高通量测序及生物信息分析

张新宇

中国科学院微生物研究所
真菌学国家重点实验室

内 容

- 基因组测序技术
- 高通量测序实验设计
- 高通量数据分析对实验室的挑战
- 数据分析策略及本地化分析平台

测序技术发展

Year

2010s

Single Molecule Sequencing:

Helicos Biosciences
Pacific Biosciences
Nanopole

Polony sequencing:

Roche (454) pyrosequencing
Illumina (Solexa) sequencing
ABI (SOLiD) sequencing

2000s

The first of the "next-generation" sequencing technologies

1990s

Lynx Therapeutics' Massively Parallel Signature Sequencing

RNA sequencing

Sanger sequencing: Chain-termination method

Maxam–Gilbert sequencing

1970s

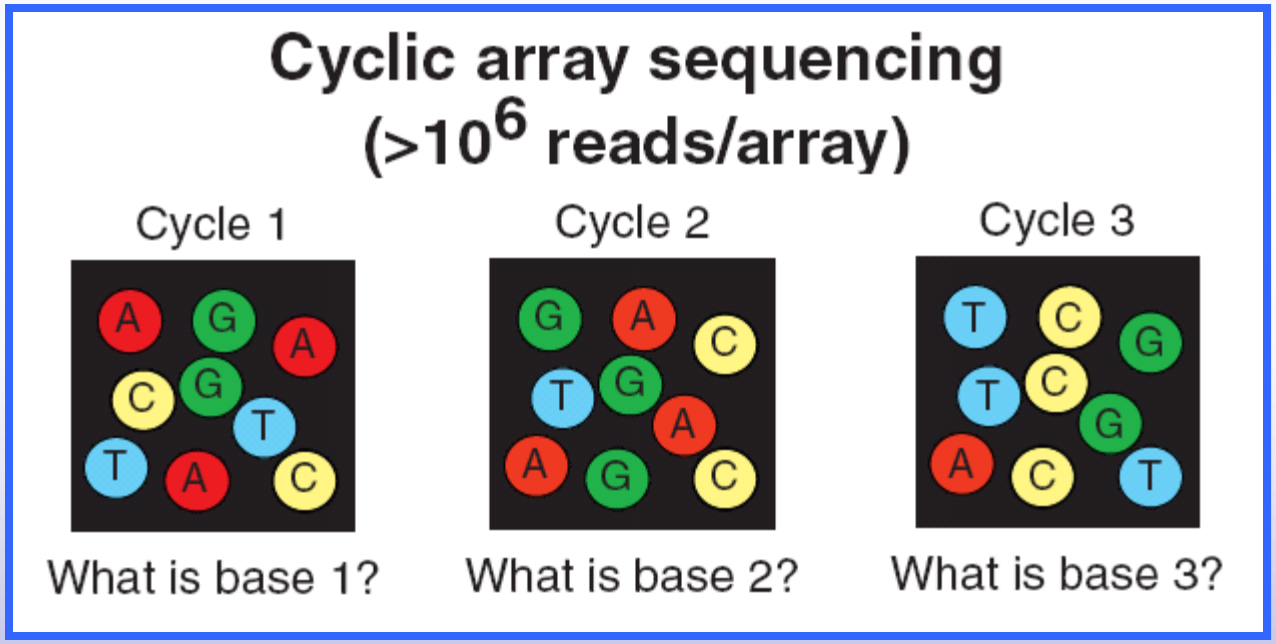
Traditional sequencing

Next-generation sequencing (NGS)

Polony 测序技术的高度并行特点

Sanger

Polony



第三代测序（单分子测序）技术特点

HeliScope：

Direct sequencing of DNA molecules:

no amplification stage

DNA fragments are attached to array

Potential benefits:

higher throughput, less errors

测序技术性能价格比较

	Read length	*Sequencing Technology*	*Throughput (per run)*	*Cost (1mbp)*
Sanger	~800bp	Sanger	400Kbp	500$
454	~500bp	Polony	500Mbp	60$
Solexa (PE)	2*100bp	Polony	20Gbp	2$
SOLiD (PE)	2*75bp	Polony	40Gbp	2$
Helicos	30-35bp	Single molecule	25Gbp	1$
SMRT	10000bp	Single molecule		
Nanopole	infinite	Single molecule		

Itai Sharon, 2009, *Next Generation Sequencing*

Ability

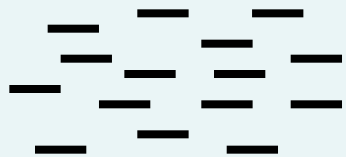
Price

内 容

- 基因组测序技术
- **高通量测序实验设计**
- 高通量数据分析对实验室的挑战
- 数据分析策略及本地化分析平台

NGS基因组测序---高通量、短序列

Genome



Short fragments of DNA

AC..GC TG..GT TC..CC
CG..CA
TT..TC
TG..AC
AC..GC GA..GC
CT..TG
GT..GC AC..GC AC..GC
AA..GC AT..AT
TT..CC

Short DNA sequences

ACGTGGTAA CGTATACAC TAGGCCATA
GTAATGGCG CACCCTTAG
TGGCGTATA CATA...

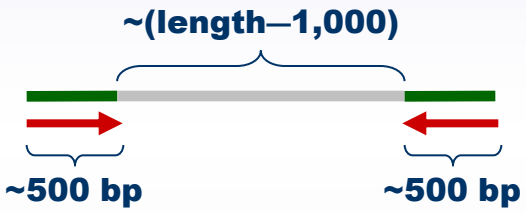
ACGTGGTAATGGCGTATACACCCTTAGGCCATA

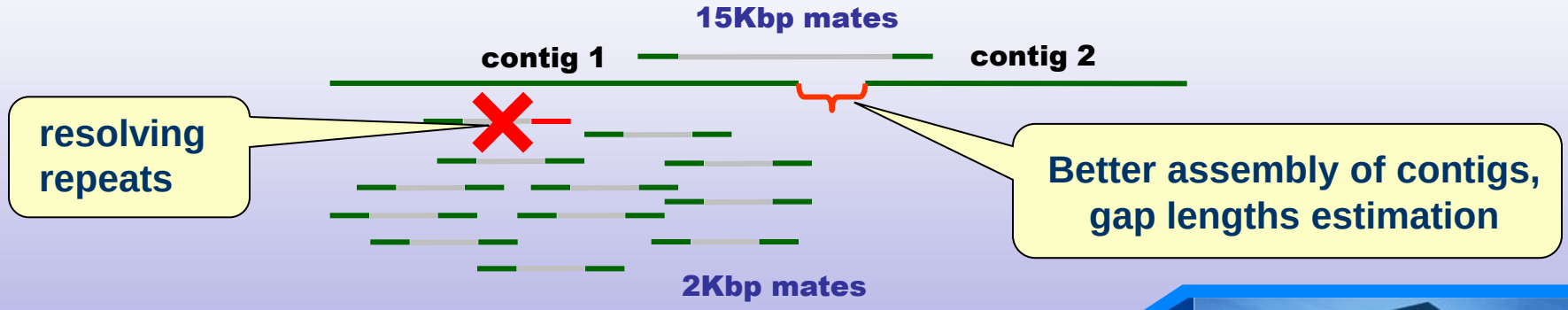
ACGTGACCGGTACTGGTAACGTACA
CCTACGTGACCGGTACTGGTAACGT
ACGCCTACGTGACCGGTACTGGTAA
CGTATACACGTGACCGGTACTGGTA
ACGTACACCTACGTGACCGGTACTG
GTAACGTACGCCTACGTGACCGGTA
CTGGTAACGTATACCTCT...

Sequenced genome

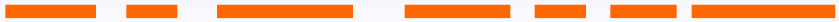
PE测序文库与基因组组装

Cut DNA to larger pieces (2Kbp, 15Kbp) and sequence both ends of each piece (Fleischmann et al., 1994)

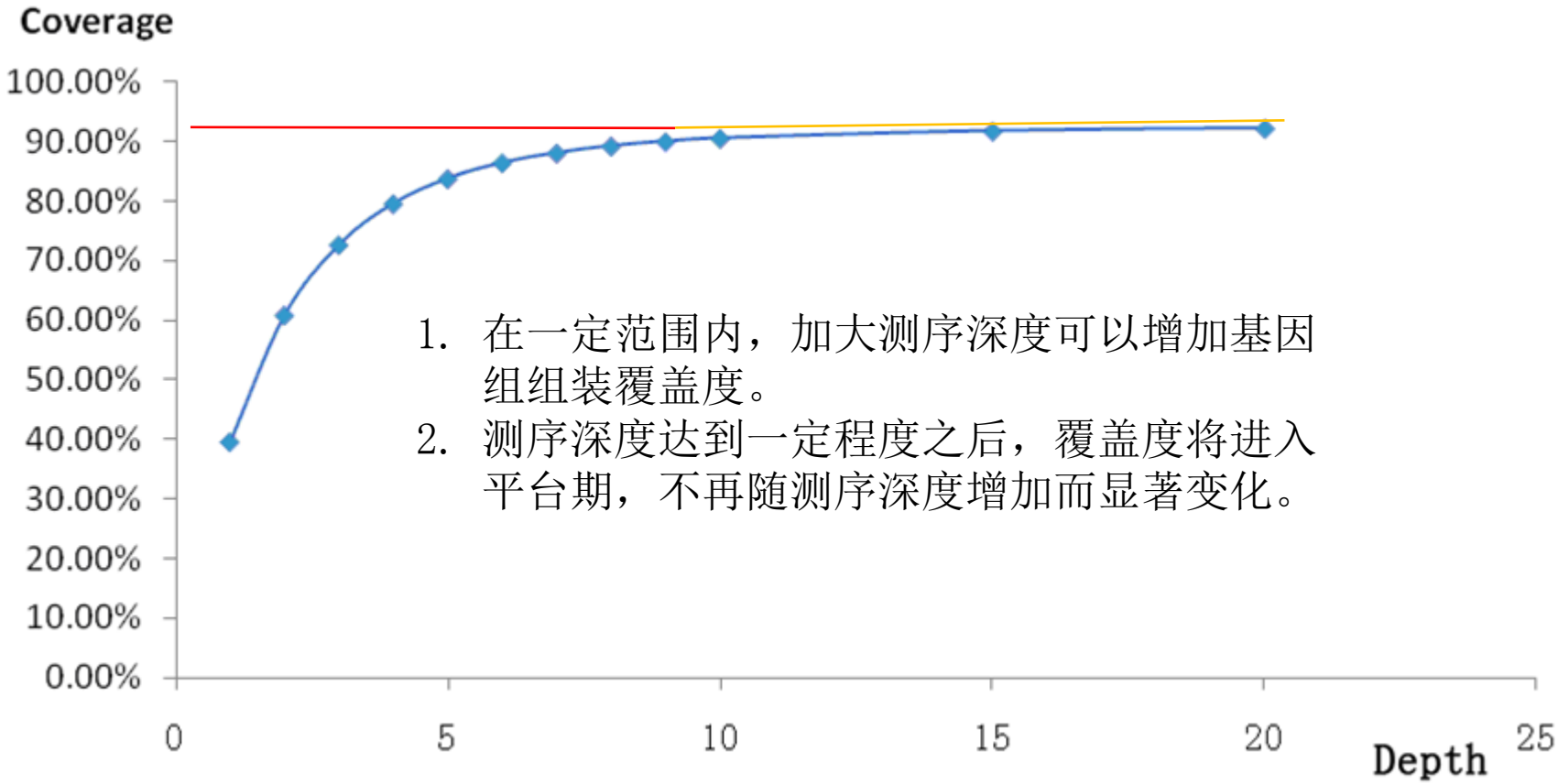




测序深度与组装效果

	Input	Output
Low coverage:	 A few pieces to assemble	 many contigs, many gaps
High coverage:	 many pieces to assemble	 a few contigs, a few gaps

测序深度与组装效果

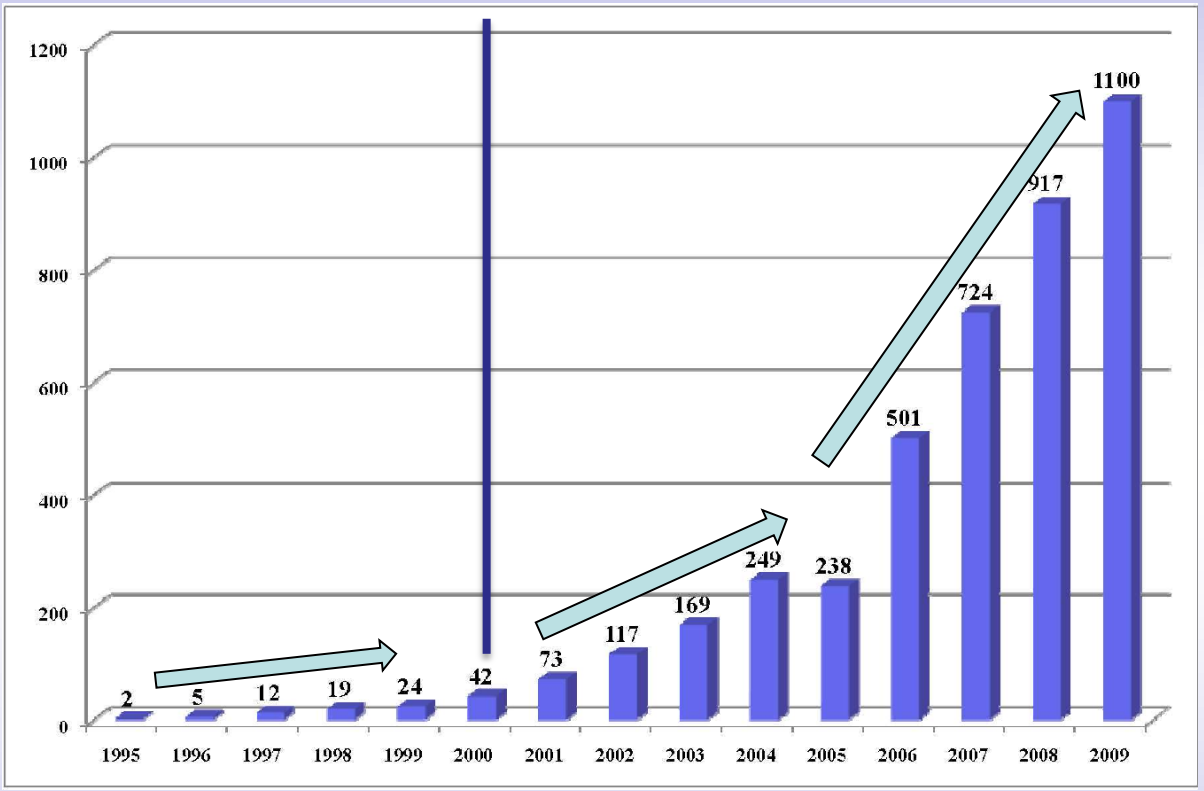


1. 在一定范围内，加大测序深度可以增加基因组组装覆盖度。
2. 测序深度达到一定程度之后，覆盖度将进入平台期，不再随测序深度增加而显著变化。

内 容

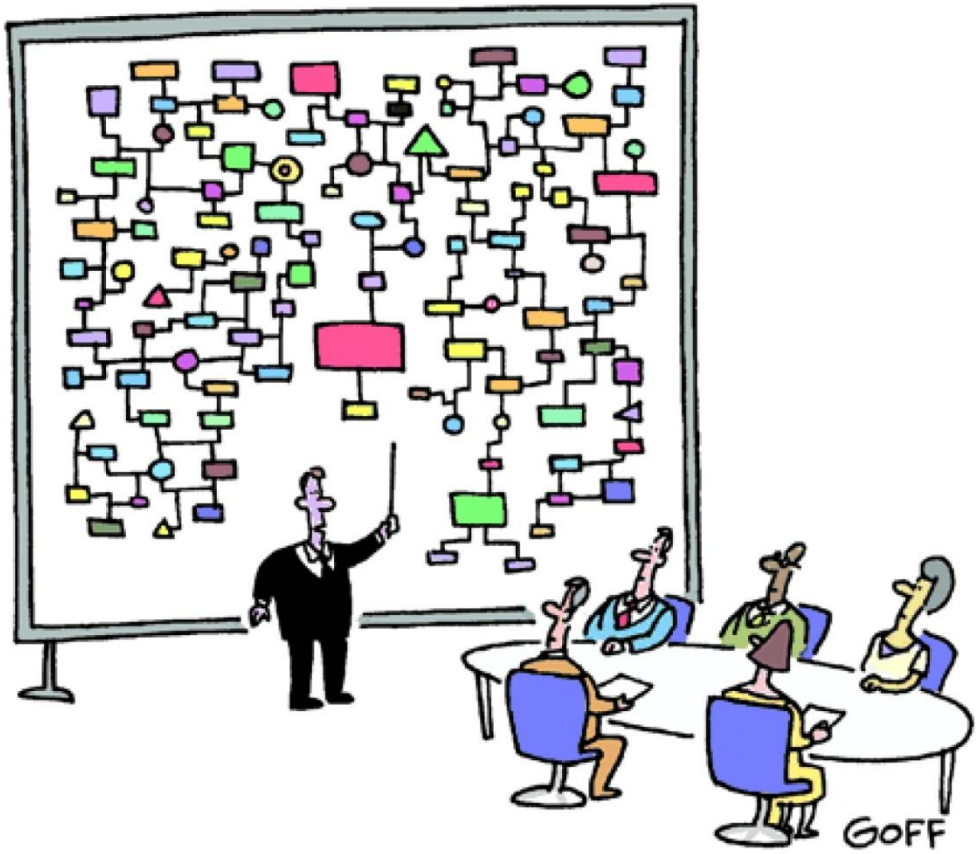
- 基因组测序技术
- 高通量测序实验设计
- **高通量数据分析对实验室的挑战**
- 数据分析策略及本地化分析平台

