Li, Karsten Kristiansen, Jun Wang. SOAP: short oligonucleotide alignment program. **Bioinformatics**. **2008** 24: 713-714

- SOAP2:

Ruiqiang Li, Chang Yu, Yingrui Li, Tak-Wah Lam, Siu-Ming Yiu, Karsten Kristiansen, Jun Wang. SOAP2: an improved ultrafast tool for short read alignment. **Bioinformatics**. **2009**

- SOAPsnp:

Ruiqiang Li, Yingrui Li, Xiaodong Fang, Huanming Yang, Jian Wang, Karsten Kristiansen, Jun Wang. SNP detection for massively parallel whole genome resequencing. **Genome Research**. **2009**

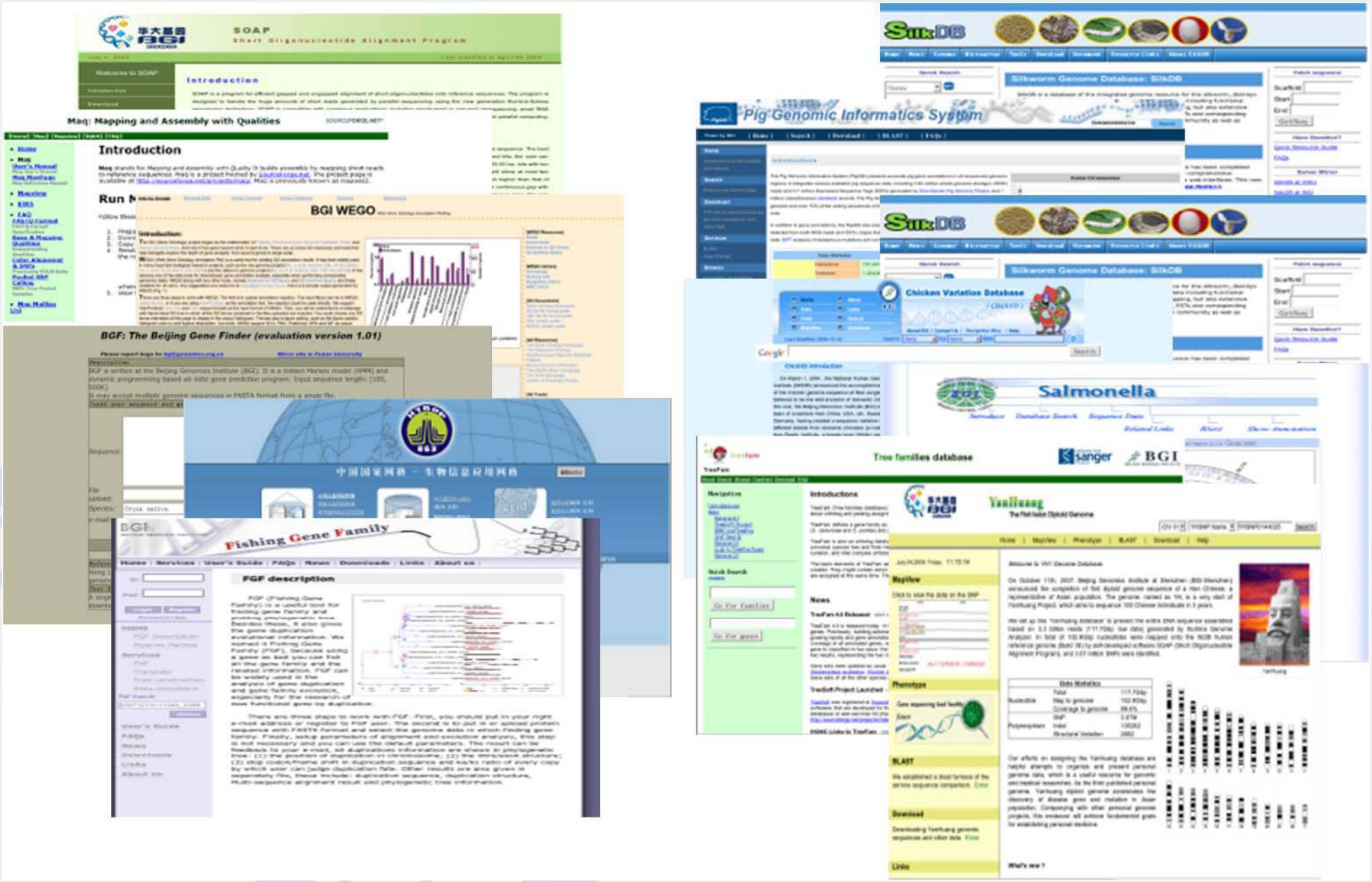
- SOAPindel, SOAPsv:

is coming …

- SOAPdenovo:

