Analysis of genomic data

Longjiang Fan (樊龙江)

Prof. of crop genetics and breeding
Institute of Crop Science
College of Agriculture and Biotechnology, Zhejiang University
http://cab.zju.edu.cn/nxx/

Prof. of bioinformatics
Institute of Bioinformatics and IBM Computational Lab, Zhejiang University
http://ibi.zju.edu.cn
James D. Watson Institute of Genome Sciences, Zhejiang University
http://wigs.zju.edu.cn

The Bioinplant Lab
(Bioinformatics and Plant Genomics Lab)
http://ibi.zju.edu.cn/bioinplant/

教学内容

	主题内容	教学人员
1	绪论（基因组数据及其分析内容）	樊龙江
2	基因组拼接与注释	樊龙江/贾磊
3	转录组表达及其数据分析（RNA-SEQ）	褚琴洁
4	表观基因组（非编码RNA）及其数据分析	褚琴洁
5	植物群体基因组变异与选择分析	叶楚玉/吴东亚
6	单细胞组学分析	樊龙江/陈洪喻
7	研究进展文献与交流	樊龙江/叶楚玉

教学目标

- 掌握大规模基因组数据分析内容、方法和工具，能够进行实战分析

- 《基因组数据分析》是基因组学和生物信息学结合的一门课程
  - 基因组学：生物学
  - 生物信息学：方法学

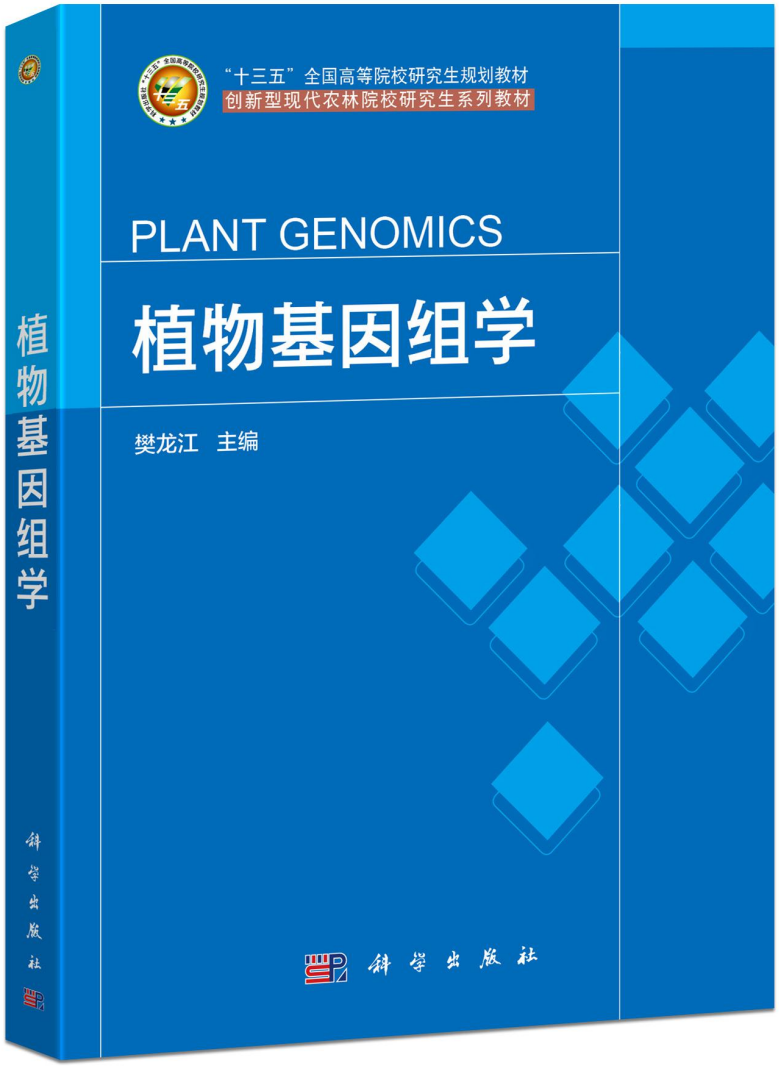
课程成绩

- 五个主题最近一年（2020）文献综述：分析内容及其方法
  - 基因组测序与拼接
  - 群体基因组
  - 转录组（RNA-SEQ）
  - 表观组（非编码RNA）
  - 单细胞组

- 任选一个主题
- 要求中文，综述内容包括3页（不包括文献）

参考教材

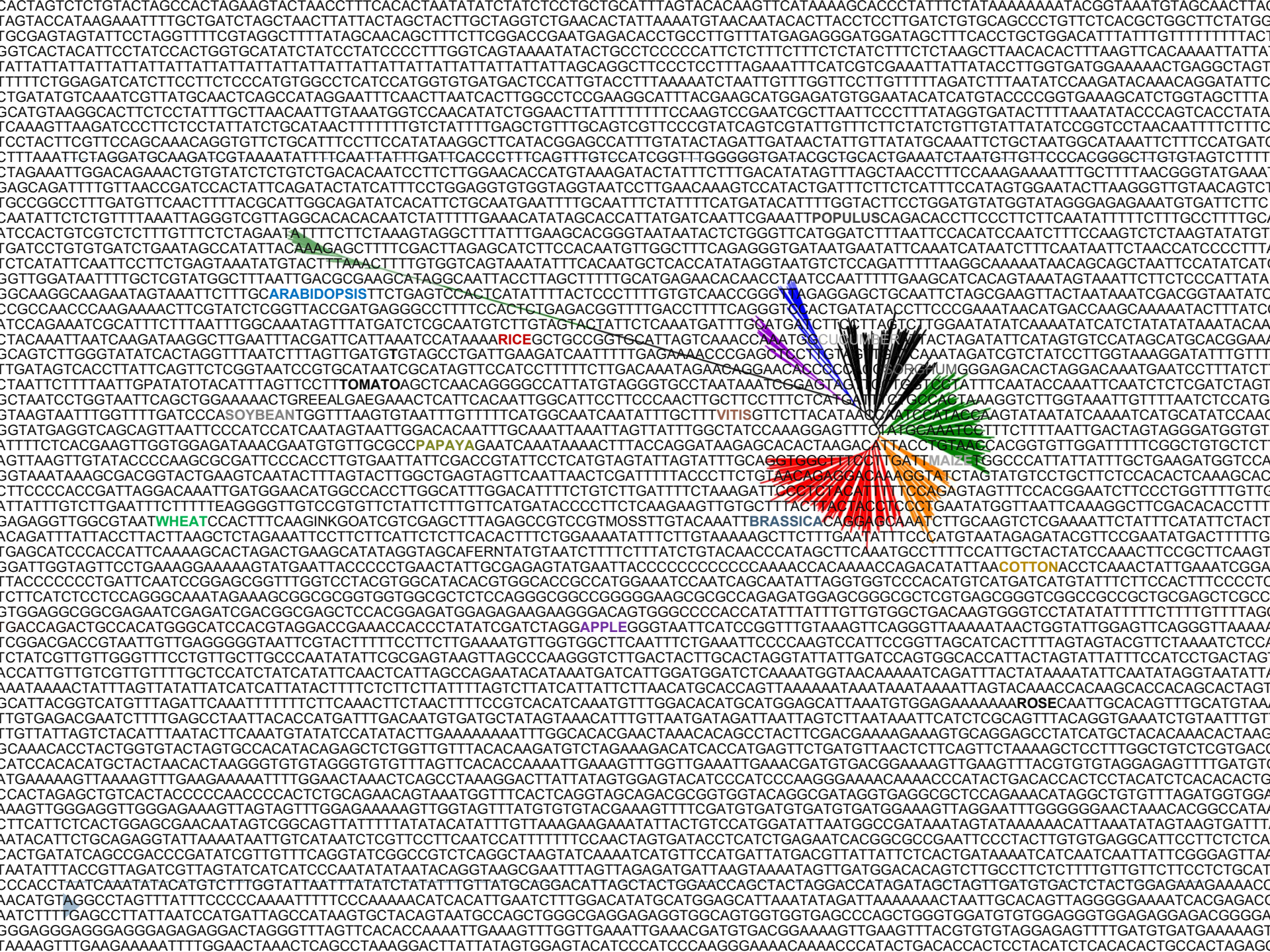
- 《植物基因组学》，樊龙江主编，科学出版社，2020
- 《生物信息学》，樊龙江主编，浙江大学出版社，2017
  （第二版，2021年，科学出版社）
- 其他：
  - 《GENOMES 4》, 2017, Brown T. A.
  - 《基因组学》，2016，杨焕明


“十三五”全国高等院校研究生规划教材
创新型现代农林院校研究生系列教材
PLANT GENOMICS
植物基因组学
樊龙江 主编
科学出版社
植物基因组学
科学出版社

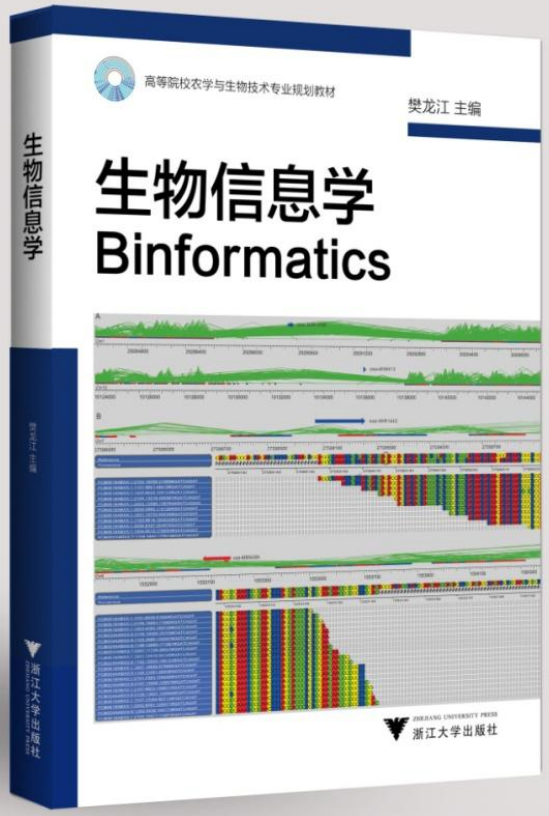
《植物基因组学》简要目录

详细目录见PDF

章节	内容
序	
前言	
	第一篇 总论
第1-1章	绪论
第1-2章	植物基因组测序与拼装
第1-3章	植物基因组构成
第1-4章	植物基因组转录
第1-5章	植物基因组表观遗传修饰
第1-6章	植物基因组进化
第1-7章	植物群体基因组
第1-8章	植物单细胞基因组
第1-9章	植物三维基因组
第1-10章	植物合成基因组
第1-11章	叶绿体基因组
第1-12章	植物线粒体基因组
第1-13章	植物功能基因组学技术
第1-14章	组学数据育种利用
第1-15章	植物基因组数据资源
	第二篇 各论
第2-1章	拟南芥基因组
第2-2章	水稻基因组
第2-3章	玉米基因组
第2-4章	小麦基因组
第2-5章	大豆基因组
第2-6章	棉花基因组
第2-7章	油菜基因组
第2-8章	蔬菜基因组
第2-9章	林果花草基因组
第2-10章	非维管束植物基因组
附录1	植物基因组测序物种清单
附录2	主要中英文术语释义及其索引


