上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

The rice pan-genome: gene presence/absence variation derived from >3000 rice genomes

Chaochun Wei （韦朝春）

Department of Bioinformatics and Biostatistics

School of Life Sciences and Biotechnology

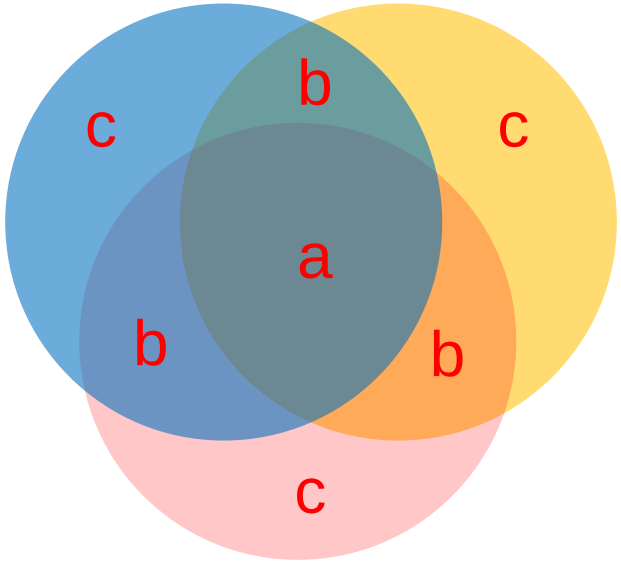
Shanghai Jiao Tong University

Korea, Japan, China Bioinformatics Symposium

June 21, 2017

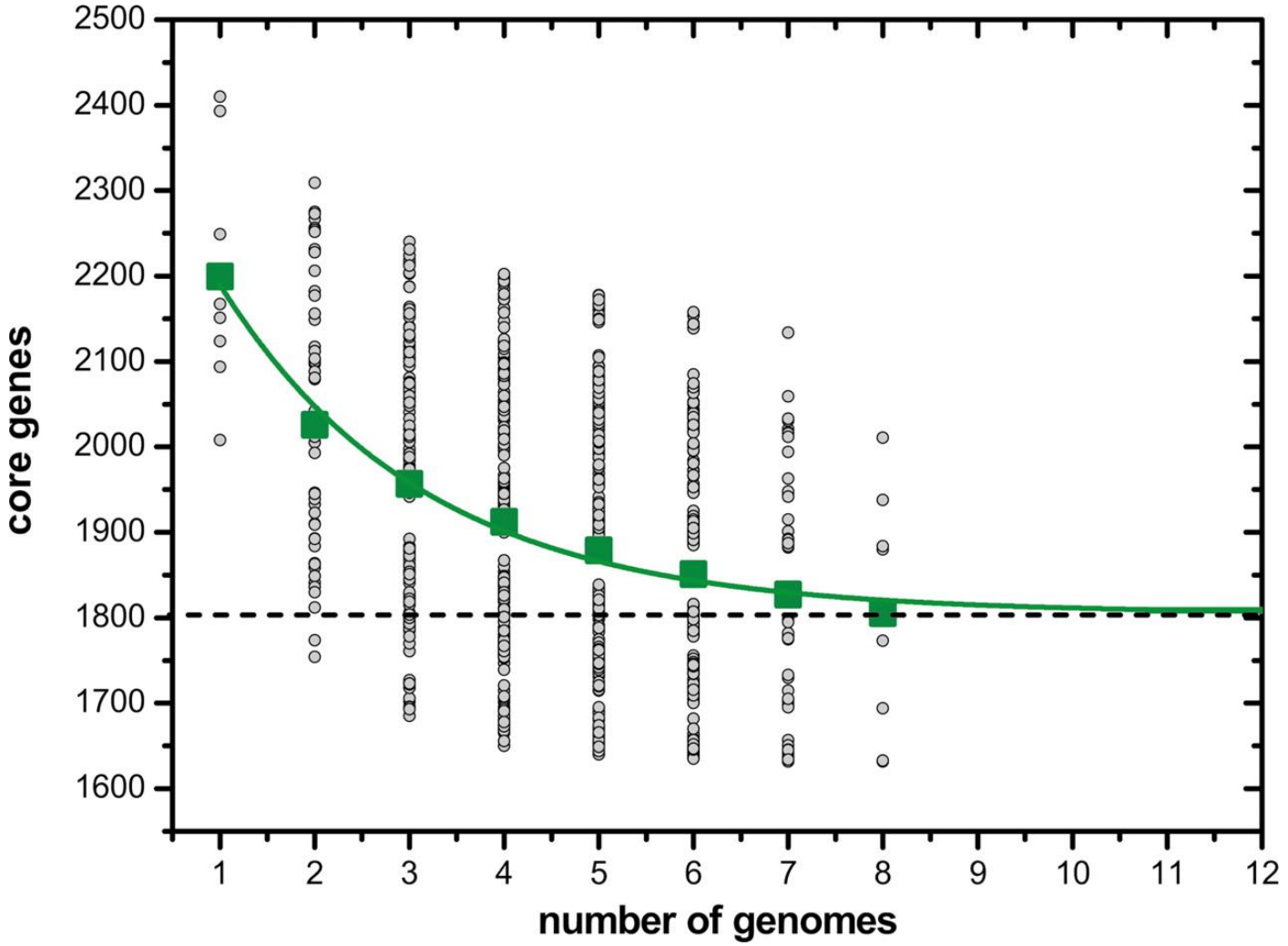
Pan-genome?

