A Gold Standard Reference Genome for Cucumber (*Cucumis sativus* L.)

Speaker: Hongbo Li
Group2 member: Xin Zhang, Qi Shen & Qian Hou

提纲

- 研究背景
- 从头测序和序列装配
- 基因组结构注释，基因功能预测
- 全文结论

提纲

- 研究背景
- 从头测序和序列装配
- 基因组结构注释，基因功能预测
- 全文结论

黄瓜具有重要的研究价值

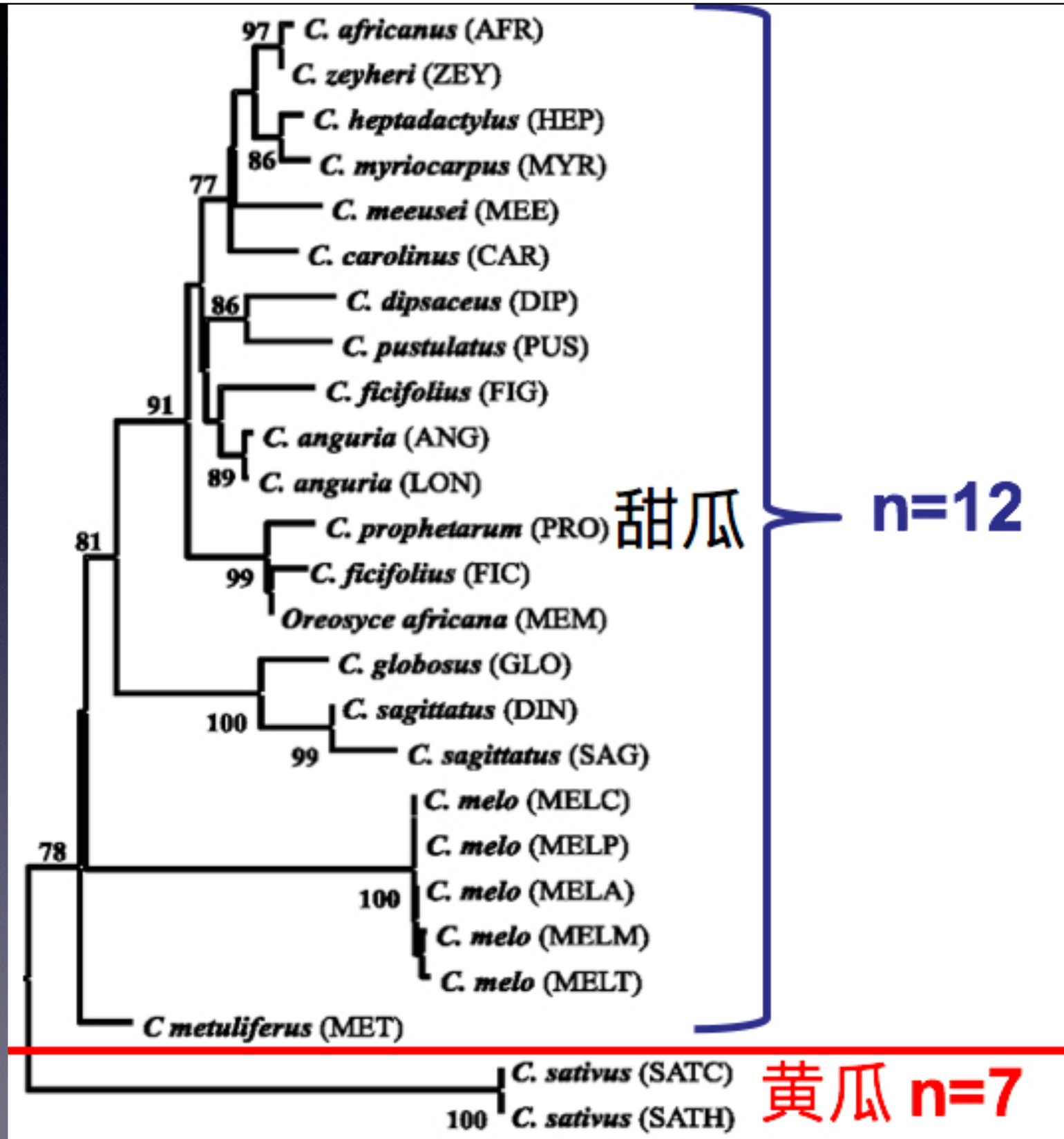
- 面积：**115万公顷** 产量：**6195万吨** (2016年)
- 是性别决定、果实发育与品质、维管束发育研究的模式物种