POPULUS
ARABIDOPSIS
RICE
TOMATO
SOYBEAN
VITIS
PAPAYA
MAIZE
WHEAT
BRASSICA
COTTON
APPLE
ROSE

《生物信息学》教材

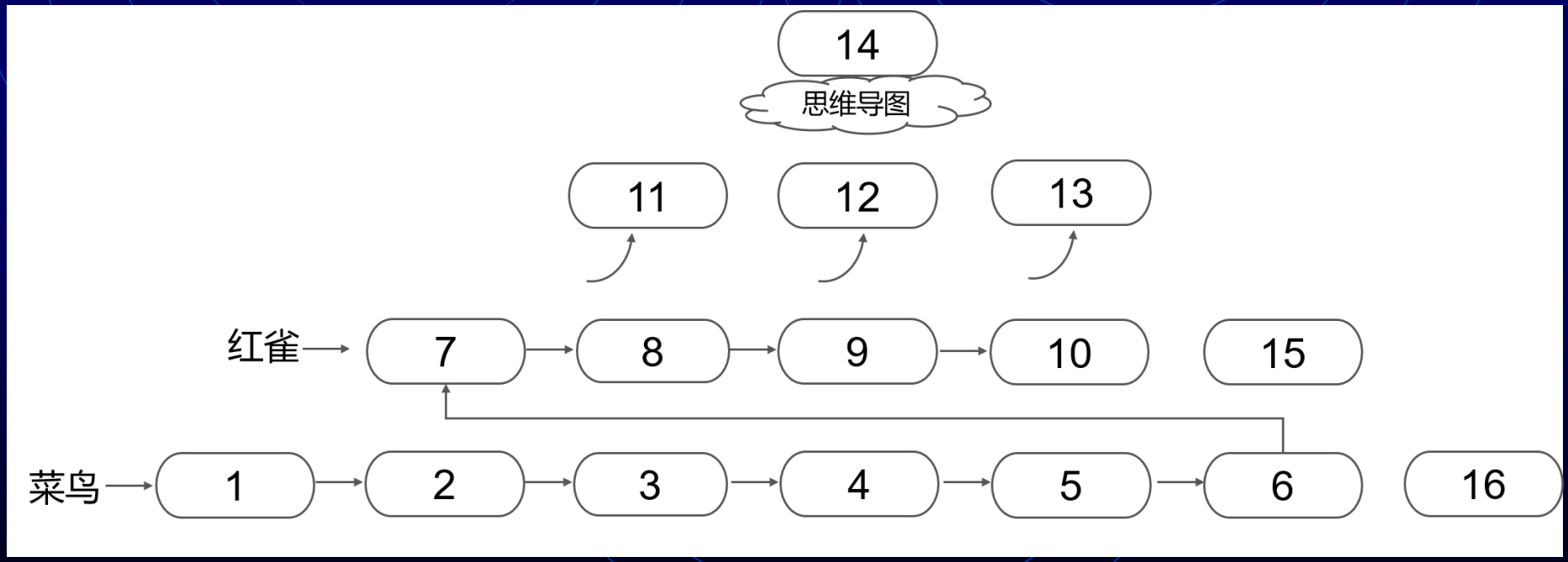


- 全书共分为四部分：
  - 生物信息学基础（第一篇）
  - 高通量测序数据分析（第二篇）
  - 生物信息学外延与交叉（第三篇）
  - 生物信息学资源与实践（第四篇）

《生物信息学》（第二版）

2021年，科学出版社

- 各章阅读指南



详细目录见WORD

致敬经典

“历史与人物”篇中讲述的主要人物

Paulien Hogeweg	Jonathan Rothberg	Margaret Dayhoff	Walter Goad	Saul Needleman
Michael Waterman	David Lipman	Russell Doolittle	Masatoshi Nei	Craig Venter
Pavel Pevzner	Andrey Markov	Leonard Baum	Geoffrey Hinton	Yoshua Bengio
Carl Woese	Leroy Hood	Victor Ambros	Thomas Bayes	Yann LeCun
Gustave Malecot	Columbus Cockerham	Guido van Rossum	Stephen Altschul	Linus Pauling

1. Introduction

- 1.1 Genomic data
  - Genome and genome sequencing
  - Targets of high-throughput sequencing
- 1.2 Analysis of genomic data
  - Roadmap of genome study
  - Analysis of plant genomic data

1.1 Genomic data

Genome and genome sequencing

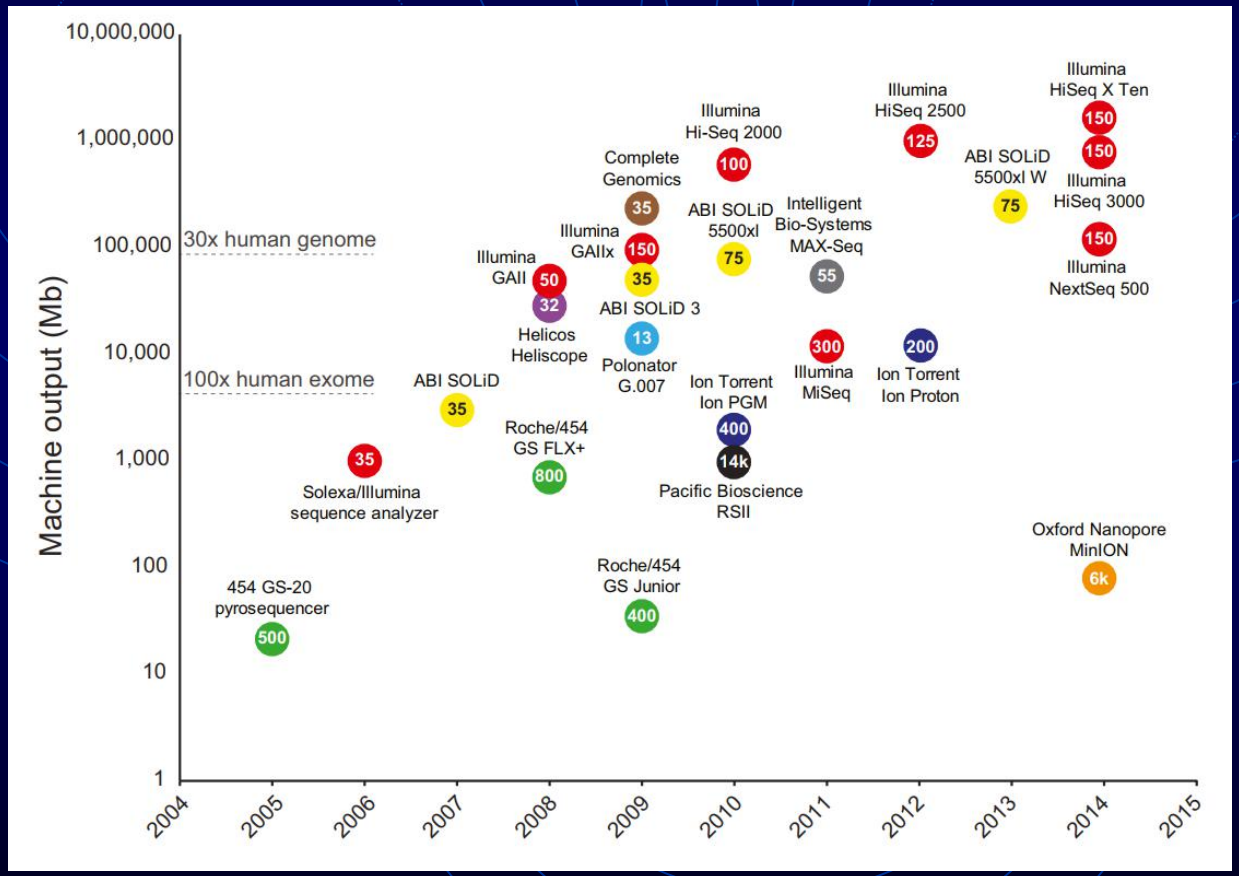
- 一个生物体的基因组是指一套染色体中的完整的DNA序列。例如，生物个体体细胞中的二倍体由两套染色体组成，其中一套DNA序列就是一个基因组。基因组一词可以特指整套核DNA（例如，核基因组），也可以用于包含自己DNA序列的细胞器基因组，如粒线体基因组或叶绿体基因组。当人们说一个有性生殖物种的基因组正在测序时，通常是指测定一套常染色体和两种性染色体的序列，这样来代表可能的两种性别。即使在只有一种性别的物种中，“一套基因组序列”可能也综合了来自不同个体的染色体。
- 对相关物种全部基因组性质的研究通常被称为基因组学。1920年，德国汉堡大学植物学教授汉斯.温克勒（Hans Winkler）首次使用基因组这一名词。该学科与遗传学不同，后者一般研究单个或一组基因的性质。
- 基因组包括基因和非编码DNA