- **Single genome**
- **Pan-genome***
  - **pan, from the Greek word παν, meaning whole.**
  - ➢ **core genome [a]**
  - ➢ **dispensable genome [b]**
  - ➢ **unique genes [c]**



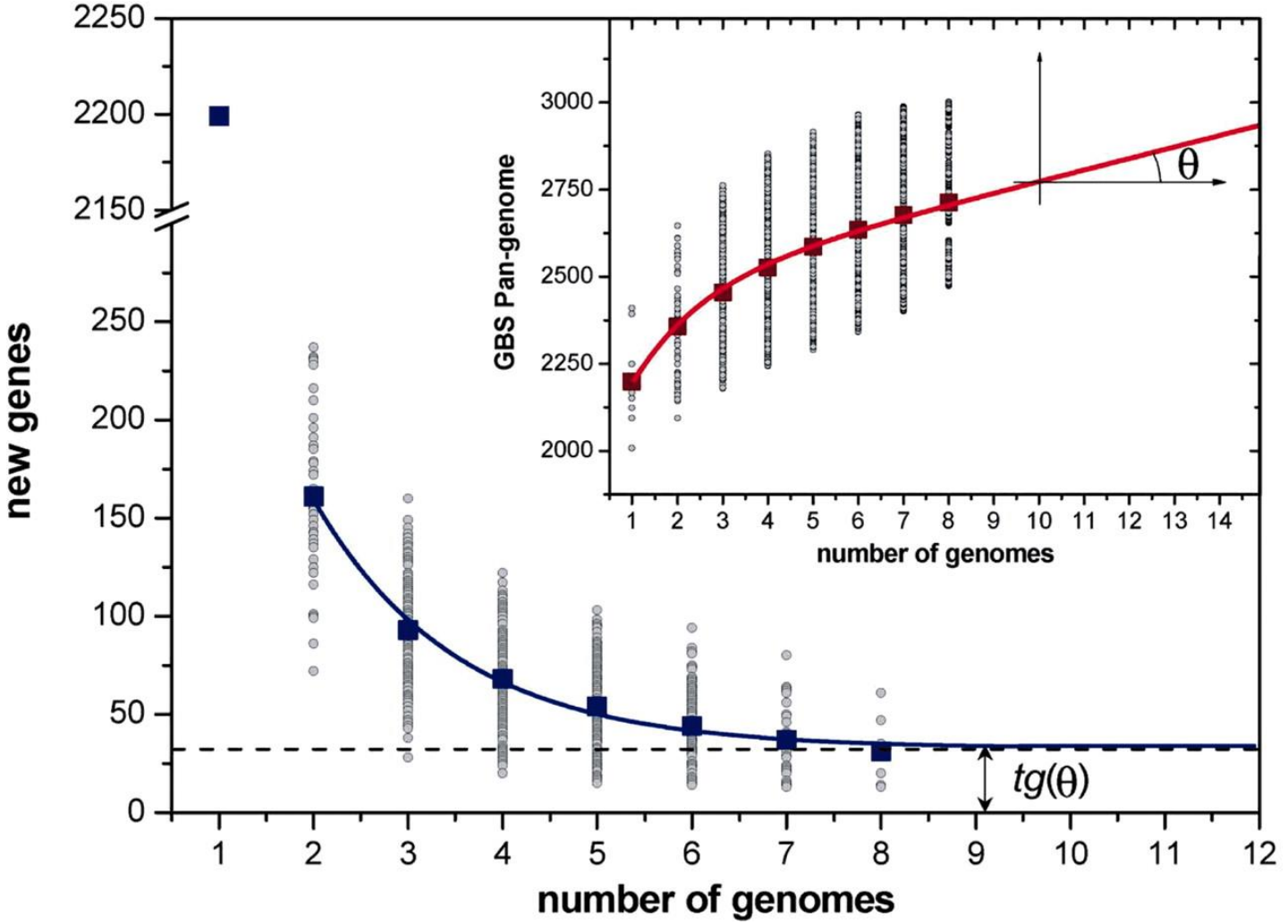
* Tettelin, Hervé, et al. "Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial “pan-genome”.“*PNAS* 102.39 (2005): 13950-13955.

GBS core genome.



