Vol 449|18 October 2007|doi:10.1038/nature06258 nature

ARTICLES

A second generation human haplotype map of over 3.1 million SNPs

The International HapMap Consortium*

72个单位， 数百人

为什么要选这篇文章？

- 影响巨大
- 关系密切

内 容

- **Introduction**
- **Construction of the Phase II HapMap**
- **The use of the Phase II HapMap in association studies**
- **New insights into linkage disequilibrium structure**
- **The distribution of recombination**
- **Natural selection**
- 作物的单倍型作图

Haplotype

The term is a contraction of *haploid genotype*, is a symbolic representation of a specific combination of linked alleles in a cluster of related genes

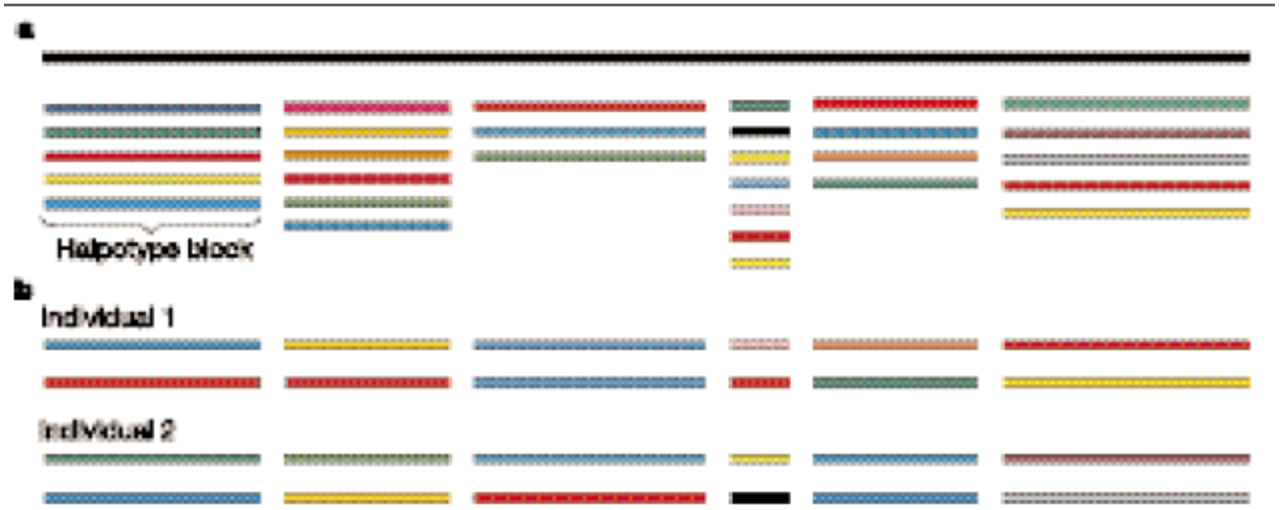
Dictionary of Genetics

单倍型：“单倍体基因型”的缩写，意思为一个基因簇中连锁等位基因的特定组合，也称单元型、单体型

A T G

C T A

Haplotype block (单倍型区段)



- Each human chromosome is made up of regions, called ‘haplotype blocks’, which are stretches of DNA sequence where three to seven variants account for most of the variation found among humans
- On average 5.5 haplotypes for every block
- Size： an average of around 10Kb
- Not all of the human genome may have a clearly definable haplotype-
- Block structure

Svante Pääbo, Nature, 2003

Haplotype map（单倍型作图）

The catalogue of haplotypes for every block makes up the ‘haplotype map’

Svante Pääbo, Nature, 2003

确定单倍型在基因组中（染色体上）的相对位置与距离

The Phase I of International HapMap Project

Launched in 2002

The objective was to genotype at least one common SNP every 5 kb across the euchromatic portion of the genome in 270 individuals from four geographically diverse populations

YRI： Yoruba in Ibadan, Nigeria
CEU： northern and western European ancestry living in Utah from the Centre d'Etude du Polymorphisme Humain (CEPH)
CHB： Han Chinese individuals
JPT： Japanese individuals in Tokyo

The Phase Ⅰ of International HapMap Project

Approximately 1.3 million SNPs were genotyped in Phase I of the project, and a description of this resource was published in 2005

Phase II HapMap

- Characterizes over 3.1 million human SNPs genotyped in 270 individuals from four geographically diverse populations
- Includes 25–35% of common SNP variation
- The map is estimated to capture untyped common variation with an average maximum r^2 of between 0.9 and 0.96 depending on population.

