黄瓜基因组测序、拼接及功能基因挖掘

Sequencing and Assembly of the Cucumber Genome and Identification of Its Functional Genes

李宏博

lihongbo_solab@163.com

主要内容

- 研究背景
- 黄瓜基因组的测序和拼接
- 利用黄瓜基因组挖掘功能基因

主要内容

- 研究背景
- 黄瓜基因组的测序和拼接
- 利用黄瓜基因组挖掘功能基因

黄瓜具有重要的研究价值

- 面积：**115万公顷** 产量：**6195万吨** (2016年)
- 是性别决定和维管束发育研究的模式物种

植物生命之树 & 重要农作物

Tree of plant kingdom

algae 藻

mosses 苔藓

Monilifor-mopses 藓

Cycadopsida 苏铁纲

Ginkgopsida 银杏纲

Gnetopsida 买麻藤纲

Coniferopsida 松柏纲

Gymnosperm 裸子植物

Amborellales 五叶梅目

Nymphaeales 睡莲目

Austrobaileyales 木兰藤目

basal angiosperms 古被子植物

Chloranthales 金粟兰目

Piperales 胡椒目

Canellales 白樟目

Magnoliales 木兰目

Laurales 樟目

magnoliids 木兰类

Acorales 菖蒲目

Alismatales 泽泻目

Poales 禾本目

Commelinales 鸭跖草目

Zingiberales 姜目

Arecales 槟榔目

Pandanales 露兜树目

Liliales 百合目

Asparagales 天门冬目

Diosc-oreales 薯蓣目

Ceratop-hyllales 金鱼藻目

Basal eudicots 分支双子叶植物

unclassified eudicots 真双子叶植物目

monocots 单子叶植物

Vitales 葡萄目

Geraniales 牻牛儿苗目

Myrtales 桃金娘目

Brassicales 十字花目

Malvales 锦葵目

Sapindales 无患子目

Malpighiales 金虎尾目

rosids 蔷薇分支

Rosales 蔷薇目

Fabales 豆目

Cucurbitales 葫芦目

Cornales 山茱萸目

Caryophyllales 石竹目

Ericales 杜鹃花目

Asterales 菊目

Apiales 伞形目

Dipsacales 川续断目

Aquifoliales 冬青目

Solanales 茄目

Lamiales 唇形目

Gentianales 龙胆目

Garryales 绞木目

asterids 菊分支

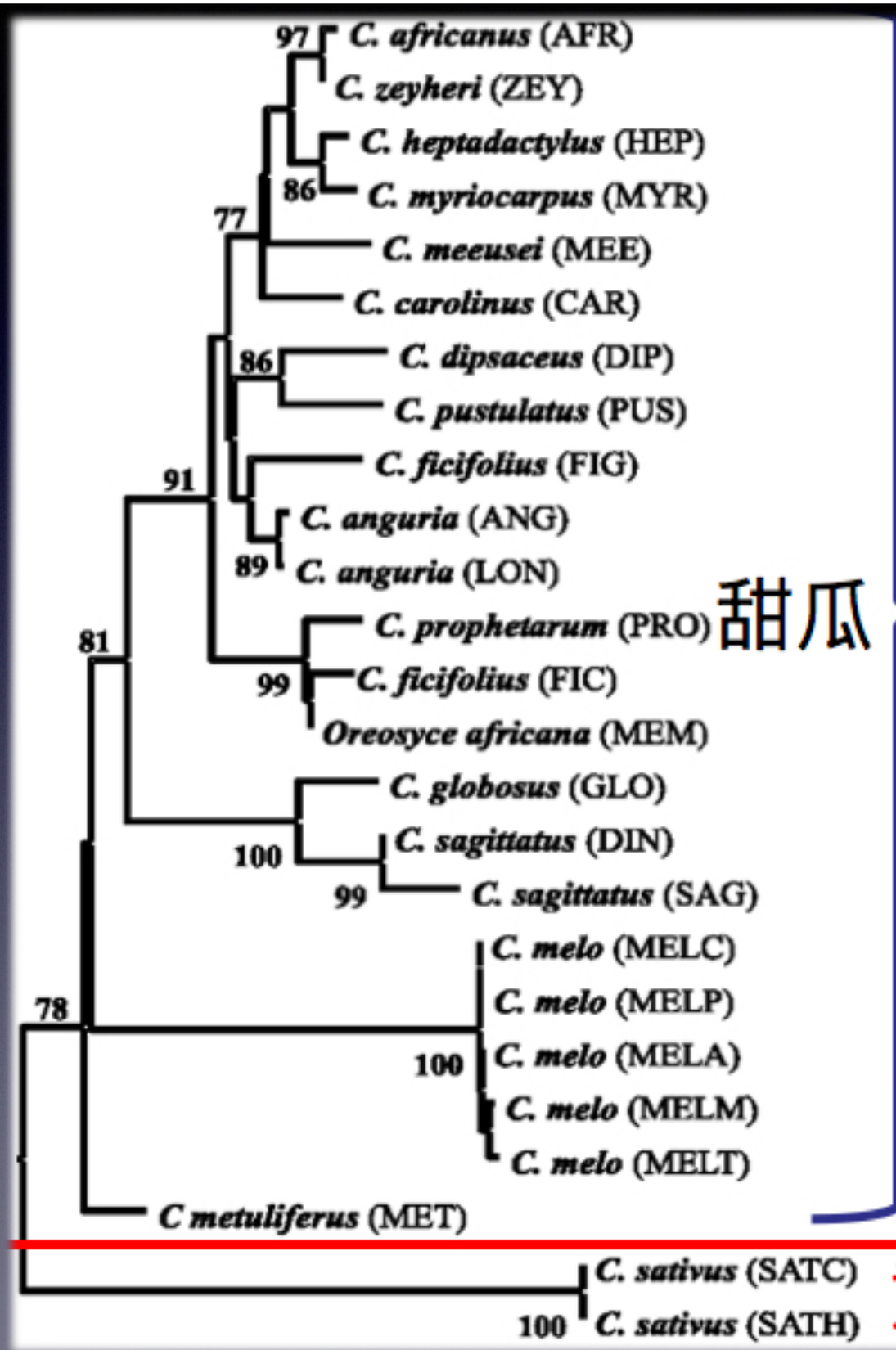
eudicots 双子叶植物

黄瓜遗传基础狭窄是结构性障碍

黄瓜是葫芦科甜瓜属中唯一单倍体染色体数目为7的物种，其余均为12，与其它近缘种基本没有基因交流。

↓

开发标记困难，正向遗传研究体系落后，制约了遗传育种研究。



基因组可突破遗传背景狭隘的瓶颈

针对黄瓜遗传基础狭窄这一结构性问题，借助**生物信息学**手段，利用基因组序列，快速构建**基因组图谱**，并探索黄瓜基因组的**变异规律**，为重要农艺性状基因克隆和分子设计育种服务。

主要内容

- 研究背景
- 黄瓜基因组的测序和拼接
- 利用黄瓜基因组挖掘功能基因

国际黄瓜基因组计划

CuGI

Cucumber Genome Initiative

国际黄瓜基因组计划

中国农业科学院蔬菜花卉研究所 IVF CAAS

华大基因 BGI

Diversity Arrays Technology Pty Ltd

High-throughput genome profiling

UC DAVIS

UNIVERSITY OF CALIFORNIA

WAGENINGEN UR

For quality of life

BEIJING NORMAL UNIVERSITY 1902

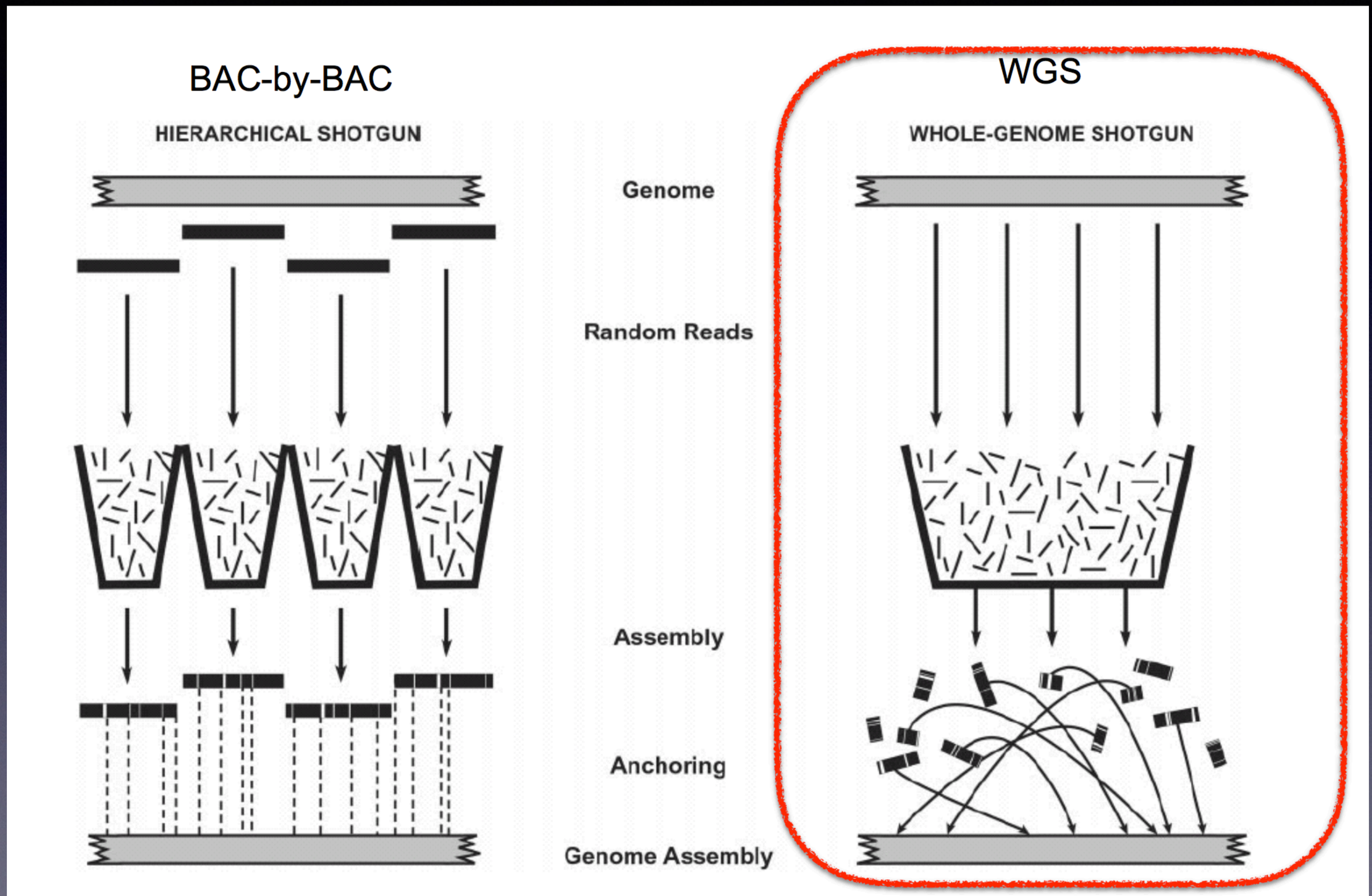
中国农业大学 CHINA AGRICULTURAL UNIVERSITY 1905