Hervé Tettelin et al. PNAS 2005;102:13950-13955

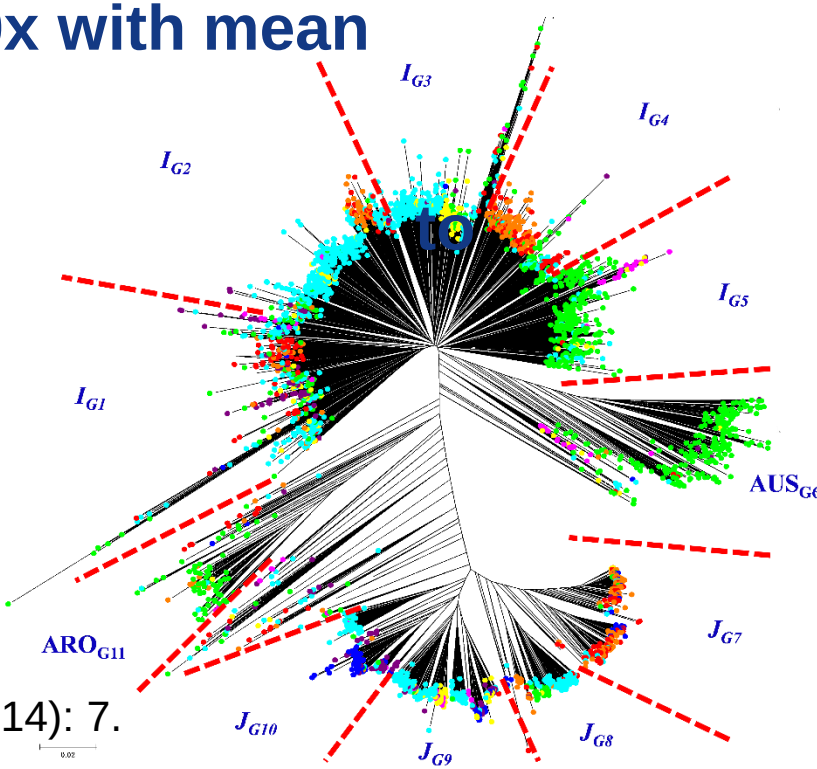
GBS pan-genome (An open pan-genome!)



Hervé Tettelin et al. PNAS 2005;102:13950-13955

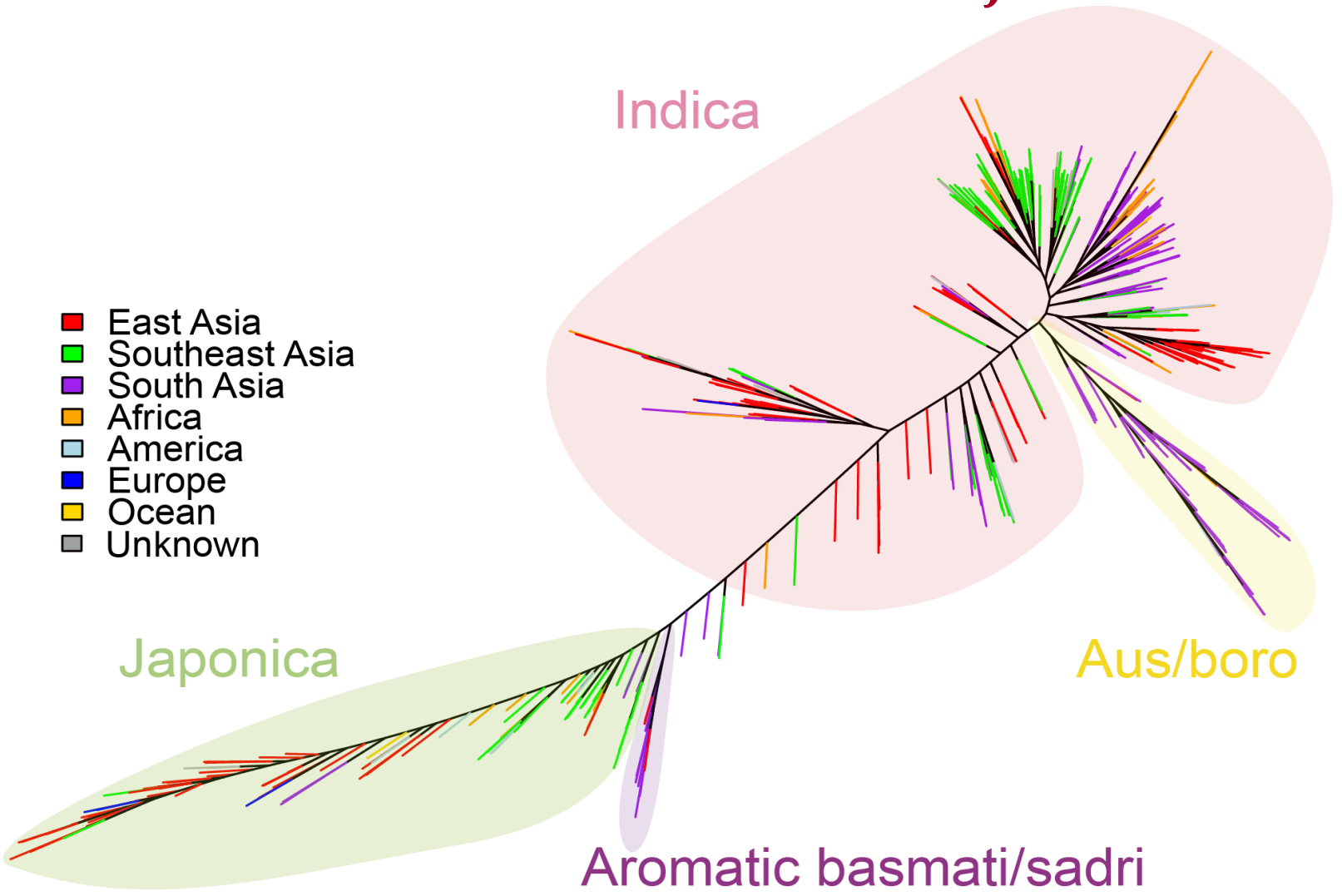
3K Rice Genome Project*

- 3,024 sequenced rice genomes
- From 89 countries of 6 continents
- Total sequenced bases: 14.7 Tbps
- Sequencing depth: from 3 to 60x with mean 14.3±6.3x
- Mapping coverage: from 74.5% to 98.4% with mean 92%