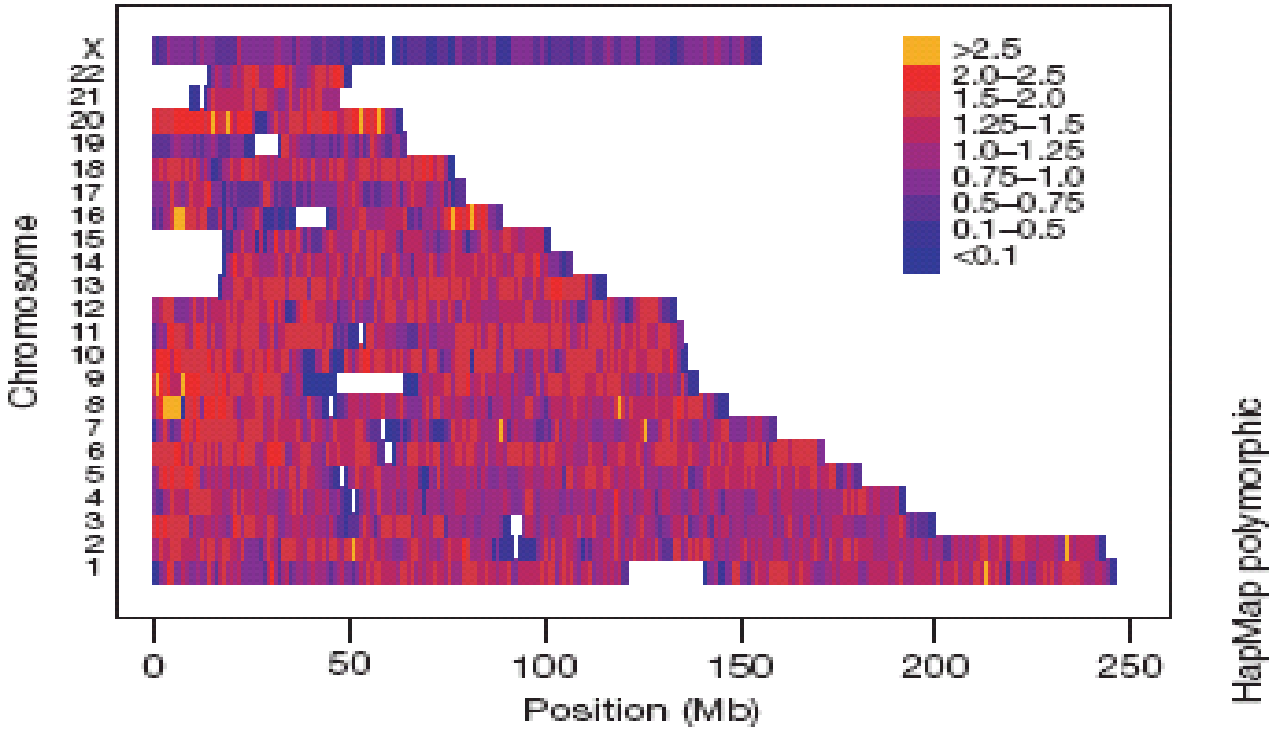
内 容

- Introduction
- **Construction of the Phase II HapMap**
- The use of the Phase II HapMap in association studies
- New insights into linkage disequilibrium structure
- The distribution of recombination
- Natural selection
- 作物的单倍型作图