WISCONSIN

Cornell University

全基因组测序策略

两种大规模基因组测序策略



PNAS **March 19, 2002 vol. 99 no. 6 3712-3716**

基因组测序和拼接

基本原理

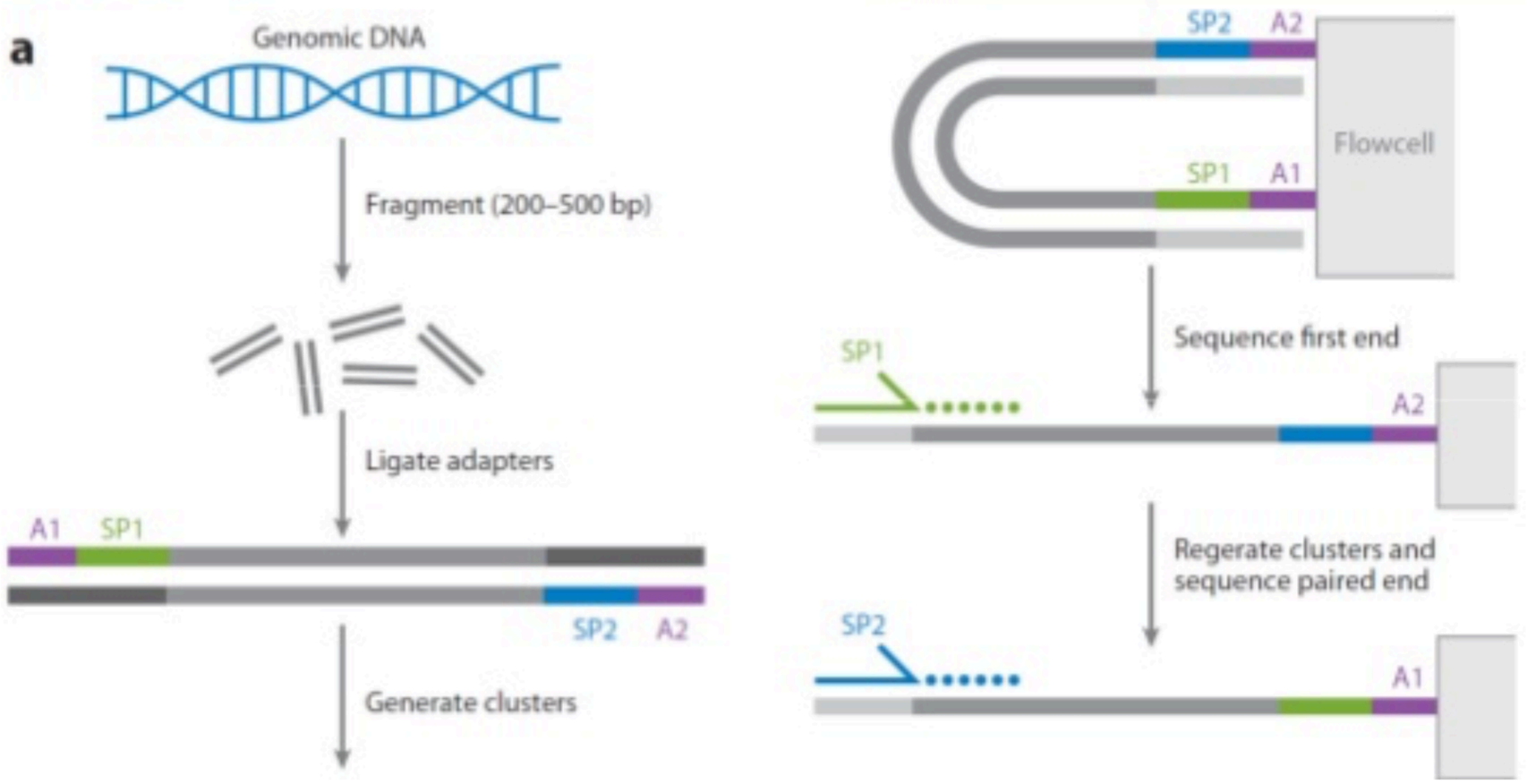
基因组测序方法

- 传统的Sanger测序 (一代测序)
- 新一代测序技术 (NGS) (Illumina, 454)
  - 单端测序 (Single-end)
  - 双端测序 (Paired-end, Mate-pair)
  - 特殊建库方式测序 (BioNano, 10X, Hi-C)
- 第三代测序技术 (PacBio, Nanopore)

基因组测序方法

- 传统的Sanger测序 (一代测序)
- 新一代测序技术 (NGS) (Illumina, 454)
  - 单端测序 (Single-end)
  - 双端测序 (Paired-end, Mate-pair)
  - 特殊建库方式测序 (BioNano, 10X, Hi-C)
- 第三代测序技术 (PacBio, Nanopore)

Illumina双端测序建库原理

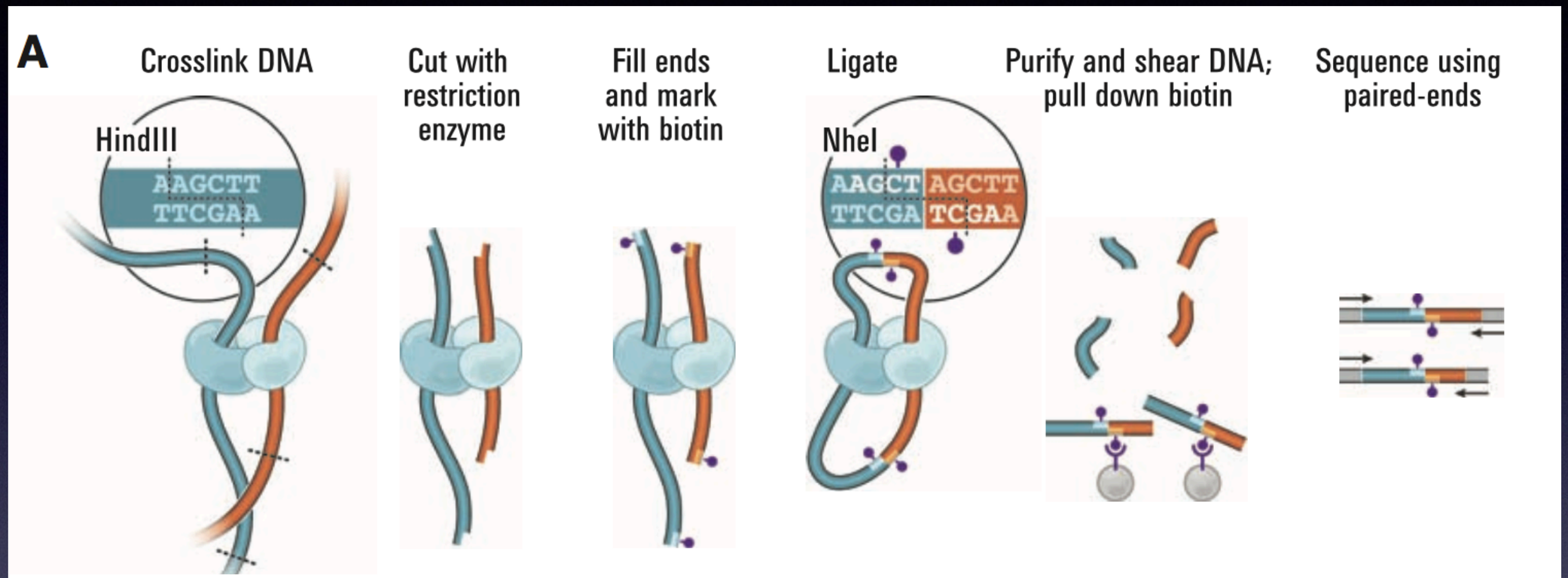


Reference Sequence：

下机数据为成对reads，一条正向，一条反向

不同插入片段长度的文库可用于构建scaffold

Hi-C测序原理



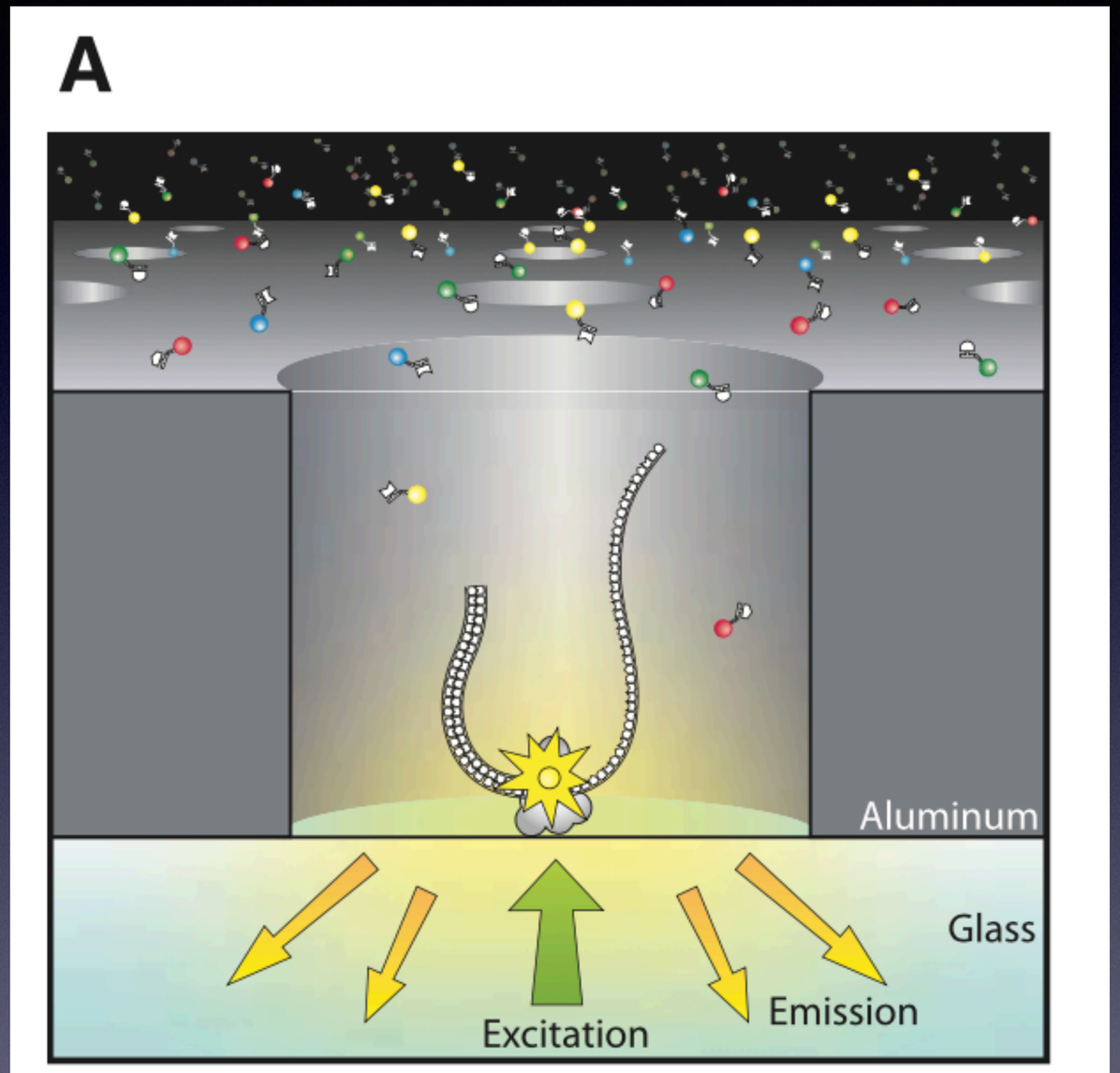
细胞培养 2.5 X 10^7 → 固定 → 酶切 → 末端修复、加生物素标记 → 连接酶连接 → 基因组DNA提取、打断