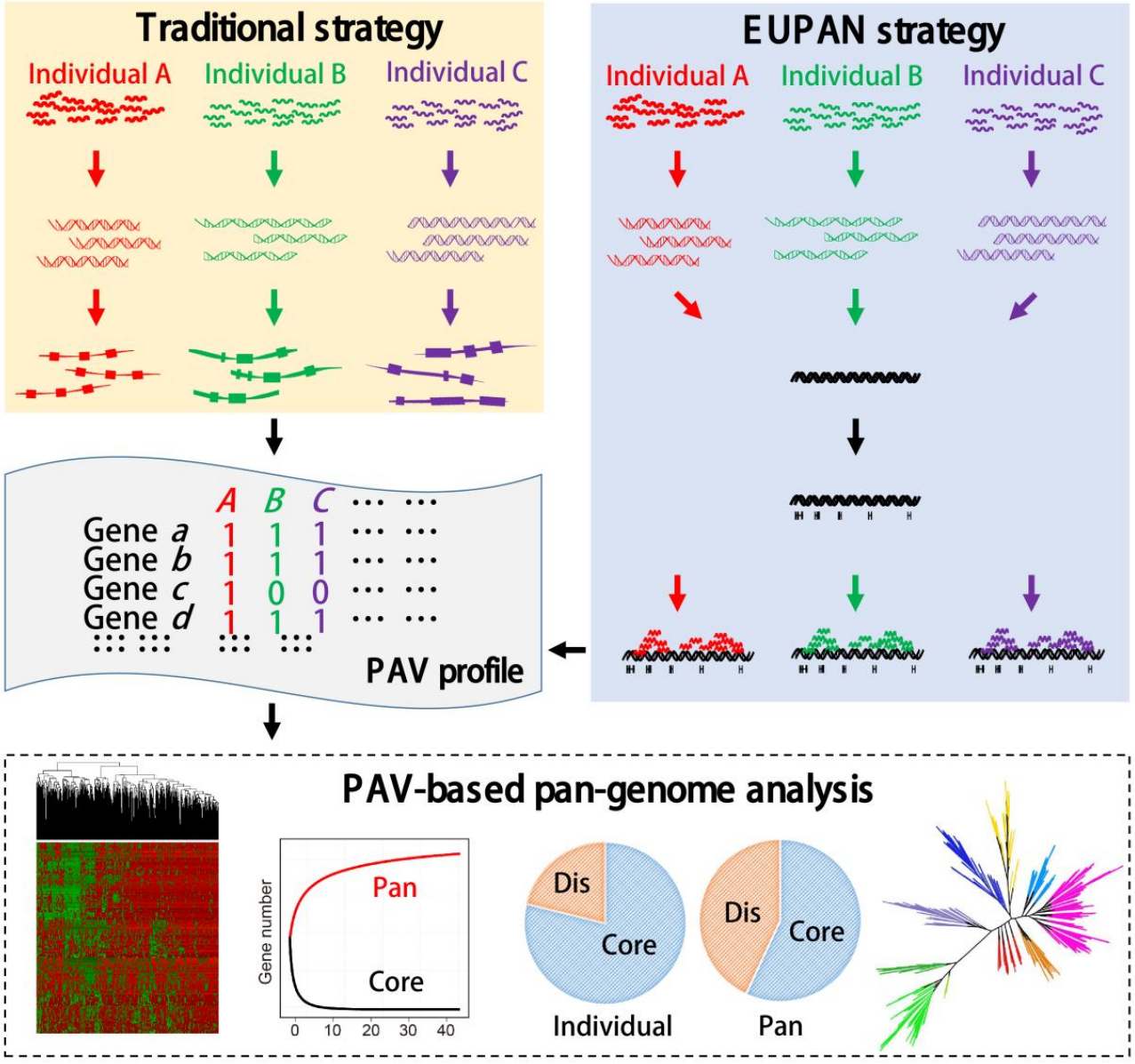
* “The 3,000 rice genomes project.” *GigaScience* 3 (2014): 7.