基因组测序方兴未艾



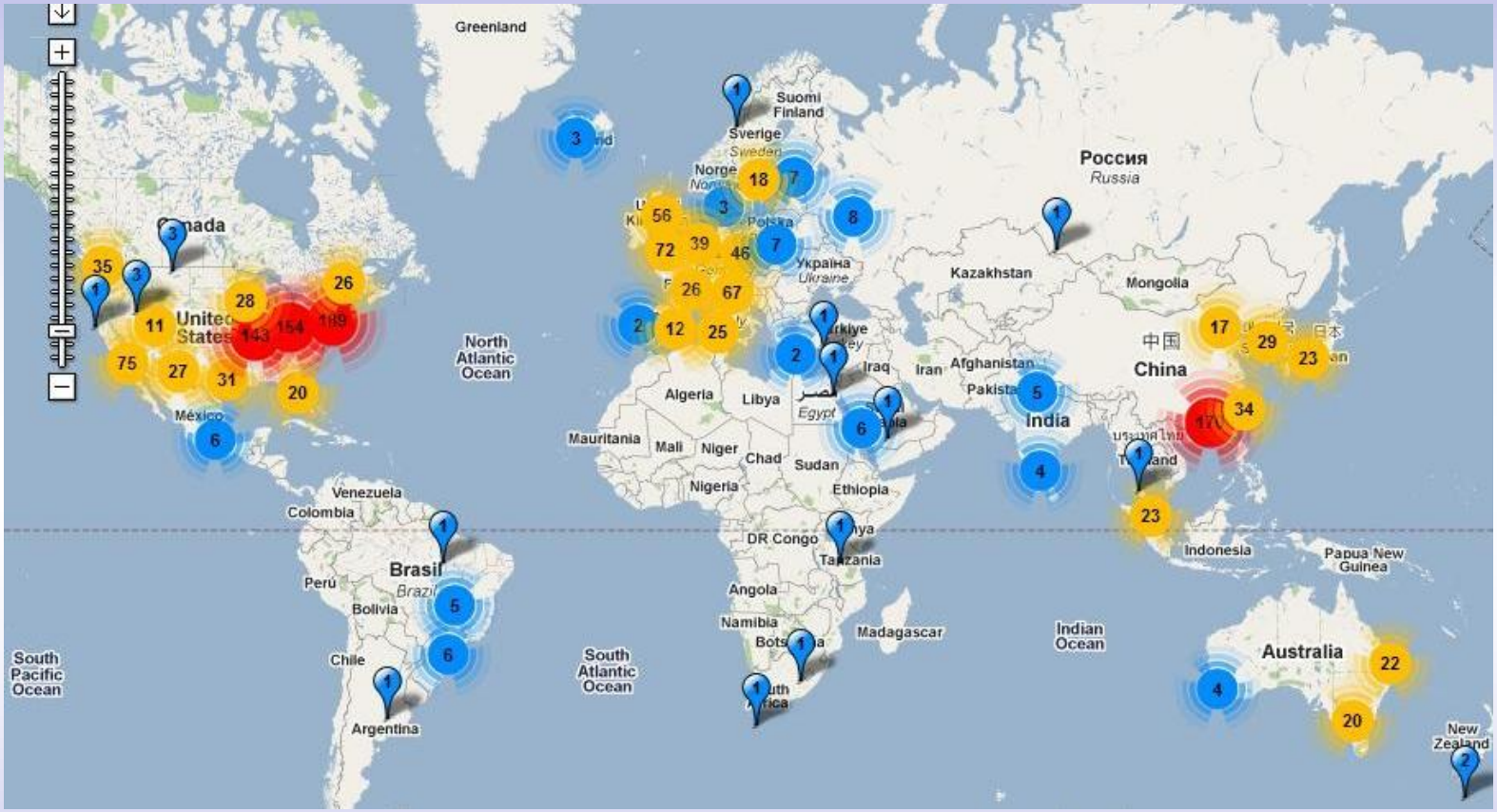
The number of completely sequenced genomes by years.

高通量数据分析对实验室的挑战



- 数据管理
- 数据解读
- 分析策略设计
- 分析方法选择
- 需要编程序？
- 键盘上的Linux

世界测序仪分布---海量数据的来源



世界测序仪分布图（全世界1981台，中国200台，仅次美国，排名第二！）。中国BGI以166台排名世界基因组中心第一，第二位的broad institute有101台。

http://omicsmaps.com/stats

高通量数据分析对实验室的挑战

高通量数据分析对实验室的挑战

生物信息学数据分析与皇帝的新装 精选

已有 2848 次阅读 2012-6-17 02:43 | 系统分类:观点评述 | 关键词:生物信息学 数据分析

本文是主要面向生命科学研究人员的一篇科普性文章，旨在探讨生物学数据分析背后的真与假，对与错，难与易的问题。

二十一世纪的前十年是生物芯片技术的天下，然而不经意间，高通量测序风起云涌，二代测序、三代测序（统称为下一代测序，简称NGS）接踵而来，并造就了中国BGI的崛起。高通量技术的迅猛发展也使得其价格的跌势丝毫不亚于中国股市。两年前一个真菌基因组的测序费用在三十万到五十万元人民币，而今只需区区十多万元甚至更低，如果考虑到国际上人民币升值而国内通货膨胀和CPI高涨的因素，其相对价格是折上加折。作为一个直接效应，起码在实验室水平NGS已经平民化，科研水平的竞争也逐步地从数据的获取转化为数据的分析和解读。

近几年，为了获得先机，实验室往往先抢地盘，花重金对新的物种测序，然而，测序公司对数据只能提供最基本的分析，只有极少数课题组能够通过代价不菲的合作获得对数据的专业解读，更多的数据被困在实验室的电脑里边，无法转化为科研成果。眼看着数据资源的优势在逐步丧失，研究人员往往采取两条途径寻求帮助，一是求助于专业公司，二是让学生着手分析数据。

科学网博文探讨高通量数据分析存在的问题，认为专业平台建设、平台共享和密切合作是有效解决方案。

内 容

- 基因组测序技术
- 高通量测序实验设计
- 高通量数据分析对实验室的挑战
- **数据分析策略及本地化分析平台**

基因组数据分析策略

1. Functional genomics

- ……
- Phylogenetic -tree
- Homology
- Genome-features
- Bio-function
- Annotation
- Assembly
- QC

NGS data

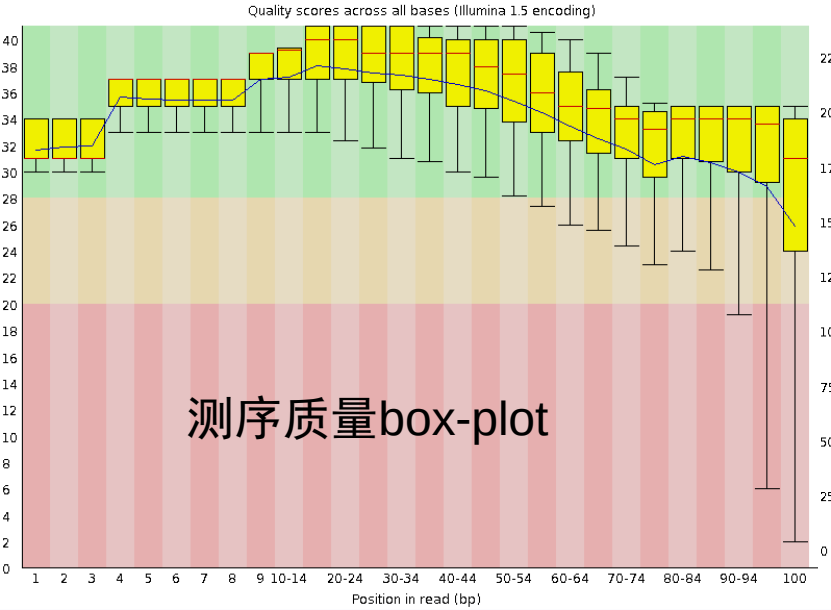
NC AN AF MO MAA MAC AO … Meta genomics

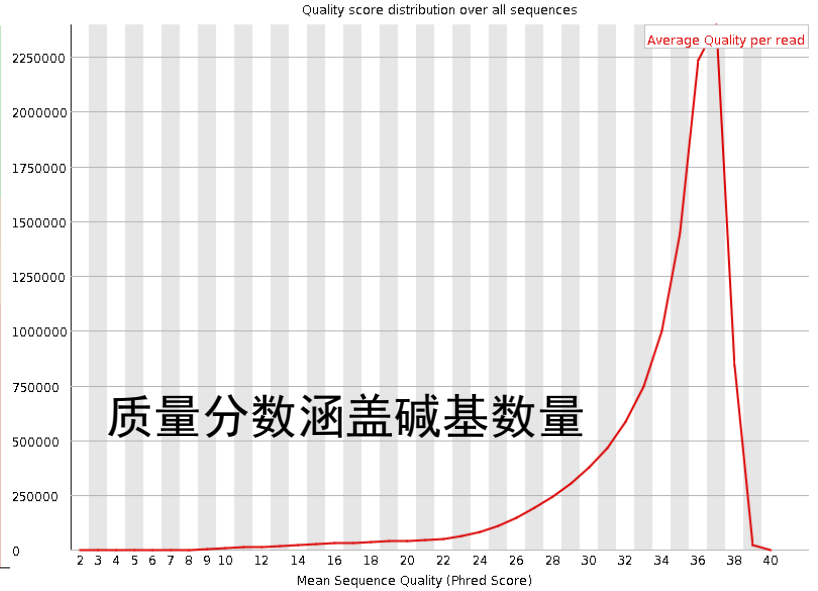
2. Comparative genomics

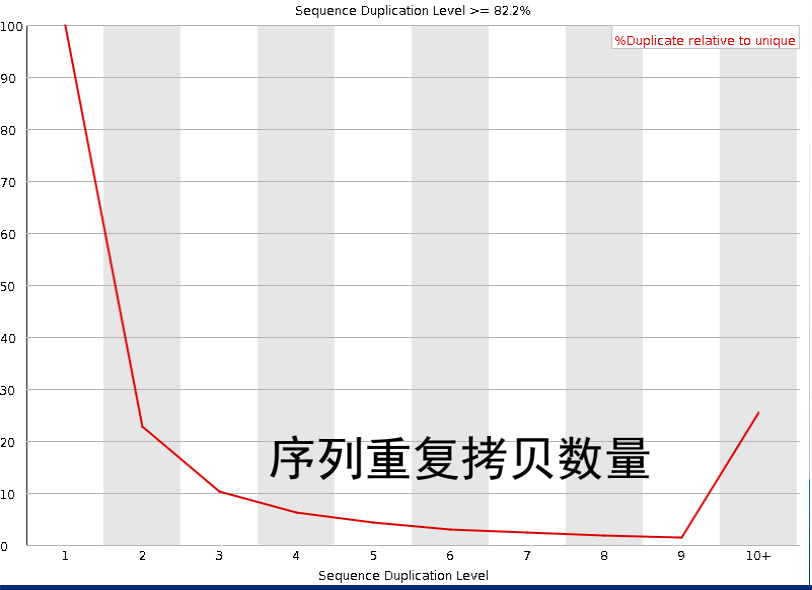
- Genomics
- Transcriptomics
- Proteomics
- Metabonomics

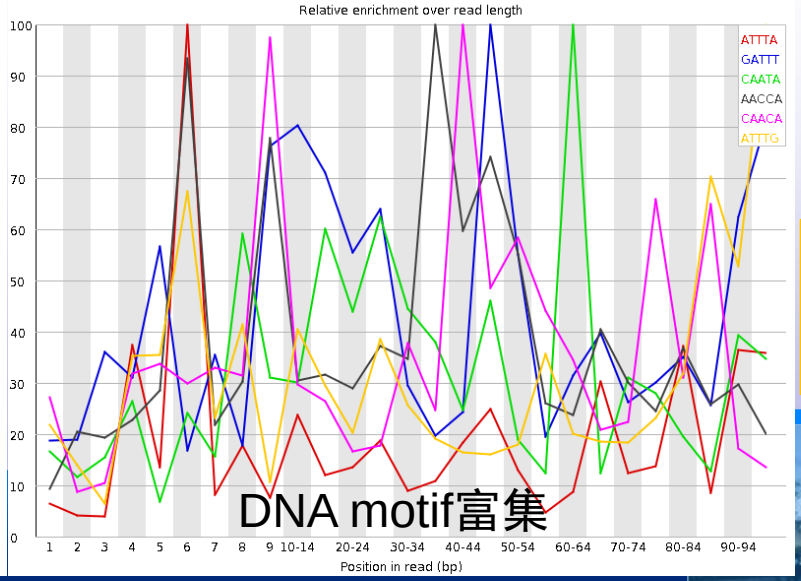
3. System biology

测序质量检测



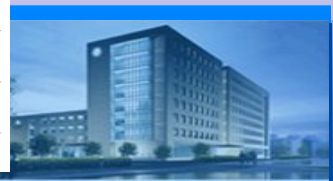






本地生物信息学平台

分析软件：
fastqc



有效tag数量检测

通过reference alignment评估有效tag数量：

reads processed: 27844233

reads with at least one reported alignment: 22239985 (**79.87%**)

reads that failed to align: 5516023 (19.81%)

reads with alignments suppressed due to -m: 88225 (0.32%)

Reported 26873982 alignments to 1 output stream(s)