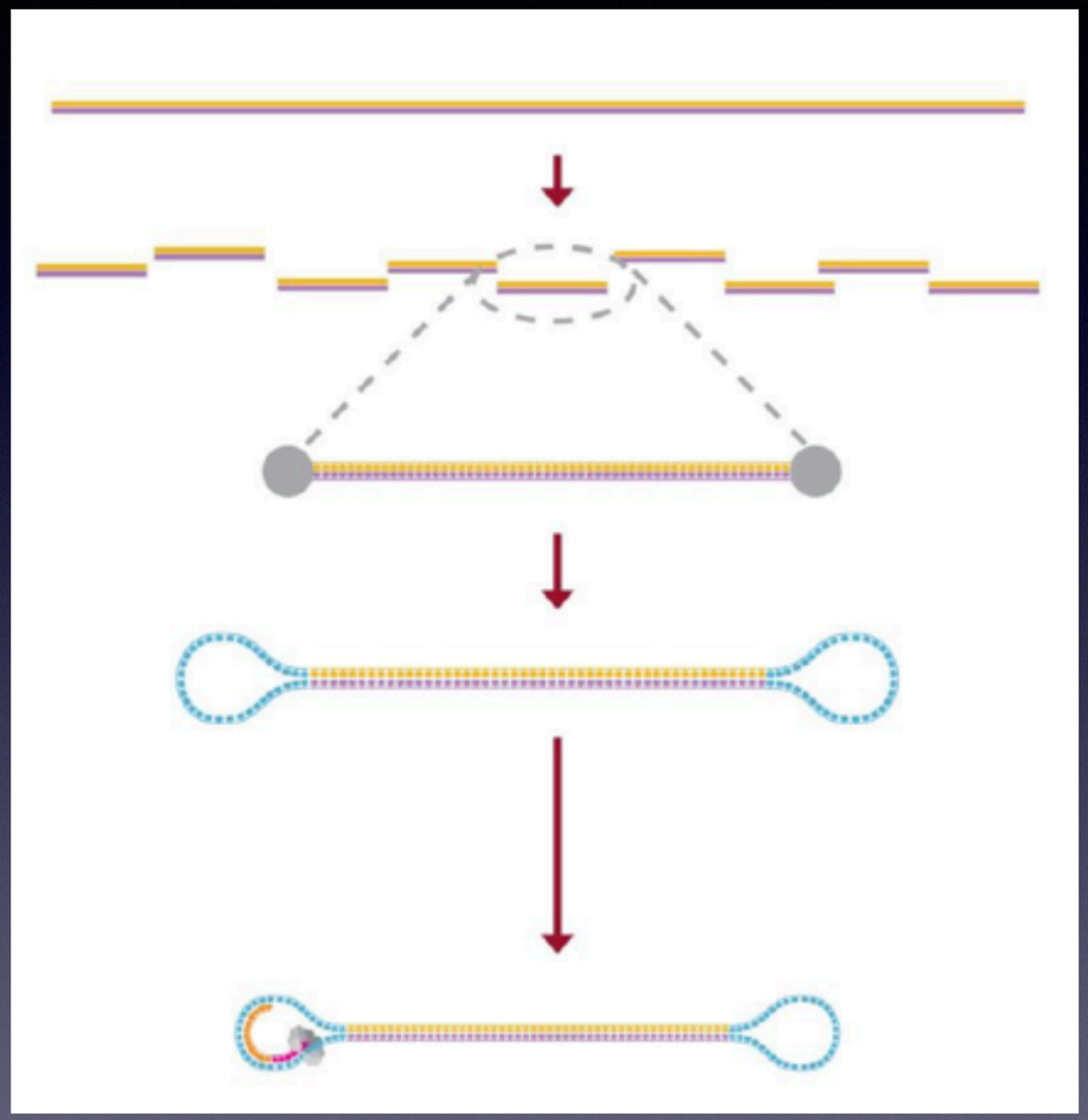
捕获基因组远距离片段的互作信息

可用于辅助基因组拼接，鉴定结构变异等

Erez et al., 2009, Science

PacBio测序原理

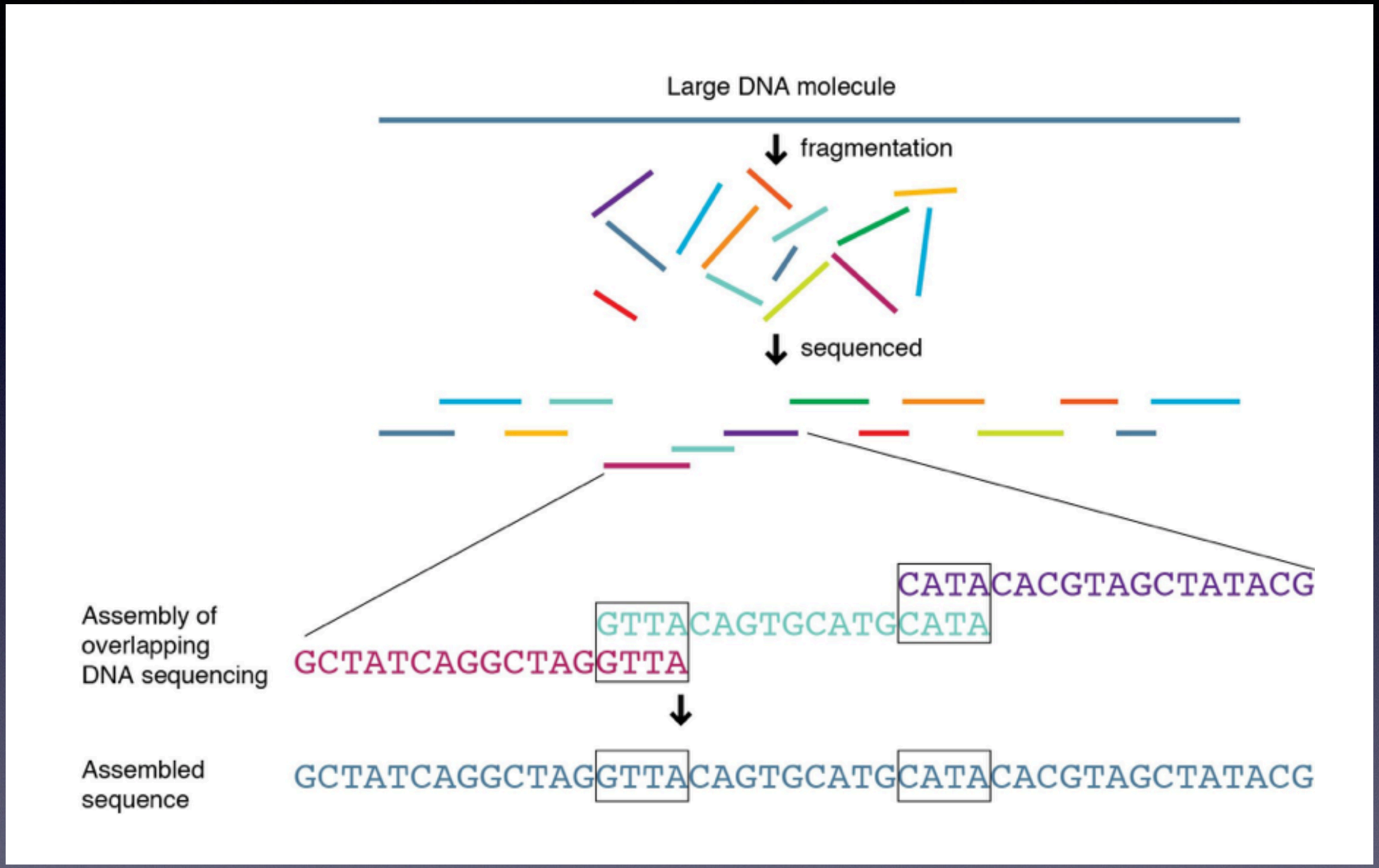




三代测序读长长，但错误率高 (~15%)

John et al., 2009, Science

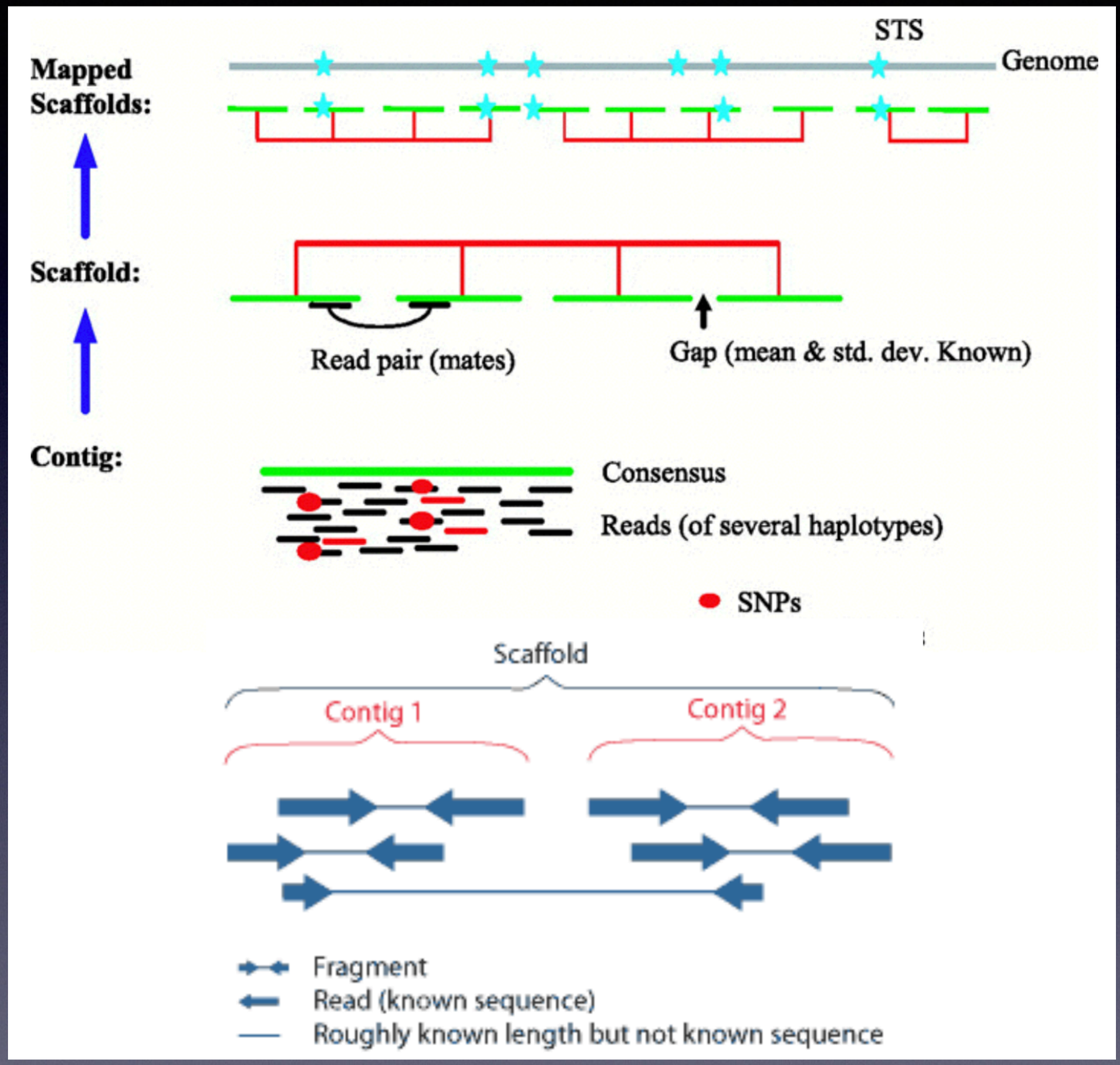
基因组拼接原理



将较短的测序读序列 (reads)

依据重叠 (overlap) 关系组装成重叠群 (contig)