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。

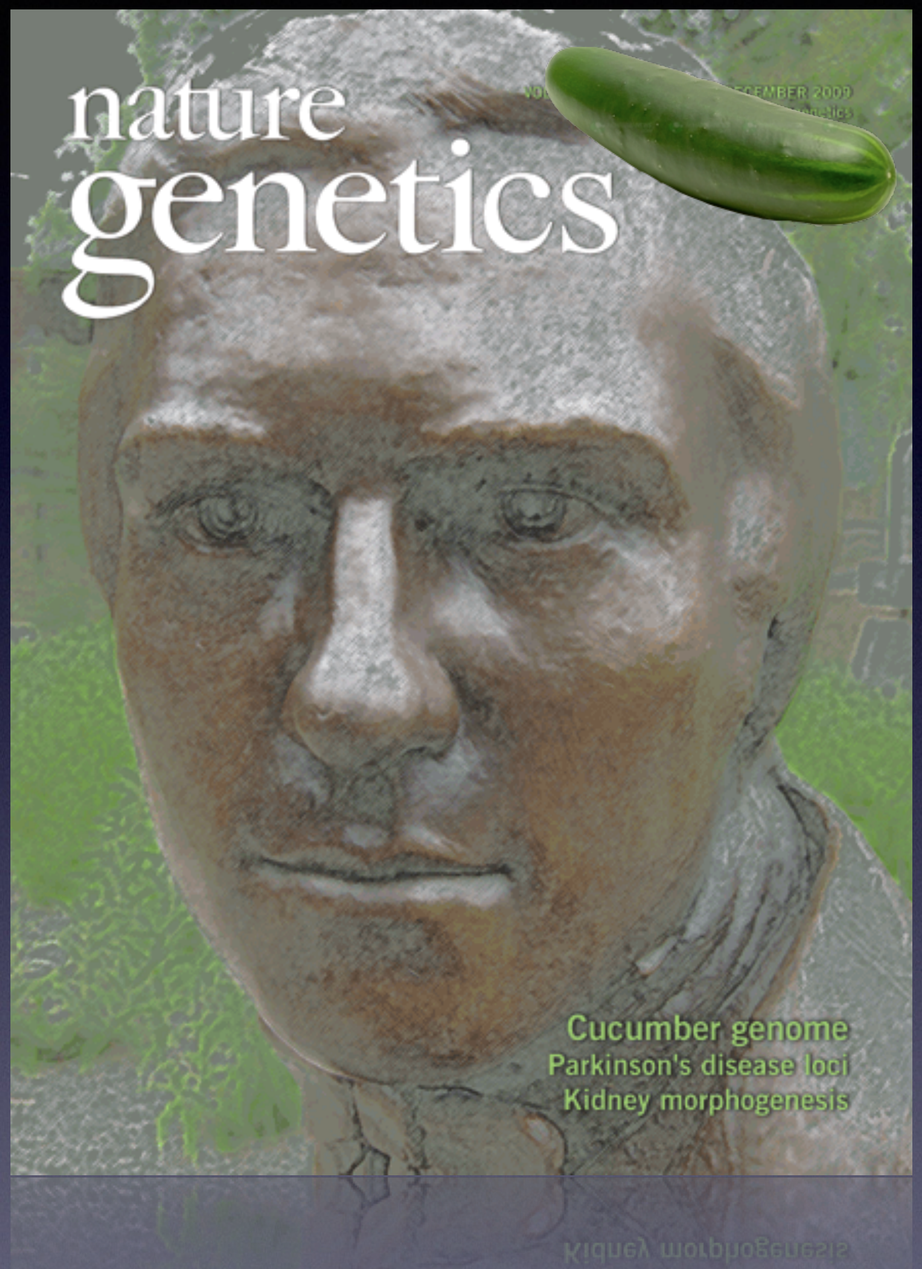


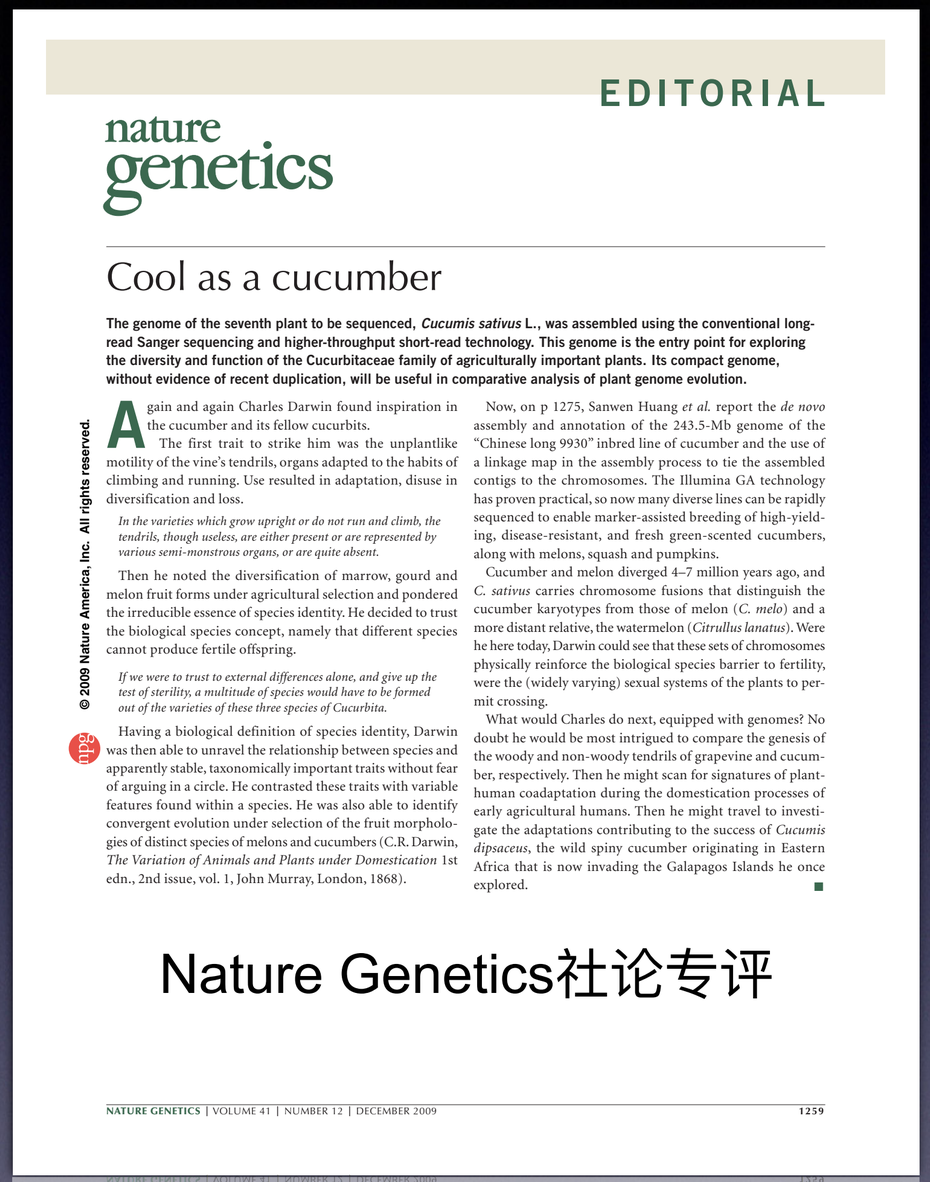
基因组是一个物种

所有遗传信息的总和

是生命体最基本的存在

第一个蔬菜作物——黄瓜基因组





EDITORIAL

nature genetics

Cool as a cucumber

The genome of the seventh plant to be sequenced, *Cucumis sativus* L., was assembled using the conventional long-read Sanger sequencing and higher-throughput short-read technology. This genome is the entry point for exploring the diversity and function of the Cucurbitaceae family of agriculturally important plants. Its compact genome, without evidence of recent duplication, will be useful in comparative analysis of plant genome evolution.

Again and again Charles Darwin found inspiration in the cucumber and its fellow cucurbits.

The first trait to strike him was the unplantlike motility of the vine's tendrils, organs adapted to the habits of climbing and running. Use resulted in adaptation, disuse in diversification and loss.

In the varieties which grow upright or do not run and climb, the tendrils, though useless, are either present or are represented by various semi-monstrous organs, or are quite absent.

Then he noted the diversification of marrow, gourd and melon fruit forms under agricultural selection and pondered the irreducible essence of species identity. He decided to trust the biological species concept, namely that different species cannot produce fertile offspring.

If we were to trust to external differences alone, and give up the test of sterility, a multitude of species would have to be formed out of the varieties of these three species of Cucurbita.

Having a biological definition of species identity, Darwin was then able to unravel the relationship between species and apparently stable, taxonomically important traits without fear of arguing in a circle. He contrasted these traits with variable features found within a species. He was also able to identify convergent evolution under selection of the fruit morphologies of distinct species of melons and cucumbers (C.R. Darwin, *The Variation of Animals and Plants under Domestication* 1st edn., 2nd issue, vol. 1, John Murray, London, 1868).