Ruiqiang Li, Hongmei Zhu, Jue Ruan, et al. De novo assembly of the human genomes with massively parallel short read sequencing. **Genome Research. 2009 (Accepted)**

The Resources:
Software and Database Development



DNA, RNA

- Short reads alignment
- Genome/transcriptome assembly
- Genome/functional annotation
- Mutant identification (including CNV, SV, etc)
- Transcriptome expression profiling
- Transcriptome fusion events
- smallRNA-pipeline

BGI's bioinformatics

- Huge amount of data
- Basic tools
- Pipeline to analysis large scale data
- Huge CPU, Memory usage
- A small group for Algorithm development

- Few software for advanced analysis
- Collaborated with the best people

Not only Genomics/Transcriptomics but also Proteomics

genetic central dogma

DNA<->RNA->Protein

But Proteomics is much more complex than genomics

BGI's activities?

传统研究模式 假说导向

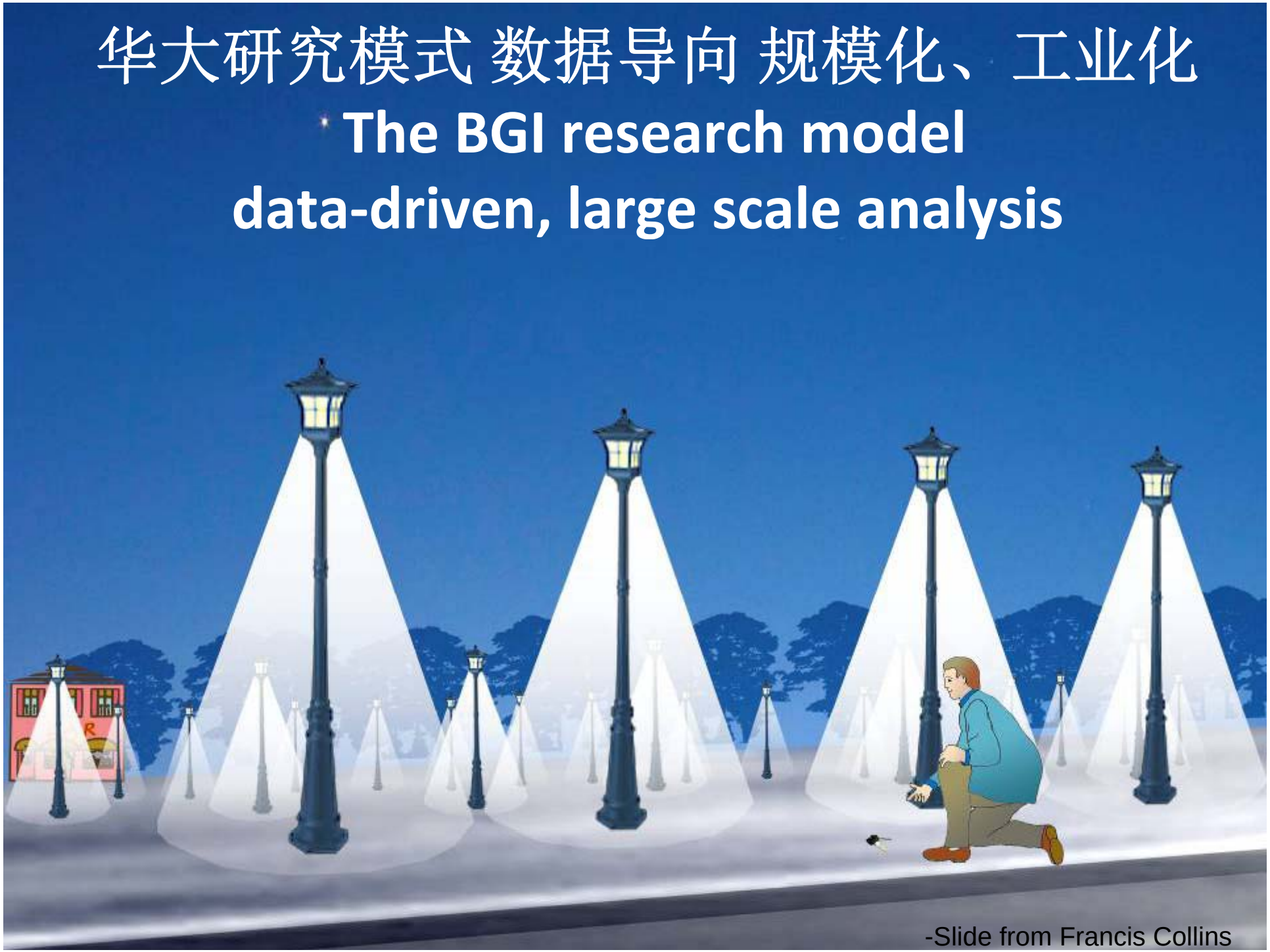
The candidate gene approach hypothesis-driven