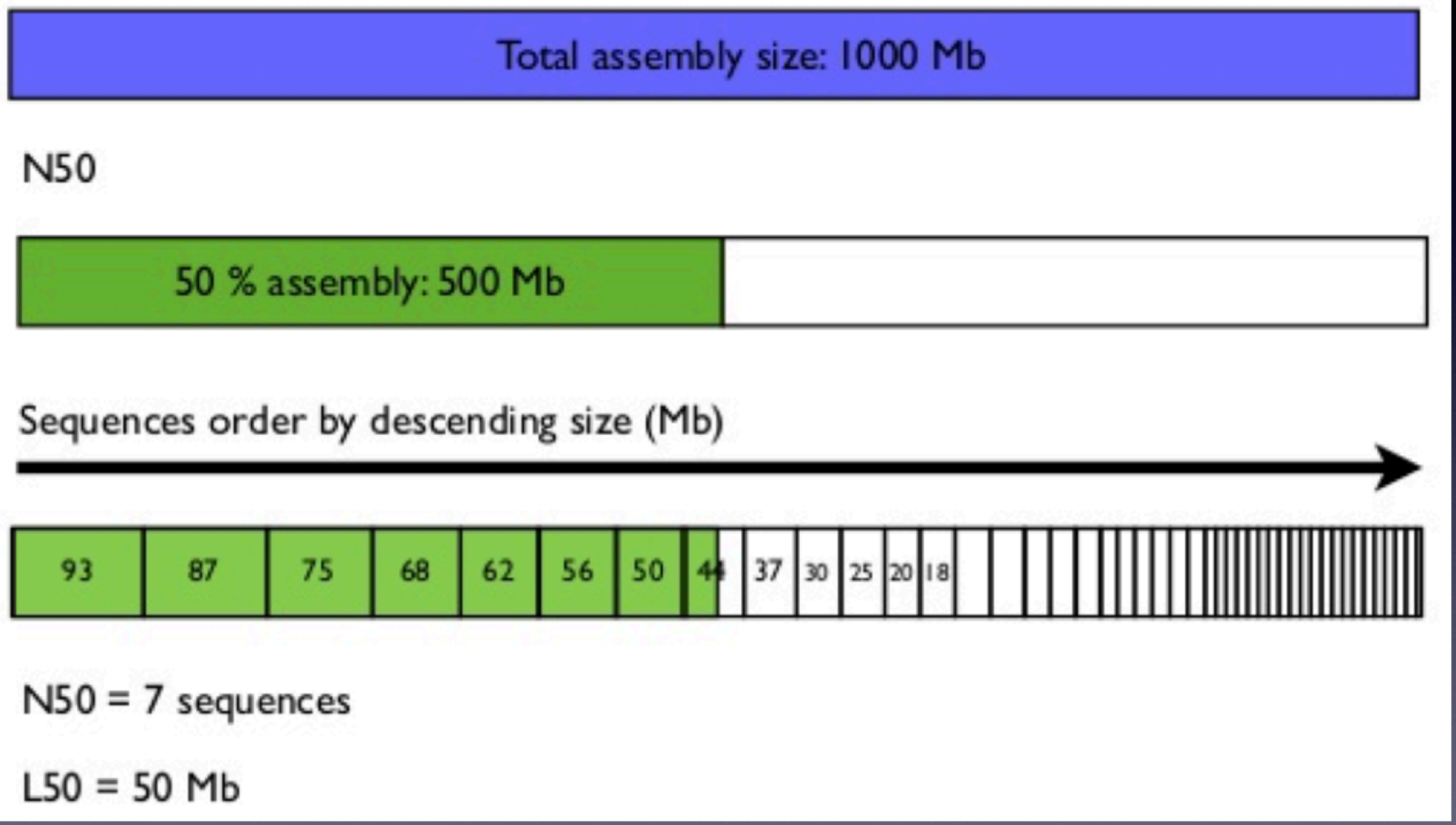
基因组拼接原理



将重叠群 (contig) 依据成对读序列间的”跨度”关系

连接成 scaffold (super-scaffold), 利用遗传图谱, Hi-C等信息锚定成染色体

N50:评估基因组拼接连续性的指标



拼接连续性越好，“gap”越少，预测出的完整基因越多

黄瓜基因组图谱的构建

基因组测序材料

华北类型

9930

Chinese long

已发表黄瓜参考基因组

ARTICLES

nature genetics

The genome of the cucumber, *Cucumis sativus* L.

Sanwen Huang[1,19], Ruiqiang Li[2,3,19], Zhonghua Zhang[1,19], Li Li[2,19], Xingfang Gu[1,19], Wei Fan[2,19], William J Lucas[4,19], Xiaowu Wang[1], Bingyan Xie[1], Peixiang Ni[2], Yuanyuan Ren[2], Hongmei Zhu[2], Jun Li[2], Kui Lin[5], Weiwei Jin[6], Zhangjun Fei[7], Guangcun Li[8], Jack Staub[9], Andrzej Kilian[10], Edwin A G van der Vossen[11], Yang Wu[5], Jie Guo[5], Jun He[1], Zhiqi Jia[1], Yi Ren[1], Geng Tian[2], Yao Lu[2], Jue Ruan[2,12], Wubin Qian[2], Mingwei Wang[2], Quanfei Huang[2], Bo Li[2], Zhaoling Xuan[2], Jianjun Cao[2], Asan[2], Zhigang Wu[2], Juanbin Zhang[2], Qingle Cai[2], Yinqi Bai[2], Bowen Zhao[13], Yonghua Han[6], Ying Li[1], Xuefeng Li[1], Shenhao Wang[1], Qiuxiang Shi[1], Shiqiang Liu[1], Won Kyong Cho[14], Jae-Yean Kim[14], Yong Xu[15], Katarzyna Heller-Uszynska[10], Han Miao[1], Zhouchao Cheng[1], Shengping Zhang[1], Jian Wu[1], Yuhong Yang[1], Houxiang Kang[1], Man Li[1], Huiqing Liang[2], Xiaoli Ren[2], Zhongbin Shi[2], Ming Wen[2], Min Jian[2], Hailong Yang[2], Guojie Zhang[2,12], Zhentao Yang[2], Rui Chen[2], Shifang Liu[2], Jianwen Li[2], Lijia Ma[2,12], Hui Liu[2], Yan Zhou[2], Jing Zhao[2], Xiaodong Fang[2], Guoqing Li[2], Lin Fang[2], Yingrui Li[2,12], Dongyuan Liu[2], Hongkun Zheng[2,3], Yong Zhang[2], Nan Qin[2], Zhuo Li[2], Guohua Yang[2], Shuang Yang[2], Lars Bolund[2,16], Karsten Kristiansen[17], Hancheng Zheng[2,18], Shaochuan Li[2,18], Xiuqing Zhang[2], Huanming Yang[2], Jian Wang[2], Rifei Sun[1], Baoxi Zhang[1], Shuzhi Jiang[1], Jun Wang[2,17], Yongchen Du[1] & Songgang Li[2]

Cucumber is an economically important crop as well as a model system for sex determination studies and plant vascular biology. Here we report the draft genome sequence of *Cucumis sativus* var. *sativus* L., assembled using a novel combination of traditional Sanger and next-generation Illumina GA sequencing technologies to obtain 72.2-fold genome coverage. The absence of recent whole-genome duplication, along with the presence of few tandem duplications, explains the small number of genes in the cucumber. Our study establishes that five of the cucumber's seven chromosomes arose from fusions of ten ancestral chromosomes after divergence from *Cucumis melo*. The sequenced cucumber genome affords insights into traits such as its sex expression, disease resistance, biosynthesis of cucurbitacin and 'fresh green' odor. We also identified 366 gene clusters related to phloem function. The cucumber genome provides a valuable resource for developing elite cultivars and for studying the evolution and function of the plant vascular system.

The botanical family Cucurbitaceae, commonly known as cucurbits and gourds, includes several economically important cultivated plants, such as cucumber (*C. sativus* L.), melon (*Cucumis melo* L.), watermelon (*Citrullus lanatus* (Thunb.) Matsum. & Nakai) and squash and pumpkin (*Cucurbita* spp.). Agricultural production of cucurbits ranks 9 million hectares of land and yields 184 million tons of vegetable fruits and seeds annually (http://faostat.fao.org). The cucurbit family also displays a rich diversity of sex expression, and the cucumber has evolved as a primary model system for sex determination studies[1]. The cucurbits are also model plants for the study of vascular biology, as both xylem and phloem sap can be readily collected for studies of long-distance signaling events[2,3].

Despite the agricultural and biological importance of cucurbits, knowledge of their genetics and genome is currently very limited. We have therefore sequenced and assembled the genome of the domestic cucumber, *C. sativus* var. *sativus* L.

All previous plant genome sequences have been derived using traditional Sanger technology[4–9]. The recent development of

[1]Key Laboratory of Horticultural Crops Genetic Improvement of Ministry of Agriculture, Sino-Dutch Joint Lab of Horticultural Genomics Technology, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China. [2]BGI-Shenzhen, Shenzhen, China. [3]Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense, Denmark. [4]Department of Plant Biology, College of Biological Sciences, University of California, Davis, California, USA. [5]College of Life Sciences, Beijing Normal University, Beijing, China. [6]National Maize Improvement Center of China, Key Laboratory of Crop Genetic Improvement and Genome of Ministry of Agriculture, Beijing Key Laboratory of Crop Genetic Improvement, China Agricultural University, Beijing, China. [7]Boyce Thompson Institute and USDA Robert W. Holley Center for Agriculture and Health, Cornell University, Ithaca, New York, USA. [8]High-Tech Research Center, Shandong Academy of Agricultural Sciences, Jinan, China. [9]US Department of Agriculture, Agricultural Research Service, Vegetable Crops Research Unit, Department of Horticulture, University of Wisconsin, Madison, Wisconsin, USA. [10]Diversity Arrays Technology, Canberra, Australia. [11]Wageningen UR Plant Breeding, Wageningen, The Netherlands. [12]The Graduate University of Chinese Academy of Sciences, Beijing, China. [13]High School Affiliated to Renmin University of China, Beijing, China. [14]Division of Applied Life Science (BK21 and WCU program), PMBBRC and EB-NCRC, Gyeongsang National University, Jinju, Republic of Korea. [15]National Engineering Research Center for Vegetables, Beijing, China. [16]Institute of Human Genetics, University of Aarhus, Aarhus, Denmark. [17]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [18]South China University of Technology, Guangzhou, China. [19]These authors contributed equally to this work. Correspondence should be addressed to Y.D. (yongchen.du@mail.caas.net.cn), S.H. (huangsanwen@caas.net.cn), Jun Wang (wangj@genomics.org.cn) or Songgang Li (lisg@genomics.org.cn).

Li *et al. BMC Genomics* 2011, **12**:540
http://www.biomedcentral.com/1471-2164/12/540

BMC Genomics

RESEARCH ARTICLE Open Access

RNA-Seq improves annotation of protein-coding genes in the cucumber genome

Zhen Li[1†], Zhonghua Zhang[2†], Pengcheng Yan[1], Sanwen Huang[2], Zhangjun Fei[3] and Kui Lin[1*]

Abstract

Background: As more and more genomes are sequenced, genome annotation becomes increasingly important in bridging the gap between sequence and biology. Gene prediction, which is at the center of genome annotation, usually integrates various resources to compute consensus gene structures. However, many newly sequenced genomes have limited resources for gene predictions. In an effort to create high-quality gene models of the cucumber genome (*Cucumis sativus* var. *sativus*), based on the EVidenceModeler gene prediction pipeline, we incorporated the massively parallel complementary DNA sequencing (RNA-Seq) reads of 10 cucumber tissues into EVidenceModeler. We applied the new pipeline to the reassembled cucumber genome and included a comparison between our predicted protein-coding gene set and the published set.

Results: The reassembled cucumber genome annotated with RNA-Seq reads from 10 tissues, has 23, 248 identified protein-coding genes. Compared with the published prediction in 2009, approximately 8, 700 genes reveal structural modifications and 5, 2[illegible] genes only appear in the reassembled cucumber genome. All the related results, including genome sequences and annotations, are available at http://cmb.bnu.edu.cn/Cucumis_sativus_v20/.

Conclusions: We conclude that RNA-Seq greatly improves the accuracy of prediction of protein-coding genes in the reassembled cucumber genome. The comparison between the two gene sets also suggests that it is feasible to use RNA-Seq reads to annotate newly sequenced or less-studied genomes.

Background

As new sequencing technologies develop, thousands of eukaryotic genomes across all kingdoms of life will be sequenced during the next decade [1,2], and this trend will spark an improvement in our knowledge of evolutionary biology and functional genomics. Genome annotation is a stepping stone to bridge the gap between genomic sequences and the biology of organisms [3]. It can be stated that the quality of genome annotations represents the value of genome sequences.

Gene prediction, within the process of genome annotation, is a complex endeavor. In eukaryotic species, it is usually carried out by integrating multiple sources of evidence [4], such as complementary DNA (cDNA), proteins in closely related species, and *de novo* predictions [5]. Representing the integral sequences of messenger RNAs (mRNAs), full-length cDNAs (FL-cDNAs) are recognized as the gold-standards for discovering and annotating gene structures in eukaryotic genomes [5,6]. Additionally, even incomplete cDNAs, i.e. expressed sequence tags (ESTs), provide more accurate evidence than other sources. Nevertheless, until recently, the sequencing of cDNA was a laborious and capital-intensive task.