分析软件：**Bowtie 和 TopHat**

测序质量控制（QC）

过滤低质量和无意义序列对于后续分析结果的准确性具有重要意义：

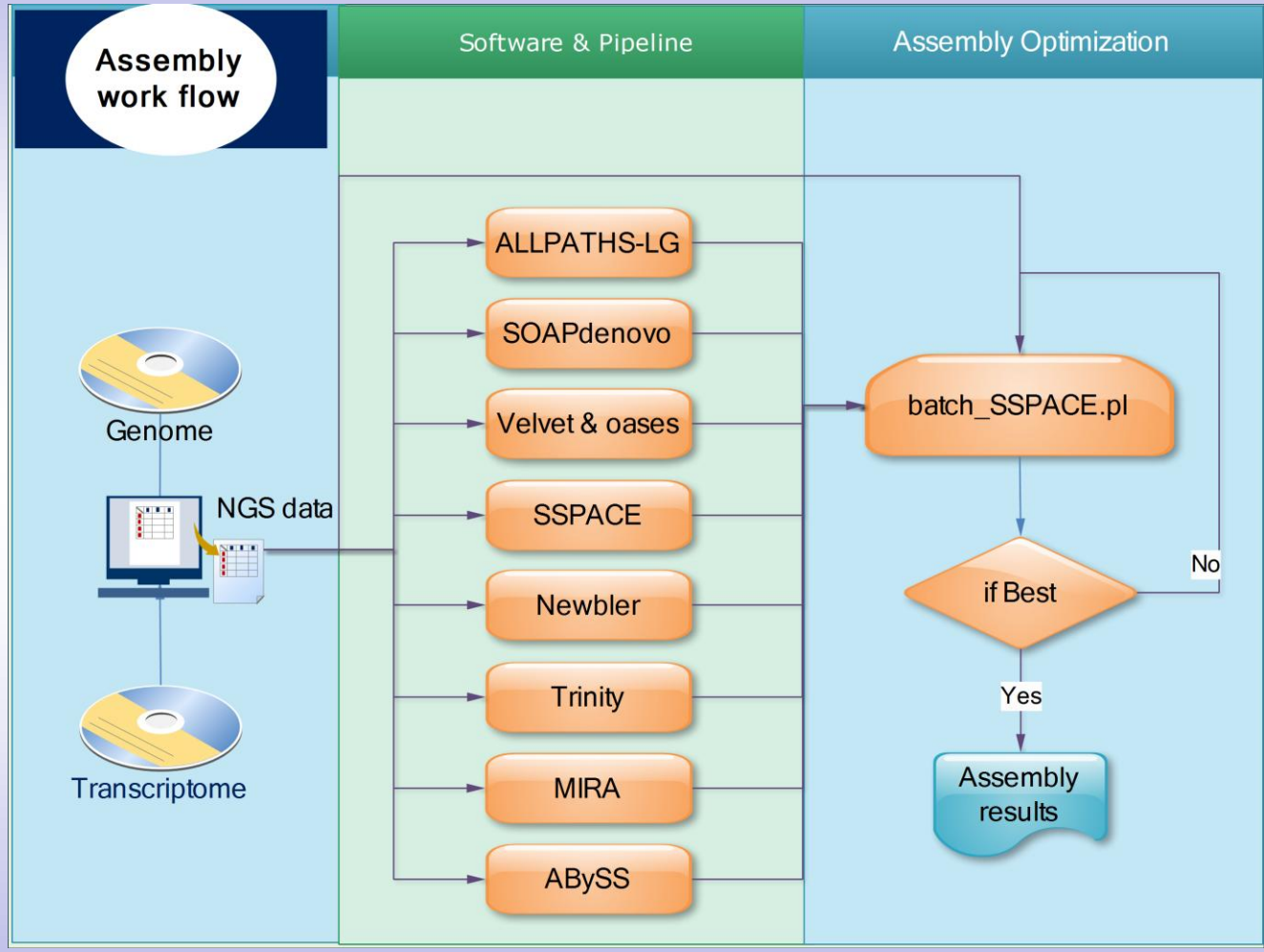
- Filter low quality reads
- Filter or trim adapter reads
- Filter PCR duplication reads
- Remove contaminate reads(mitochondrion or other)
- Split tandem repeat reads (di or three-nucleotides) [option]
- Filter low frequency Kmer reads(Corrector)

分析软件：
fastqc，
fastq_quality_filter

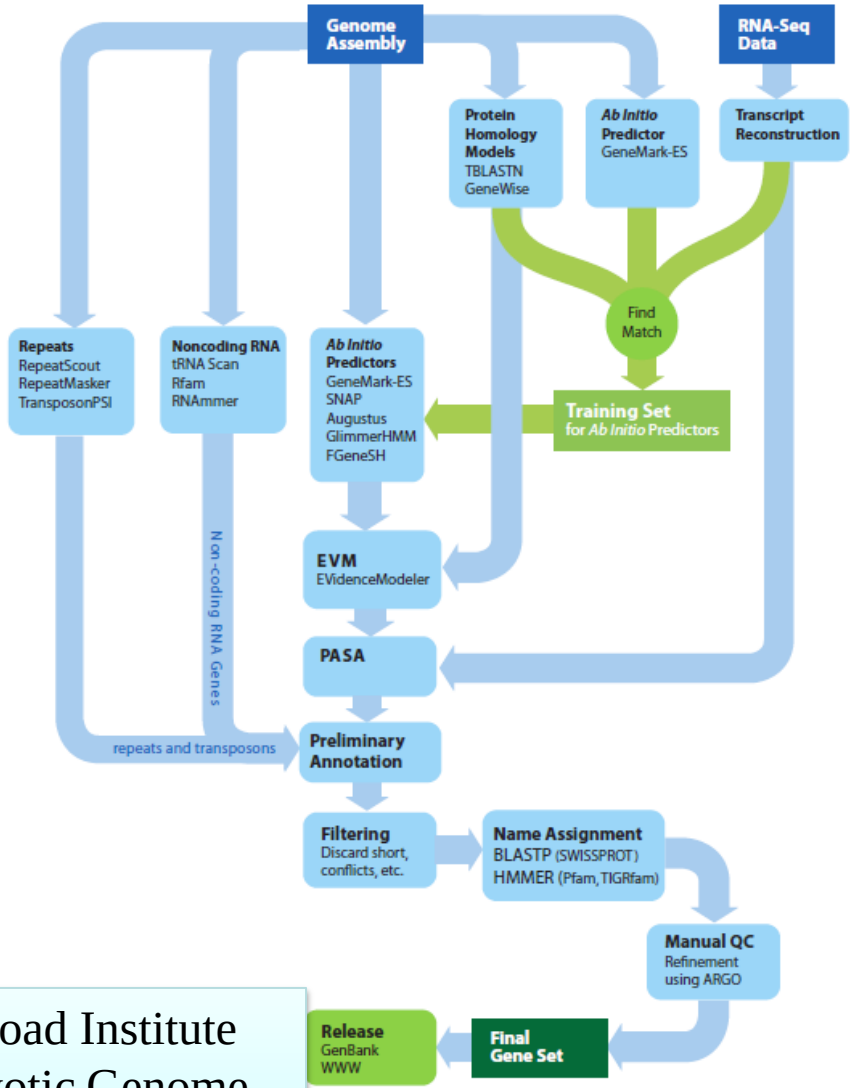
序列组装（Assembly）

组装流程

- Contig assembly:
  Short reads: Solexa, Solid
  Long reads: Sanger, 454 reads
  Hybrid reads: short + long reads
- Scaffolding
- Gap fixing



基因注释（Gene annotation）

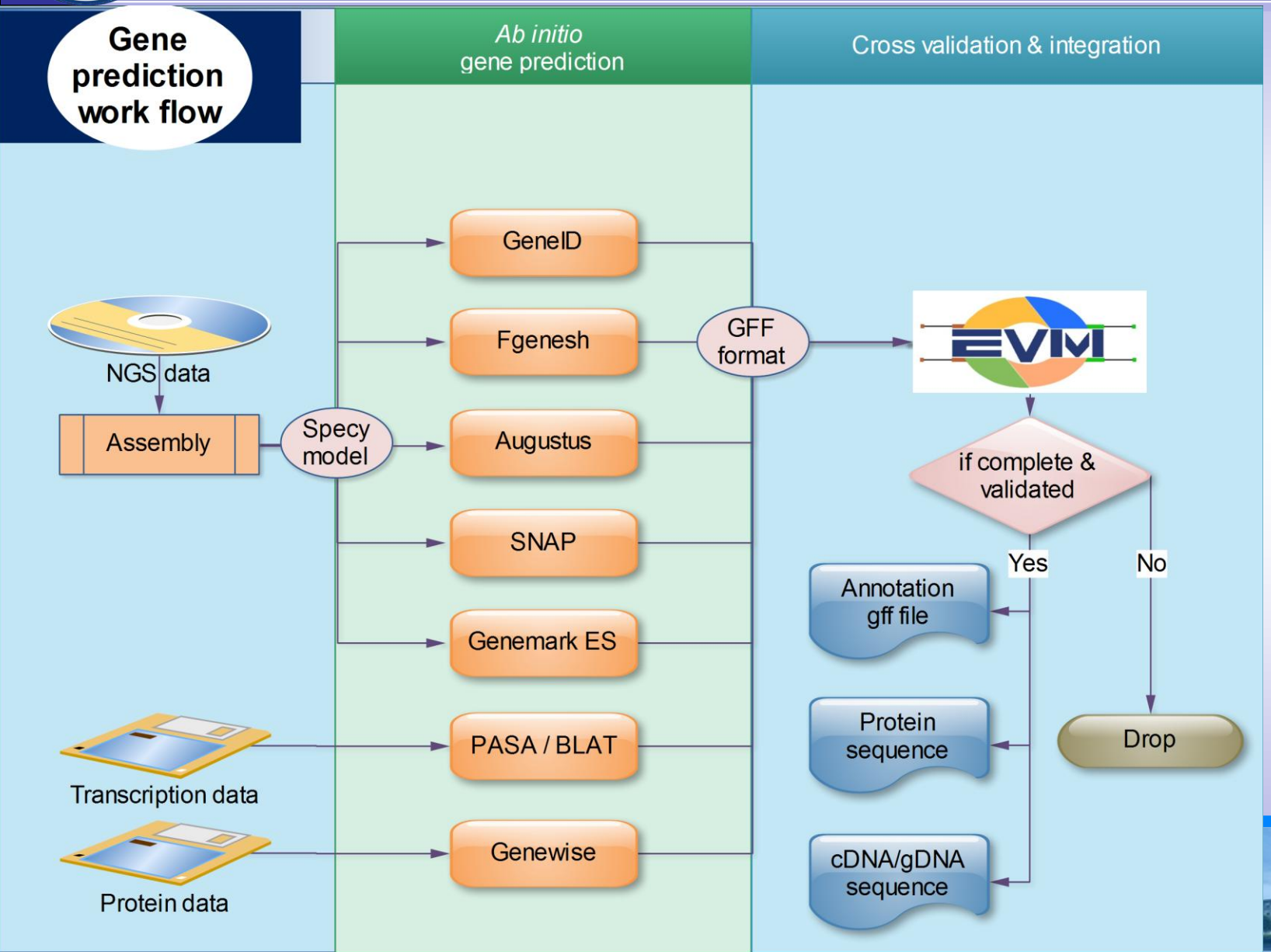


The Broad Institute Eukaryotic Genome Annotation Pipeline.

- 作为功能基因组分析的早期步骤，基因预测结果对后续深入分析的准确性影响很大
- 选取多个合适的软件和model进行从头预测
- 尽可能利用基因表达产物（mRNA, protein）协助CDS预测
- 加权整合，得到最优结果

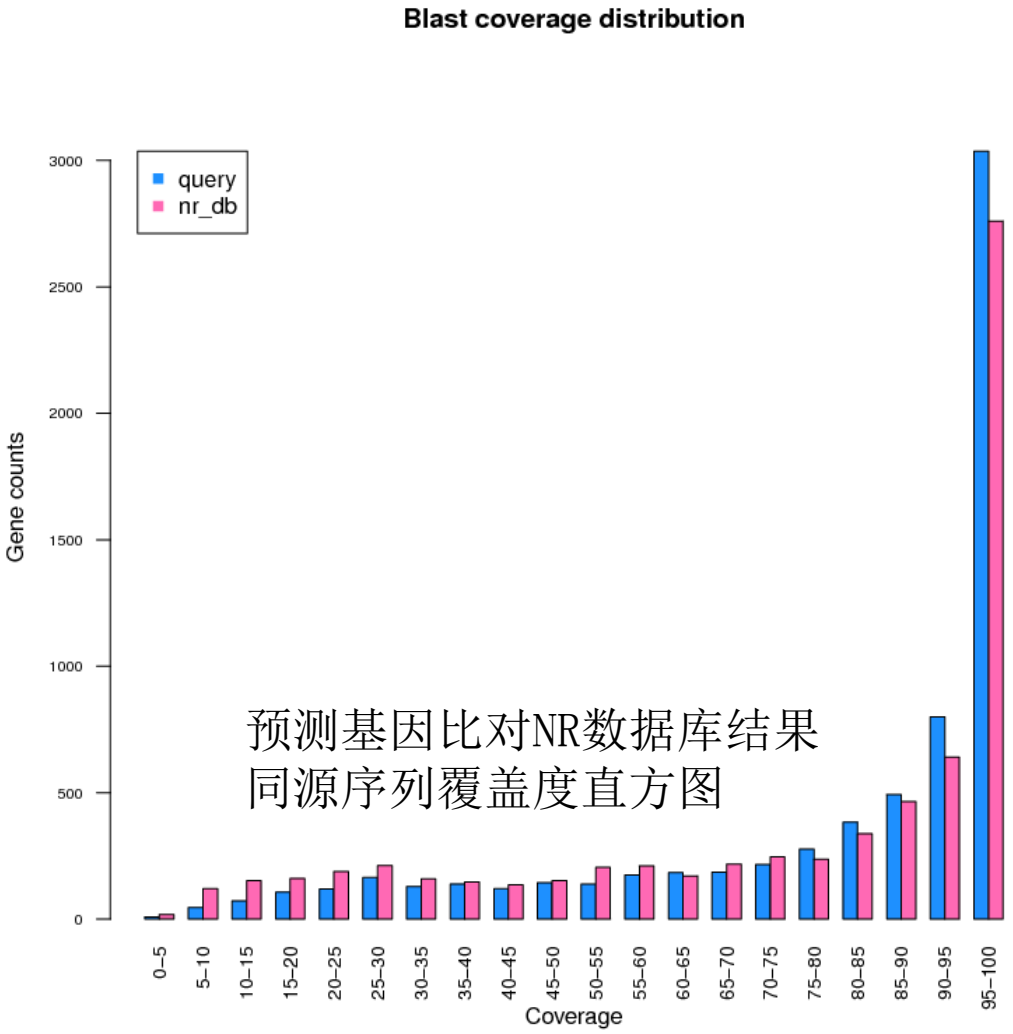
Haas, B. J., Q. Zeng, et al. (2011). "Approaches to Fungal Genome Annotation." Mycology **2(3): 118-141.**