Now, on p 1275, Sanwen Huang *et al.* report the *de novo* assembly and annotation of the 243.5-Mb genome of the "Chinese long 9930" inbred line of cucumber and the use of a linkage map in the assembly process to tie the assembled contigs to the chromosomes. The Illumina GA technology has proven practical, so now many diverse lines can be rapidly sequenced to enable marker-assisted breeding of high-yielding, disease-resistant, and fresh green-scented cucumbers, along with melons, squash and pumpkins.

Cucumber and melon diverged 4–7 million years ago, and *C. sativus* carries chromosome fusions that distinguish the cucumber karyotypes from those of melon (*C. melo*) and a more distant relative, the watermelon (*Citrullus lanatus*). Were he here today, Darwin could see that these sets of chromosomes physically reinforce the biological species barrier to fertility, were the (widely varying) sexual systems of the plants to permit crossing.

What would Charles do next, equipped with genomes? No doubt he would be most intrigued to compare the genesis of the woody and non-woody tendrils of grapevine and cucumber, respectively. Then he might scan for signatures of plant-human coadaptation during the domestication processes of early agricultural humans. Then he might travel to investigate the adaptations contributing to the success of *Cucumis dipsaceus*, the wild spiny cucumber originating in Eastern Africa that is now invading the Galapagos Islands he once explored. ■

Nature Genetics社论专评

黄瓜基因组打开了探究瓜类作物多样性和功能基因的大门，它没有近期的全基因组复制，有利于研究植物基因组进化。研究发现黄瓜染色体是多次融合产生的，形成了与甜瓜和西瓜明显不同的染色体核型。如果达尔文得知这个结果，他将明白染色体核型差异强化了相邻物种之间的生殖隔离。

4 个已发布黄瓜参考基因组

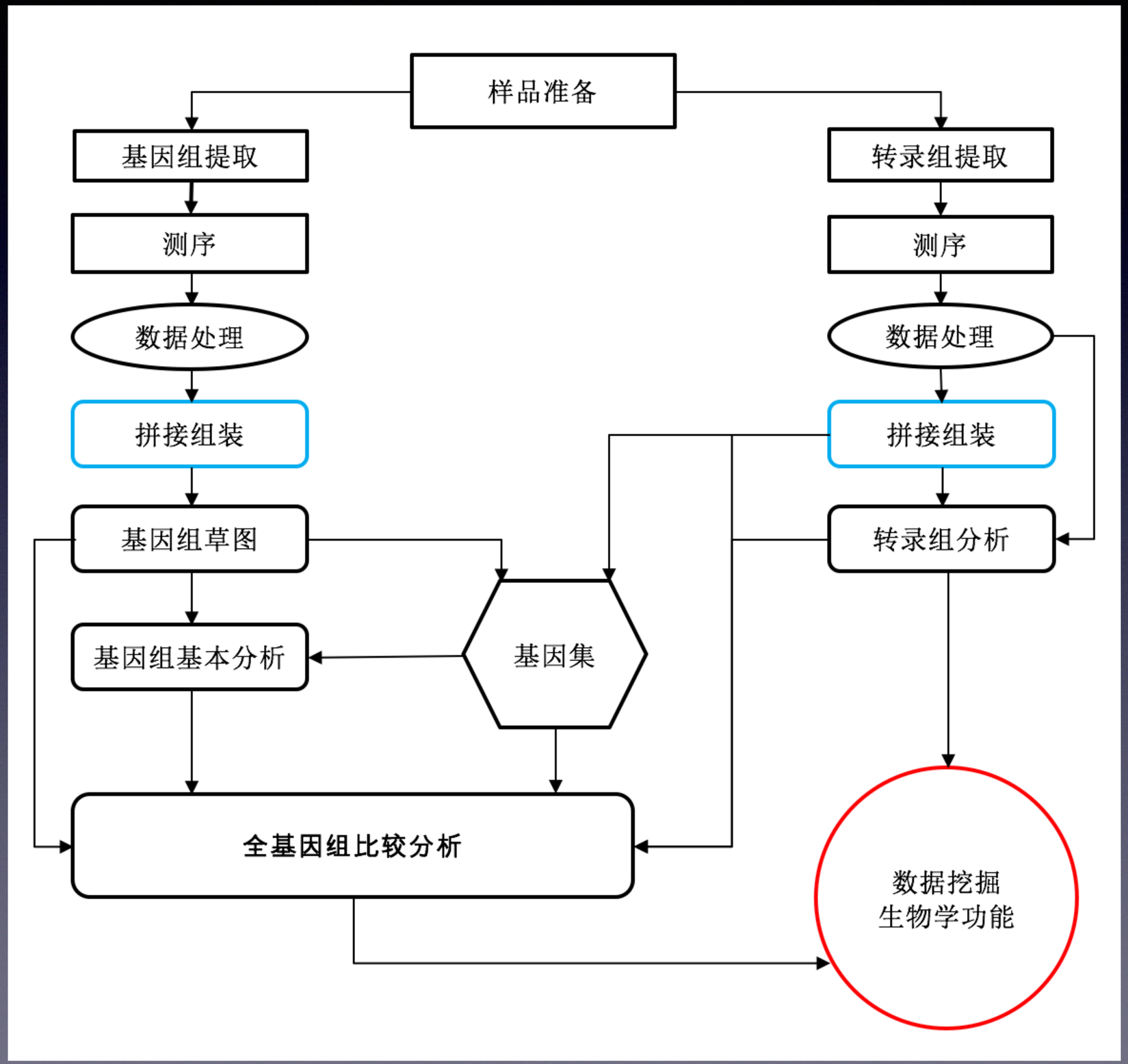
	9930 V2.0	Gy14	B10	PI 183967
亚群	East-Asian	Eurasian	Eurasian	Indian
总长 (Mb)	196.5	203.1	224.0	204.8
Contig N50 (kb)	37.9	/	27.1	119.1
Scaffold N50 (Mb)	1.4	0.9	2.3	4.2