Thanks to the massively parallel cDNA sequencing (RNA-Seq) technologies [7], scientists can obtain cDNA fragments from transcriptomes with reasonably complete coverage in a reduced time scale and at a lower cost [8]. With its informative content, RNA-Seq is expected to revolutionize the prediction of genes [9]. RNA-Seq has been used to improve the genome annotations, including: (i) correcting predicted gene structures [10]; (ii) detecting new alternative splicing isoforms [11]; and (iii) discovering new genes and new transcripts [12,13]. However, most of these applications focused on species with well-annotated genomes, such as human, mouse, yeast, *Arabidopsis thaliana*, and rice. Among these studies, Trapnell, Williams and Pertea *et al*. and Guttman, Garver and

* Correspondence: linkui@bnu.edu.cn
† Contributed equally
[1]College of Life Sciences, Beijing Normal University, 19 Xinjiekouwai Street, Beijing, 100875, China
Full list of author information is available at the end of the article

LOW CONTIGUITY

9930 V1.0
Total length: 243.5Mb
Contig N50: 19.8kb

9930 V2.0
Total length: 196.5Mb
Contig N50: 37.9kb

黄瓜基因组拼接结果

PacBio
Contig

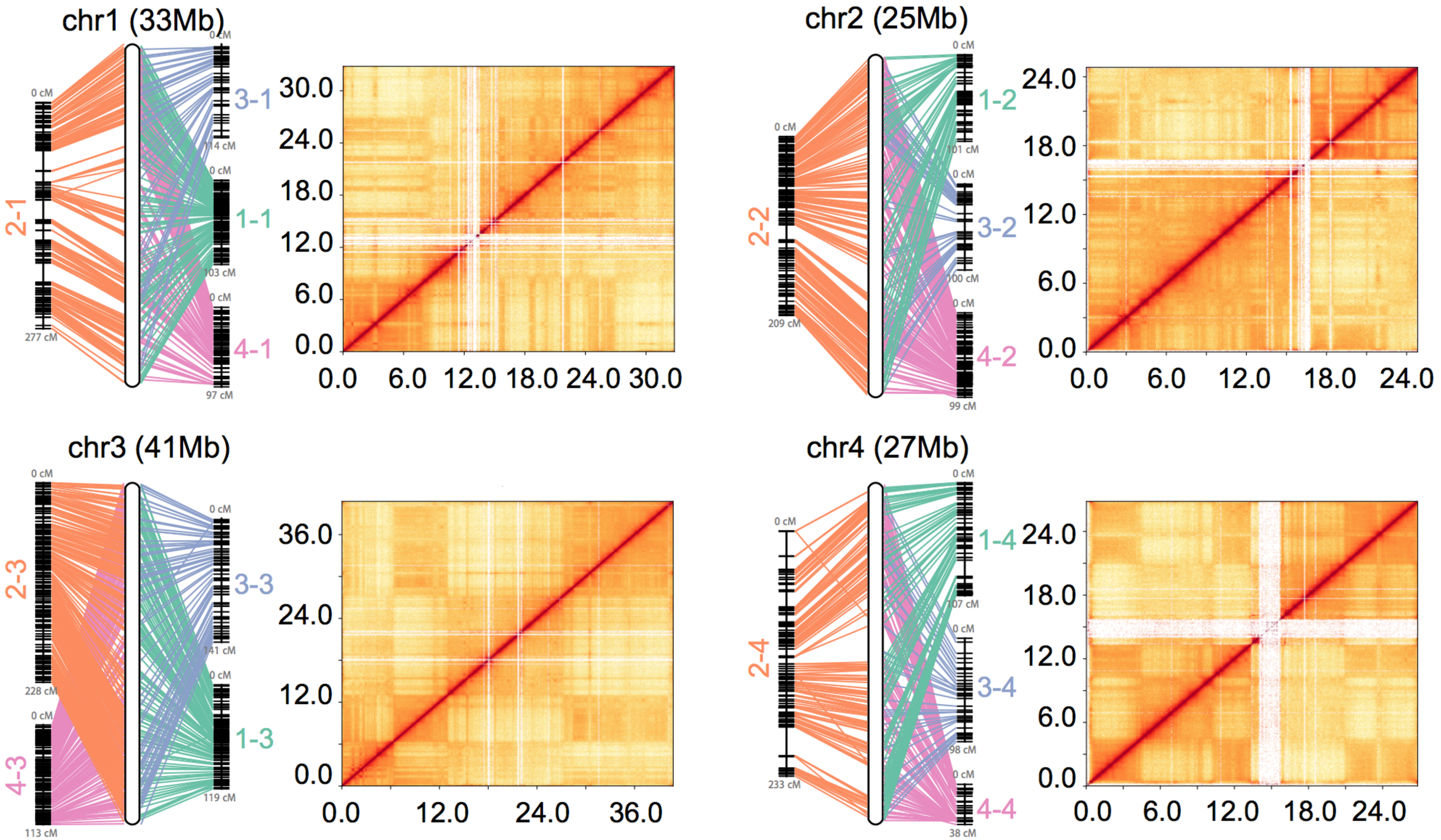
Illumina
Correcting

10X
Scaffold

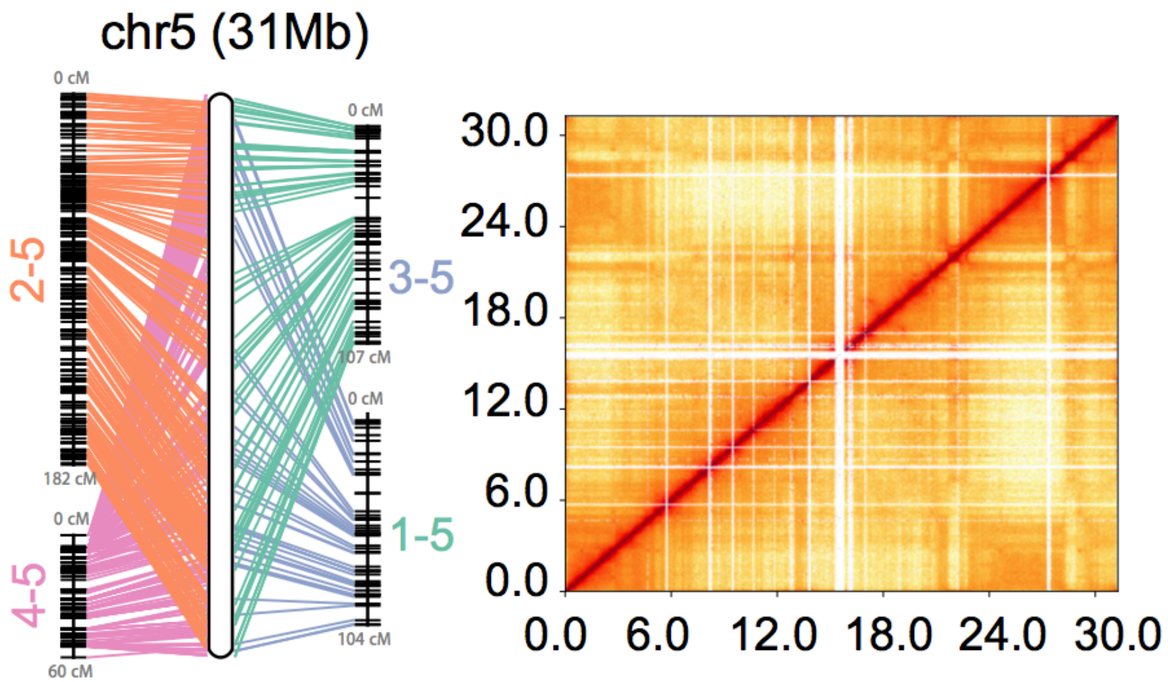
Hi-C
Super-scaffold

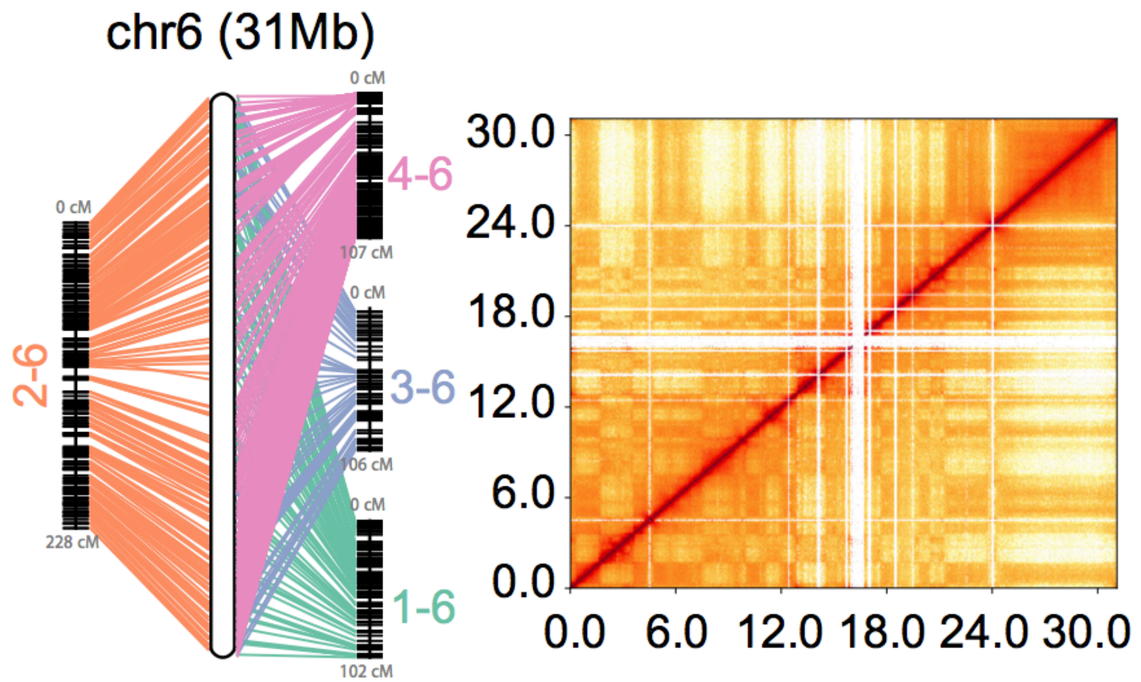
	Contig	Scaffold	Super-scaffold
总长 (Mb)	226.2	226.2	226.6
Gap 长度 (kb)	0	0.2	35.2
Contig 个数	174	-	-
Contig N50 (Mb)	8.9	-	-
Scaffold 个数	-	157	85
Scaffold N50 (Mb)	-	11.5	31.1
Largest Scaffold (Mb)	-	21.7	40.9