基因预测（Gene annotation）

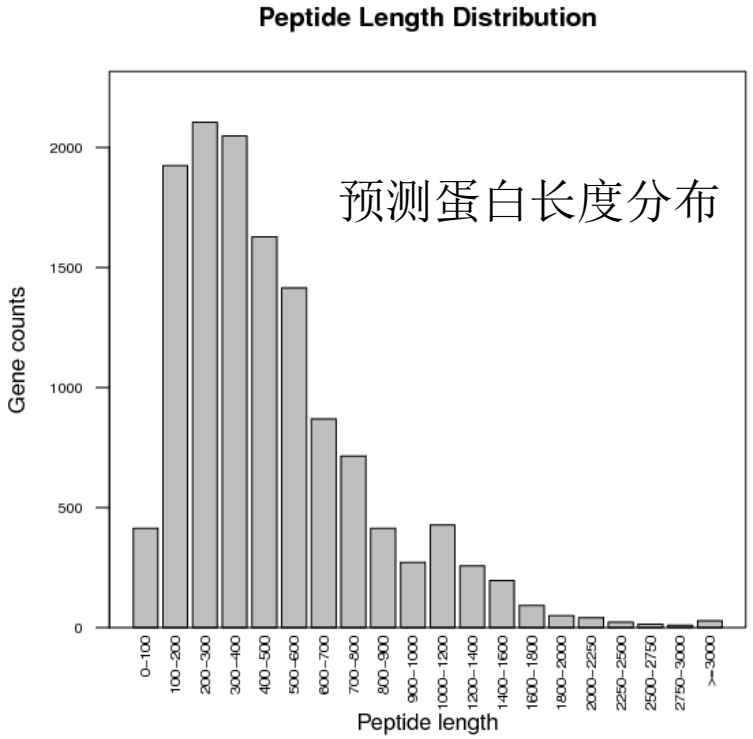


本地生物信息学平台

基因预测（Gene annotation）



预测基因比对NR数据库结果
同源序列覆盖度直方图



预测蛋白长度分布

基因预测结果准确度
和敏感度评估

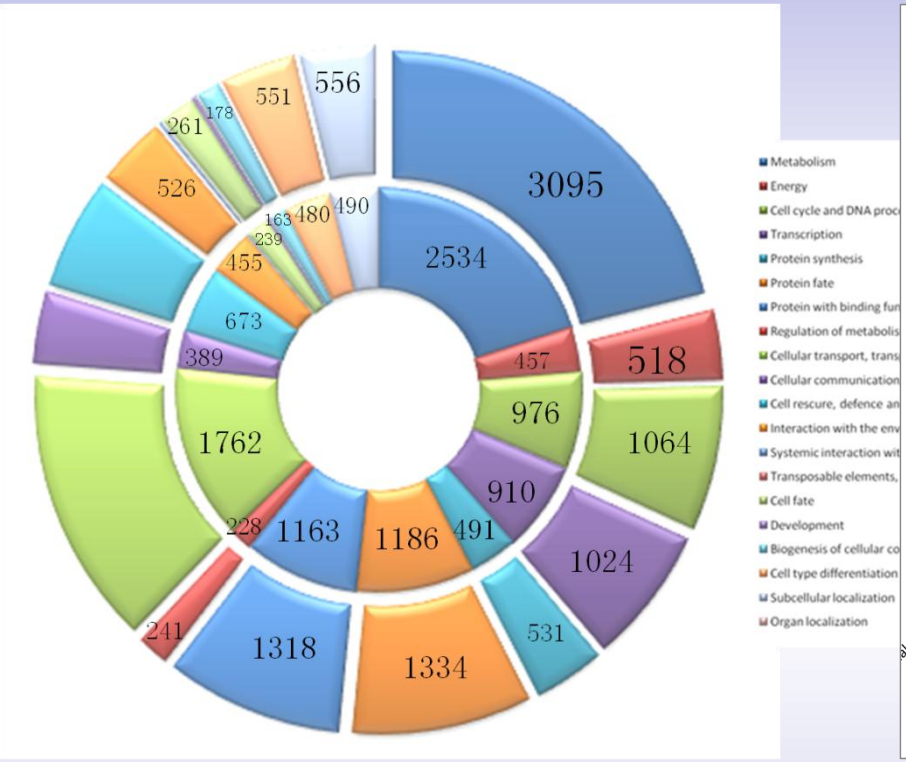
基因功能（Gene function）

从多个角度全面分析和注释基因功能，其中比较重要的分析包括：

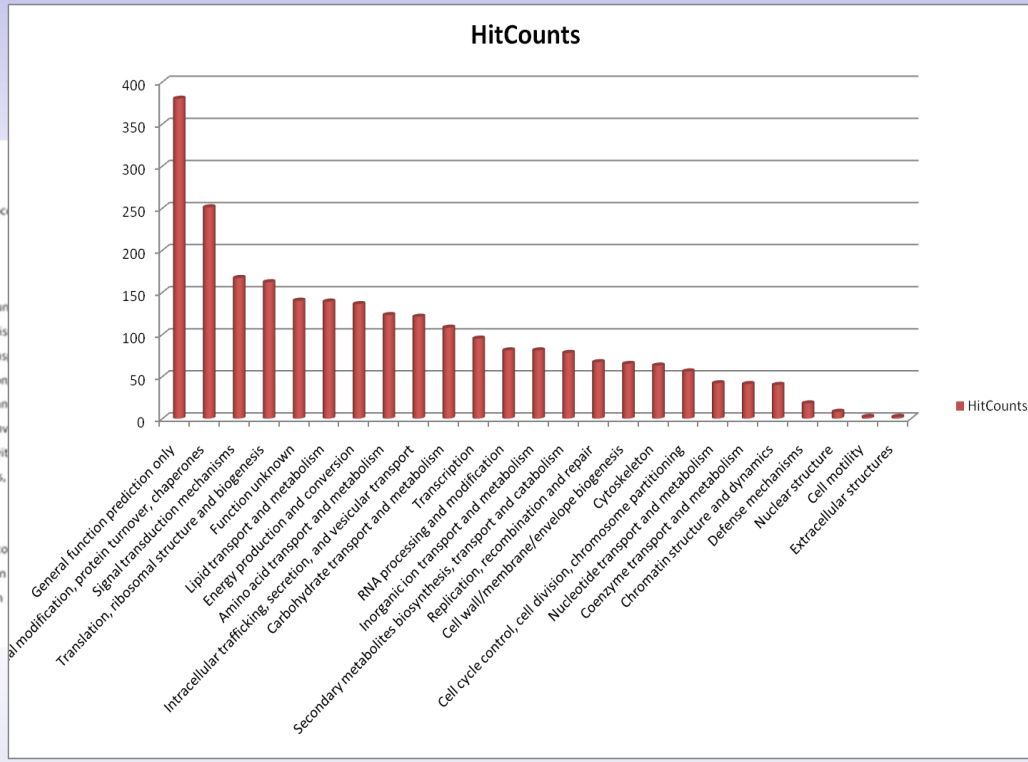
- KEGG信号通路
- GO基因本体论
- InterProScan 蛋白质功能结构域分析
- TF转录因子
- SingalP 信号肽