这 4 个参考基因组的低完整度和连续性阻碍了他们在瓜类作物比较基因组学中的应用

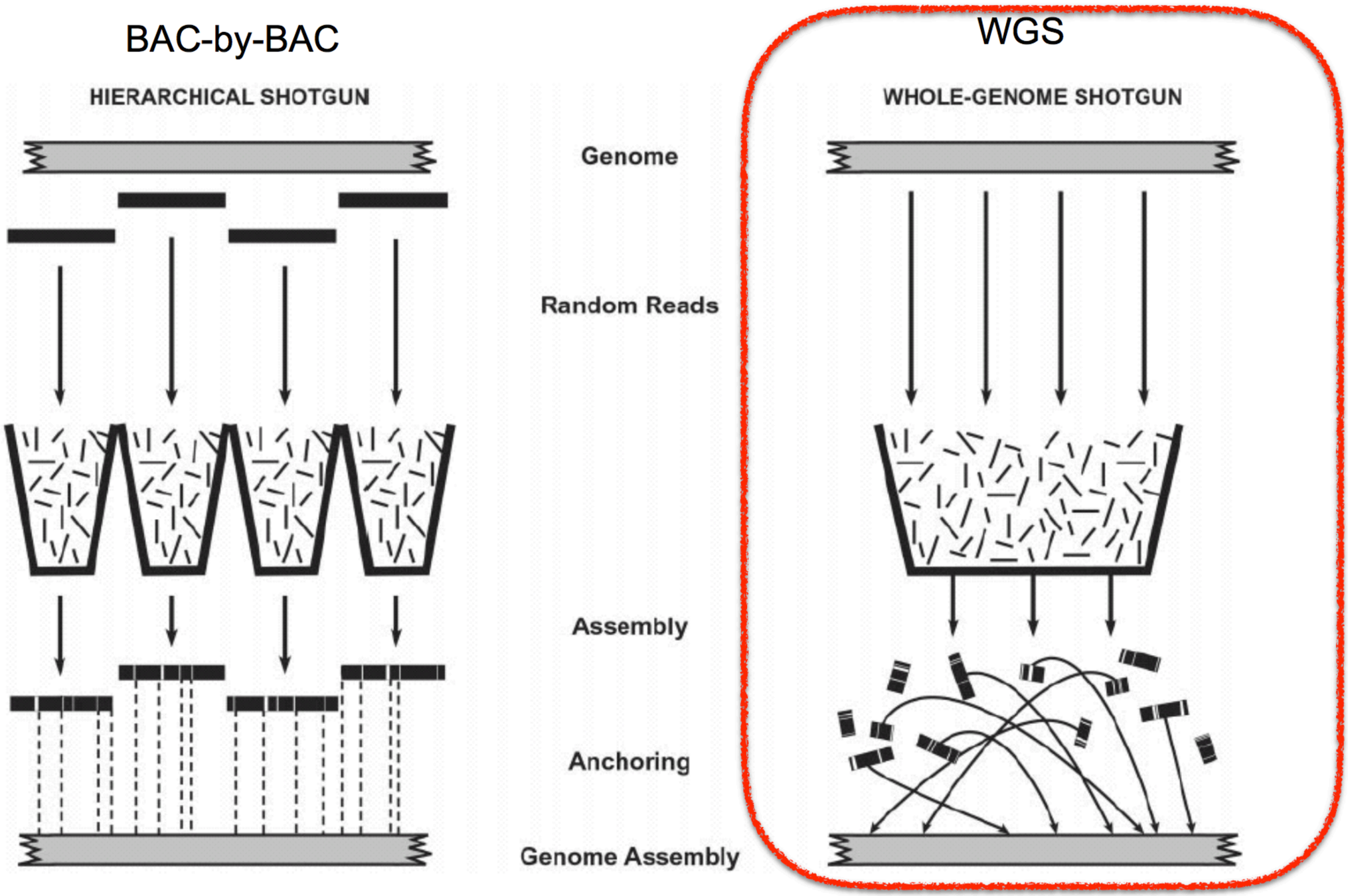
提纲

- 研究背景
- 从头测序和序列装配
- 基因组结构注释，基因功能预测
- 全文结论

基因组从头测序技术路线

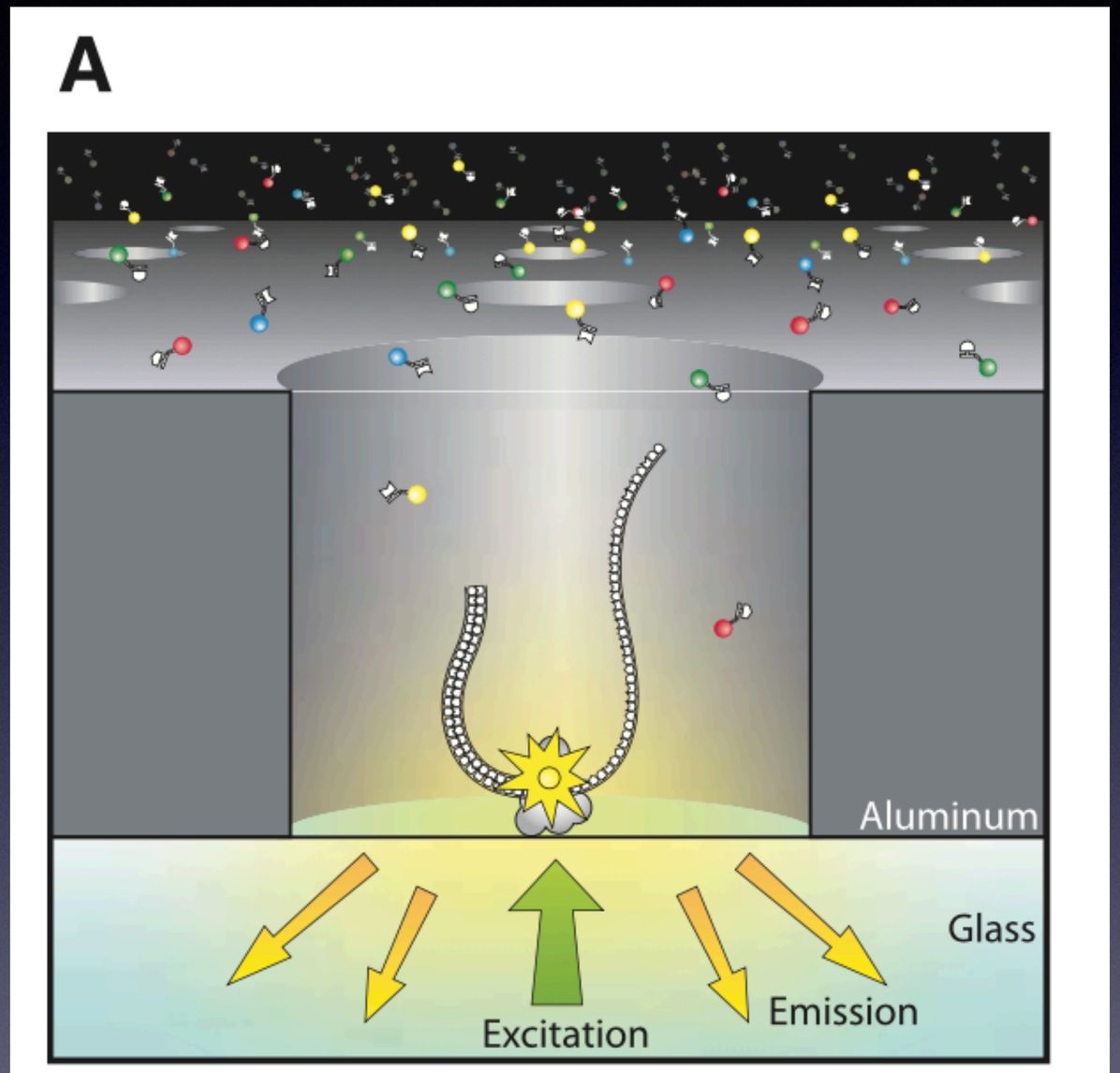


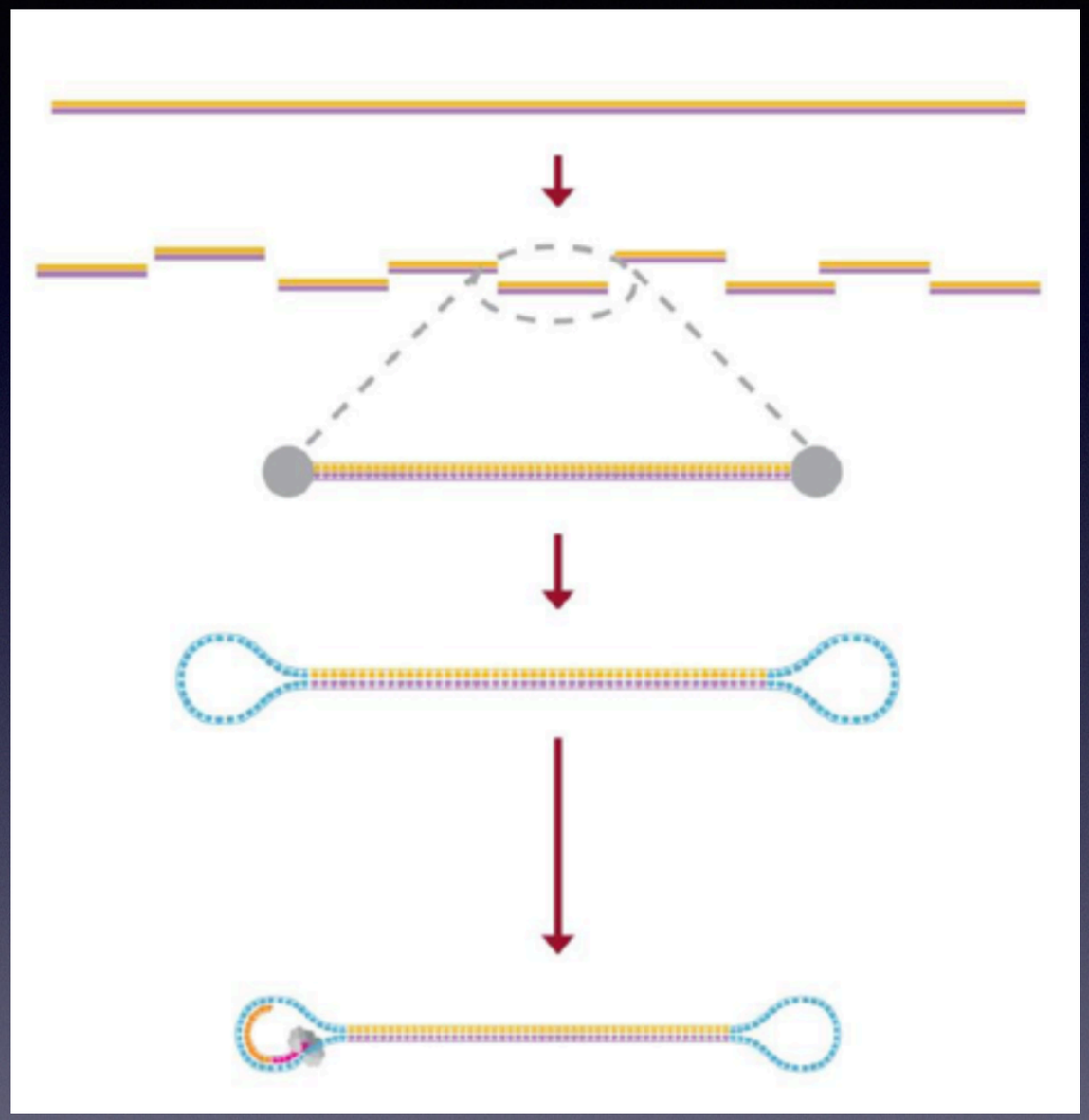
从头测序策略



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

基因组测序技术

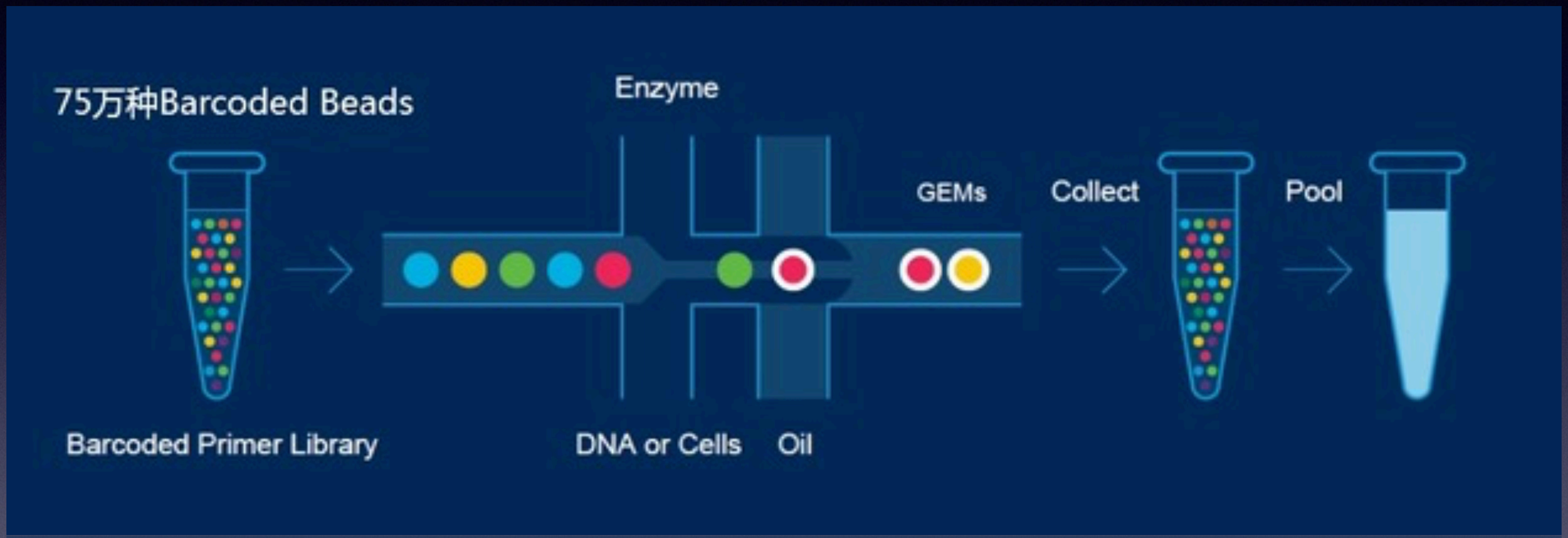




PacBio 单分子实时测序 (三代测序)