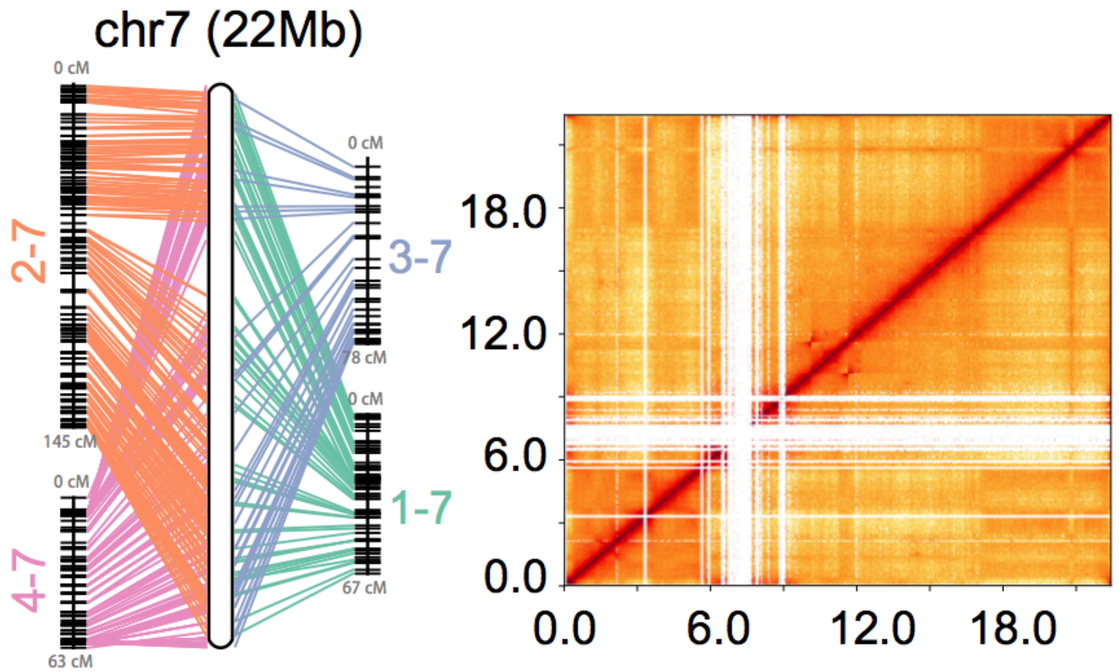
遗传图谱和Hi-C数据整合物理图谱



遗传图谱和Hi-C数据整合物理图谱







基因组注释

24,317

蛋白编码基因

90.8%

BUSCO*

93.3Mb

重复序列

40.8%

占全基因组

*BUSCO: 有胚植物单拷贝保守基因集

主要内容

- 研究背景
- 黄瓜基因组的测序和拼接
- 利用黄瓜基因组挖掘功能基因

黄瓜果肉颜色控制基因的克隆

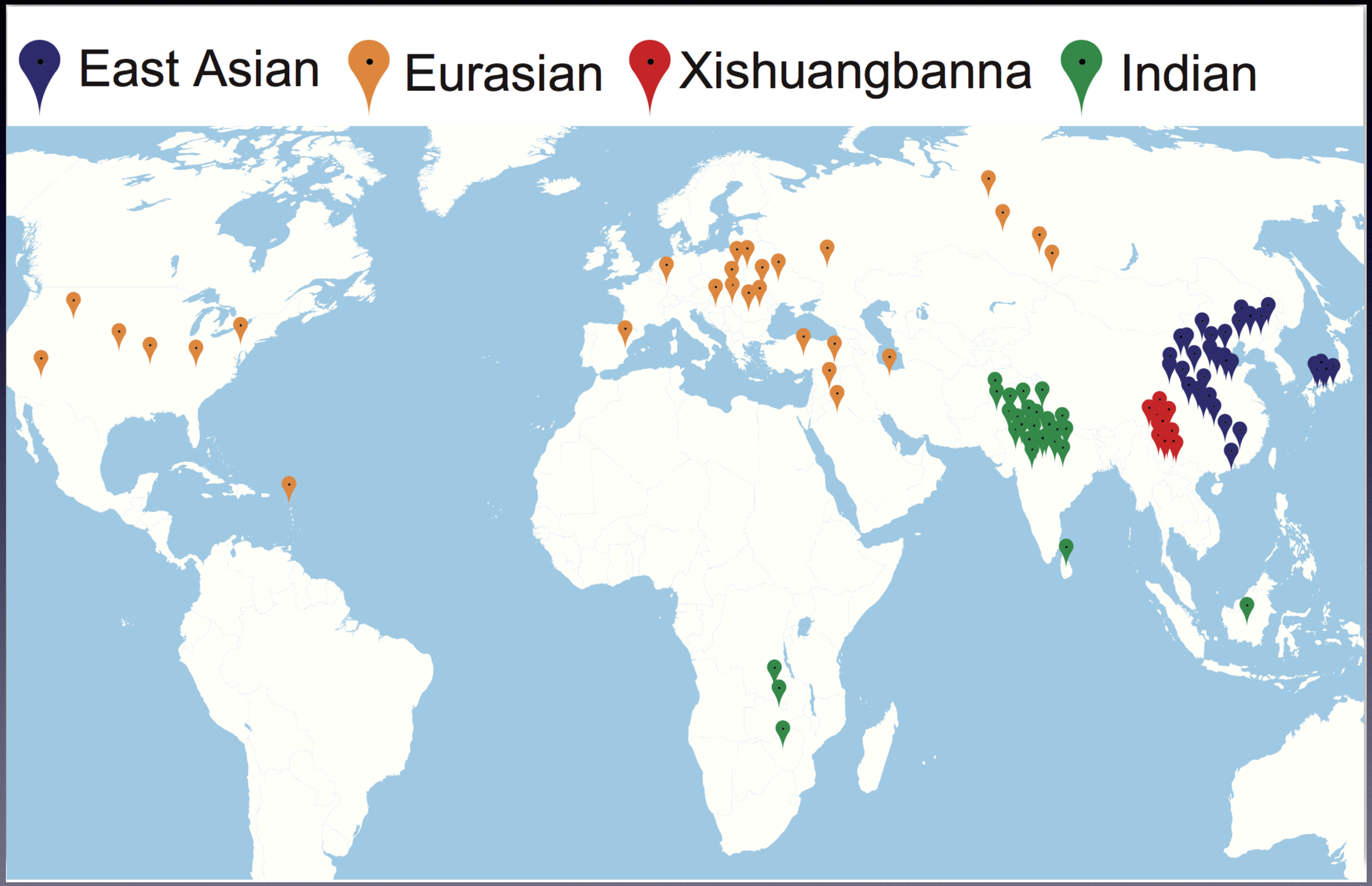
包含115份材料的核心种质资源库








包含115份材料的核心种质资源库



A variation map

115

core germplasm lines, representing 80% diversity

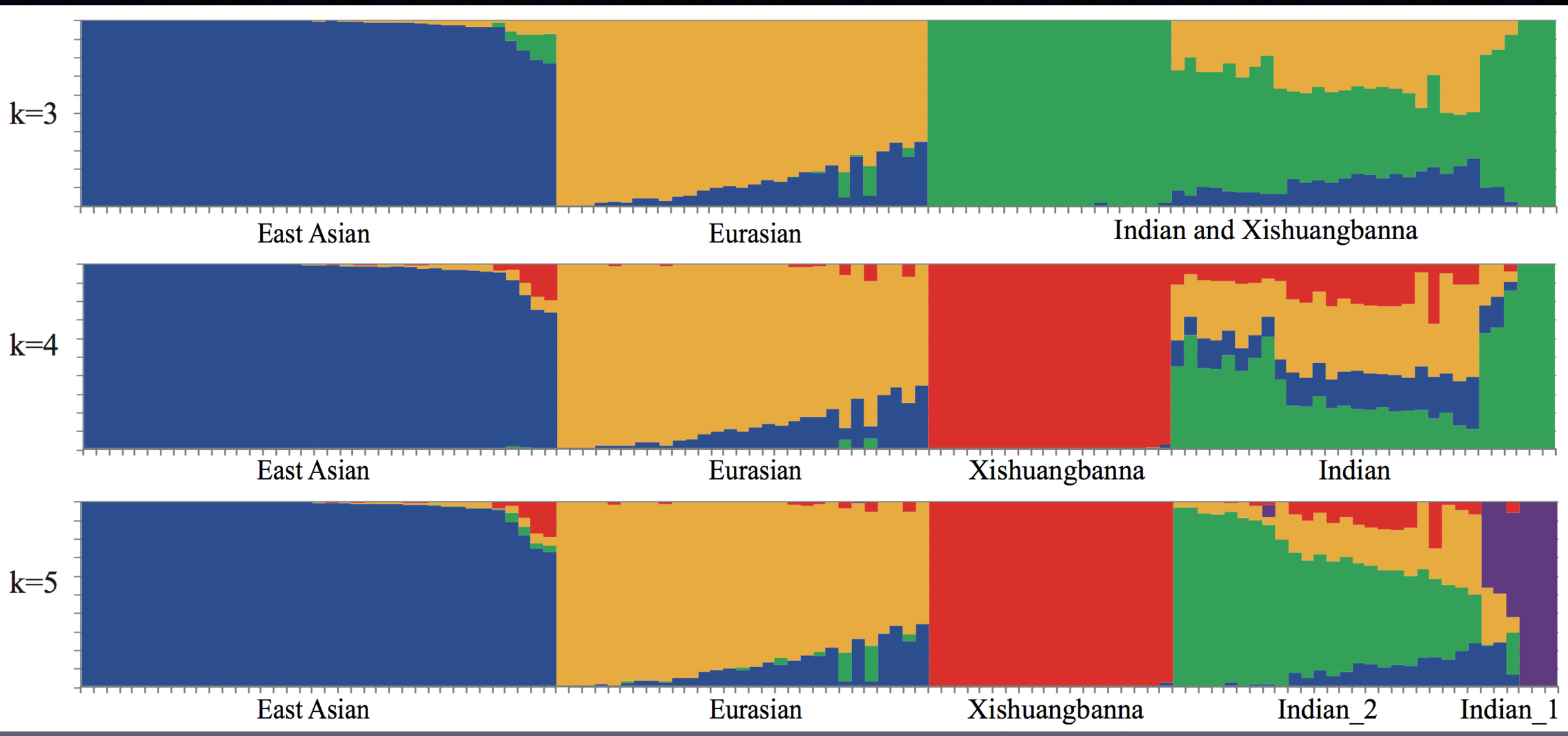
18X

sequencing depth

3.6M

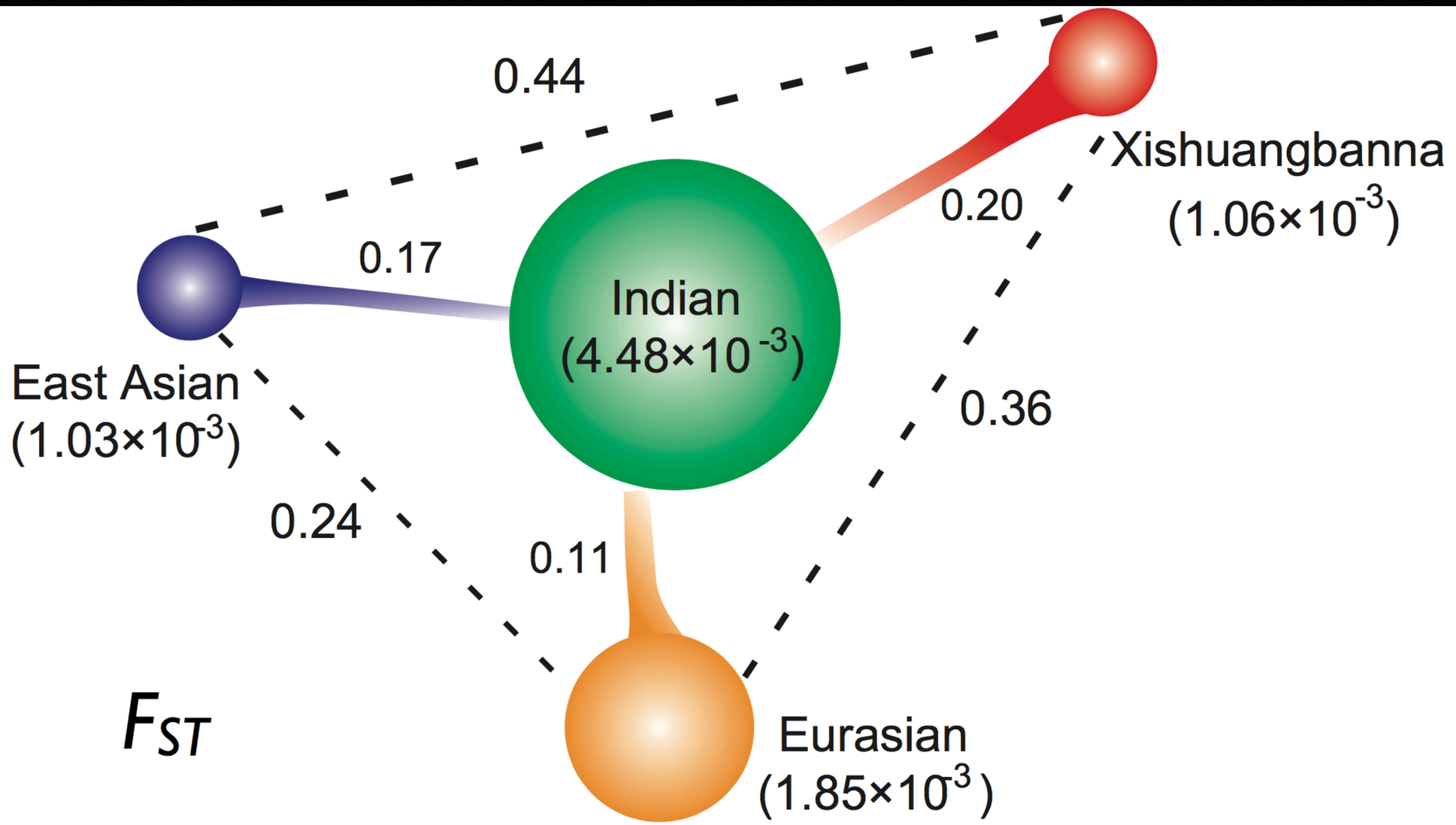
variants (SNP, small InDel)