本地生物信息学平台

基因功能（Gene function）



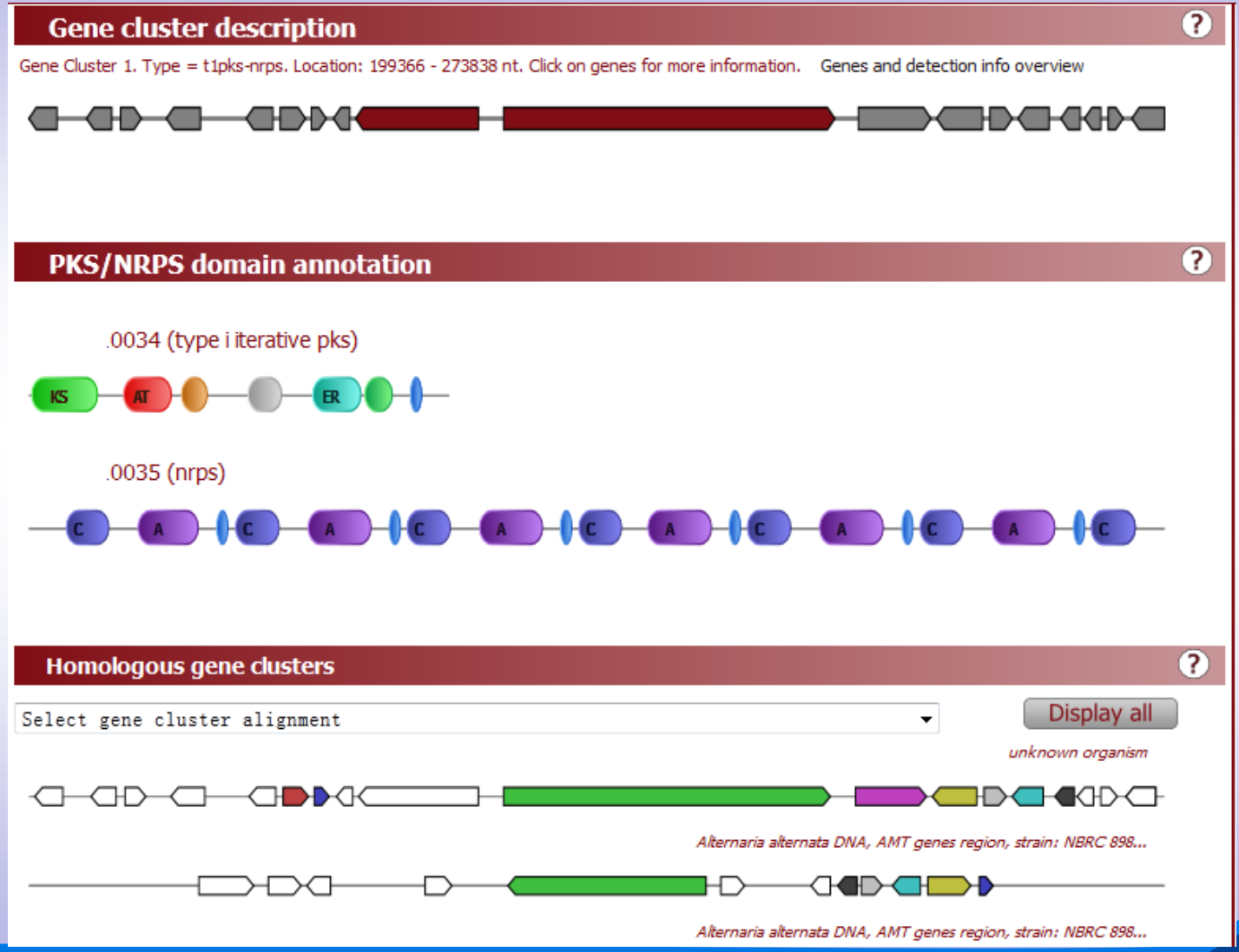
Funcat



KOG

基因注释分类图示

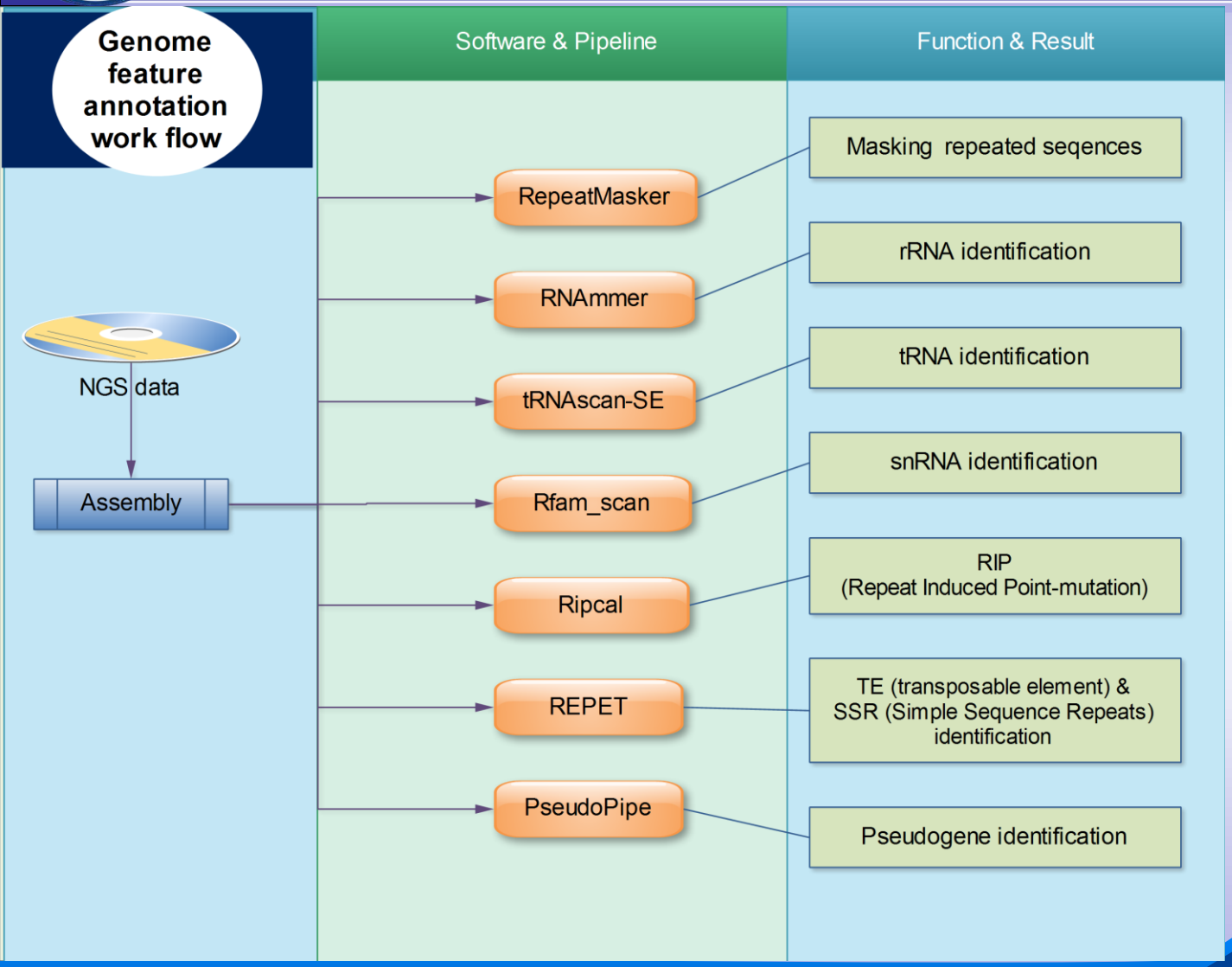
次级代谢基因簇---antiSmash分析



- 通过PKS/NRPS功能结构域的隐马模型(HMM)检索分析次级代谢关键基因
- 基于同源基因簇保守性及蛋白质功能分析次级代谢相关基因簇

软件：smurf，antiSmash，hmmpfam

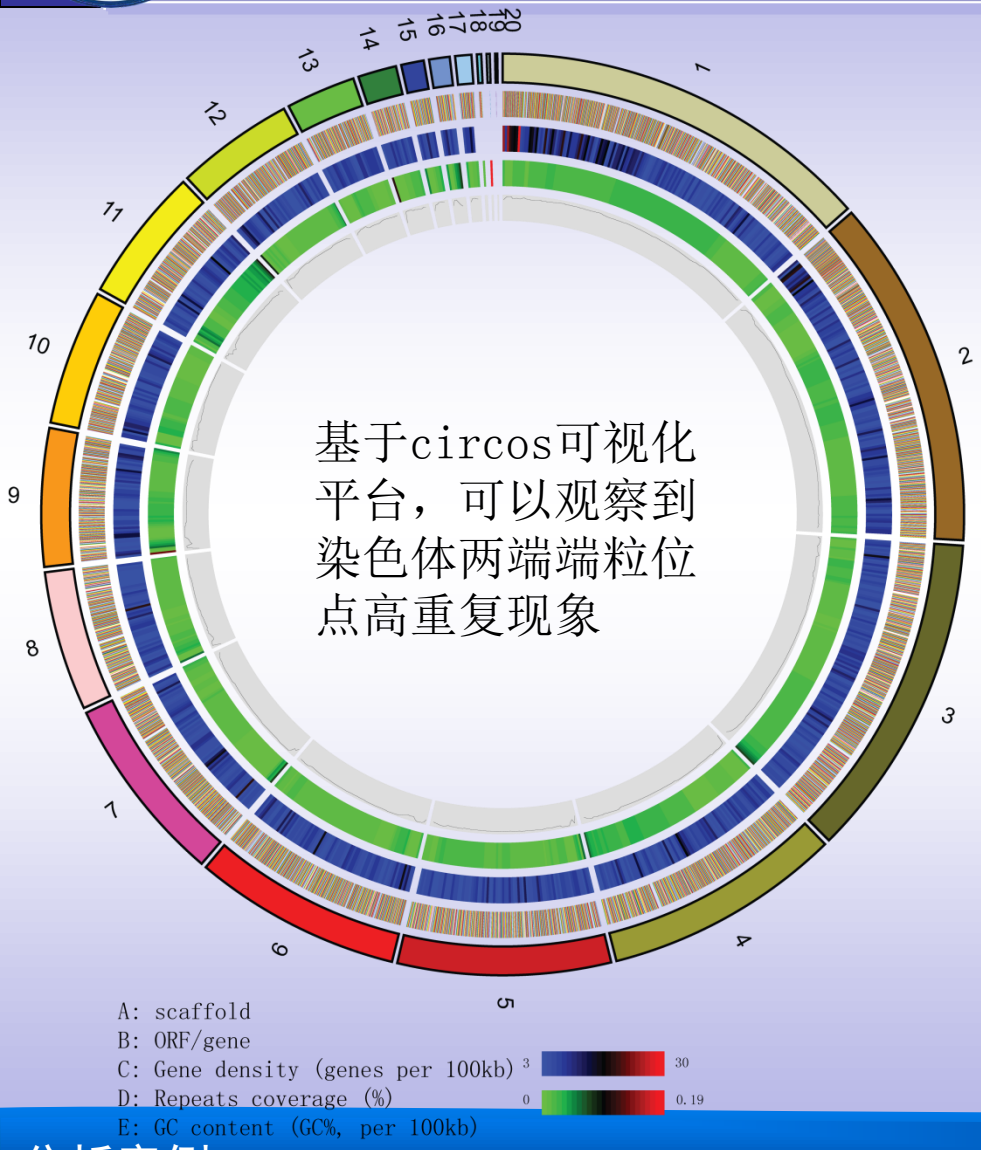
基因组特征（Genome feature）



除了预测编码基因，基因组注释还包括以下分析：

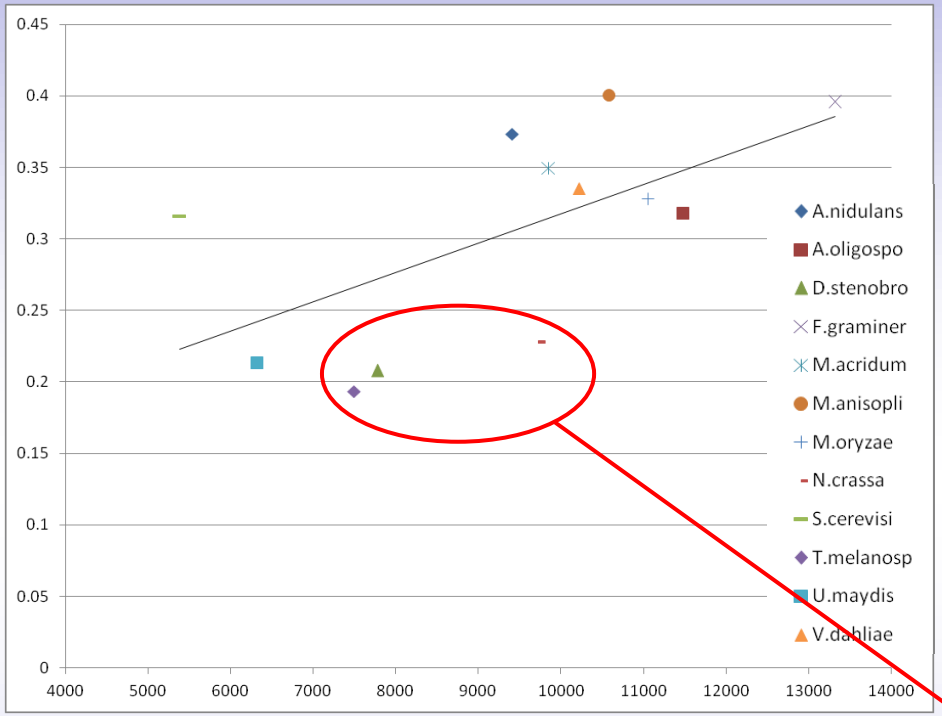
- 重复序列
- 转座子
- Non-coding RNA
- RIP (真菌)
- 假基因
- 基因组，基因，外显子，内含子等GC含量
- ……

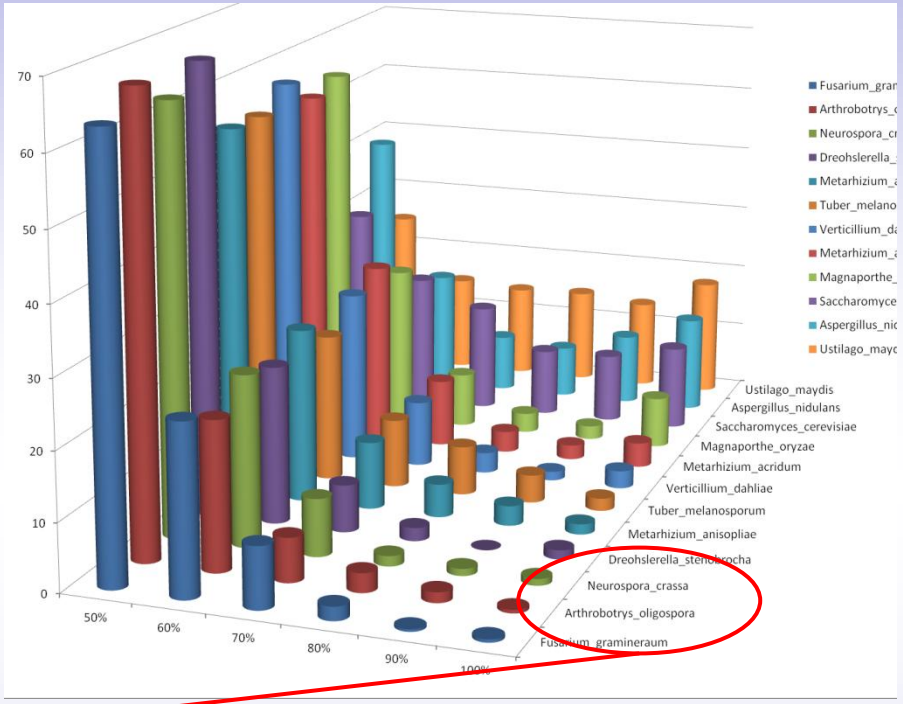
基因组特征（Genome feature）



Features	Data
Assembly Size(Mb)	37.5
Scaffold N50 (kb)	178
Coverage(fold)	78
G+C content (%)	46.01
GC Exonic (%)	51.73
GC Intronic (%)	47.05
Repeat rate(%)	1.68
Protein-coding genes	9405
Gene density (per Mbp)	250.8
Exons per genes	2.53
tRNAs	72
rRNAs	19
SM(Secondary Metabolism)genes	28
TE	15%

RIP (Repeat Induced Point-mutation)

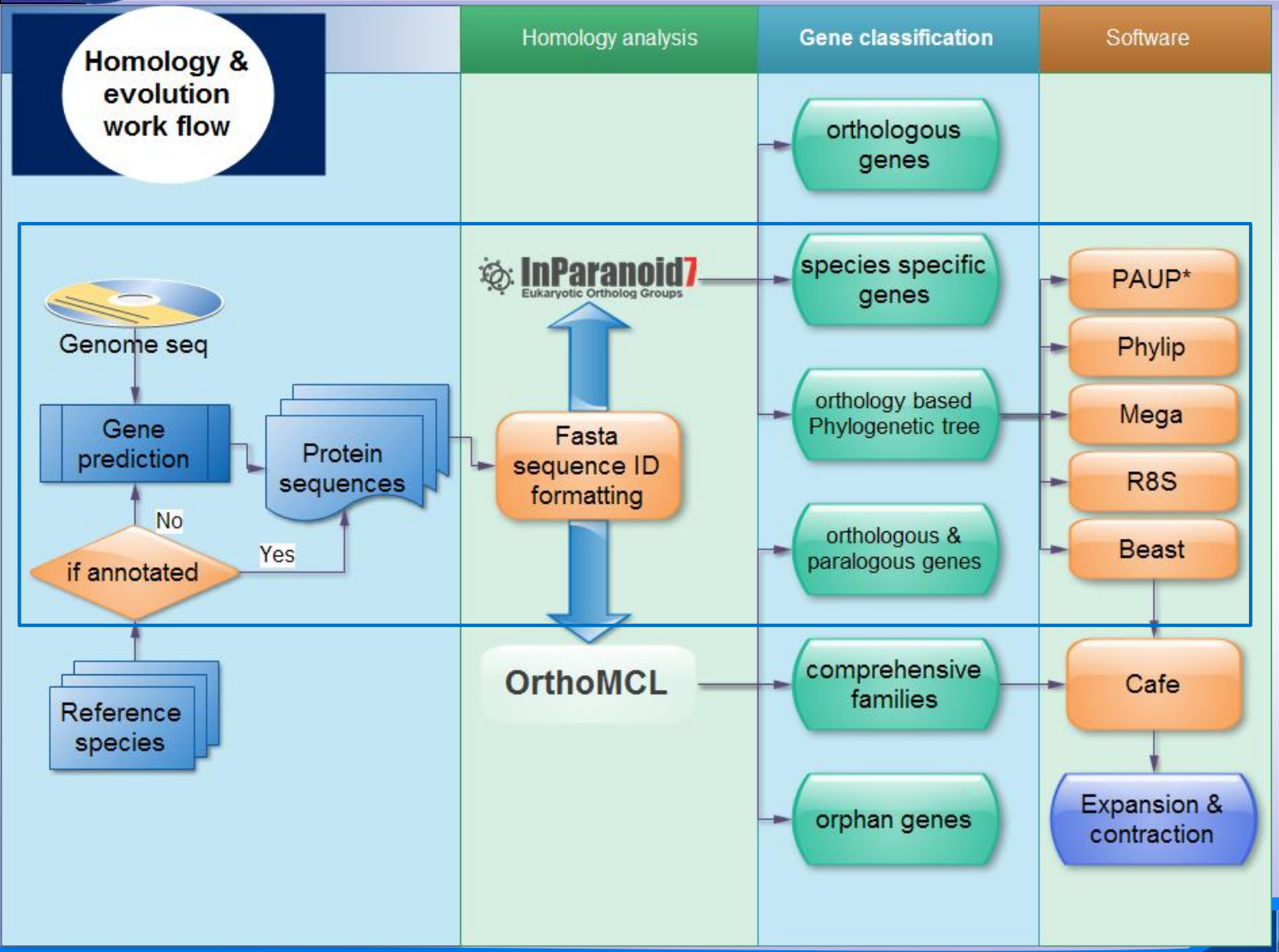




RIP程度与多基因家族基因数量与相似度的关系

RIP要求最少400bp和80%以上的序列相似度，并通过点突变抑制序列复制。因此，在RIP程度较高的真菌基因组中相对包含较少比例的多基因家族基因；物种蛋白序列在all-vs-all blast比对结果中，80%以上相似度比对数量大幅减少。

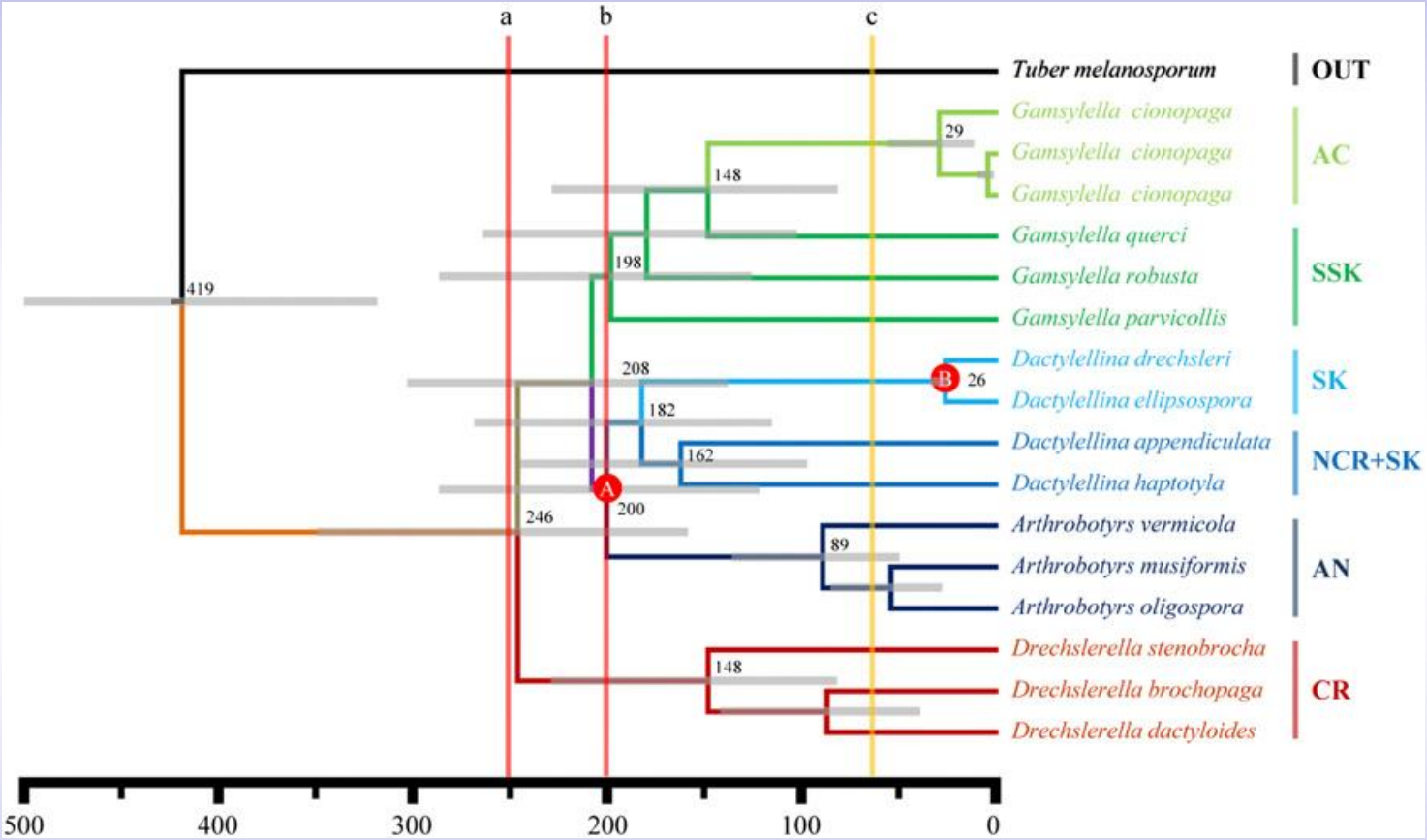
基于全基因组同源基因的系统发育树



对所有物种基因序列进行同源性分析，筛选所有物种共同具有的全部直系同源基因，并以之构建系统发育树

优点：可以避免单个或者少量基因进化速率不同造成的系统误差

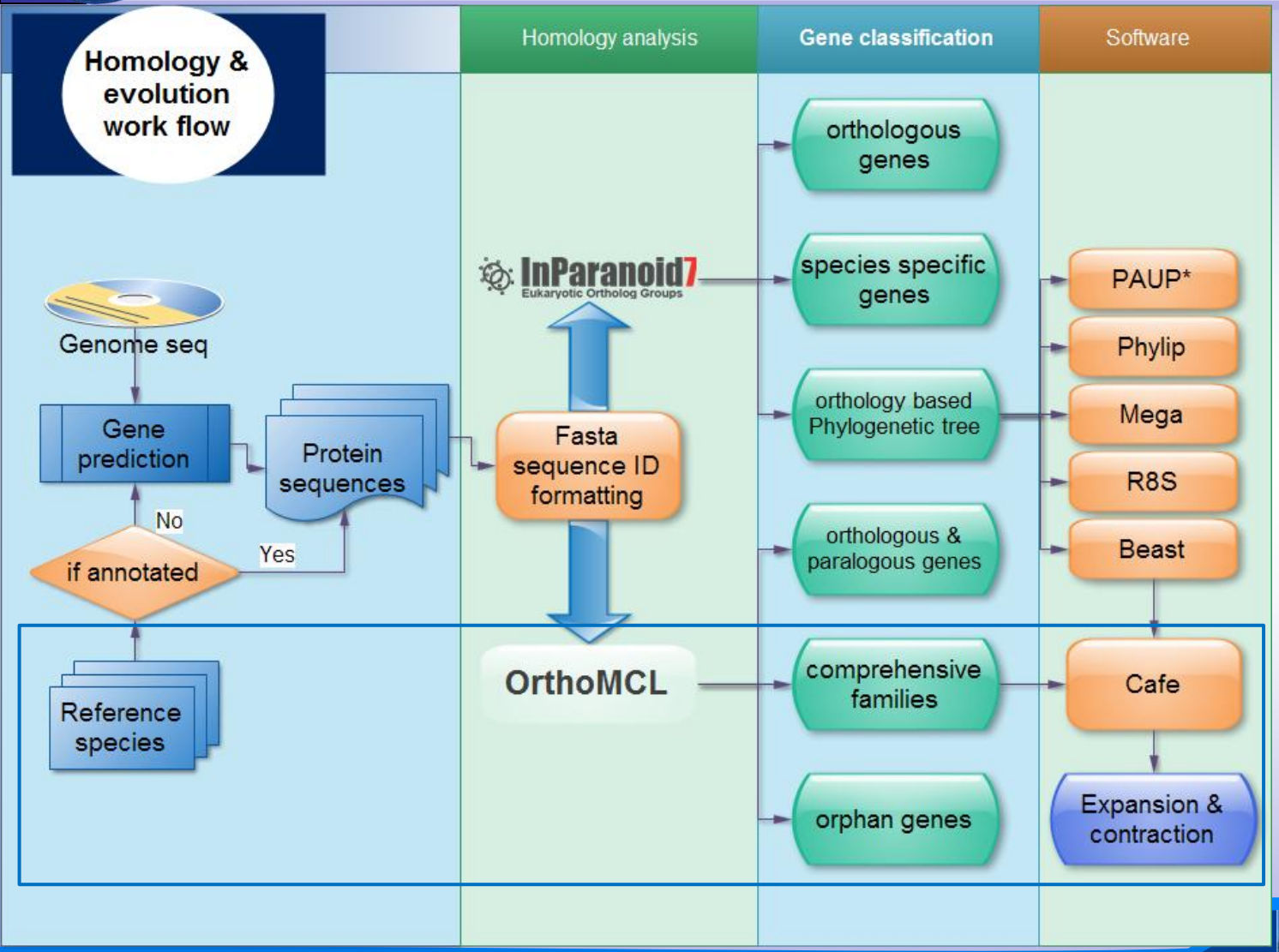
基于全基因组同源基因的系统发育树



- 基于1367 直系同源基因构建的系统发育树;
- 采用3个已知的进化时间点估算捕食线虫菌进化时间root node (300–500 Mya), node A (100–500 Mya), and node B (24–500 Mya).