Type	Organism	Size	Note
Virus	Bacteriophage MS2	3.5kb	First sequenced RNA-genome[8]
Virus	SV40	5.2kb	[9]
Virus	Phage Φ-X174	5.4kb	First sequenced DNA-genome[10]
Virus	HIV	9.7kb	
Virus	Phage λ	48kb	
Virus	Mimivirus	1.2Mb	Largest known viral genome
Bacterium	*Haemophilus influenzae*	1.8Mb	First genome of a living organism sequenced, 1995[12]
Bacterium	*Carsonella ruddii*	160kb	Smallest non-viral genome.[13]
Bacterium	*Buchnera aphidicola*	600kb	
Bacterium	*Wigglesworthia glossinidia*	700Kb	
Bacterium	*Escherichia coli*	4.6Mb	[14]
Bacterium	*Solibacter usitatus* (strain Ellin 6076)	10Mb	Largest known Bacterial genome
Amoeboid	*Polychaos dubium* (*"Amoeba" dubia*)	670Gb	Largest known genome.[16] (Disputed [15])
Plant	*Arabidopsis thaliana*	157Mb	First plant genome sequenced, December 2000.[17]
Plant	*Genlisea margaretae*	63Mb	Smallest recorded flowering plant genome, 2006.[17]
Plant	*Fritillaria assyrica*	130Gb	
Plant	*Populus trichocarpa*	480Mb	First tree genome sequenced, September 2006
Plant	*Pieris japonica* (Japanese-native)	150Gb	Largest plant genome known
Moss	*Physcomitrella patens*	480Mb	First genome of a bryophyte sequenced, 2008.[18]
Yeast	*Saccharomyces cerevisiae*	12.1Mb	First eukaryotic genome sequenced, 1996[19]
Fungus	*Aspergillus nidulans*	30Mb	
Nematode	*Caenorhabditis elegans*	100Mb	First multicellular animal genome sequenced, 1998[20]
Nematode	*Pratylenchus coffeae*	20Mb	Smallest animal genome known[21]
Insect	*Drosophila melanogaster* (fruit fly)	130Mb	[22]
Insect	*Bombyx mori* (silk moth)	530Mb	
Insect	*Apis mellifera* (honey bee)	236Mb	
Insect	*Solenopsis invicta* (fire ant)	480Mb	[23]
Fish	*Tetraodon nigroviridis* (puffer fish)	390Mb	Smallest vertebrate genome known
Mammal	*Homo sapiens*	3.2Gb	

Genome sequencing

- Wikipedia: Full genome sequencing (FGS)
  - also known as whole genome sequencing (WGS), complete genome sequencing, or entire genome sequencing
- Sequencing technology
  - Sanger method
    - ABI3730: 700-900bp per read
  - high-throughput approaches
    - Illumina Geome Analyzer II System/ HiSeq 2000
    - 454/Roche GS FLX
    - PacBio
    - Nanopore
    - ...

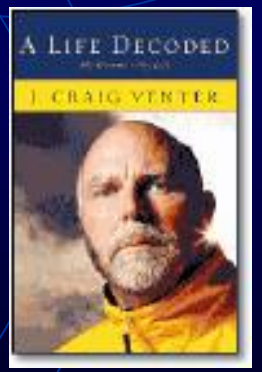
高通量测序技术出现年代及其测序通量变化趋势（Reuter 等，2015）



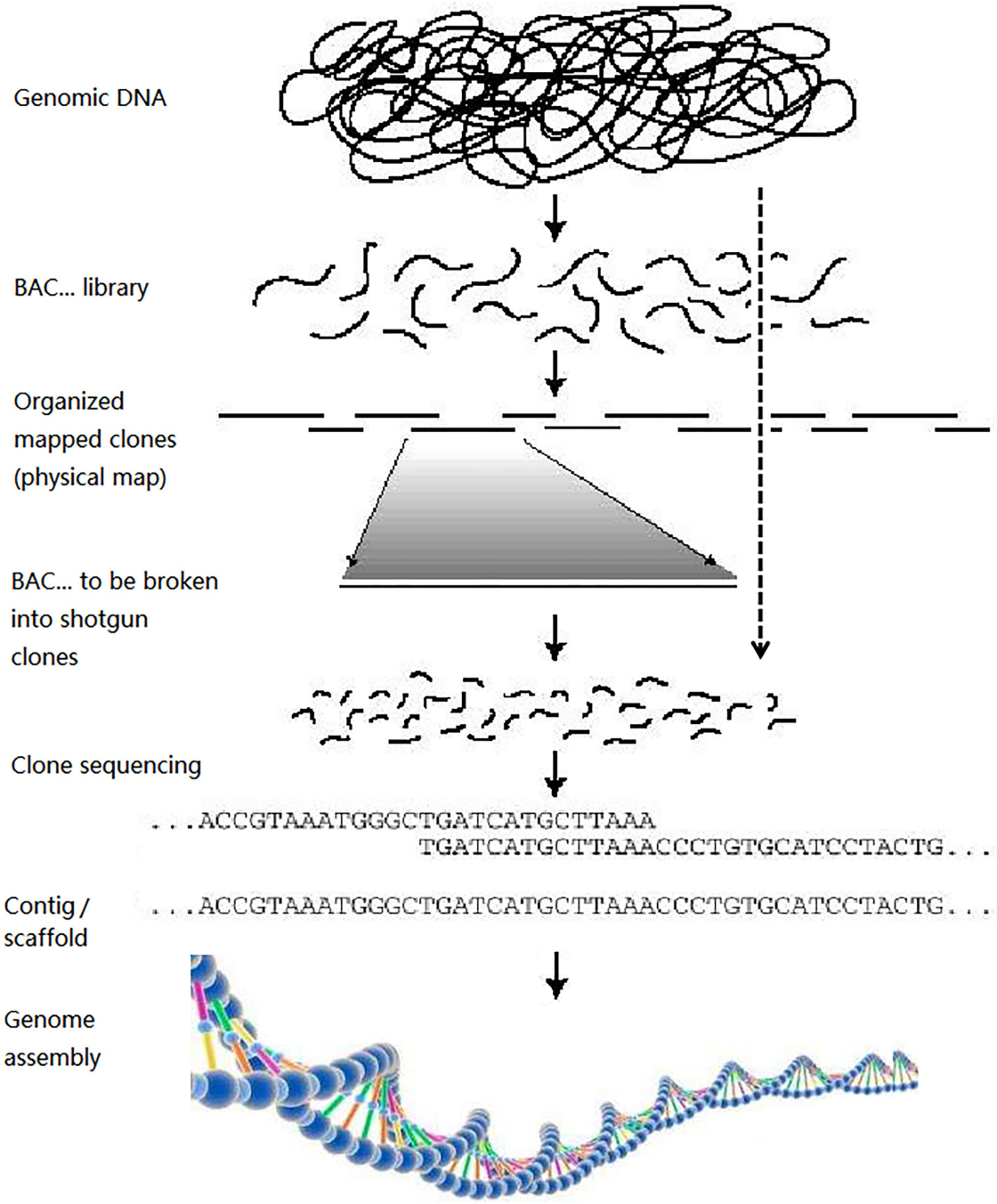
基因组测序的策略
How to sequence a genome?

两个基因组测序策略：

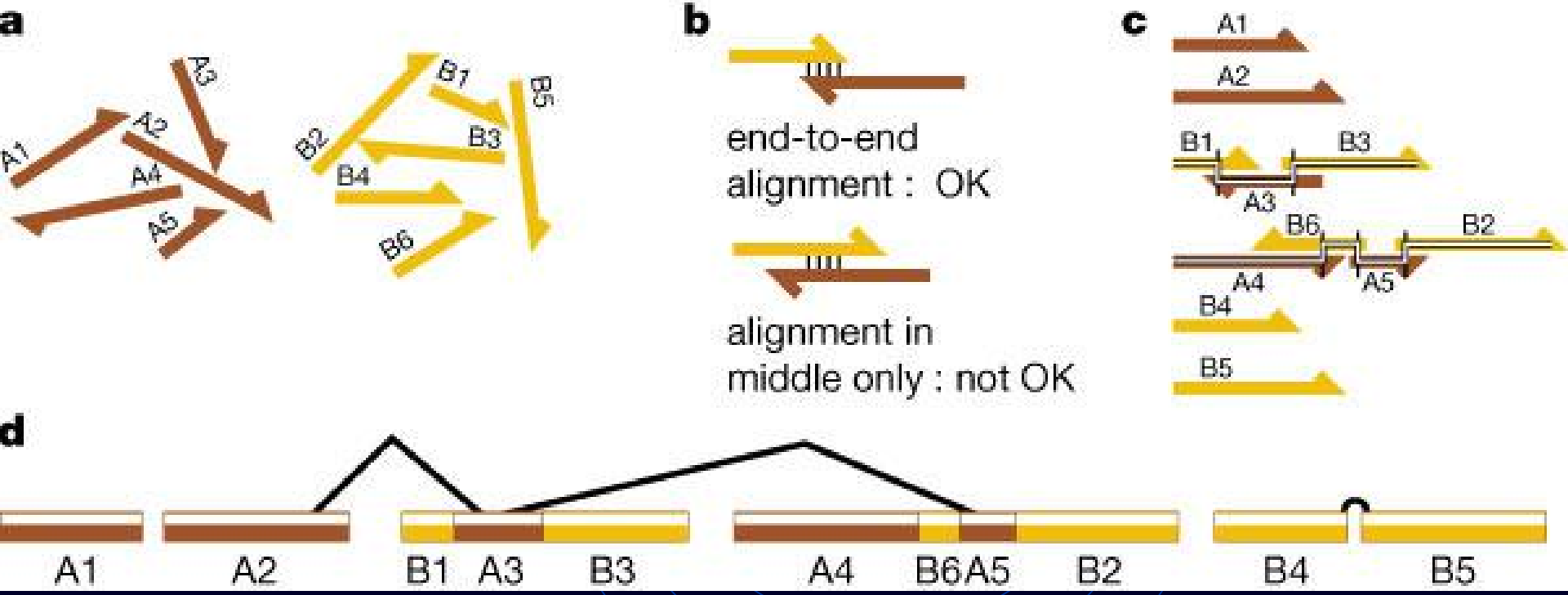
逐步克隆方法
（clone by clone)
全基因组鸟枪方法
（whole genome shotgun，WGS)