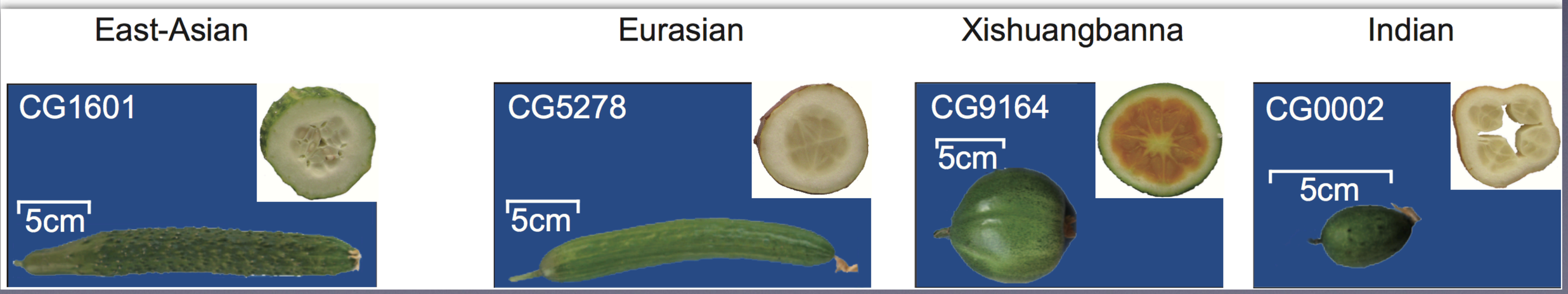
群体结构



印度类群为野生类群

群体分化和特征


0.44
0.17
0.20
0.24
0.11
0.36
East Asian
(1.03×10-3)
Indian
(4.48×10-3)
Xishuangbanna
(1.06×10-3)
Eurasian
(1.85×10-3)
FST

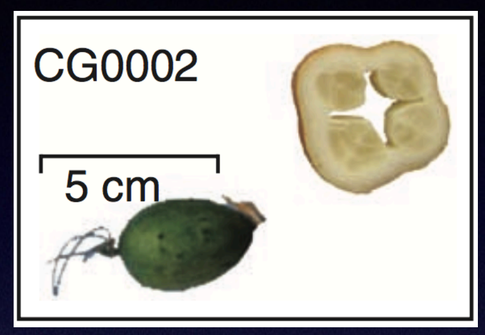

East-Asian
Eurasian
Xishuangbanna
Indian
CG1601
5cm
CG5278
5cm
CG9164
5cm
CG0002
5cm

受选择异义突变SNP的鉴定

东亚

欧美

印度



西双版纳

96份

19份

F_{ST}=1 受选择的异义突变

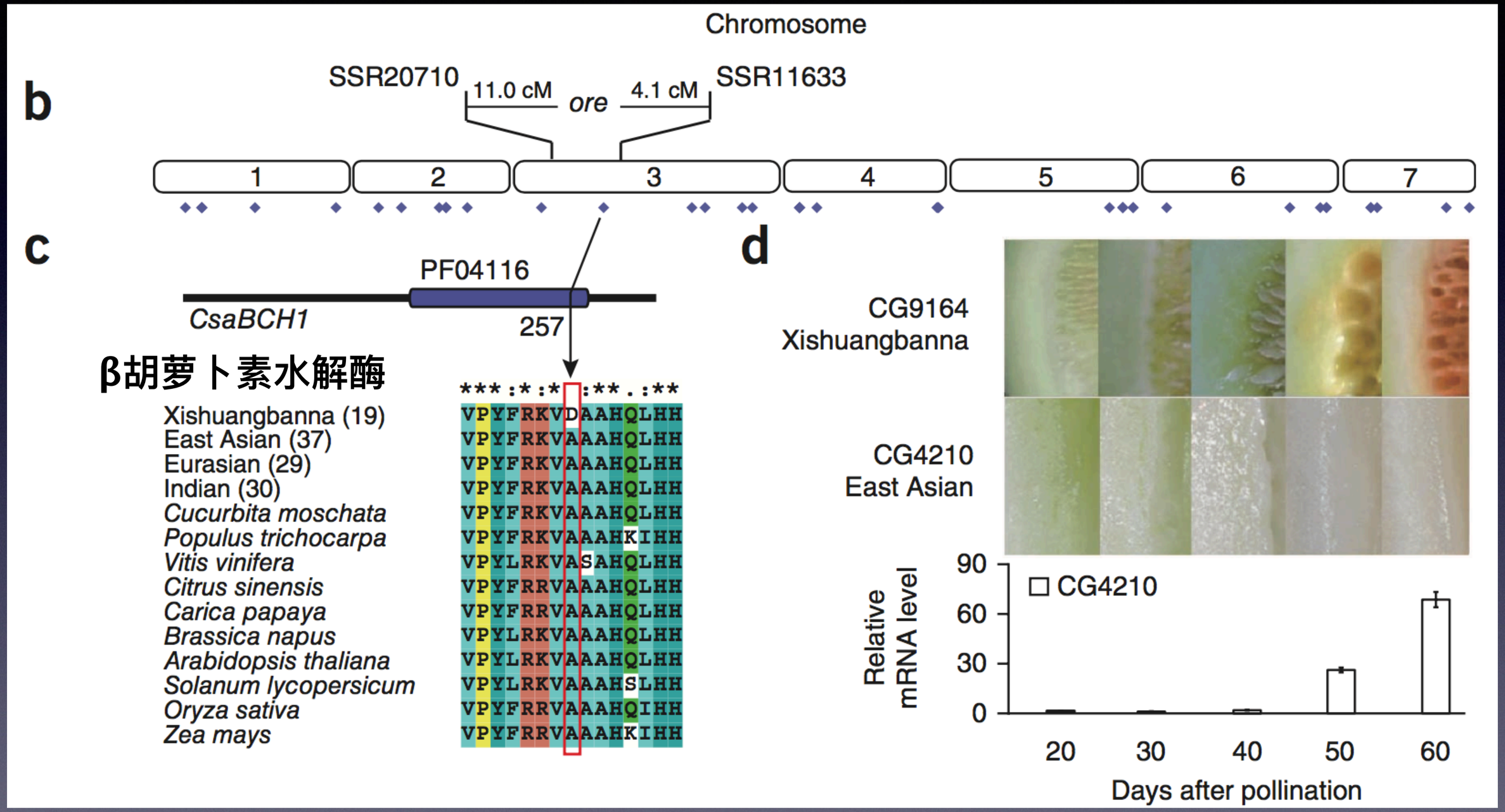
3.3M

SNP

43

SNP

果实积累β胡萝卜素*ore*基因的克隆



发现了黄瓜果实中控制**β胡萝卜素积累的**

关键基因突变，为克隆重要功能基因提供了**新的思路**

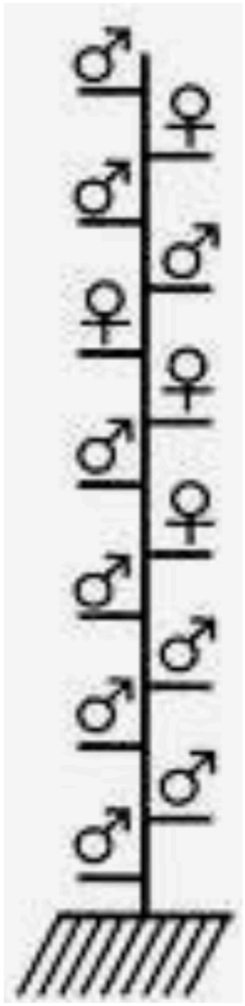
黄瓜性别决定基因*F*

雌雄同株异花与全雌植株

f Monoecious

F Gynoecious

雌性基因*F*的已有研究结果

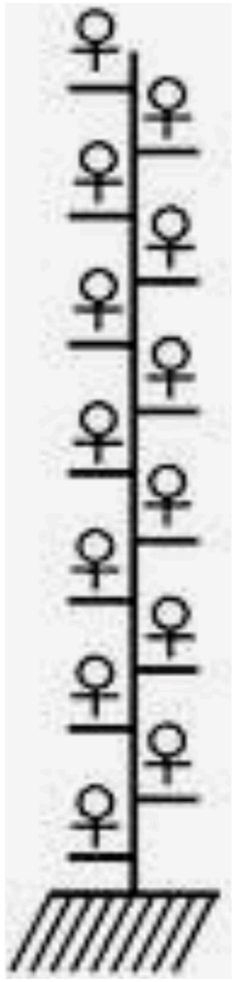


单性系
(ff*MM*)

1个*ACS1*基因

CNV产生机制？

2个ACS1基因



全雌系
(*FFMM*)

黄瓜结构变异图谱

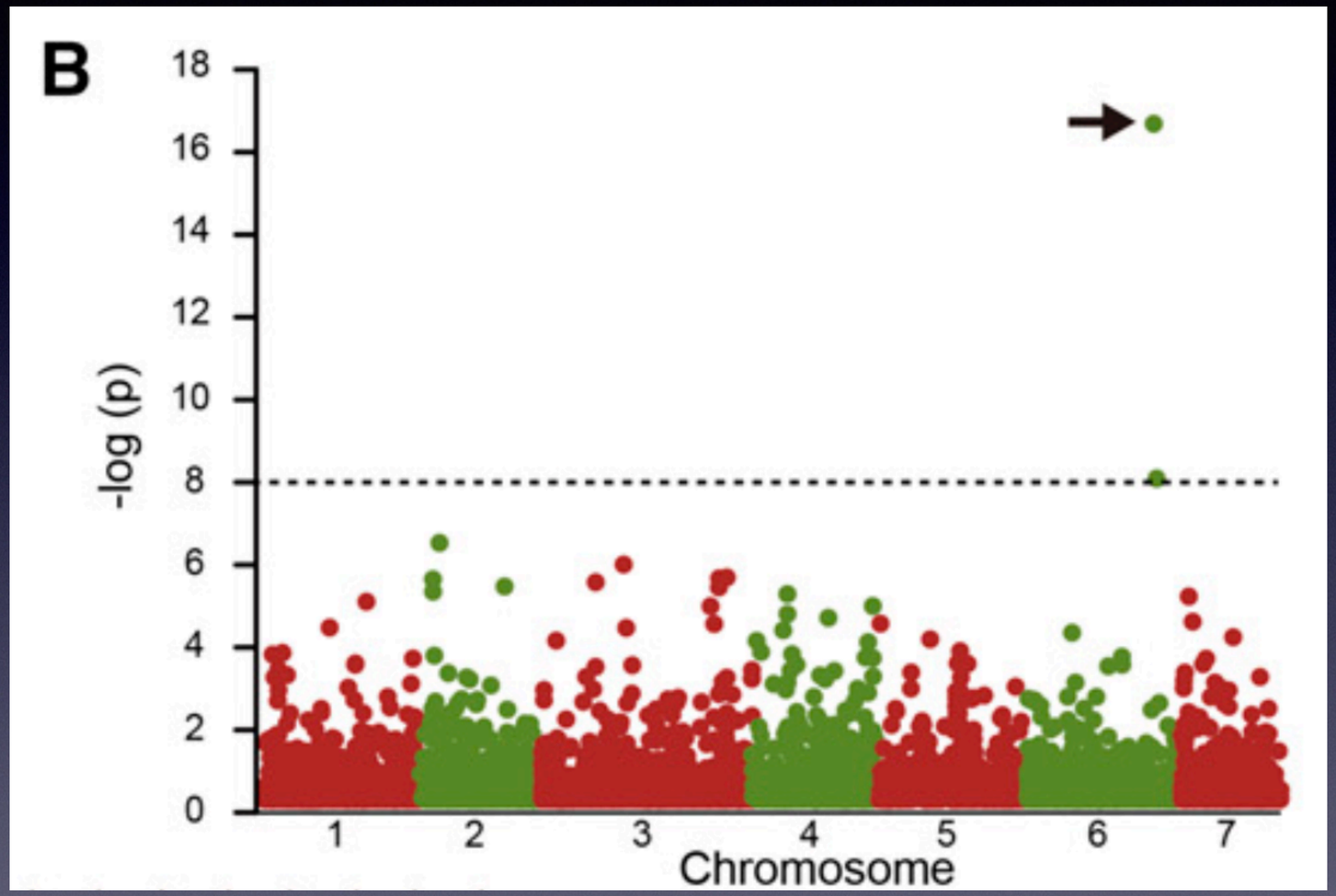
Table 1. Summary of Cucumber SVs and Their Functional Impact

	Deletions	Insertions	Duplications	Inversions	Total
No.	19,168	7,337	205	78	26,788
Size (bp)	6,501,467	679,902	4,212,813	325,747	11,751,414
Number of SVs overlapping with protein coding genes[a]					
Full CDS overlap	112 (130)[b]	0 (0)	134 (509)	15 (27)	261 (666)
Partial CDS overlap	619 (622)	226 (217)	117 (149)	17 (22)	979 (1,010)
UTR overlap[c]	266 (262)	90 (90)	7 (7)	1 (1)	364 (360)
Intron overlap[c]	2,409 (1,979)	863 (775)	3 (3)	2 (3)	3,277 (2,760)
Total	3,379 (2,952)	1,175 (1,076)	171 (668)	27 (53)	

[a]An SV can fall into multiple categories.
[b]Values in parentheses indicate the number of genes overlapping with SVs.
[c]Only genes with CDS not overlapping with SVs are counted.