Yang, E., Xu, L., Yang, Y., Zhang, X., Xiang, M., Wang, C., An, Z., Liu, X., 2012, Origin and evolution of carnivorism in the Ascomycota (fungi). *Proceedings of the National Academy of Sciences*.

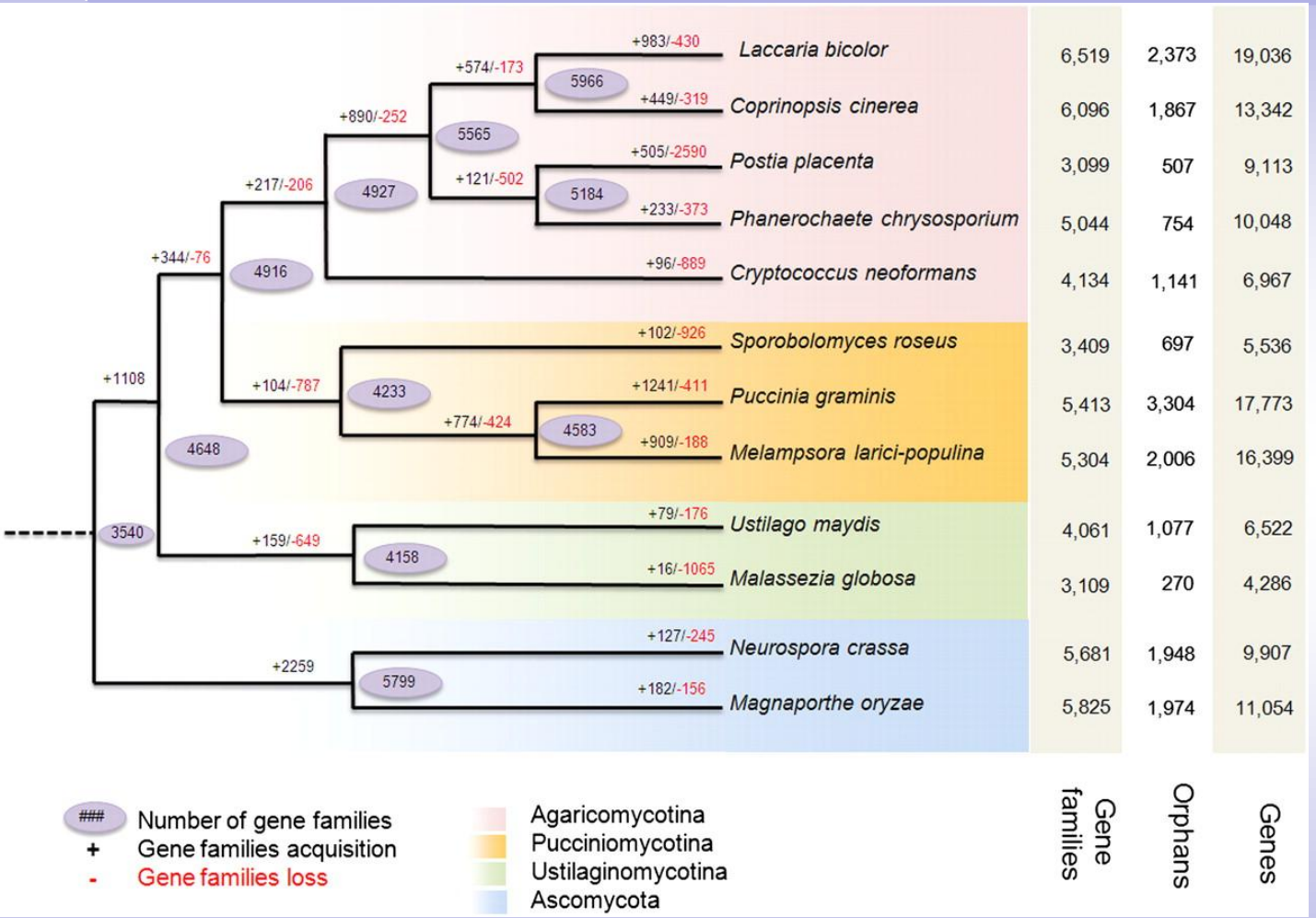
蛋白质家族进化分析



蛋白质家族进化分析用于筛选在进化过程中得到基因数量扩张（expansion or gain）或者压缩（contraction or loss）的基因家族。

常用软件：cafe, bandirate

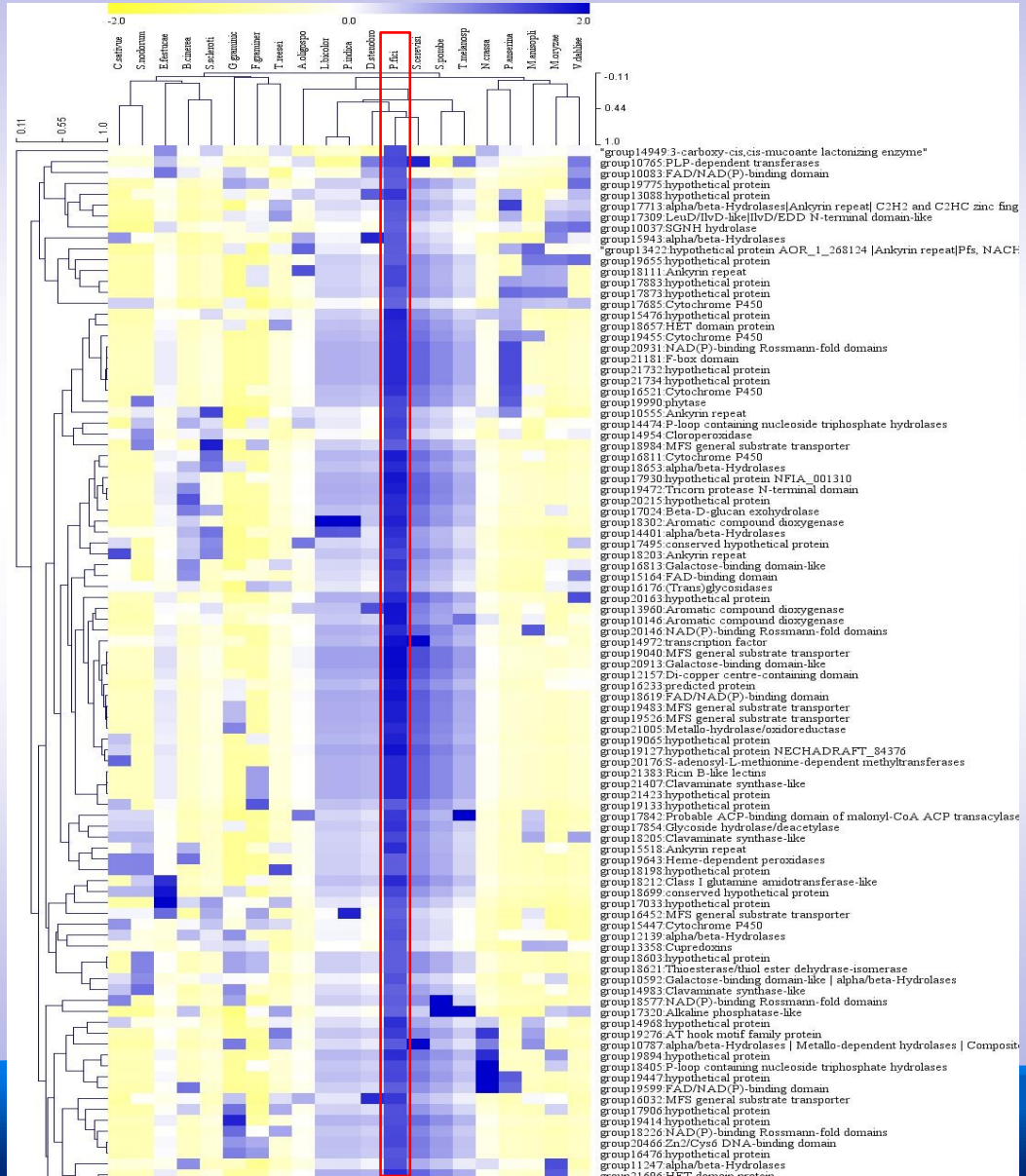
蛋白质家族进化分析



- 基因家族在进化过程中的获得和缺失往往跟其特定生物学表型相对应。
- 该分析必须在系统发育树的背景下进行，因此跟常见的差异分析不同。

Duplessis, S., C. A. Cuomo, et al. Obligate biotrophy features unraveled by the genomic analysis "of rust fungi." Proceedings of the National Academy of Sciences **108**(22): 9166.

蛋白质家族进化分析



Expansion:

MFS general substrate transporter

Aromatic compound dioxygenase

Cytochrome P450

FAD-binding domain Oxidoreductase

Ankyrin repeat

Zn_2/Cys_6 DNA-binding domain

……

分析并收集各个选定物种的蛋白质家族基因数量，采用BandiRate进行gain-loss分析，筛选扩张的蛋白质家族。对筛选数据进行Z-score标准化，采用欧氏距离算法计算节点距离，做聚类分析并作图。

分析案例

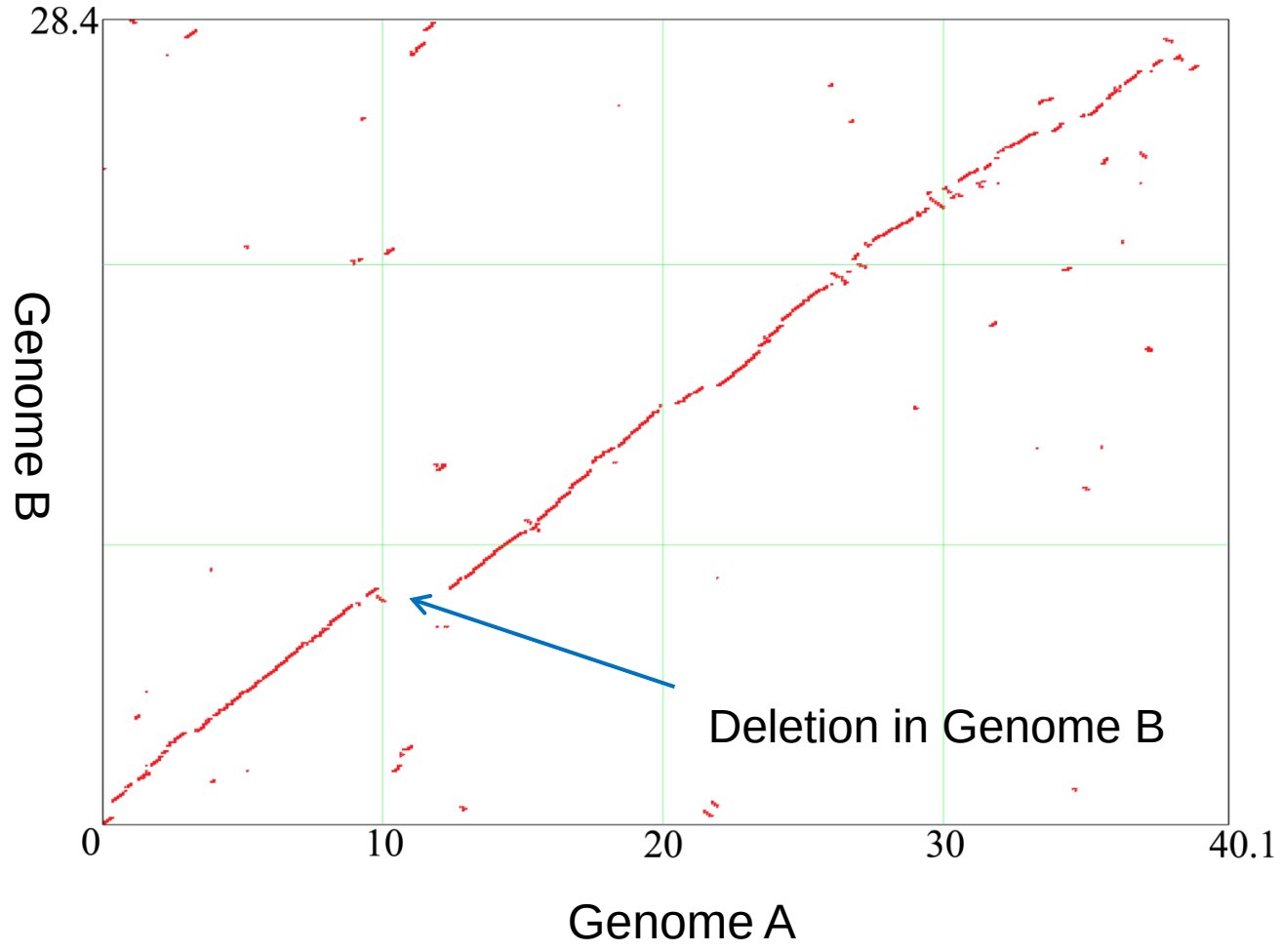
蛋白质家族/功能结构域差异分析

- Database
- Gene Ontology
- Funcat
- KEGG
- KOG/COG
- IPRSCAN
- Protein family
- PKS/NRPS
- MCL group

DB_Desc	*pvalue*	*case1*	*case2*	*control1*	*control2*	*control3*
SUBFAMILY NOT NAMED	0.003884	0	0	5	6	5
Flavodoxin, conserved site	0.005865	0	0	5	4	4
Salmonella virulence plasmid 65kDa B protein	0.008163	1	1	5	4	5
SpvB	0.008163	0	0	3	4	4
Chromo domain subgroup	0.008163	0	0	3	4	4
Myelin P0 protein	0.009852	0	0	3	4	3
Insecticide toxin TcdB middle/N-terminal	0.009852	0	0	3	4	3
Integrin alpha beta-propellor	0.009852	0	0	3	3	4
PUTATIVE UNCHARACTERIZED PROTEIN	0.009852	0	0	3	4	3