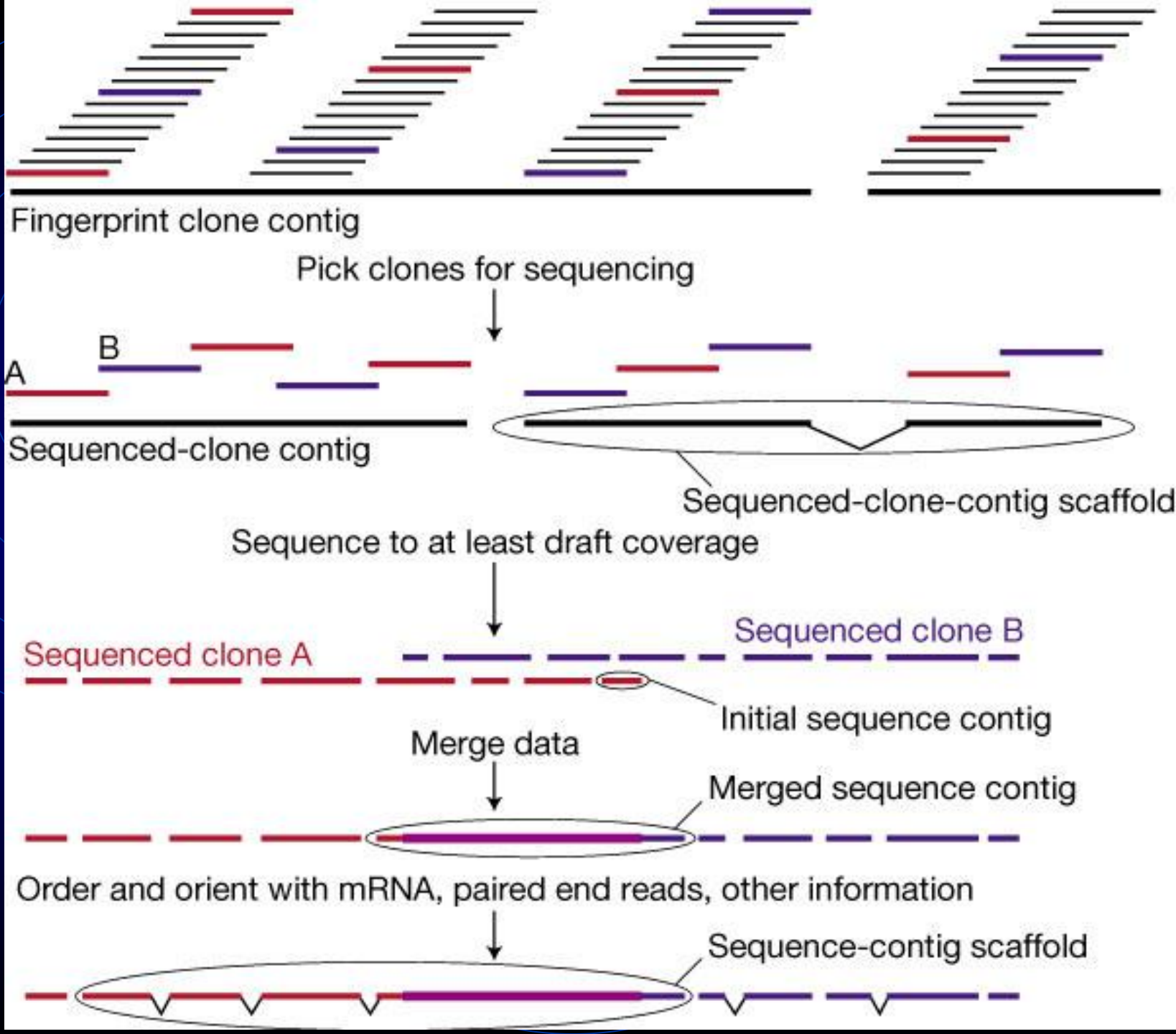
Clone by Clone Approach



The key steps in assembling individual sequenced clones into the draft genome sequence



Contig and Scaffold



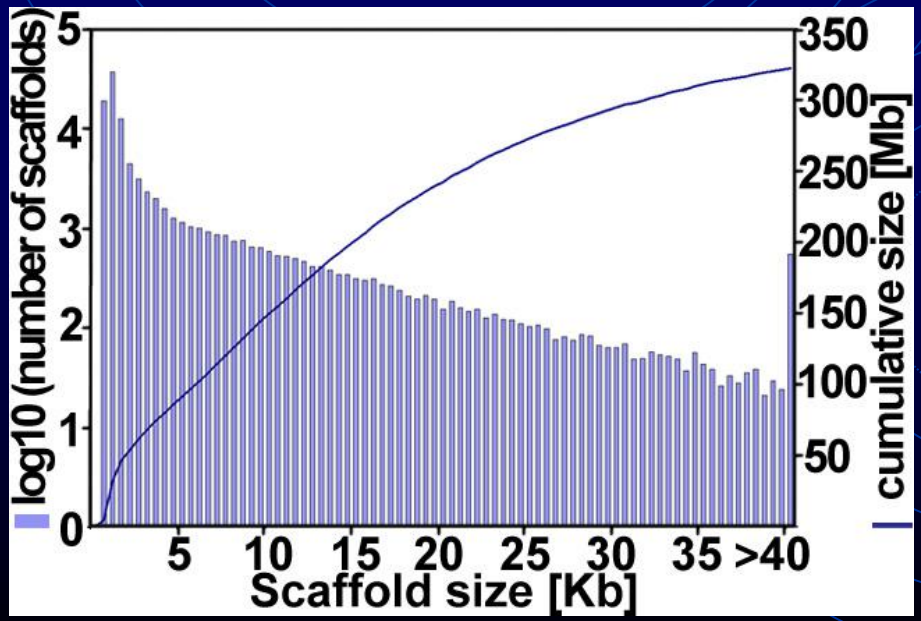
Scaftig

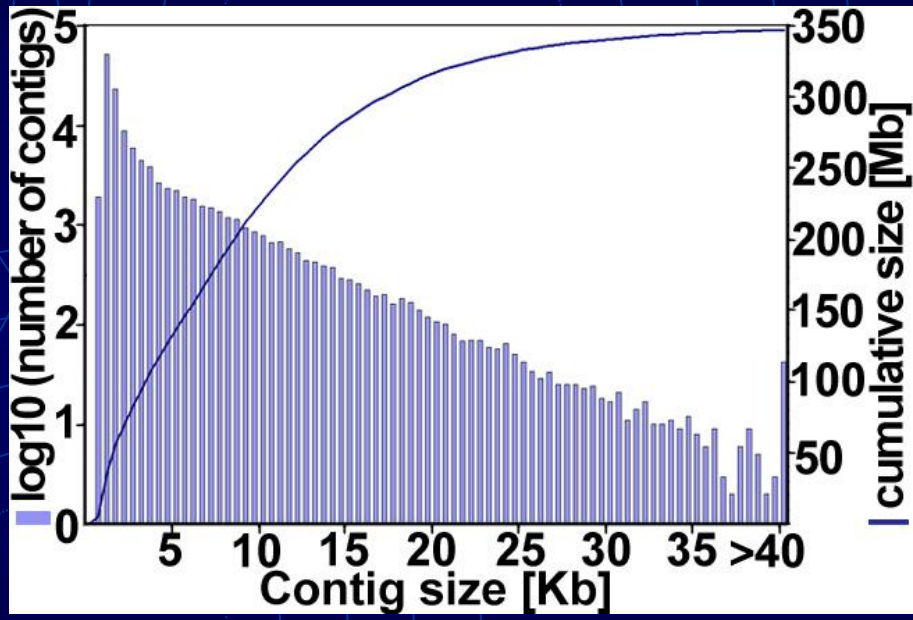
- Contigs by de Bruijn Graph
- Scaftig: a contig contains the de Bruijn contigs which were linked by read alignments and there is no gap in it
- Scaffold: a sequence contains scaftigs which were linked by PE reads and therefore gaps

Quality

Genome coverage
Functional coverage
Assembly

Contigs:127,550
(N50=6,688 bp)





Scaffolds: 102,444
(N50=11,764 bp)

Quality estimation

- Assembly quality: BAC/PAC/FOSMID clone sequencing
- Genomic coverage: flow cytometer; genome size estimation by k-mers.
- Functional coverage: traditional ESTs; reference genomes

Targets of high-throughput sequencing

- 基因组及基因组重测序
- 转录组（RNA-SEQ）
- 表达谱测序（DGE）
- 小RNA群体测序
- 蛋白质－DNA互作测序（ChIP-SEQ）
- 甲基化位点测序（MeDIP-SEQ）
- 降解组测序 **(Degradome-Seq)**
- 目标区域捕获：外显子组 (Exome-SEQ)
- 简化基因组测序：RAD-SEQ
- 宏基因组（环境样品）测序

Genome sequencing

- Genome sequencing: working draft, fine map and finished or essentially complete.
- Genome assembly is a very difficult computational problem, made more difficult because many genomes contain large numbers of identical sequences, known as *repeats*. These repeats can be thousands of nucleotides long, and some occur in thousands of different locations, especially in the large genomes of plants and animals.
- The resulting (draft) genome sequence is produced by combining the information sequenced contigs and then employing linking information to create scaffolds (PE libraries with different insertion size). Scaffolds are positioned along the physical map/genetic map of the chromosomes creating a "golden path".

Genome sequencing steps

- Genome survey: 30-50X. Estimation of genome size; hetrogenous rate (assembly results and k-mers distribution); others
- Fine map: PE libraries with different insertion sizes
- Chromosome sequences: genetic/physical maps

Genome re-sequencing

- Genome re-sequencing:
  - Deep sequencing: 30-50X
    - SNP calling/ de novo assembly
  - Germplasm survey: 10-15X
    - SNP calling

RNA-SEQ

- Whole Transcriptome Shotgun Sequencing: RNA poly(A) library; cDNA
- It can be done with a variety of platforms to test a high amount of ideas and hypotheses. For example, using the Illumina Genome Analyzer platform, recent applications include sequencing mammalian transcriptomes, ABI Solid Sequencing to profile stem cell transcriptomes or Life Science's 454 Sequencing to discover SNPs in maize. Even though each platform has its technical differences, the information gathered from each is of the same nature.