The 3000 Rice Genome Project



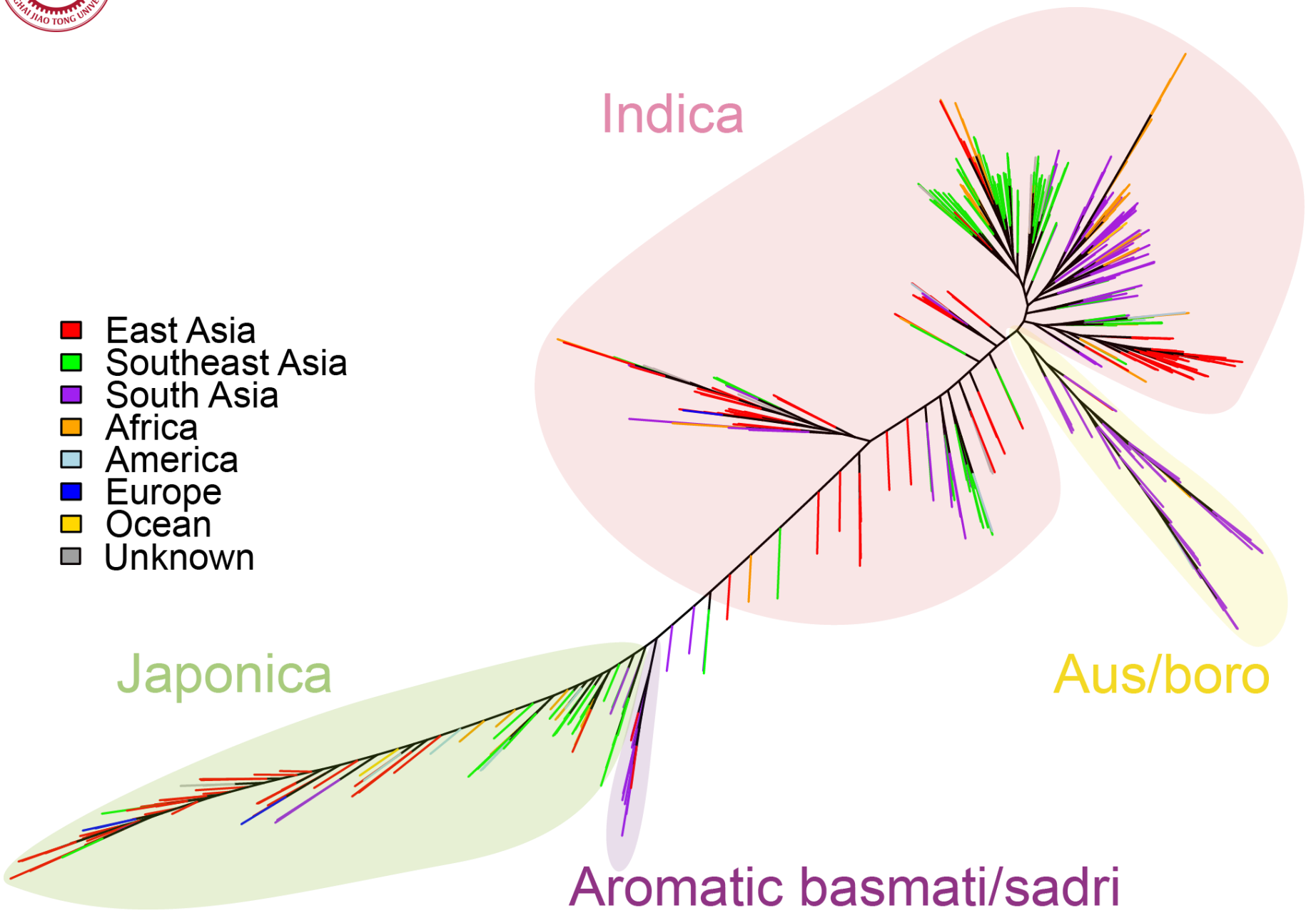
Sun, C. et al., “RPAN: Rice Pan-genome Browser for ~3,000 rice genomes”, ***Nucleic Acids Research***, 2017, 45(2): 597-605.