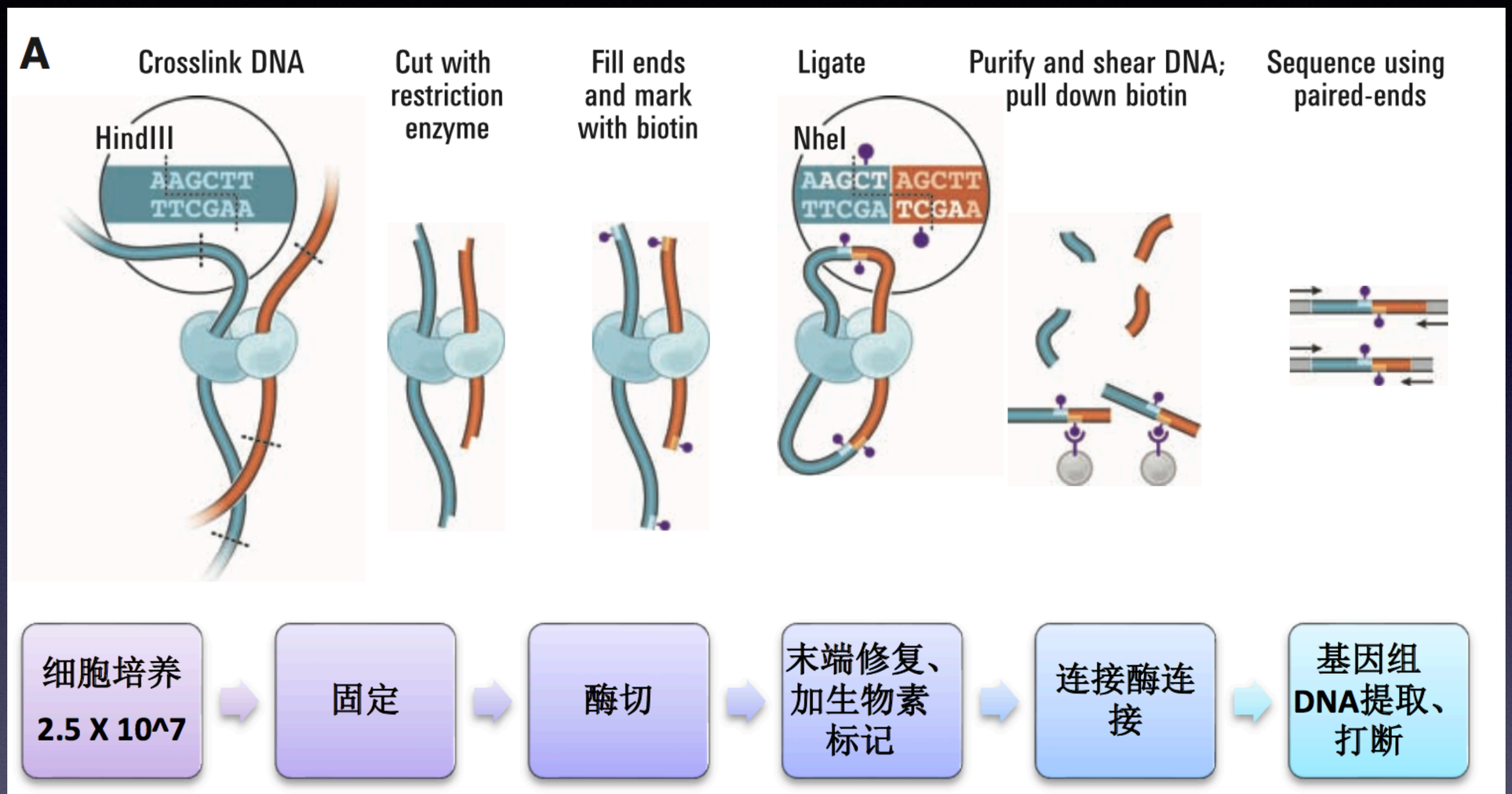
John et al., 2009, Science

基因组测序技术



10X genomics linked-reads 测序

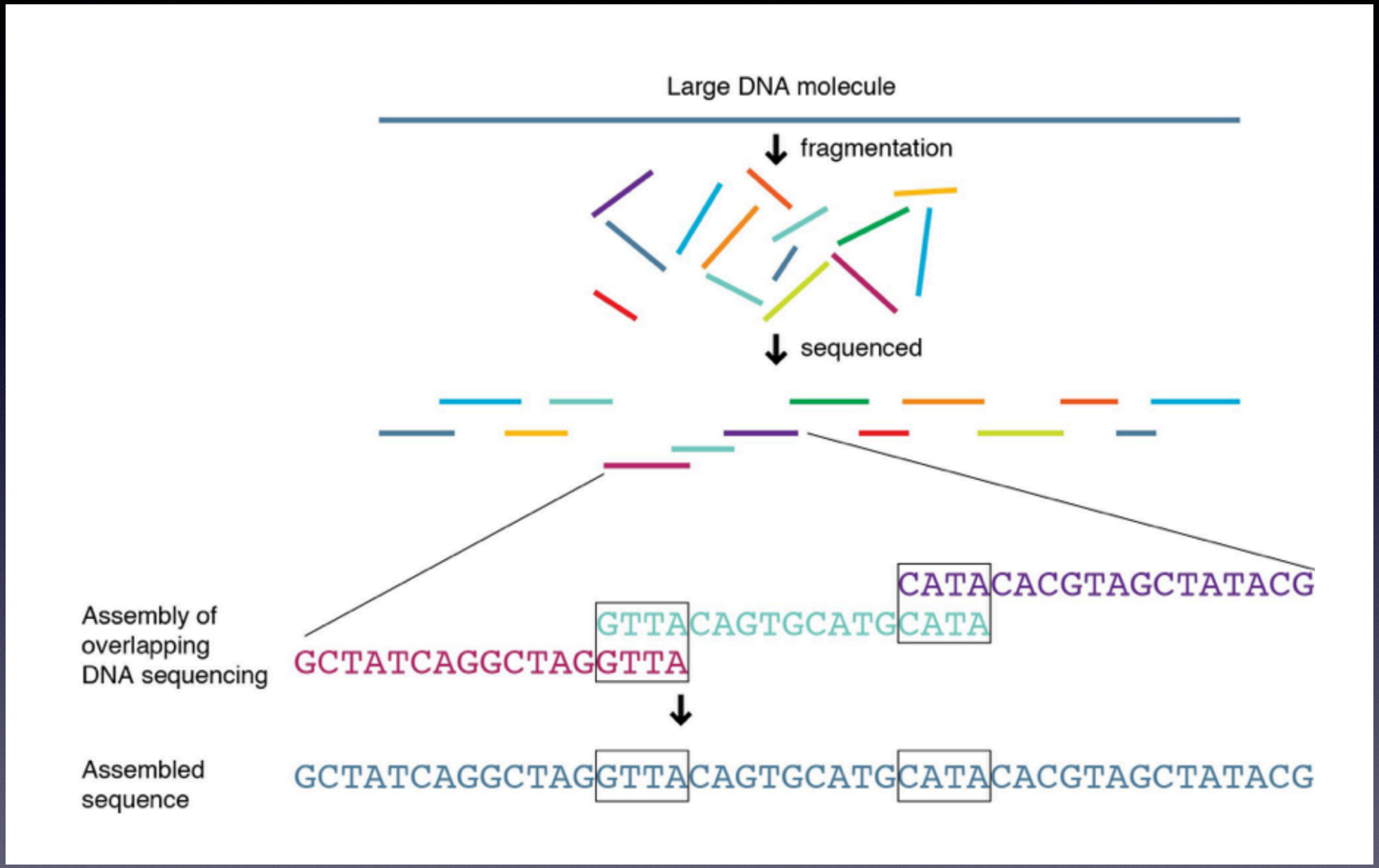
基因组测序技术



高通量染色质构象捕捉测序 (Hi-C)