Small RNAs

- Total RNA: 18-35nt
- Non-PE read

ChIP-Seq

- **ChIP-Sequencing**, also known as **ChIP-Seq**, is used to analyze protein interactions with DNA. ChIP-Seq combines chromatin immunoprecipitation (ChIP, 染色质免疫共沉淀) with massively parallel DNA sequencing to identify the cistrome of DNA-associated proteins. It can be used to precisely map global binding sites for any protein of interest

MeDIP-seq

- **Methylated DNA immunoprecipitation** (MeDIP or mDIP) is a large-scale (chromosome- or genome-wide) technique that is used to enrich for methylated DNA sequences. It consists of isolating methylated DNA fragments via an antibody raised against 5-methylcytosine (5mC). The MeDIP-seq approach, i.e. the coupling of MeDIP with next generation, short-read sequencing technologies

Degradome-Seq

- **Degradome sequencing (Degradome-Seq)**, also referred to as **parallel analysis of RNA ends (PARE)**, is a modified 5'-rapid amplification of cDNA ends (RACE) with high-throughput deep sequencing (SBS) method. Degradome sequencing provides a comprehensive means of analyzing patterns of RNA degradation
- Degradome sequencing has been use to identify microRNA (miRNA) cleavage sites and revealed many known and novel plant miRNA (siRNA) targets.

Targeted sequence capture

- 目标序列捕获测序是将感兴趣的基因组区域定制成特异性探针与基因组DNA在序列捕获芯片（或溶液）进行杂交，将目标基因组区域的DNA片段进行富集后再利用第二代测序技术进行测序的研究策略.
- NimbleGen Sequence Capture Array; Agilent SureSelect Target Enrichment System; Agilent SureSelect DNA Capture Array

Exome-SEQ

- **Exome sequencing** (also known as **targeted exome capture**) is an efficient strategy to selectively sequence the coding regions of the genome to identify novel genes associated with rare and common disorders
- In total there are about 180,000 exons found in the human genome. These protein coding regions constitute about 1% of the human genome which translates to about 30 megabases (Mb) in length. It is estimated that the protein coding regions of the human genome constitute about 85% of the disease-causing mutations.

RAD-SEQ

- **Restriction site Associated DNA (RAD) markers** are a type of genetic marker that can be used for genetic mapping.
- An important aspect of RAD markers and mapping is the process of isolating RAD tags, which are the DNA sequences that immediately flank each instance of a particular restriction enzyme site throughout the genome.
- The use of high-throughput sequencing to analyze RAD tags has been termed RAD tag sequencing, RAD-Seq.

Environmental samples

- **Metagenomics** is the study of **metagenomes**, genetic material recovered directly from environmental samples
- Traditional microbiology and microbial genome sequencing rely upon cultivated clonal cultures environmental samples.
- Early environmental gene sequencing cloned specific genes (often the 16S rRNA gene) to produce a profile of diversity in a natural sample. Such work revealed that the vast majority of microbial biodiversity had been missed by cultivation-based methods.
- Recent studies use "shotgun" Sanger sequencing or massively parallel pyrosequencing to get largely unbiased samples of all genes from all the members of the sampled communities.

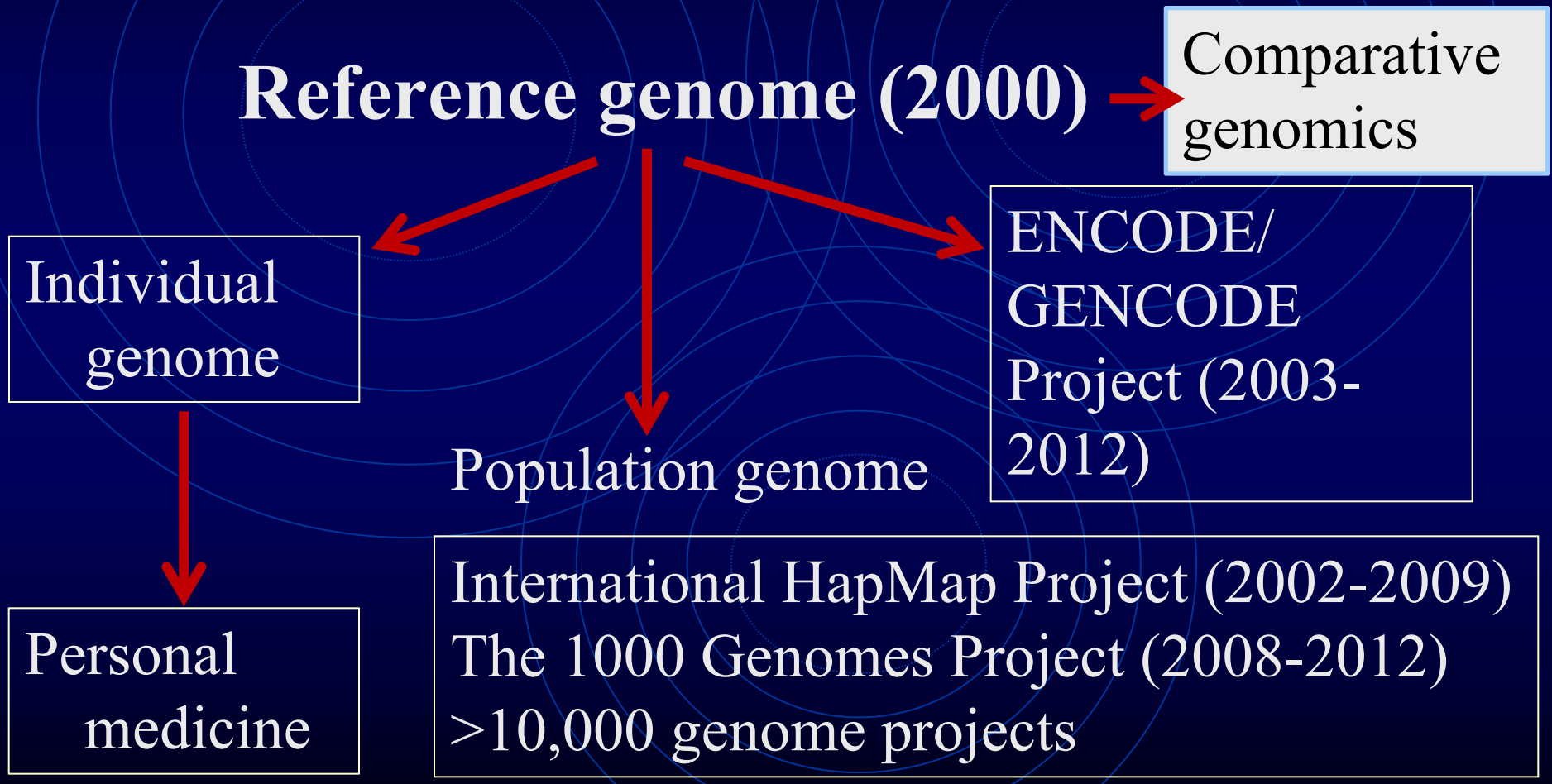
Digital gene expression profiling

- **Gene expression profiling** is the measurement of the activity (the expression) of thousands of genes at once, to create a global picture of cellular function
- DNA Microarray technology measures the relative activity of previously identified target genes. Sequence based techniques, like serial analysis of gene expression (SAGE, SuperSAGE) are also used for gene expression profiling. SuperSAGE is especially accurate and can measure any active gene, not just a predefined set. The advent of next-generation sequencing has made sequence based expression analysis an increasingly popular, "digital" alternative to microarrays

1.2 Analysis of genomic data

- *De novo* assembly of reads: big indels (>10bp); structural polymorphism; transcripts
- Sequence alignment/mapping to reference genome: genetic diversity: SNP calling, small indel etc; identification of unique genomic sites
- Sequence analysis: prediction; evolution; selection; pathway; genome size, etc.

Roadmap for human genome study



Roadmap for crop genome study

Reference genome (2002-) → Comparative genomics

pedigree genome
(maize, Lai et al. 2010)

Population genome

Full-length cDNA project
?

Rice, maize, soybean, millet
~50 rice lines (Xu et al. 2011)
~500 rice lines (Huang et al. 2009)
~1000 rice lines (Huang et al. 2012)
First haplotype map of maize (Gore et al. 2009)
~31 soybean lines (Lam et al. 2010)
~1000 millet lines (Peng et al. 2013)

Inbred line genome

Challenges in analysis of genomic data

- Assembly for complex genomes
- Assembly algorithm: accurate, fast
- Diverse RNA sequence data
- Mixed samples
- Mapping: accurate
- how to use the genomic data? *K*-mers