EUPAN: pan-genome analysis for eukaryotes



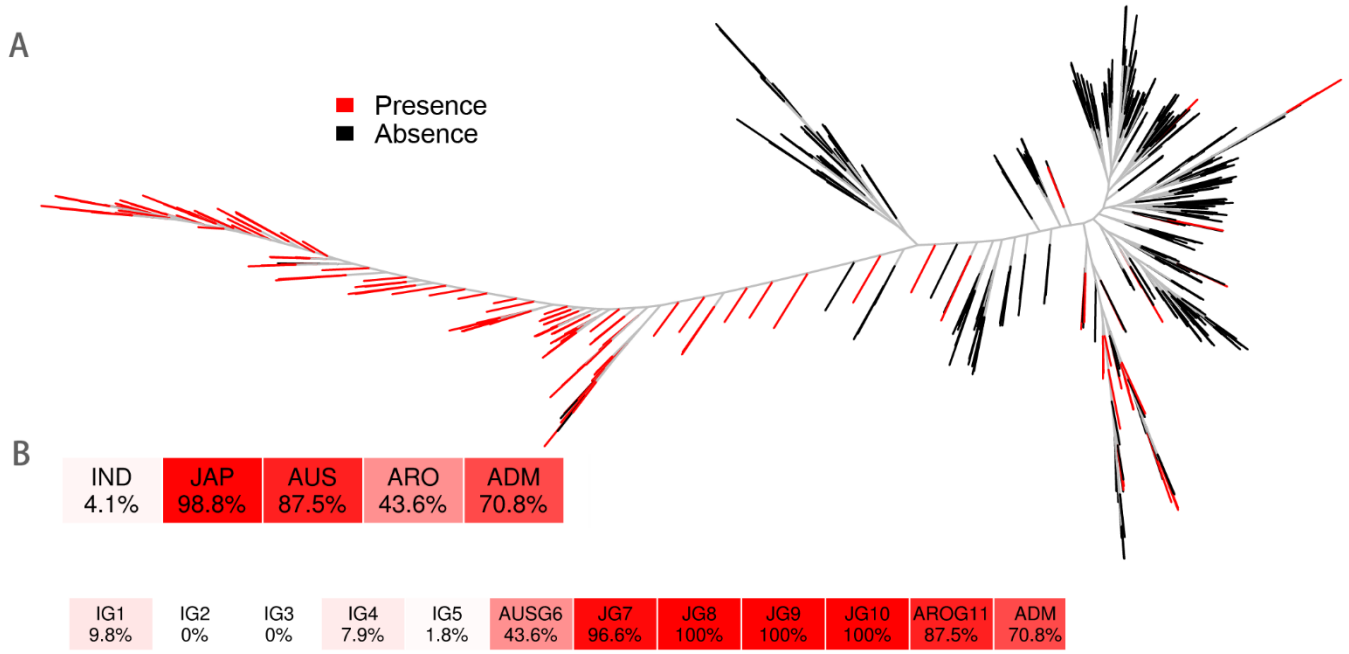
Hu, Z. et al, “EUPAN enables pan-genome studies of a large number of eukaryotic genomes”, ***Bioinformatics***, 2017

Phylogenic tree based on PAVs



Use Case (1)

- ***OsHsp23.7*** **is a gene with potentially important roles in rice acclimation to salt and drought stresses.**
- **It presents in 1,107 accessions(795 are *Japonica)*.**



Use Case (2)

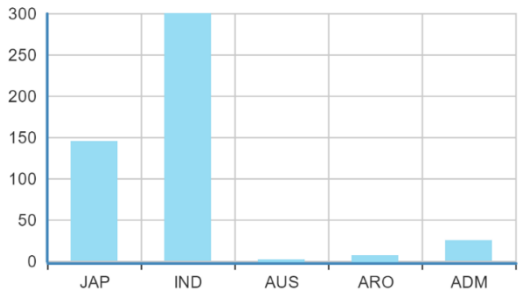
132 domestication-related protein-coding genes

- **84.8% of them are the core or candidate core genes.**

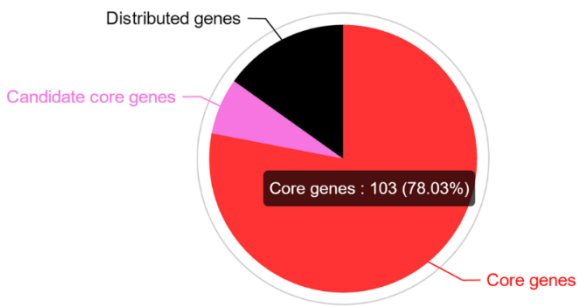
A Gene Statistics

#Total genes	132
#Core genes	103
#Candidate core genes	9
#Distributed genes	20
#Subspecies-unbalanced genes	6
#*Indica*-dominant genes	3
#*Japonica*-dominant genes	3
#Subspecies-specific genes	0
#*Indica*-specific genes	0
#*Japonica*-specific genes	0
#AUS-specific genes	0
#ARO-specific genes	0
#Subgroup-unbalanced genes	15
#*Indica*-subgroup-unbalanced genes	14
#*Japonica*-subgroup-unbalanced genes	3
#Random genes	5