Erez et al., 2009, Science

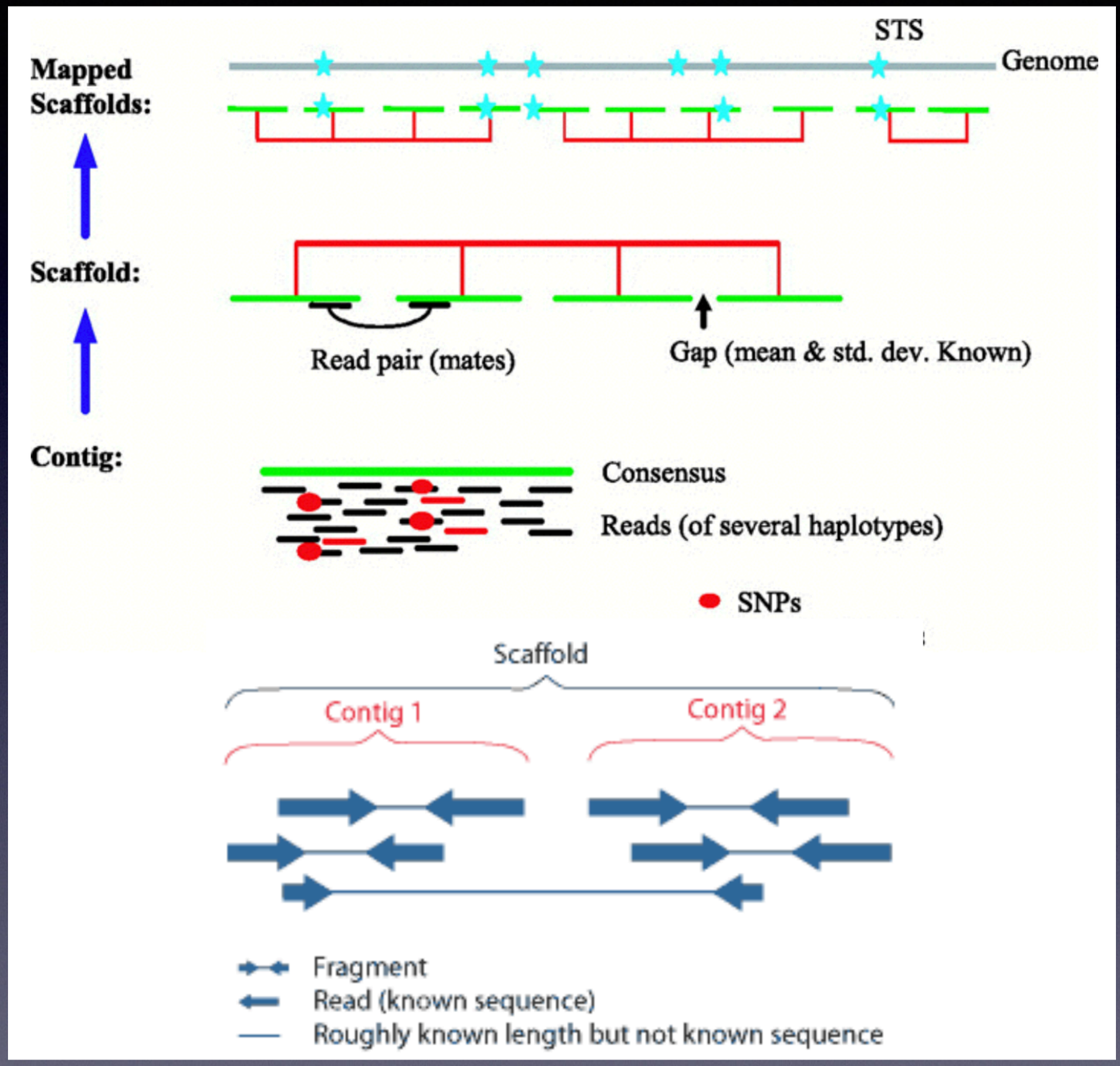
基因组拼接原理



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, 利用遗传图谱, Hi-C等信息锚定成染色体

黄瓜基因组从头装配流程

三代测序

10X 测序

Hi-C 测序

二代测序

Pilon

Canu

3d-dna

校正
contigs

arcs & LINKS

10XG
scaffolds

Hi-C
互作矩阵

4 张
遗传图谱

ALLMAPs

Super-
scaffolds

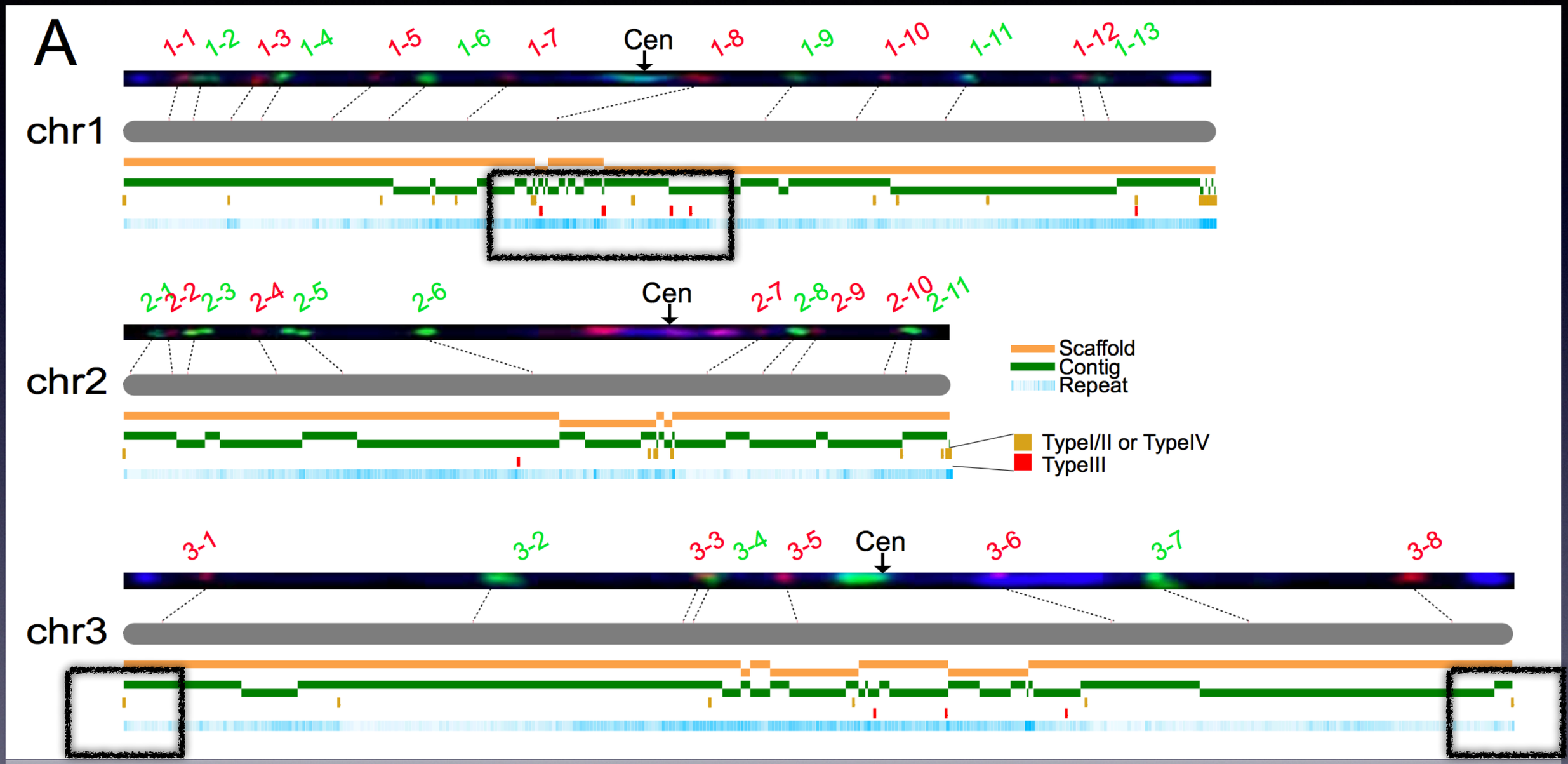
参考
基因组

“9930”
高代自交系

从头装配结果

	Contig	Scaffold	Super-scaffold
总长 (Mb)	226.1	226.1	226.6
Gap 长度 (kb)	0	1.0	52.4
Contig 个数	992	-	-
Contig N50 (Mb)	2.4	-	-
Scaffold 个数	-	892	369
Scaffold N50 (Mb)	-	7.5	31.7
Largest Scaffold (Mb)	-	18.0	40.7