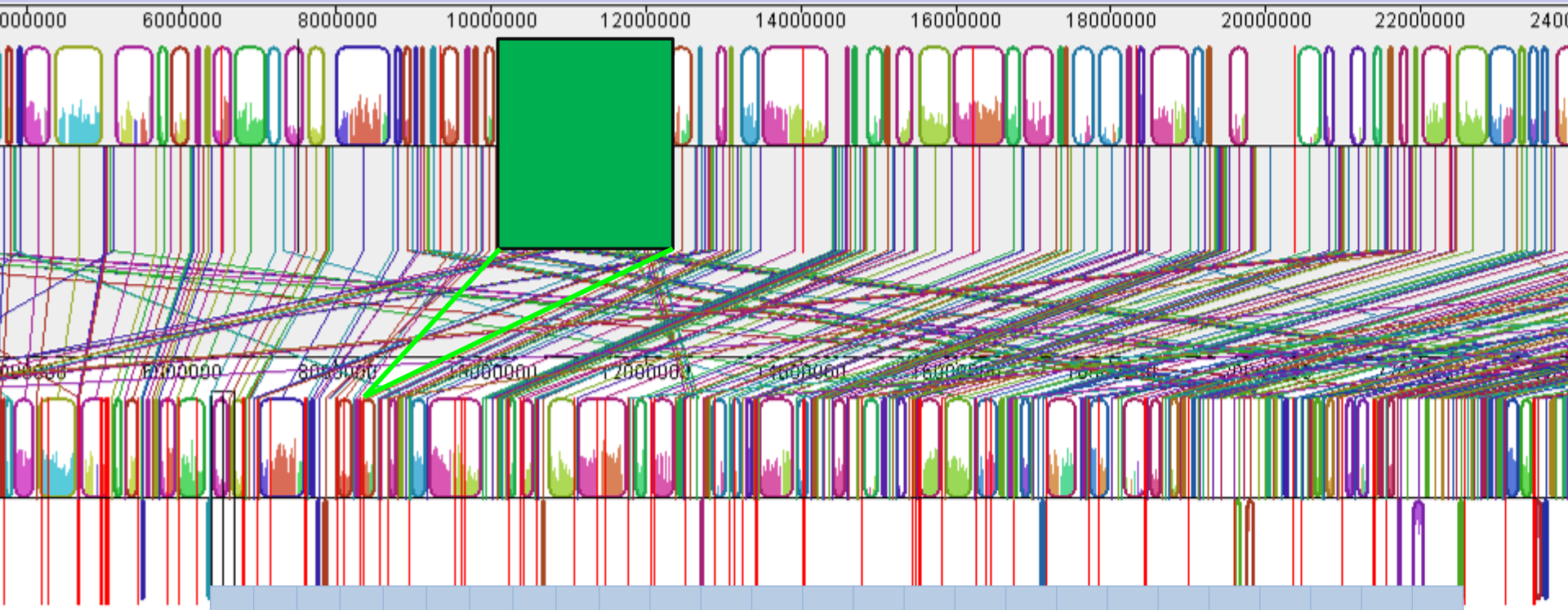
通过ttest等统计学方法进行分组差异分析，以筛选不同表型物种基因数量显著差异的蛋白质家族或者生物学功能，可应用于IprScan，GO，KOG，蛋白质家族等多个基因功能注释结果，且可对不同表型分组分别统计分析

比较基因组学---Dot Plot



- 一般用于近缘物种基因组比较
- 对组装质量要求较高
- 需要对基因组Scaffold/Chromosome序列进行相对排序以后才可以作图
- 软件： Mauve, Mummer, Argo等

比较基因组学---共线性分析



Genome A:	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Genome B:	0	-18	-19	1	6	3	-4	13	5	2	7	9	-10	-12	11	14	16	15	17	20	22	-24	21	23	25	8

共线性分析 (Locally Collinear Blocks, LCBs)

水平转移基因（HGT）--- features

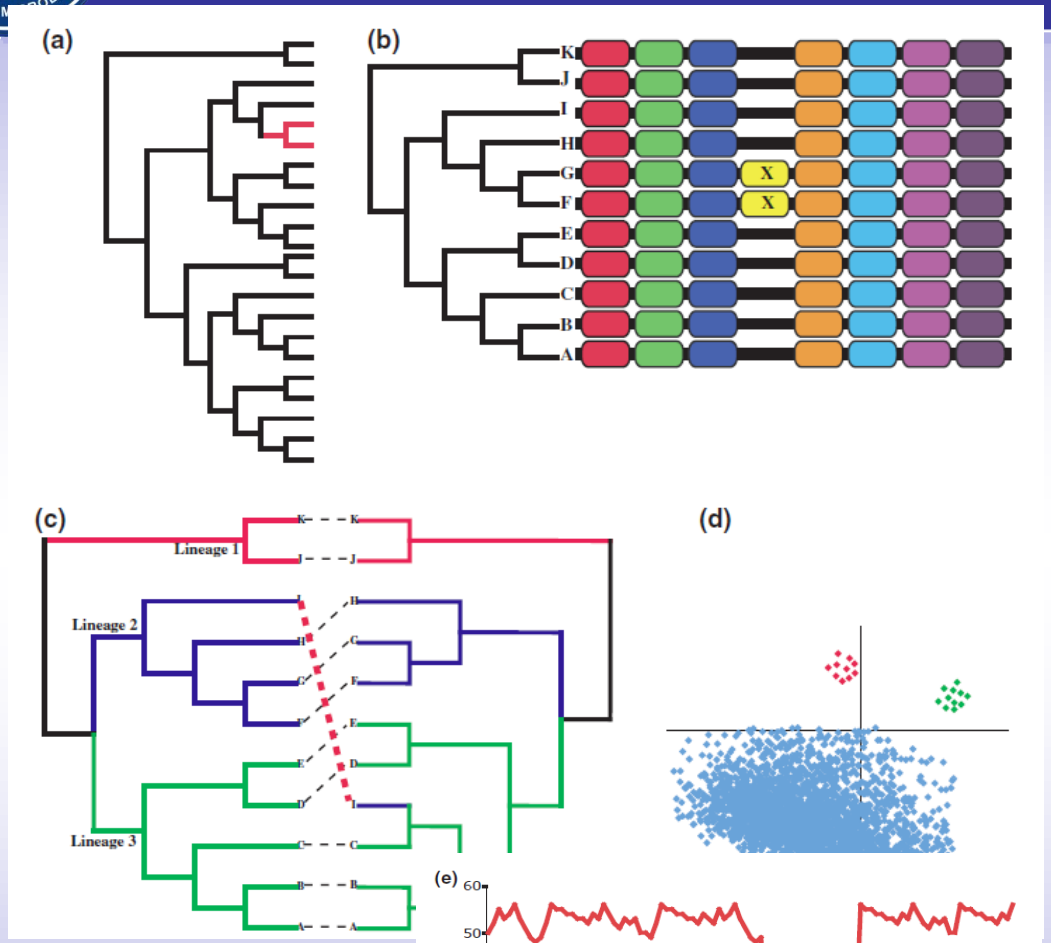
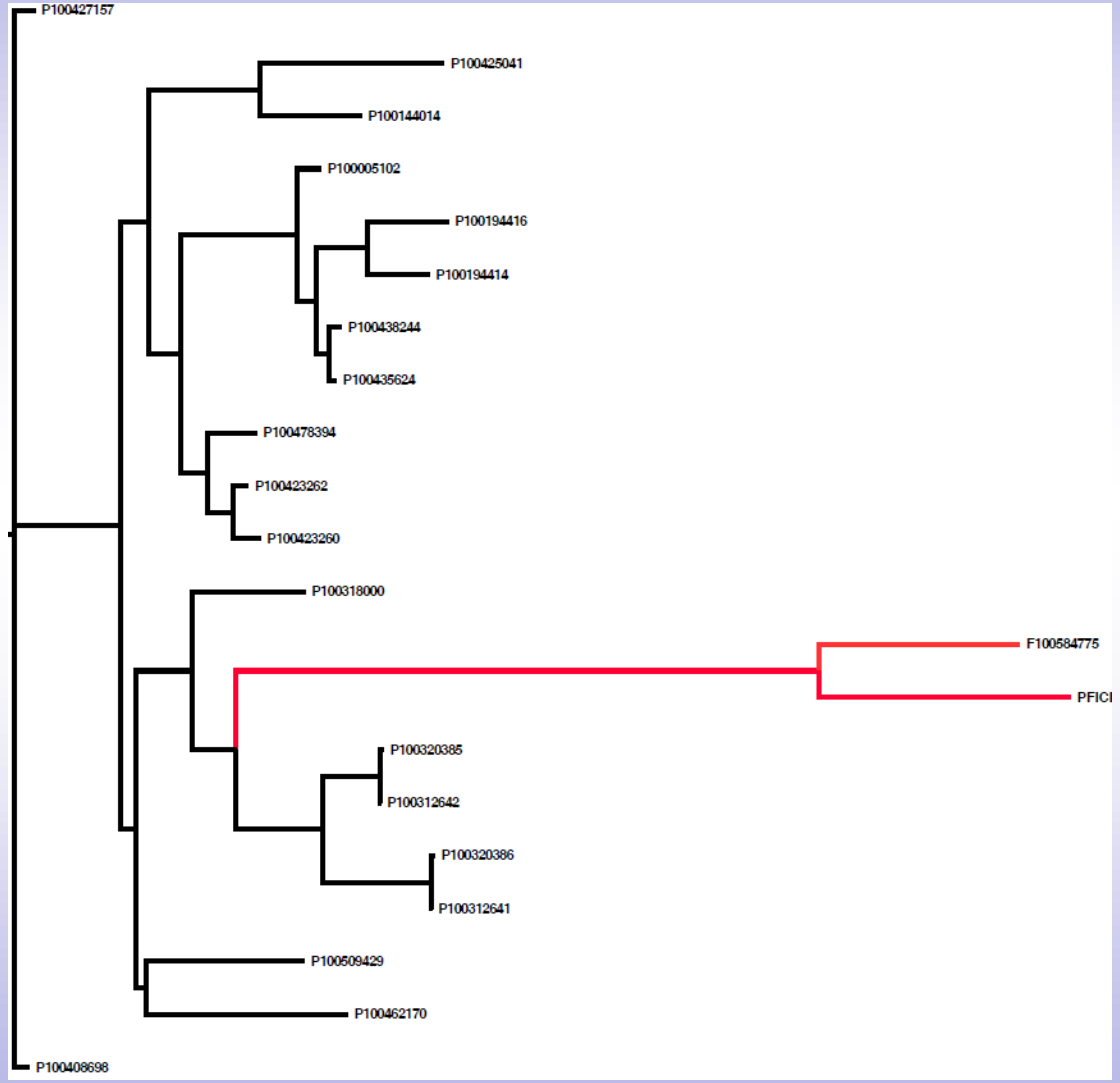


Fig. 1. Detecting incidences of HGT. (a) Patchy phyletic distribution, gene of interest is not found in closely related relatives, and orthologs can only be located in distantly related species. (b) Gene of interest located in conserved synteny block and absent from closely related species. May also indicate a gene loss but similarity-based searches can help validate if it is a potential HGT event or loss. (c) Phylogenetic inference, species gene tree on the right differs from gene tree on the left. Phylogenetic incongruence can be used to detect HGT and also determine the donor species. (d) Codon usage variation, native genes have a preferential codon usage pattern (blue dots); recently transferred genes have yet to ameliorate to their new hosts genome and still display the codon usage pattern of their cognate genome. (e) Variation in GC composition along a chromosome may indicate that alien genetic material has recently been acquired. In this case, the transferred DNA has a GC content lower than the recipient genome.

Fitzpatrick, D.A. (2012) Horizontal gene transfer in fungi, *FEMS Microbiology Letters.*

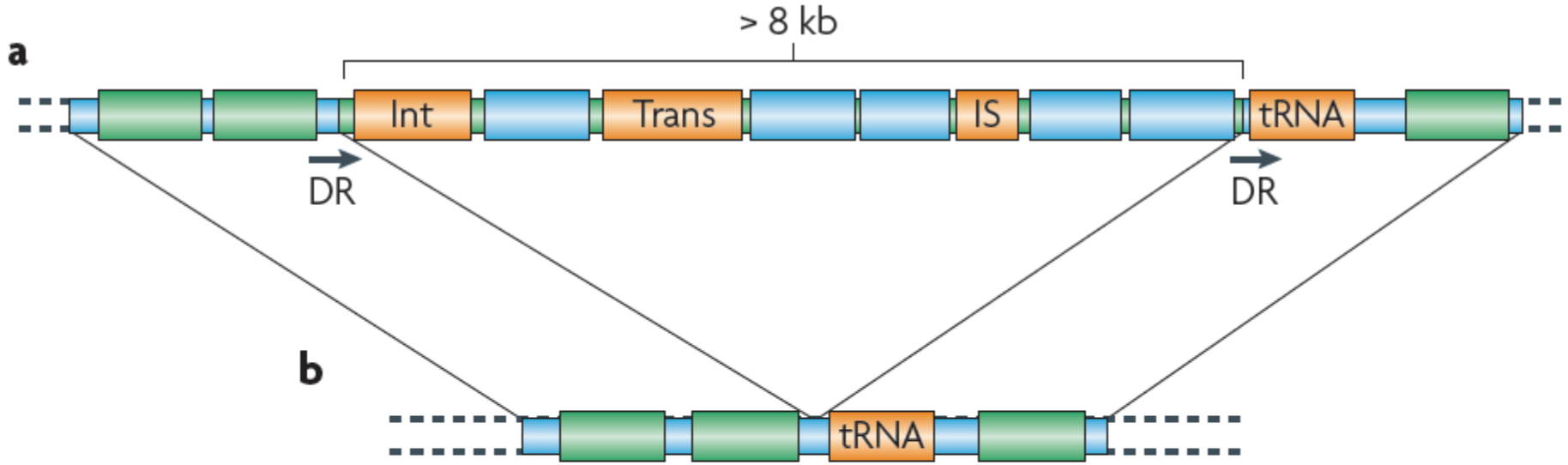
水平转移基因（HGT）--- predictions

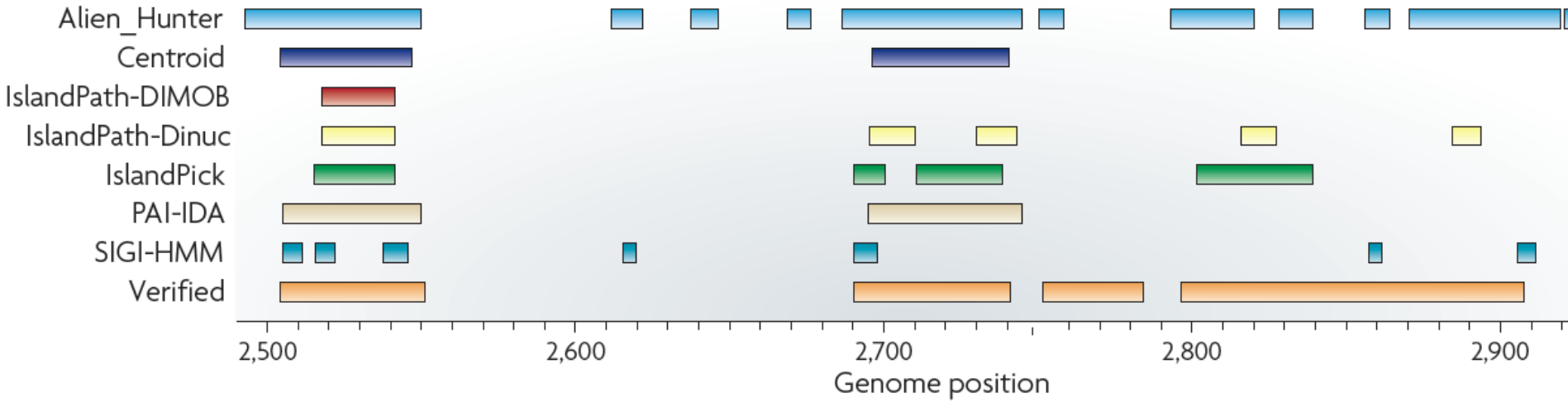


- 通过blast比对原核生物，植物，真菌，无脊椎动物，哺乳动物等基因数据库并根据相似度，覆盖度，以及匹配基因数量等筛选候选HGT
- 提取同源基因序列，进行多序列比对以及共有保守序列并做进化树。可用来筛选HGT以及水平转移方向
- 结合蛋白质功能分析确定HGT的生物学意义