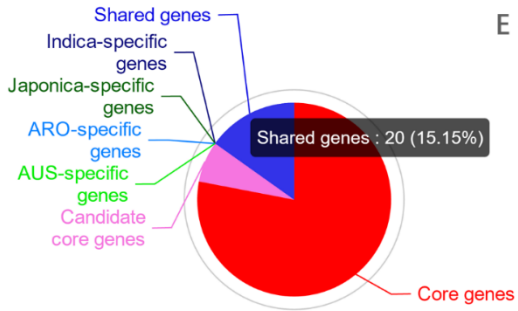
B Rice Distribution



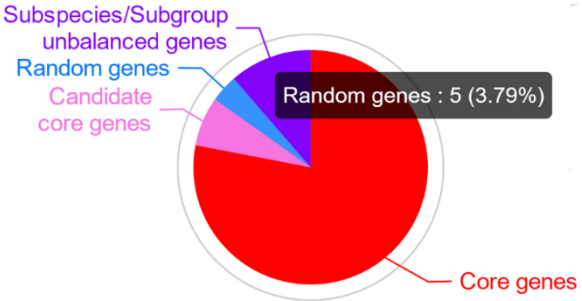
C



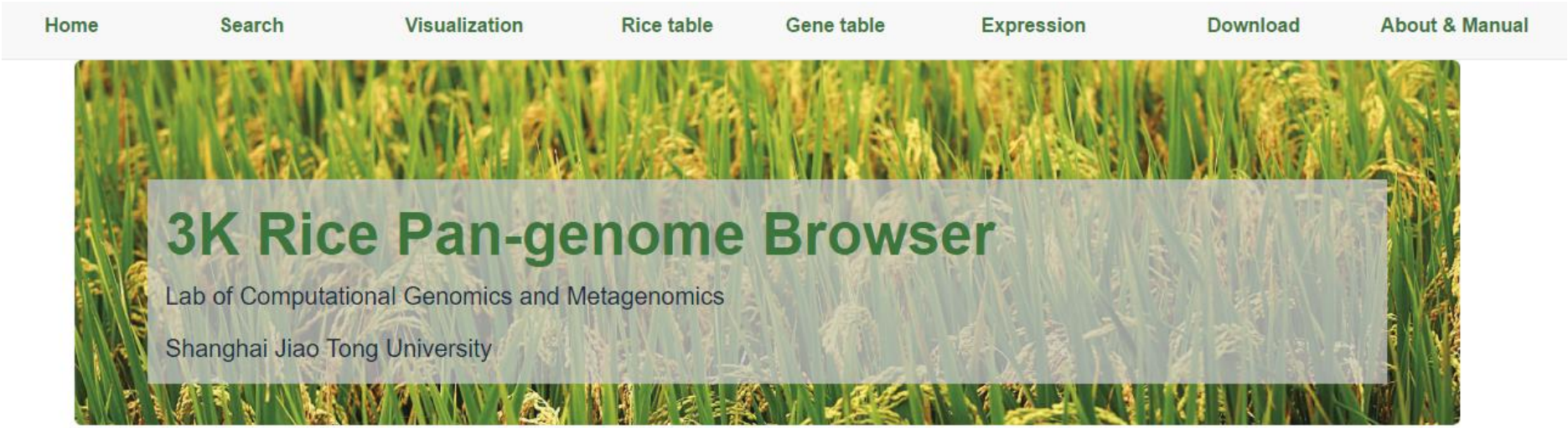
D



E



RPAN



- **http://cgm.sjtu.edu.cn/3kricedb/**

Sun, C., Hu, Z., Lu, K., Zhao, Y., Lu, J., Zheng, T., Wang, W., Shi, J., Zhang, D., Li, Z.*, **Wei, C.***, “RPAN: Rice Pan-genome Browser for ~3,000 rice genomes”, ***Nucleic Acids Research***, 2017, 45(2): 597-605

Contents in Rice Pan-genome Browser

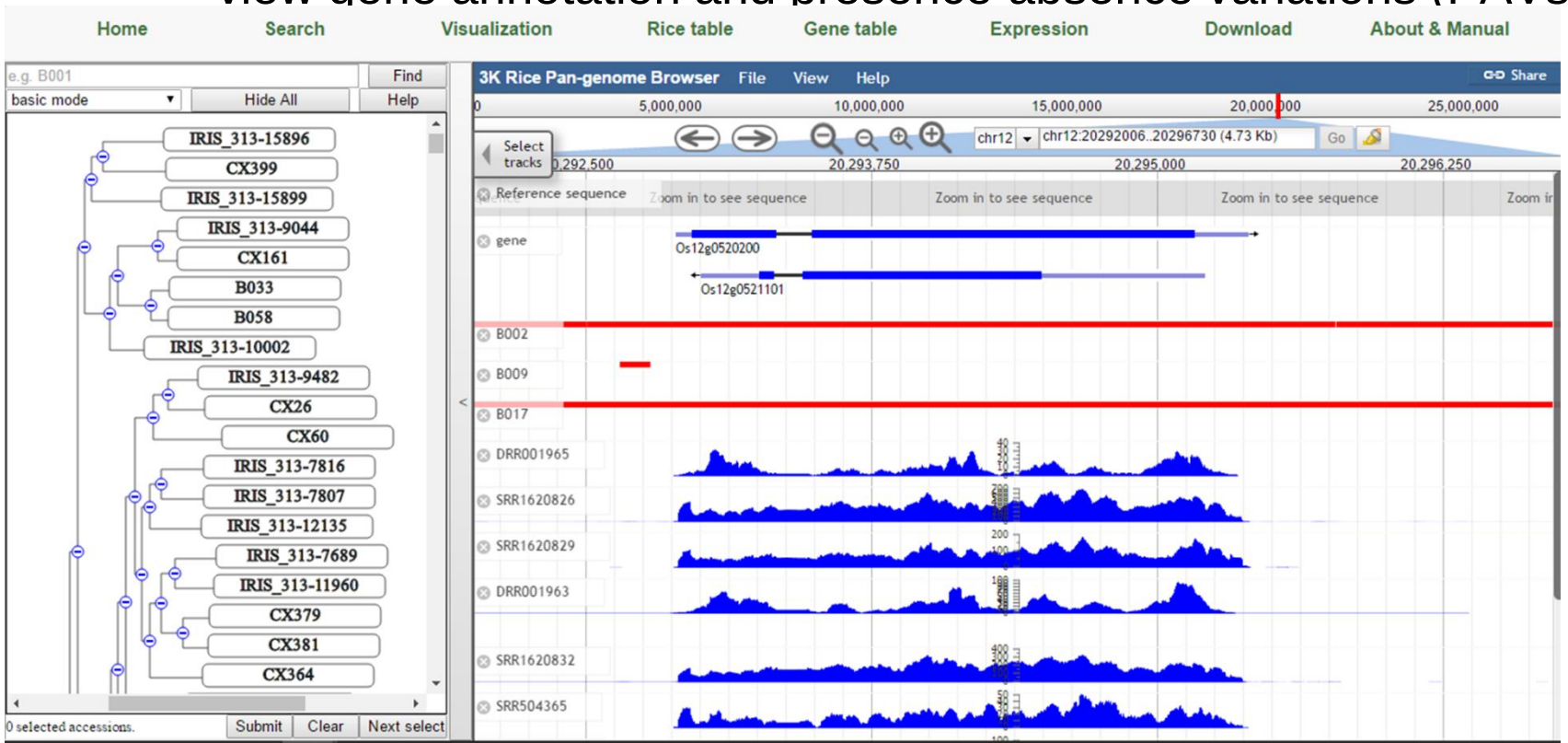
1. Basic information of the 3K rice accessions
2. Sequences and gene annotations
3. Gene presence-absence variations (PAVs)
4. Genome-wide expression profiles