植物基因组的特征

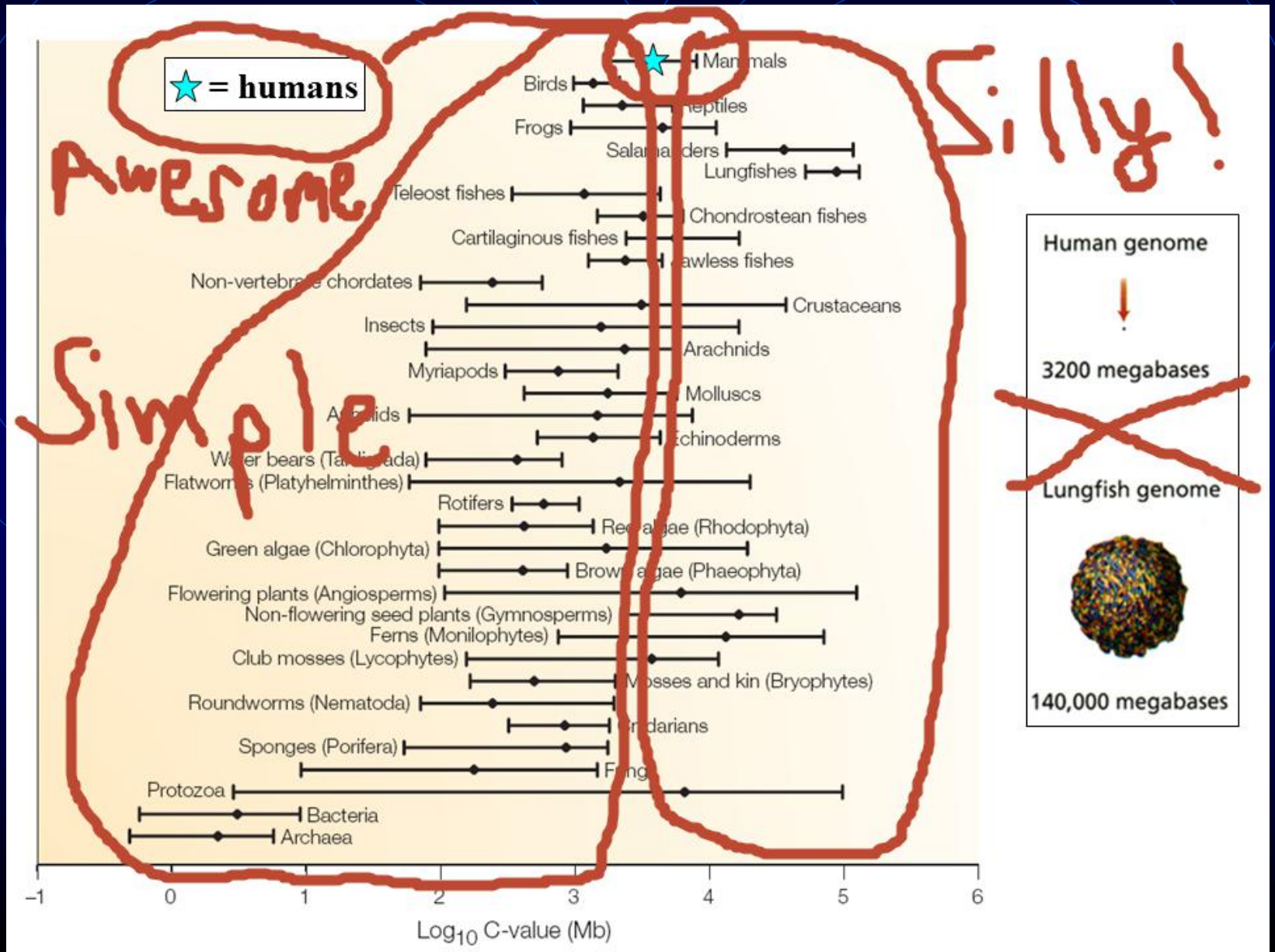
植物分类

“Genomes gone wild”
The Scientist, Jan. 2014

“Weird and wonderful, plant DNA is challenging preconceptions about the evolution of life, including our own species.”

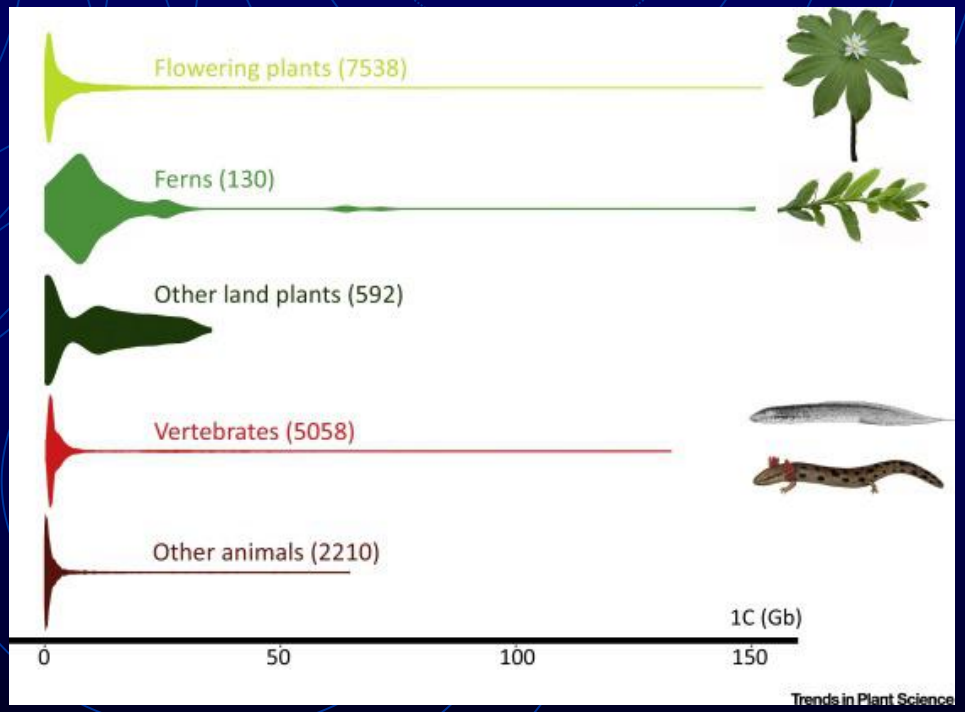
- Size matters
  - the biggest and smallest genome size in plants
- Double the fun
  - an increase in genome size is typically a consequence of one of two mechanisms: the duplication of the entire genome, or the multiplication of TEs within a genome. The former is common in plants and sults in polyploidyy. But the latter is a far more common cause of size increase in animal genomes
- Evolution in the fast lane
  - Plants don't rely only on horizontal gene transfer for new alleles; they also gain new genes through the traditional route of mutation. Animals don't normally swap genes.
  - the fastest and lowest mutation rates of genomes (chloroplast and mitochondrial) in plants

Genome size of plants



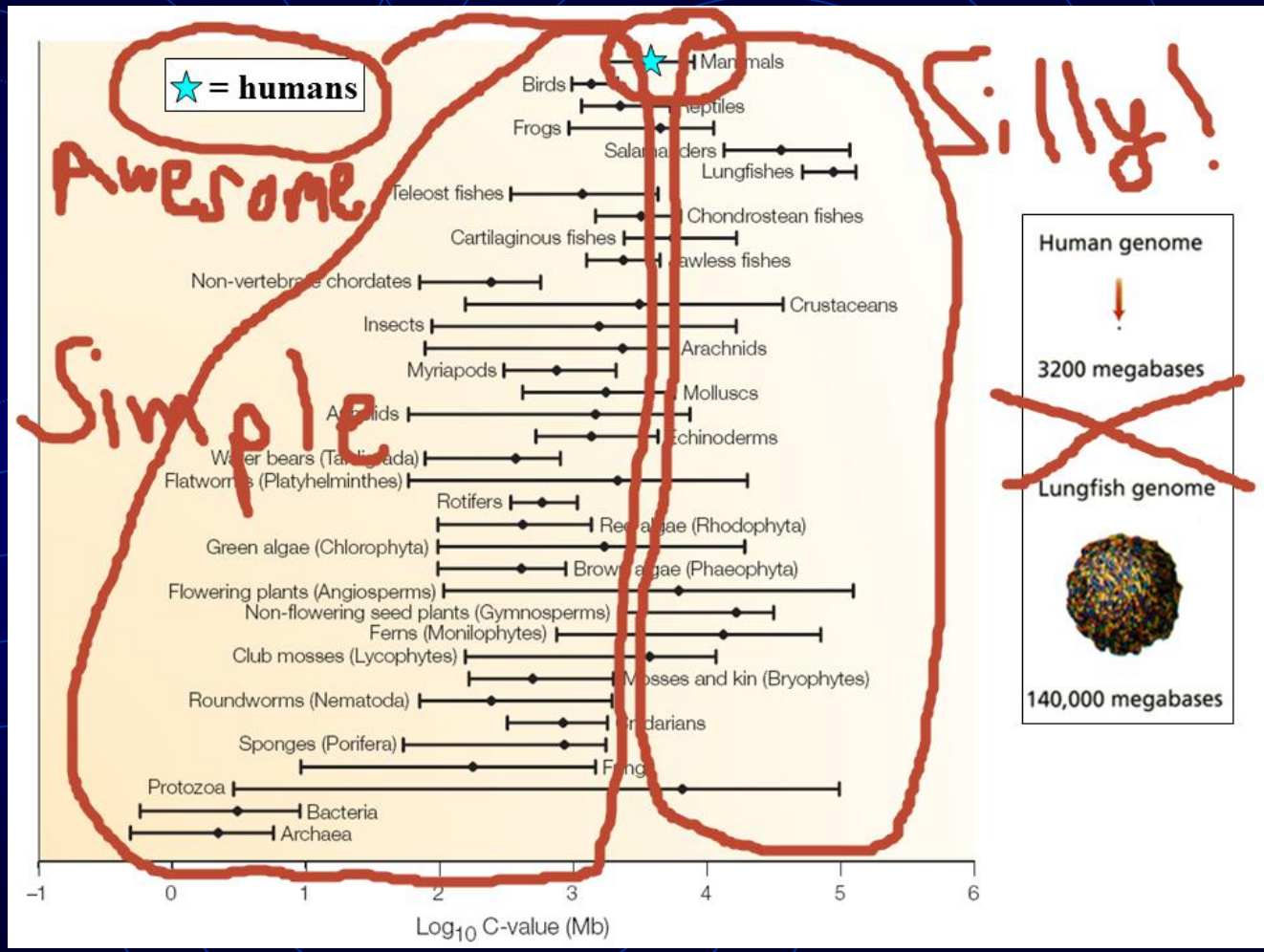
Genomes in numbers

- Sizes:
  - virus: 10^3 to 10^5 nt
  - bacteria: 10^5 to 10^7 nt
  - yeast: 1.35 x 10^7 nt
  - mammals: 10^8 to 10^{10} nt
  - plants: 10^8 to 10^{11} nt

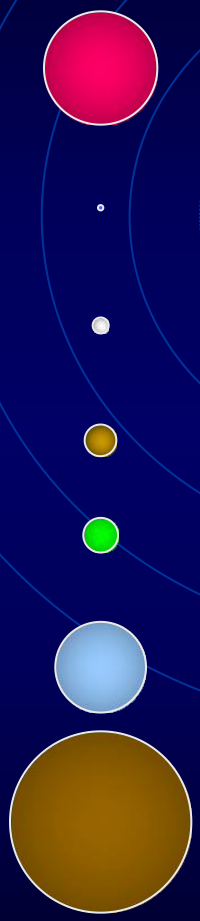


Size distribution of crop genomes

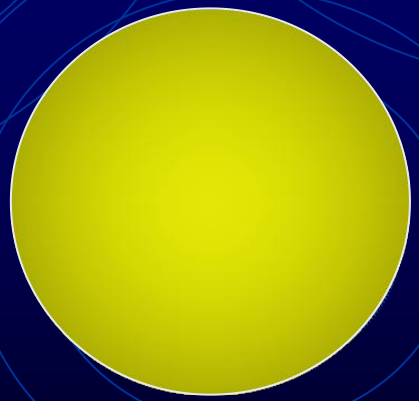
- Rice: 400Mb
- Sorghum: 670Mb
- Soybean: 1Gb
- Millet: 1 Gb
- Oilrape *(B. napus)*: 1.2Gb
- Maize: 2.5Gb
- Tobacco: 4 Gb
- Barley: 5Gb
- Wheatt: 15 Gb
- Peanut
- Sesame
- Sunflower
- Potato
- Sweet potato
- Casava
- Cotton