细胞遗传学图谱整合结果



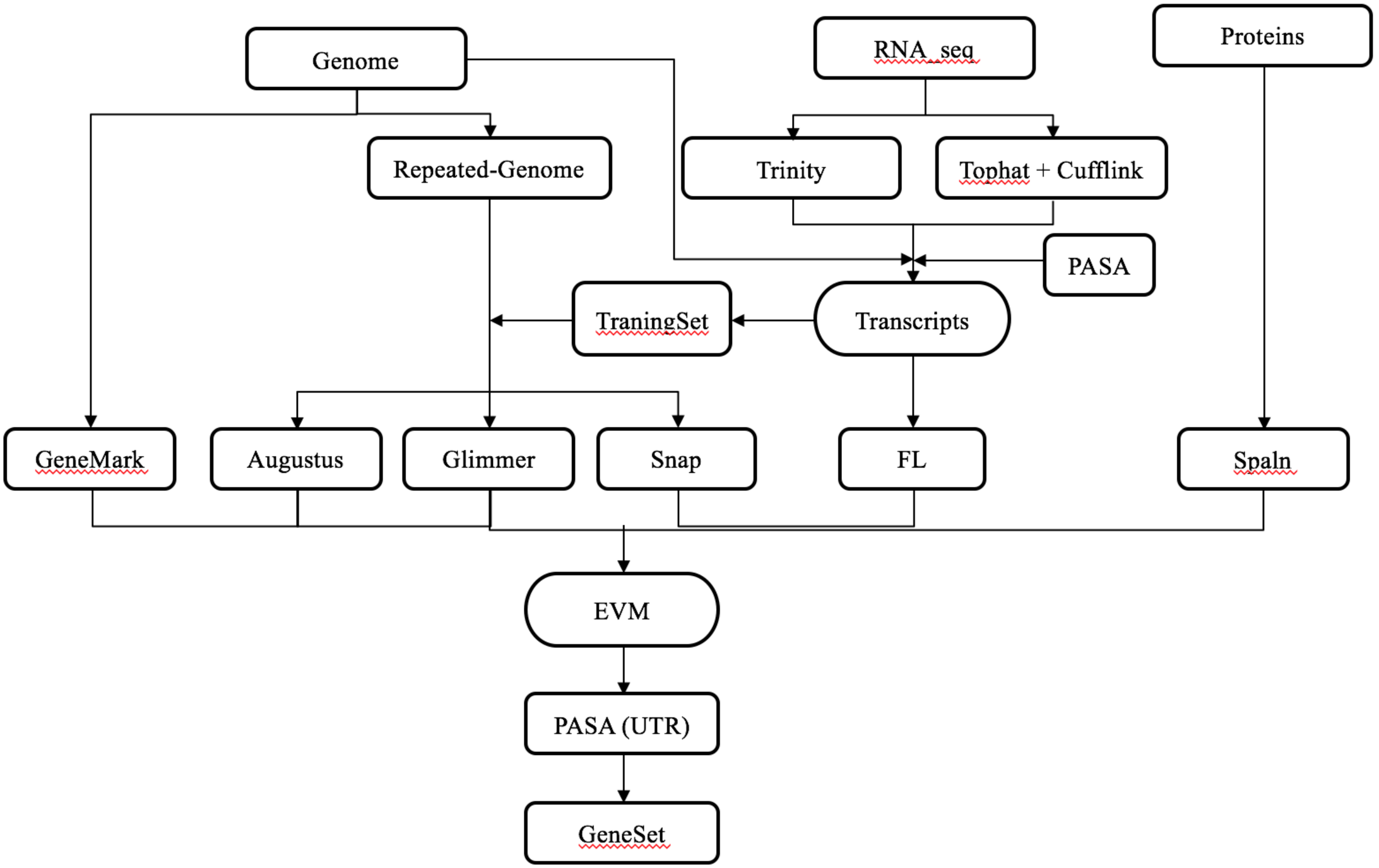
着丝粒和端粒这些高度重复区域的边缘

已经被拼接出来

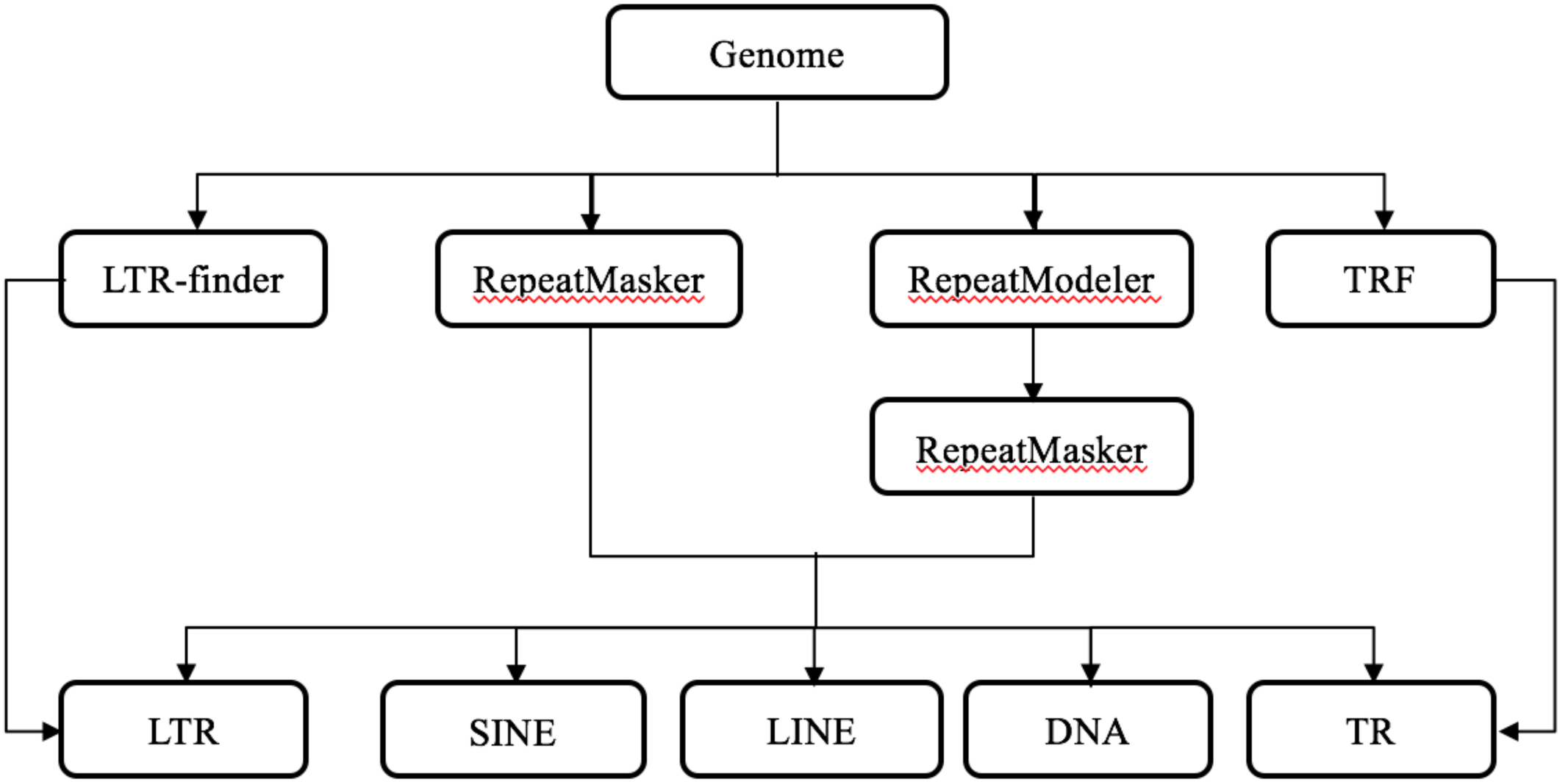
提纲

- 研究背景
- 从头测序和序列装配
- **基因组结构注释，基因功能预测**
- 全文结论

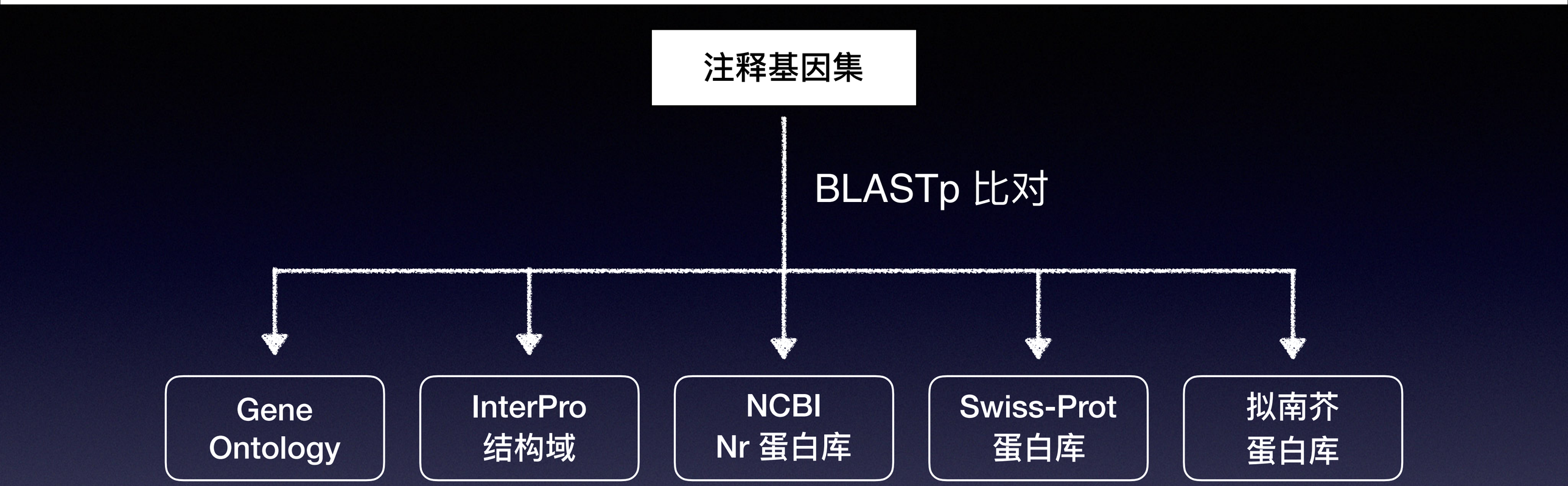
蛋白编码基因注释流程


Genome
RNA_seq
Proteins
Repeated-Genome
Trinity
Tophat + Cufflink
PASA
TraningSet
Transcripts
GeneMark
Augustus
Glimmer
Snap
FL
Spaln
EVM
PASA (UTR)
GeneSet

基因组重复注释流程


Genome
LTR-finder
RepeatMasker
RepeatModeler
TRF
RepeatMasker
LTR
SINE
LINE
DNA
TR

基因功能预测流程



#Chr	start	end	Gene_ID	GO	IPR	NR	Swissprot	Ara
chr3	18833786	18834544	Csa37_3G022450	-	-	gi|778665109|ref|XP_011648489.1|	PREDICTED: carboxypeptidase Y-like [Cucumis sati	
chr1	14575859	14576012	Csa37_1G025870	-	-	-	-	-
chr1	14580001	14580325	Csa37_1G025880	-	-	-	-	-
chr1	14584714	14585015	Csa37_1G025890	-	IPR036312	Bifunctional inhibitor/plant lipid transfer protein/seed storage helical domain		
chr1	14586255	14586487	Csa37_1G025900	-	-	-	-	-
chr1	14605680	14605869	Csa37_1G025910	-	-	-	-	-
chr1	14638683	14638999	Csa37_1G025920	-	-	gi|700210094|gb|KGN65190.1|	hypothetical protein Csa_1G263450 [Cucumis sativus] -	
chr1	14664088	14664342	Csa37_1G025930	-	-	gi|700210000|gb|KGN65096.1|	hypothetical protein Csa_1G212820 [Cucumis sativus] -	
chr1	14665887	14667219	Csa37_1G025940	GO:0016491 oxidoreductase activity;GO:0055114 oxidation-reduction process;		IPR005123 Oxoglut		
chr1	14672883	14676738	Csa37_1G025950	-	IPR026992 Non-haem dioxygenase N-terminal domain;IPR027443 Isopenicillin N synthase-like;			
chr1	14683443	14684051	Csa37_1G025960	GO:0016491 oxidoreductase activity;GO:0055114 oxidation-reduction process;		IPR005123 Oxoglut		
chr1	14687389	14687576	Csa37_1G025970	-	-	-	-	-
chr1	14691640	14693852	Csa37_1G025980	GO:0005515 protein binding;		IPR000270 PB1 domain;IPR011990 Tetratricopeptide-like helical dom		
chr1	14702023	14706408	Csa37_1G025990	-	IPR009011 Mannose-6-phosphate receptor binding domain superfamily;IPR012913 Protein OS9-l			
chr1	14711096	14711896	Csa37_1G026000	-	IPR013761 Sterile alpha motif/pointed domain superfamily;		gi|449458009|ref|XP_00414	
chr1	14718320	14719681	Csa37_1G026010	-	-	gi|449457949|ref|XP_004146710.1|	PREDICTED: uncharacterized serine-rich protein C	
chr1	14728399	14728901	Csa37_1G026020	-	-	gi|778660170|ref|XP_011655735.1|	PREDICTED: uncharacterized protein LOC101204798	
chr1	14731558	14741796	Csa37_1G026030	-	-	gi|449457947|ref|XP_004146709.1|	PREDICTED: uncharacterized protein LOC101216821	
chr1	14748671	14753072	Csa37_1G026040	GO:0004499 N,N-dimethylaniline monooxygenase activity;GO:0050660 flavin adenine dinucleotide bind				