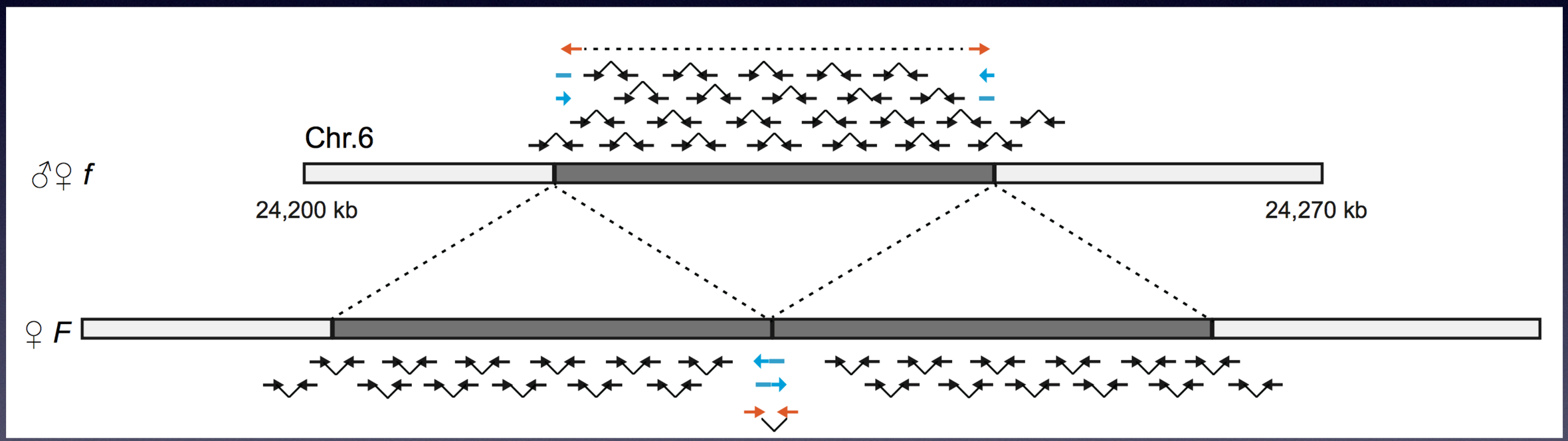
将115份核心种质资源的重测序数据
比对到9930黄瓜参考基因组上，鉴定结构变异 (SV)

GWAS鉴定出与全雌性状显著关联信号



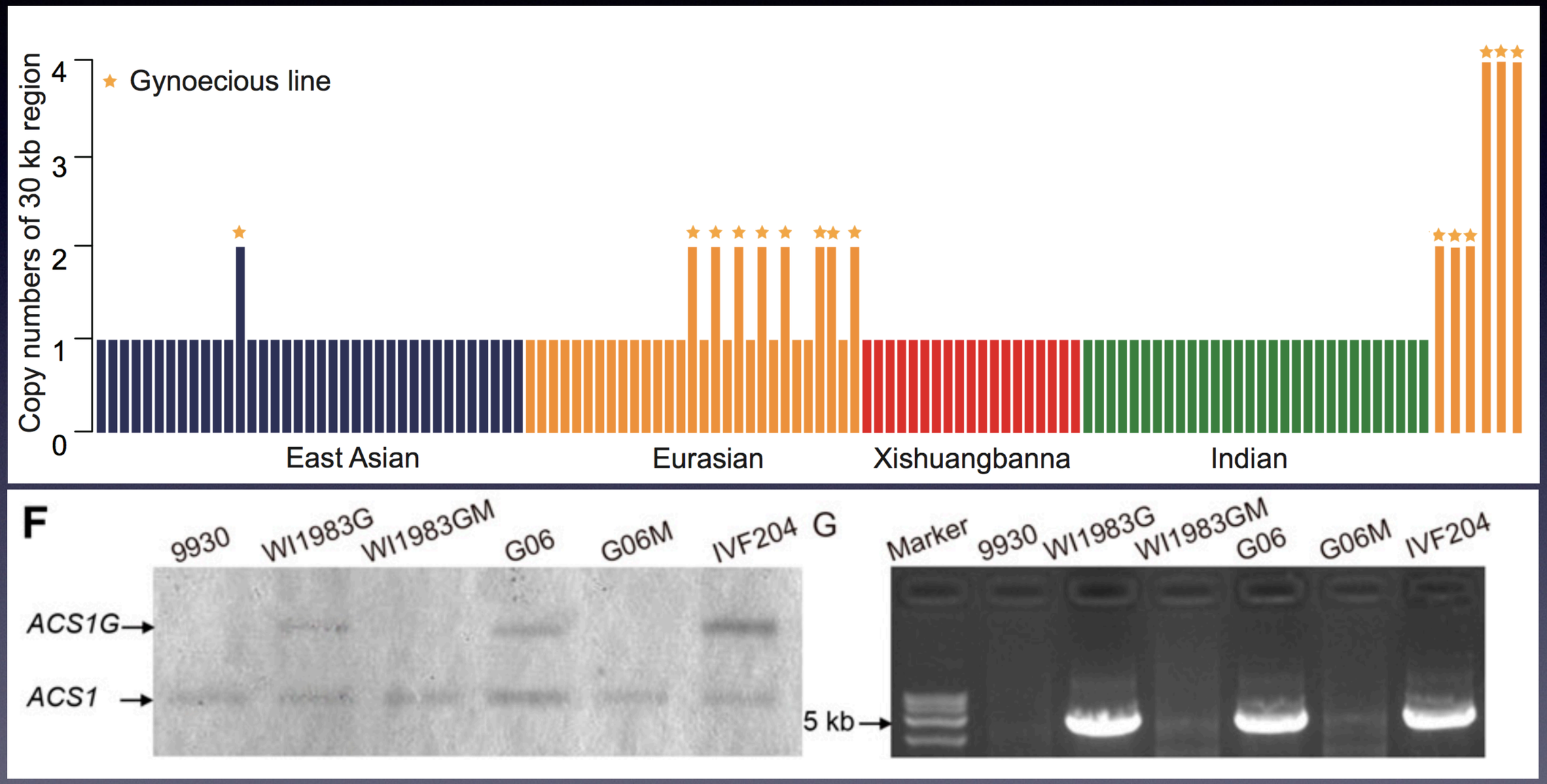
使用群体的SV数据集进行全基因组关联分析

*F*基因区域存在拷贝数变异



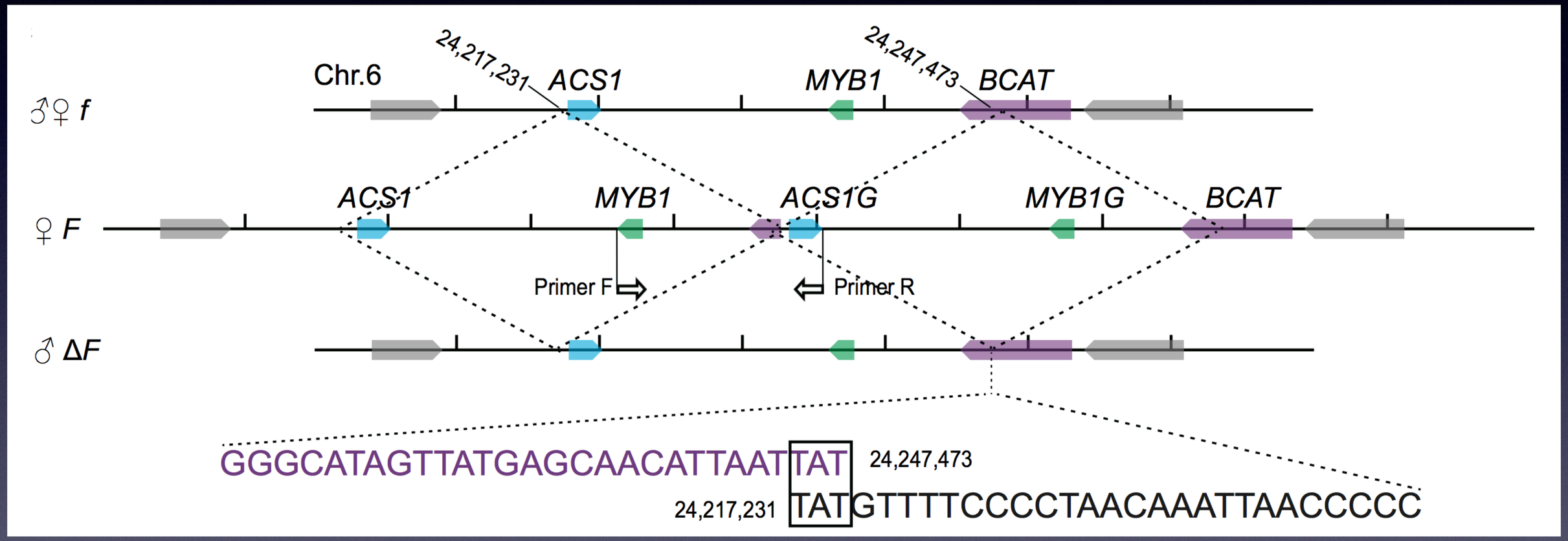
在全雌材料中存在一个30kb的重复

重复区域与全雌性状的紧密关联



所有全雌材料均含多个30kb重复

*F*基因的产生机制



*F*基因的产生涉及**三个基因的串联重复**，断点处TAT重复

暗示**微同源介导的断裂诱导复制**可能是该拷贝数变异形成的机制

总结

- 利用PacBio, Hi-C等测序技术构建了高质量的黄瓜参考基因组图谱，并完成了基因和重复序列的注释分析
- 构建了全基因组变异图谱，确定了黄瓜的群体结构、特征
- 探索了利用群体分化研究功能基因的方法，快速克隆了黄瓜果实积累β胡萝卜素基因*ore*
- 利用SV数据，揭示了黄瓜雌性基因*F*的产生机制

论文发表情况

- Huang et al., 2009. The genome of the cucumber, *Cucumis sativus* L. *Nature Genetics* 41:1275-1281.
- Li et al., 2011. RNA-Seq improves annotation of protein-coding genes in the cucumber genome. *BMC Genomics* 12:540.
- Qi et al., 2013. A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nature Genetics* 45 (12): 1510-1515.
- Zhang et al., 2015. Genome-Wide mapping of structural variations reveals a copy number variant that determines reproductive morphology in cucumber. *Plant Cell* doi:10.1105/tpc.114.135848.
- Li et al,. A gold standard reference genome for cucumber (*Cucumis sativus* L.). *In Preparation.*

致谢

张忠华
研究员

黄三文
研究员

谢谢！

恳请各位批评指正！

lihongbo_solab@163.com