Plant Genomes – Haploid Size

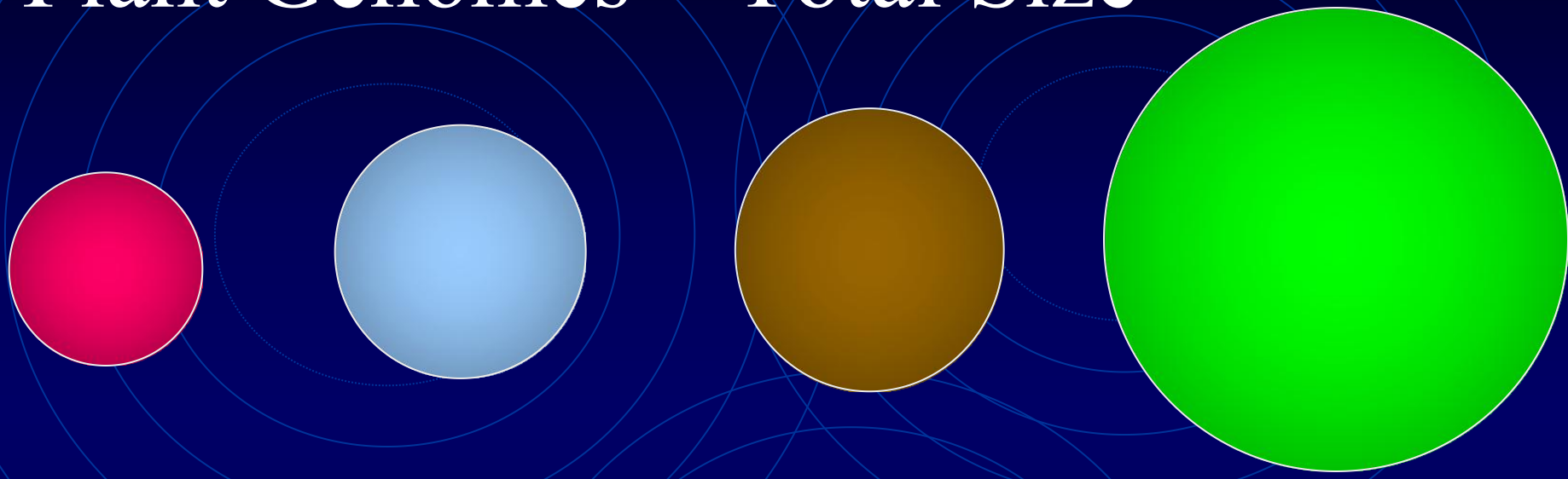


Species	Human	Arabidopsis	Sugarcane	Cotton
Size (Mbp)	3,300	155	930	2,500
Ploidy	x 2	x 2	x 8 -12	x 4



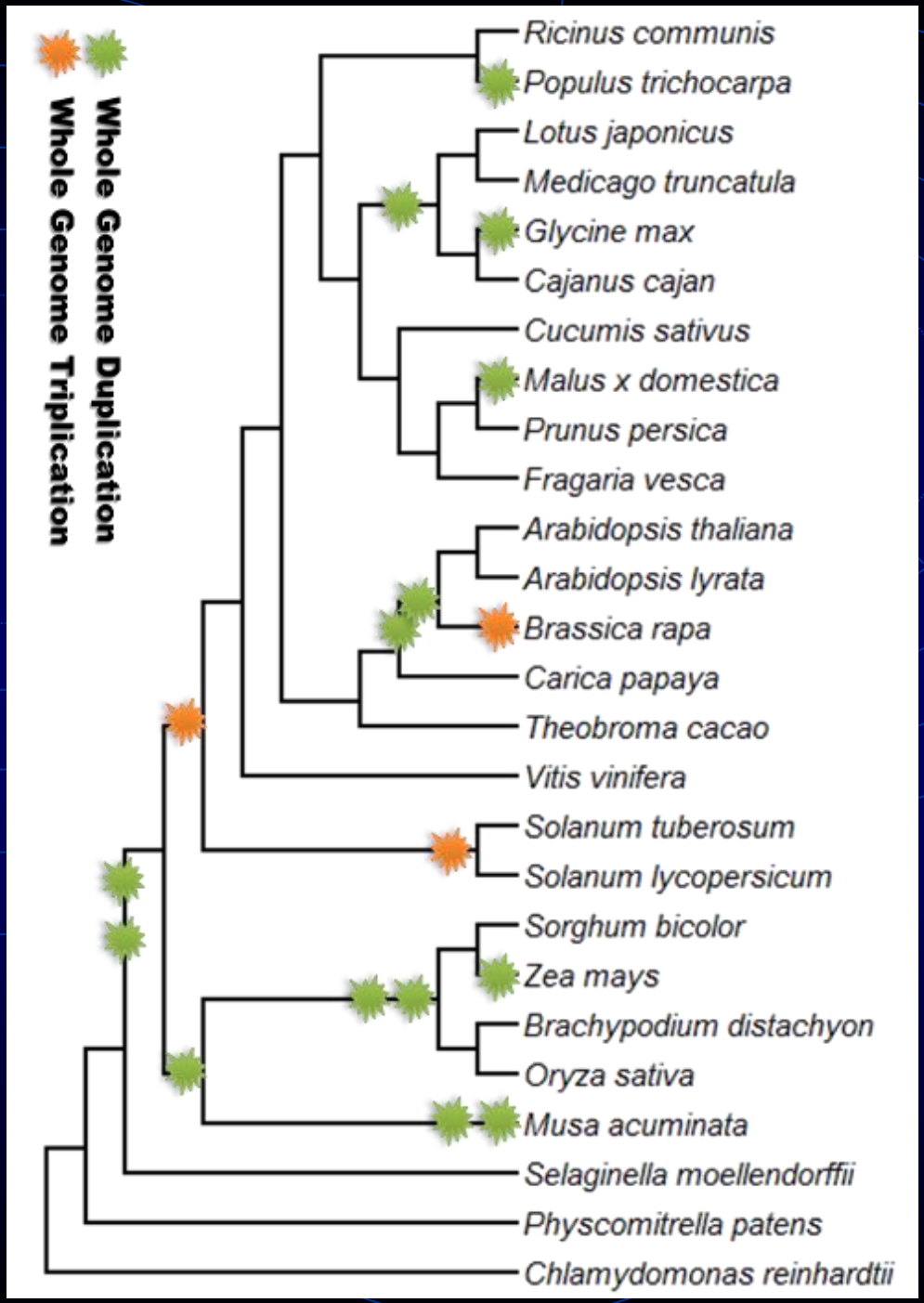
Wheat
17,000
x 6

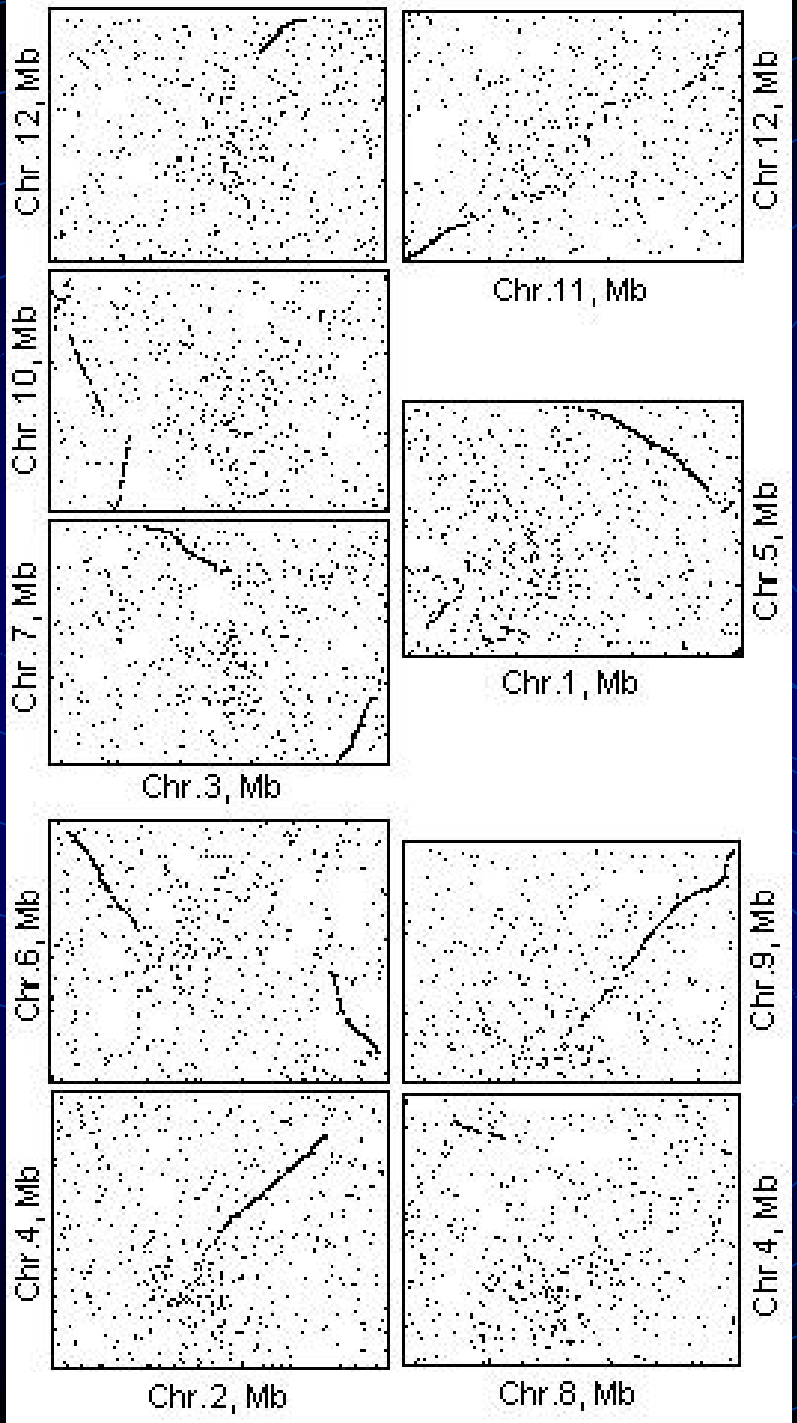
Plant Genomes – Total Size



Human Tobacco Barley wheat

Pieris japonica


Whole Genome Duplication
Whole Genome Triplication
Ricinus communis
Populus trichocarpa
Lotus japonicus
Medicago truncatula
Glycine max
Cajanus cajan
Cucumis sativus
Malus x domestica
Prunus persica
Fragaria vesca
Arabidopsis thaliana
Arabidopsis lyrata
Brassica rapa
Carica papaya
Theobroma cacao
Vitis vinifera
Solanum tuberosum
Solanum lycopersicum
Sorghum bicolor
Zea mays
Brachypodium distachyon
Oryza sativa
Musa acuminata
Selaginella moellendorffii
Physcomitrella patens
Chlamydomonas reinhardtii