Search functions in Rice Pan-genome Browser

- **Basic search functions:**
  - **Search a single gene**
    - basic information, distributions, PAVs and gene functions
  - **Search a single rice accession**
    - sequencing landscape and meta-information (source, classification, etc.)
  - **Search sequence(s)**
    - against pan-genome sequences
- **Advanced search functions:**
  - **Search multiple rice accessions**
    - shared genes
  - **Search multiple genes**
    - rice accessions where they all present

Rice Pan-genome Browser

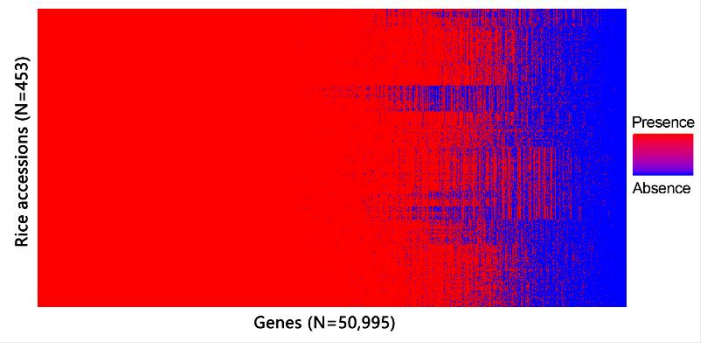
- **Visualization:**
  - **A tree browser**
    - view the phylogeny of 3K rice accessions
  - **A genome browser**
    - view gene annotation and presence-absence variations (PAVs)



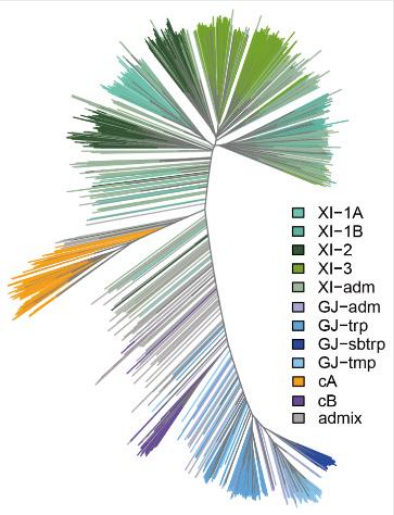
Summary

- Rice pan-genome is much bigger than the reference rice genome
  - ~80% bigger
  - ~40% more genes
  - ~40% more gene families
- Important genes can be dispensable
- Presence/absence of genes in rice accessions can be searched now at RPAN: http://cgm.sjtu.edu.cn/3kricedb

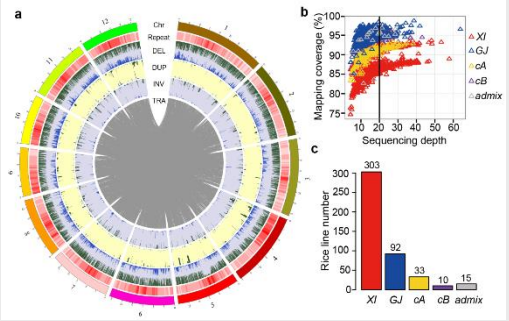
Future： Pan-genomics based genomics analysis



Pan-genome analysis
PAV



SNP



SV

Biomarker identification

Phenotype

Acknowledgement

Zhiqiang Hu
Chen Sun
Kuang-chen Lu
Jingyuan Lu
Yue Zhao

Dabing Zhang
Jianxin Shi
Xiao Cui
Rui Li

Zhikang Li （CAAS）
Wensheng Zhang （CAAS）
Tianqing Zheng （CAAS）

Fundings

- Gates Foundation
- NSFC
- 863
- 973

高性能计算中心
Center for High Performance Computing