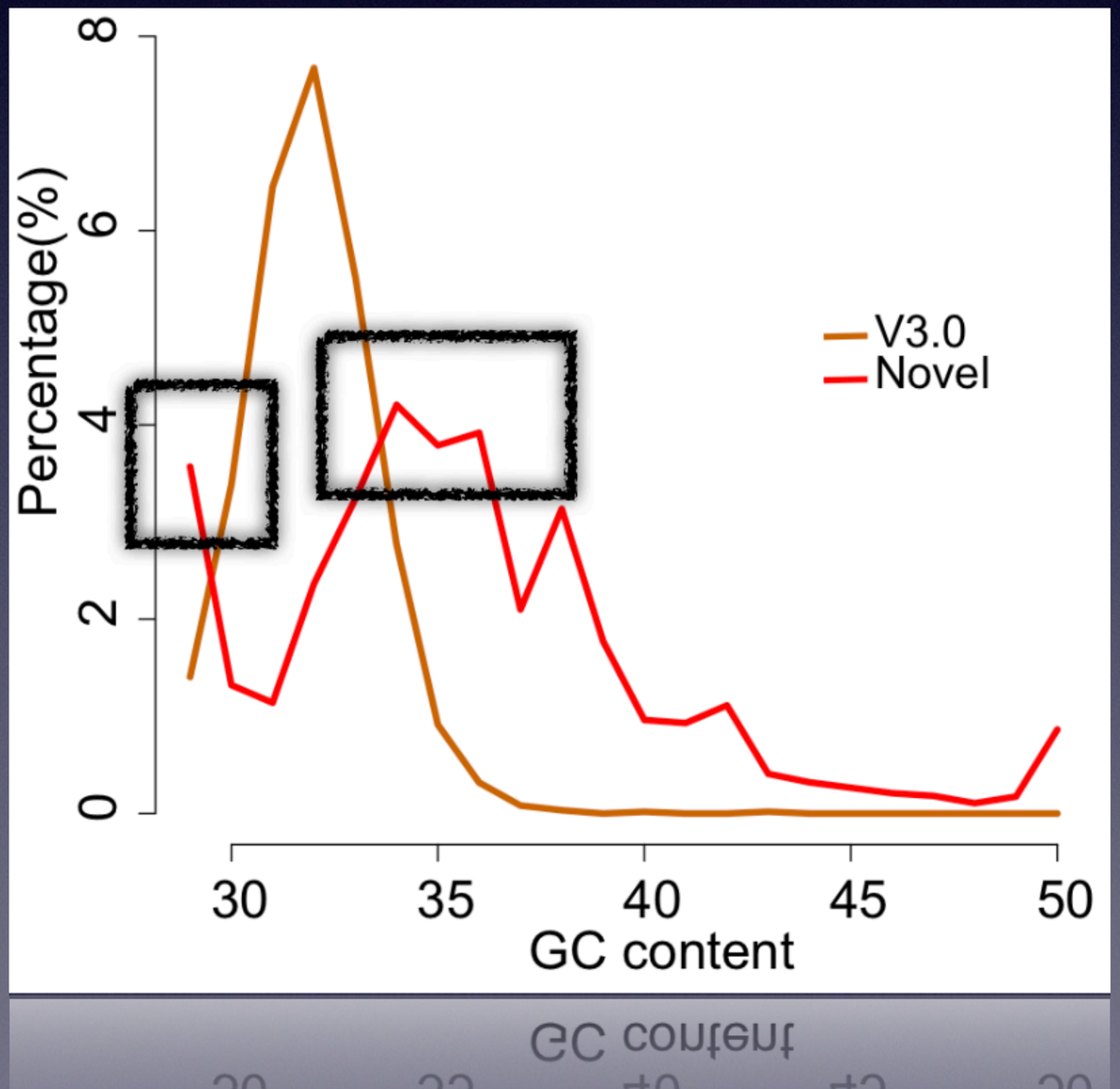
基因组注释结果

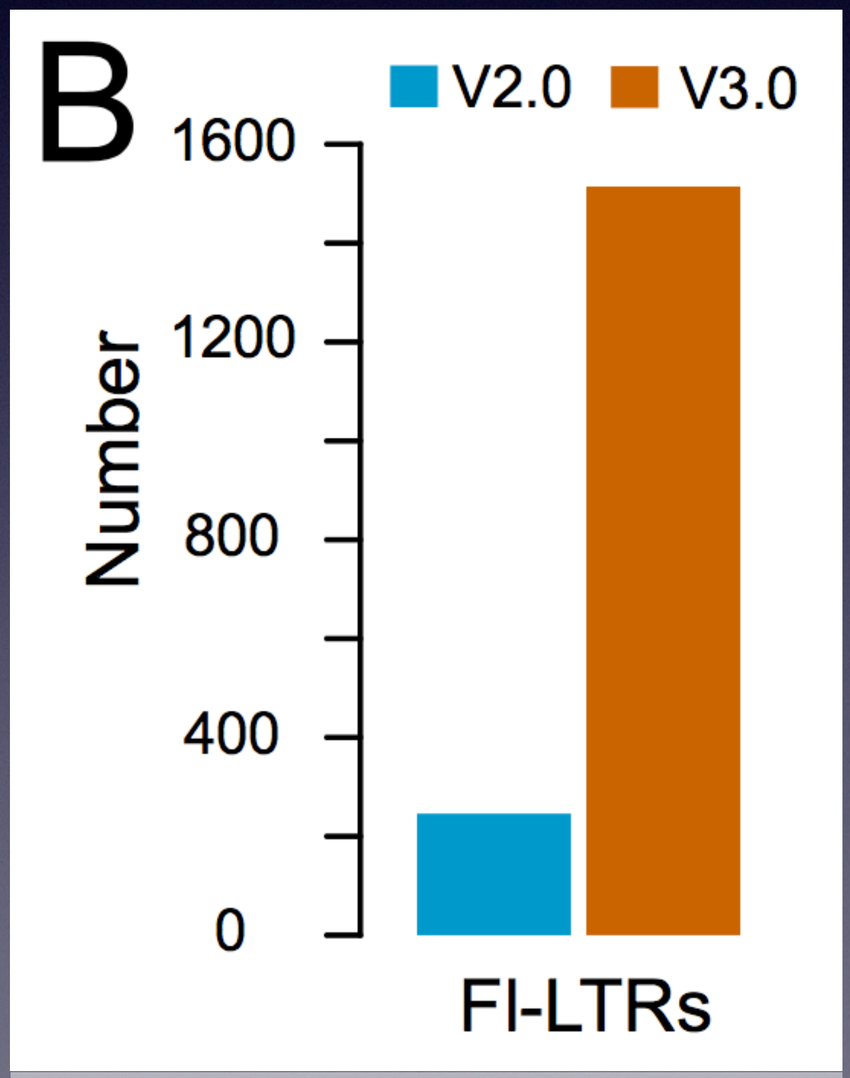
24,032
蛋白编码基因

1,112
新拼接基因

93.3Mb
重复序列

40.8%
占全基因组





提纲

- 研究背景
- 从头测序和序列装配
- 基因组结构注释，基因功能预测
- 全文结论

全文结论

- 黄瓜新版参考基因组相比上一个版本，额外拼接出了31.7 Mb序列，序列的连续性也有了显著提高：Contig N50 长度提高了 63.3 倍，Scaffold N50 则是 22.7 倍；
- 新版基因组中更新了相当多一部分全新拼接的重复序列和基因；
- 该高质量参考基因组不仅能够为黄瓜的遗传学研究提供宝贵资源，而且也有利于植物比较基因组学研究。

Acknowledgement

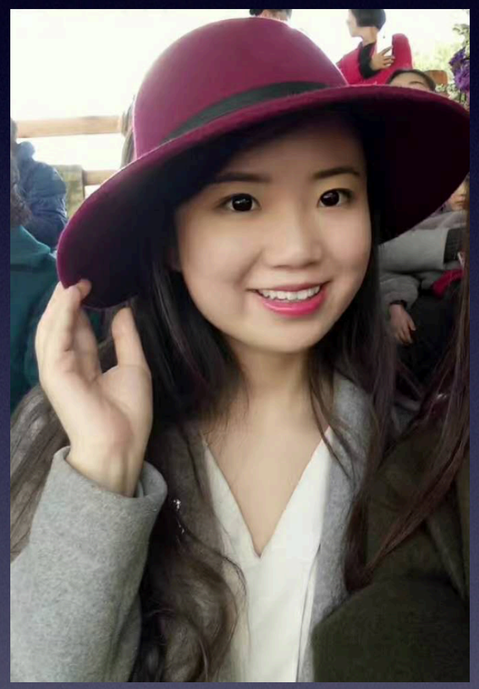
Prof. Zhonghua Zhang

Prof. Sanwen Huang

Acknowledgement

Qing Li

Master candidate

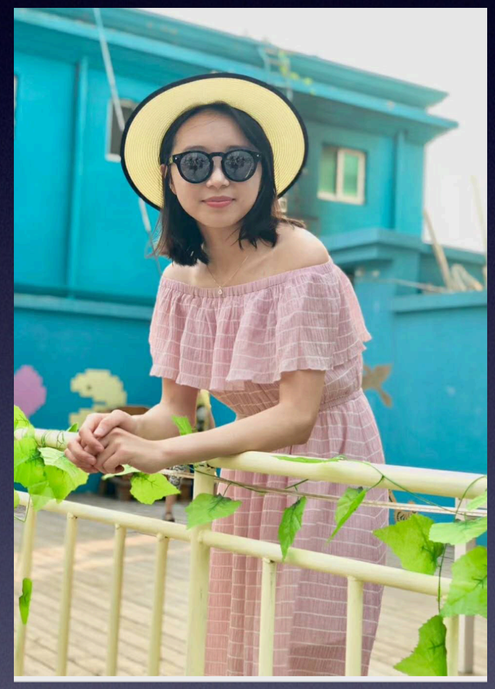


Wu Huang

Master candidate

Yuanchao Xu

Ph.D candidate



Qian Zhou

Ph.D candidate