-Slide from Francis Collins

华大研究模式 数据导向 规模化、工业化

The BGI research model

data-driven, large scale analysis



-Slide from Francis Collins

华大基因
BGI

All instruments are high-throughput, fast, cost-efficient.

Huge amount of data generated.

Supercomputing power and
800 bioinformaticians

MS-based Proteomics

- MS (Mass Spectrometry)

Orbitrap, Thermo Scientific

QTRAP 5500, AB Sciex

- ABSciex:
  - **Qtrap 5500**
  - **Qtrap 5600?**
- ABSciex:
  - MALDI-TOF/TOF
- Thermo Scientific:
  - **LTQ-Orbitrap**

- High-throughput , high accuracy and cost-efficient

Applications in BGI

- *de novo* genome assembly
- Whole genome resequencing
- Exome-capture sequencing
- Transcriptome
- Digital gene expression profiling
- MicroRNA
- Epigenome
- Metagenome/Metaproteome
- MS-based Proteome
- MS/NMR-based Metabolism

DNA Level

RNA Level

Epi

Meta-OMICS

Proteome

Metabolism

Proteomics Part

- Peptide/Protein Identification
- Quantitative Proteome
- MRM-development
- Systematic analysis combined with DNA, RNA, … information

Proteomics Focus

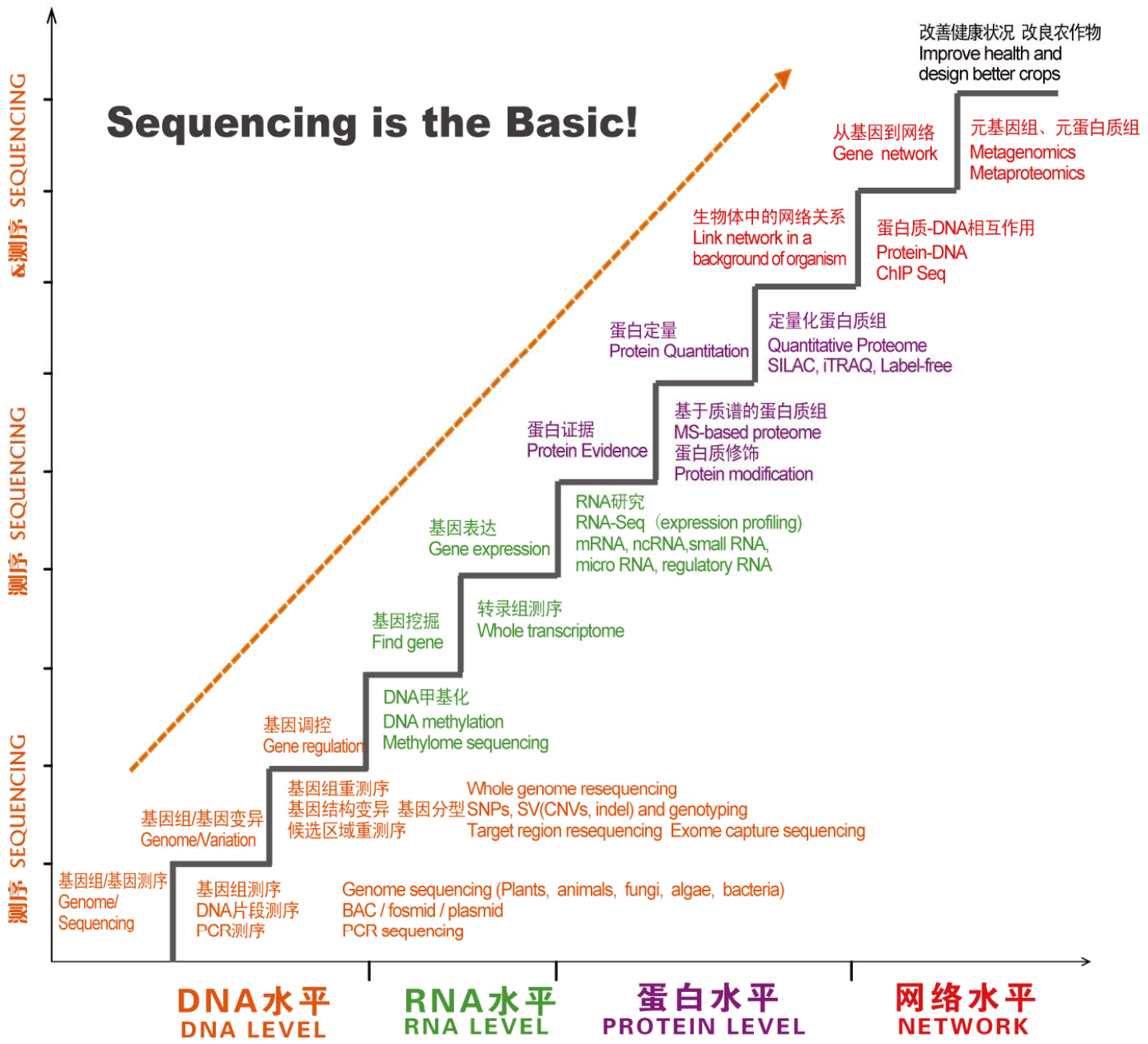
- Focus on MS data
- Upstream analysis than the further functional analysis
- Automatic analysis pipeline package
- Adapted to industrial usage/standard