软件：Blast, Gblocks, Clustalw, Phylip

基因组岛（GI）--- features



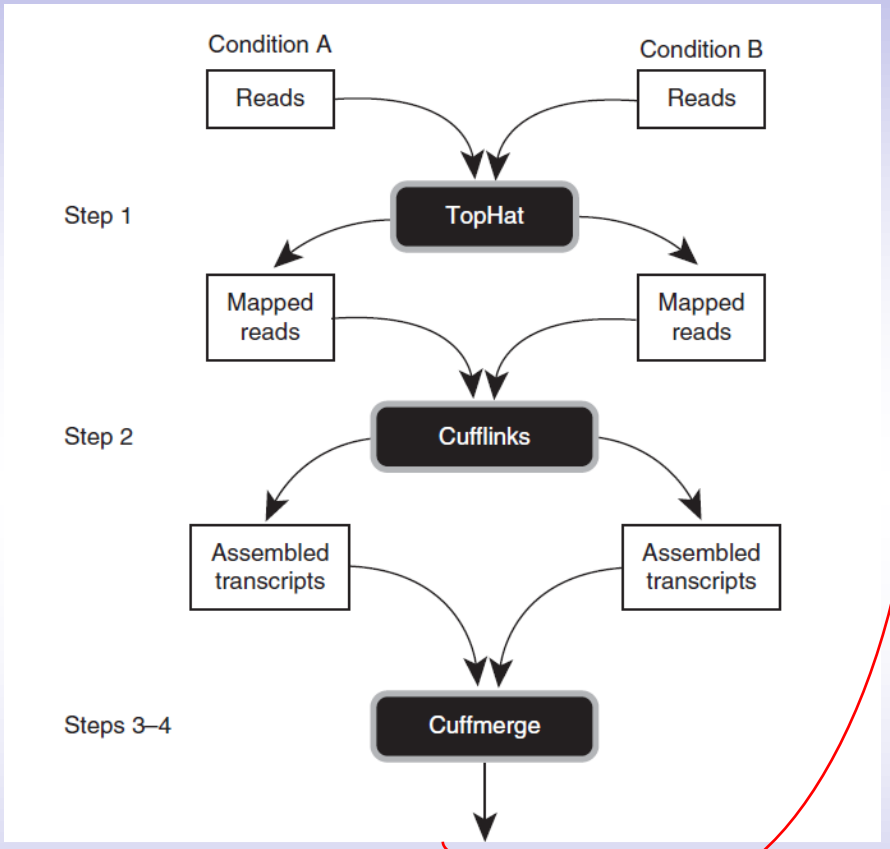


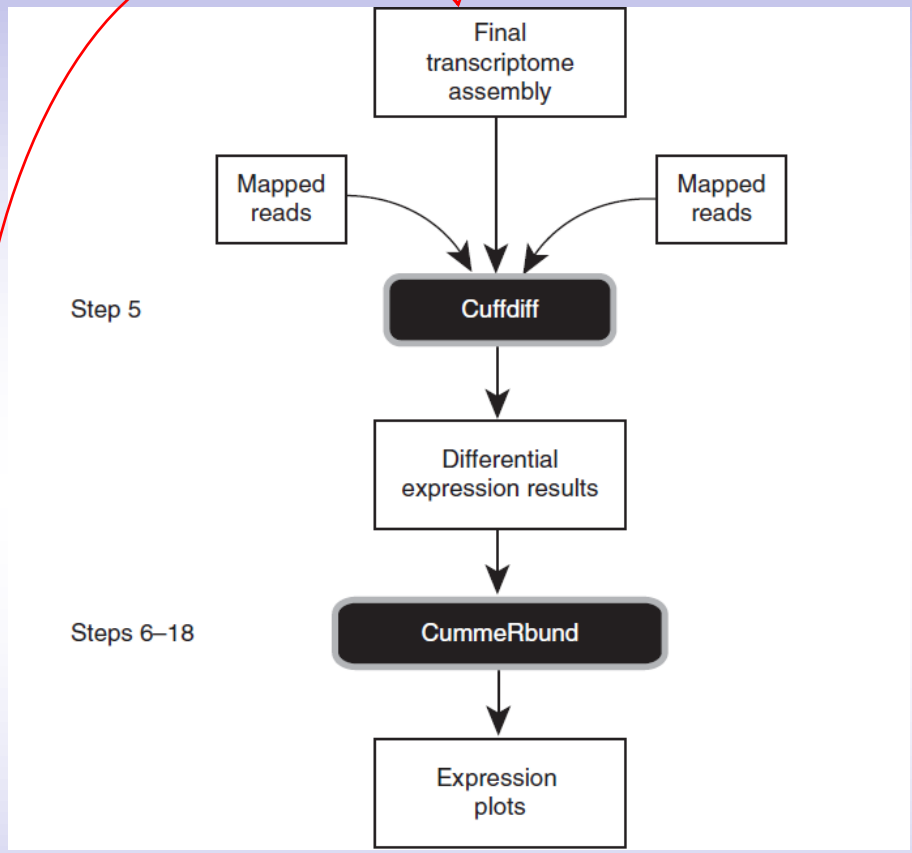
Langille, M.G.I., Hsiao, W.W.L. and Brinkman, F.S.L. (2010) Detecting genomic islands using bioinformatics approaches, *Nature Reviews Microbiology*, ***8, 373-382.***

基因组岛（GI）--- prediction

GI No.1: 7 genes	HMMPfam	superfamily	SIGNALP	WOLFPSORT
Fungi1_03412				nucl
Fungi1_03413	Stigma-specific_protein_Stig1		SignalP-noTM	
Fungi1_03417	Transcription_factor,_fungi // Zn(2)-C6_fungal-type_DNA-binding_domain	Zn(2)-C6_fungal-type_DNA-binding_domain		nucl
Fungi1_03414				
Fungi1_03411 Fungi1_04263	Serine/threonine-protein_kinase-like_domain	Protein_kinase-like_domain		
Fungi1_03415			SignalP-noTM	extr
GI No.2: 6 genes				
Fungi1_00493				nucl
Fungi1_00490	Zinc_finger,_C6HC-type	RING/U-box		nucl
Fungi1_00492	Src_homology-3_domain	Src_homology-3_domain		nucl
Fungi1_00489				
Fungi1_00491	Phosphatidylethanolamine-binding_protein_PEBP	Phosphatidylethanolamine-binding_protein_PEBP	SignalP-noTM	

转录组学数据分析

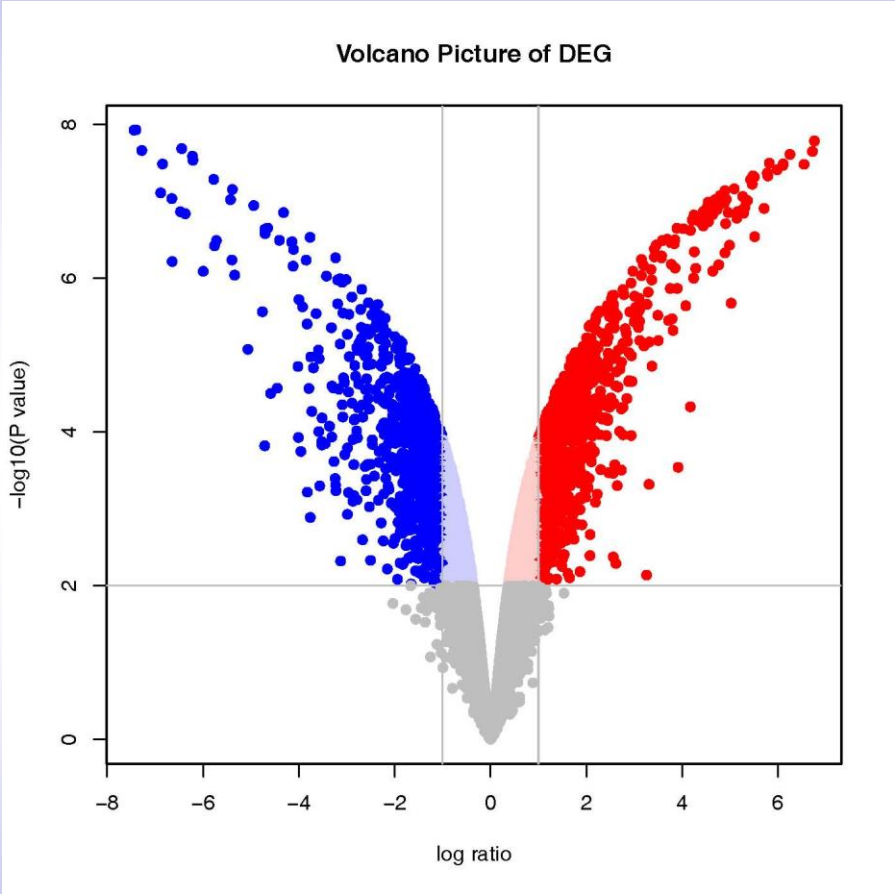




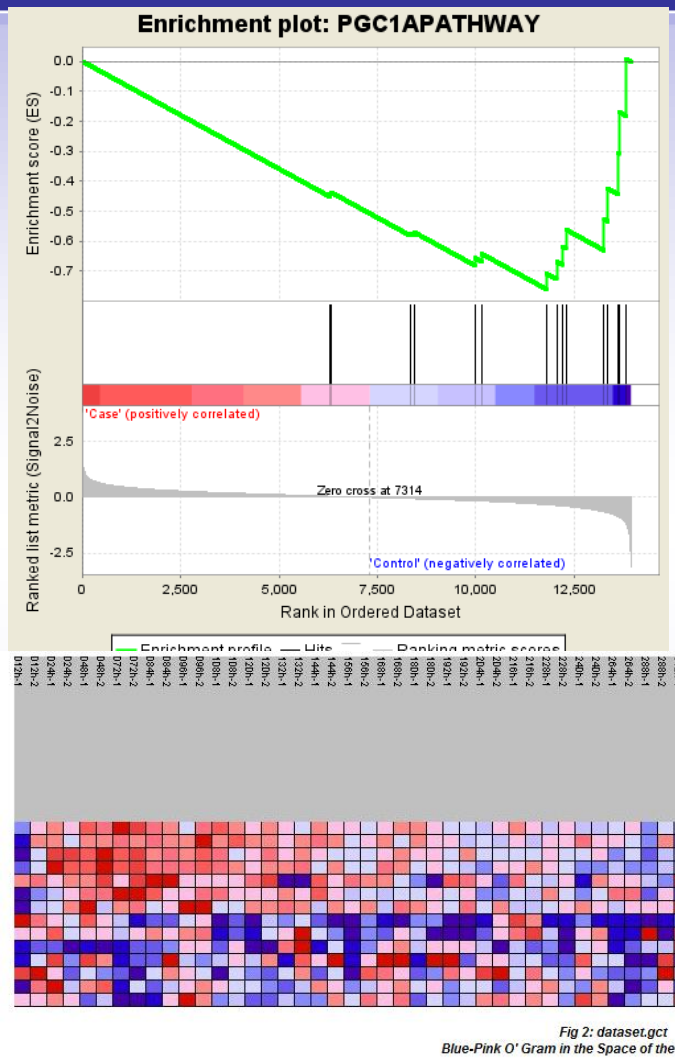
Trapnell, C., *et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, Nature Protocols,* ***7, 562-578.***

使用TopHat和Cufflinks组合流程分析转录组定量数据，一般要求有参考基因组。差异表达分析也可用NOISeq，EdgeR等

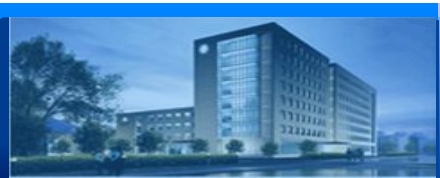
转录组学---差异表达和功能富集



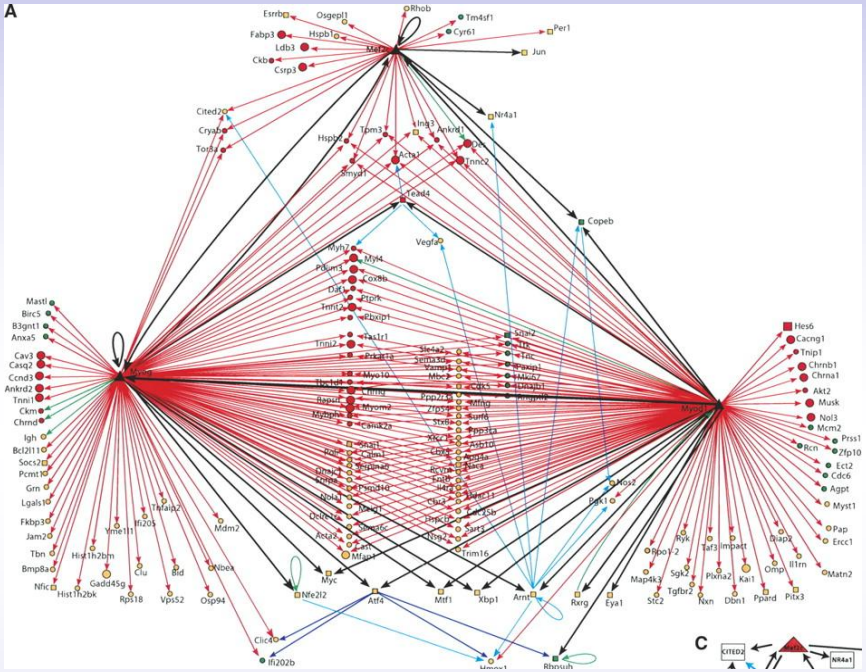
火山图，红色表示表达上调，蓝色表示表达下调



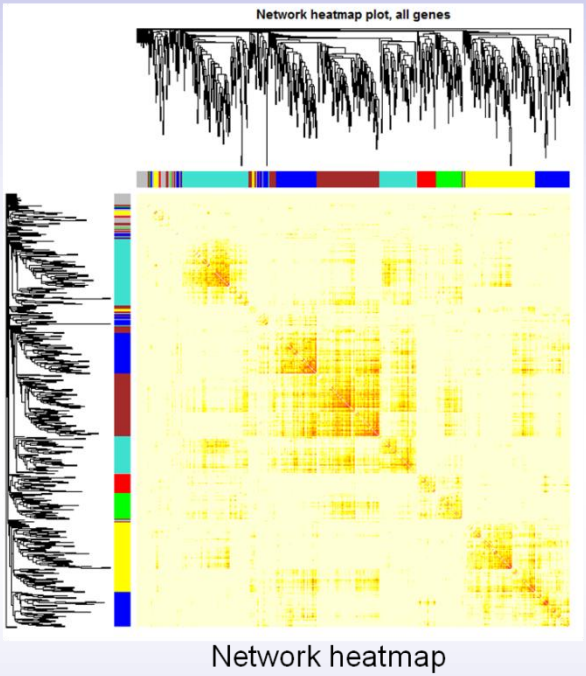
GSEA分析结果，用于筛选显著富集（激活或者抑制）的生物学功能



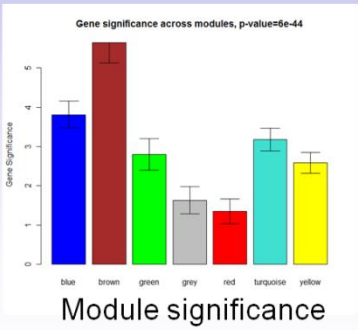
系统生物学和数据整合



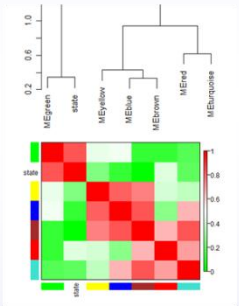
转录调控网络



Network heatmap



Module significance



Module state

共表达基因模块

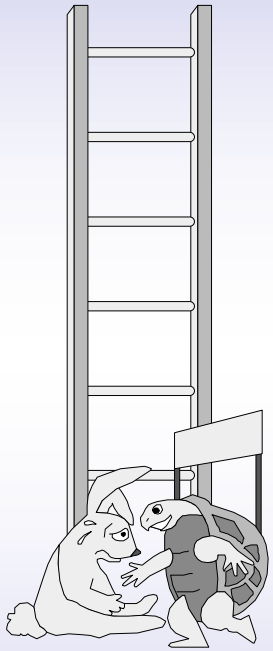
系统生物学分析旨在整合基因组、转录组、蛋白质组等不同层次信息以理解生物系统各不同部分之间的相互关系和相互作用

路漫漫其修远兮……

个性化分析

交流，生物学反馈，确立分析重点

常规分析



生物学高通量数据的准确解读理应是**生物学导向的信息学手段的正确应用**。

致 谢

感谢微生物所李彦副所长对平台工作的支持！

感谢真菌室各位领导为生信平台发展创造条件！

感谢真菌室王秀娜等同学在平台建设的文献调研方面做出的杰出贡献！

感谢本组喻浴飞、章小灵在平台建设和数据分析方面付出的辛苦劳动和不懈努力！


Thanks
for your attention!!!