Duplicated blocks based on annotation of TIGR (osa1, version1)

Reciprocal BLASTN searching (E<e-14), one-to-one paralogous gene pairs

(Zhang et al., 2005)

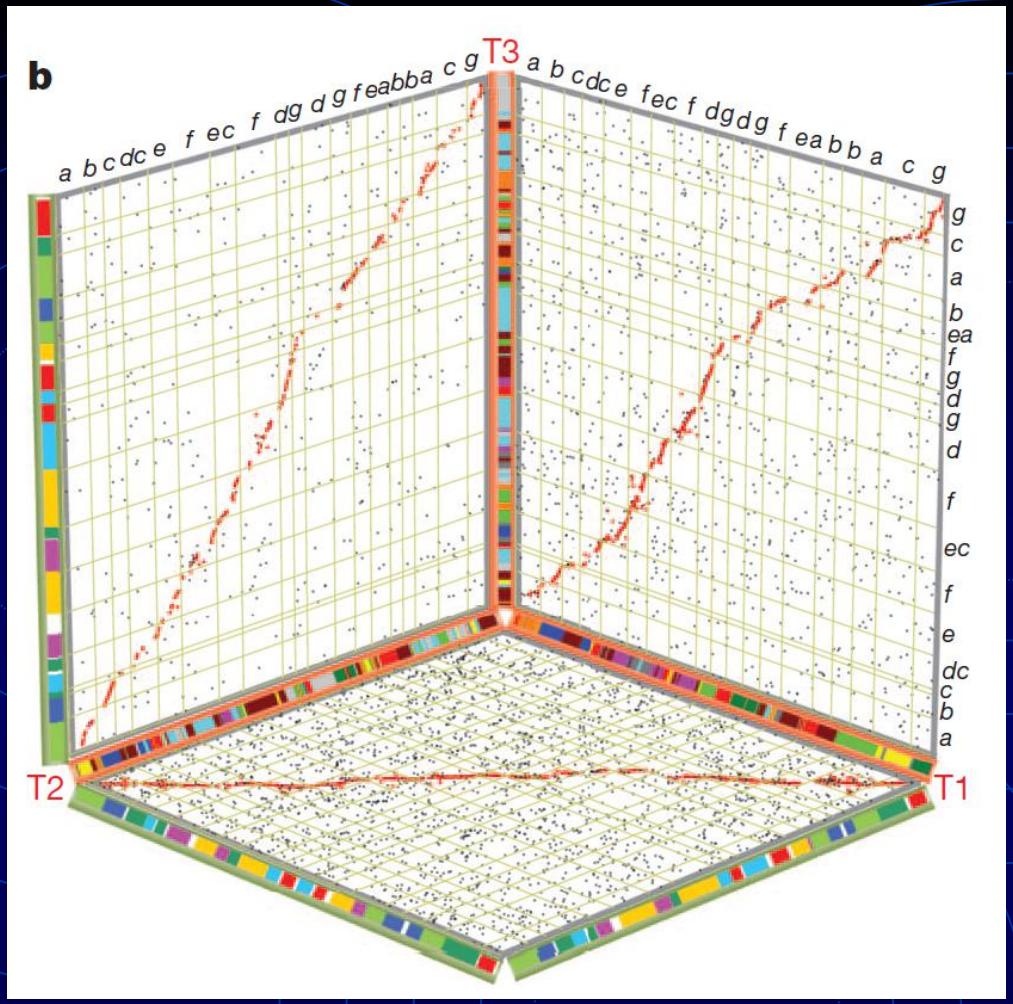
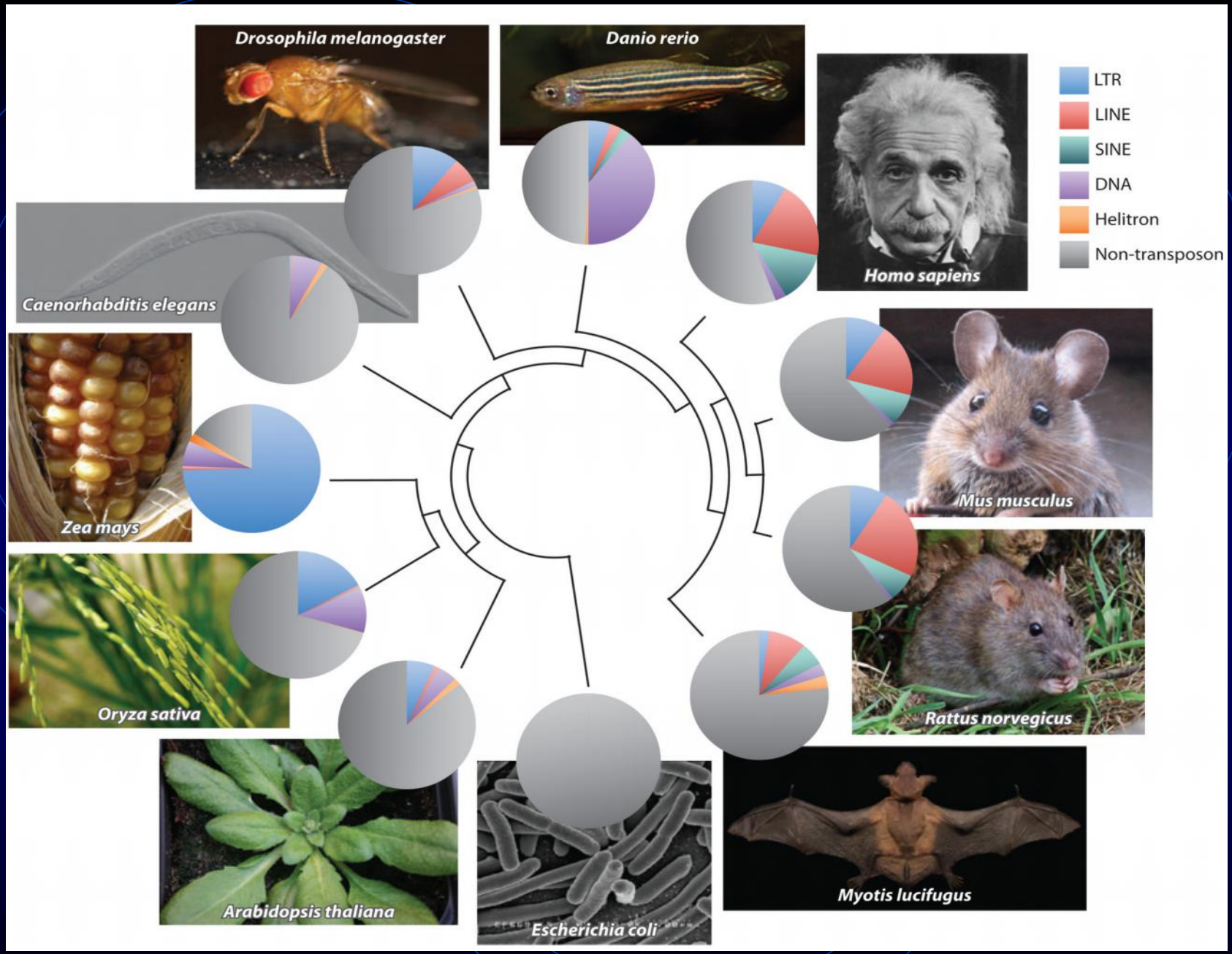


Figure 2 | The *Solanum* whole genome triplication. **a**, Speciation and polyploidization in eudicot lineages. Confirmed whole-genome duplications and triplications are shown with annotated circles, including '*T*' (this paper) and previously discovered events α, β, γ[10,11,14]. Dashed circles represent one or more suspected polyploidies reported in previous publications that need further support from genome assemblies[27,28]. Grey branches indicate unpublished genomes. Black and red error bars bracket indicate the likely timings of divergence of major asterid lineages and of '*T*', respectively. The post-'*T*' subgenomes, designated T1, T2, and T3, are further detailed in Supplementary Fig. 10. **b**, On the basis of alignments of multiple tomato genome segments to single grape genome segments, the tomato genome is partitioned into three non-overlapping 'subgenomes' (T1, T2, T3), each represented by one axis in the three-dimensional plot. The ancestral gene order of each subgenome is inferred according to orthologous grape regions, with tomato chromosomal affinities shown by red (inner) bars. Segments tracing to pan-eudicot triplication (γ) are shown by green (outer) bars with colours representing the seven putative pre-γ eudicot ancestral chromosomes[10], also coded *a*–*g*.

- **No polyploidy crop genome are available until 2014**
- Genome duplication events happened in many crops: the grass family
- Genome duplication, Gene duplication – large gene families

Gene size

- Plant
  - Average size: 2-4Kb
  - Genome duplication, Gene duplication – large gene families
- Human:
  - Average size: ?
  - The Duchenne muscular dystrophy gene, >2 Mb
  - Some genes with several hundreds of exons?

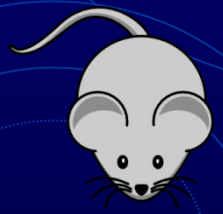


Transposon compositions in different species (Huang et al. Annu Rev Genet. 2012)

Large evolutionary distances

75 MYA

155 MYA



Chloroplast genome

- 基因组很小（~150 Kb）
- 母系遗传
- 遗传稳定（进化速率慢）
- 特殊用途

Summary

Plant Genomes – "Unique Challenges"

- Genome size
  - Large range of genome sizes
- Ploidy
  - Genome duplication
- Pan genome / kingdom
  - Large evolutionary distances
  - Plants, microbial, viral, fungal