Current Activities

- Peak recognition algorithm
- Analysis Package on Supercomputing System
  - Peptide/Protein (w mod.) identification
  - Quantitative proteomics
- Cloud computing platform
  - Integrate good softwares and provide webservice

Current Projects

- DNA, RNA, Protein
  - Correlation on three levels
  - Bring MS-proteome into all plant/animal species
- Target proteomics
  - smallRNA NGS + MRM for regulated proteins
  - Body fluid studies (Urine and Blood)
  - For Candidate gene (DNA sequencing + Proteome)
- Novel modification discovery


Sequencing is the Basic!
改善健康状况 改良农作物
Improve health and
design better crops
从基因到网络
Gene network
元基因组、元蛋白质组
Metagenomics
Metaproteomics
生物体中的网络关系
Link network in a
background of organism
蛋白质-DNA相互作用
Protein-DNA
ChIP Seq
蛋白定量
Protein Quantitation
定量化蛋白质组
Quantitative Proteome
SILAC, iTRAQ, Label-free
蛋白证据
Protein Evidence
基于质谱的蛋白质组
MS-based proteome
蛋白质修饰
Protein modification
基因表达
Gene expression
RNA研究
RNA-Seq（expression profiling)
mRNA, ncRNA,small RNA,
micro RNA, regulatory RNA
基因挖掘
Find gene
转录组测序
Whole transcriptome
基因调控
Gene regulation
DNA甲基化
DNA methylation
Methylome sequencing
基因组重测序
基因结构变异 基因分型
候选区域重测序
Whole genome resequencing
SNPs, SV(CNVs, indel) and genotyping
Target region resequencing Exome capture sequencing
基因组/基因变异
Genome/Variation
基因组/基因测序
Genome/
Sequencing
基因组测序
DNA片段测序
PCR测序
Genome sequencing (Plants, animals, fungi, algae, bacteria)
BAC / fosmid / plasmid
PCR sequencing
测序 SEQUENCING
测序 SEQUENCING
&测序 SEQUENCING
DNA水平
DNA LEVEL
RNA水平
RNA LEVEL
蛋白水平
PROTEIN LEVEL
网络水平
NETWORK

华大的生物信息

- 高精准、海量的数据是基础
- 高效能的大型计算机是保障
- 核心算法小组+流程开发与优化小组
- 自由交流的氛围保证了员工的快速成长

- 和最优秀的生物业内专家合作
- 和最优秀的生物信息算法专家合作
- 实战、实战、实战

华大基因
BGI

Thanks!
Welcome to BGI Shenzhen!

ZHANG Yong (zhangy@genomics.org.cn)