Phase II of the HapMap Project

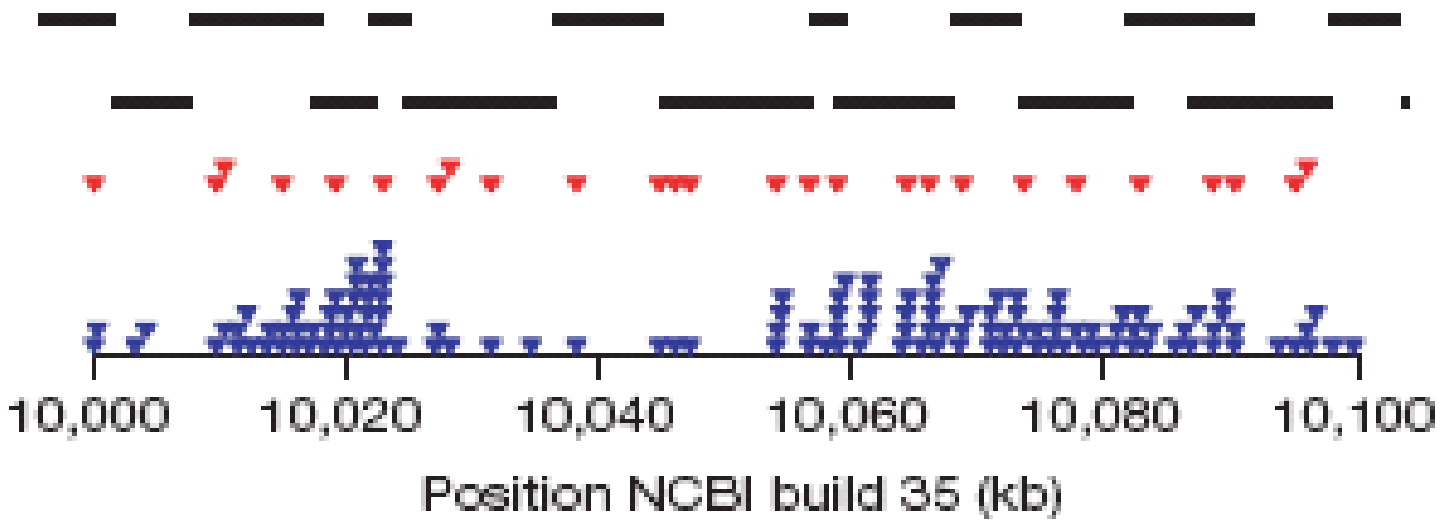
- A further 2.1 million SNPs were successfully genotyped
- HapMap：SNP density of approximately one per kilobase
- Contain approximately 25–35% of all the 9–10 million common SNPs

SNP density in the Phase II HapMap



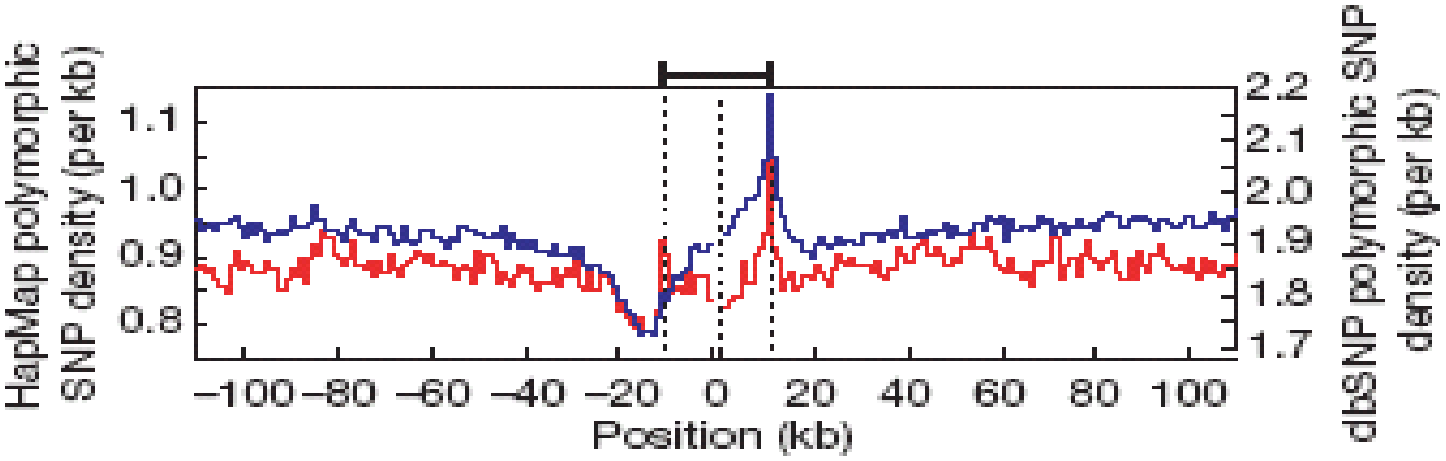
SNP density across the genome. Colours indicate the number of polymorphic SNPs per kb in the consensus data set. Gaps in the assembly are shown as white.

SNP density in the Phase II HapMap



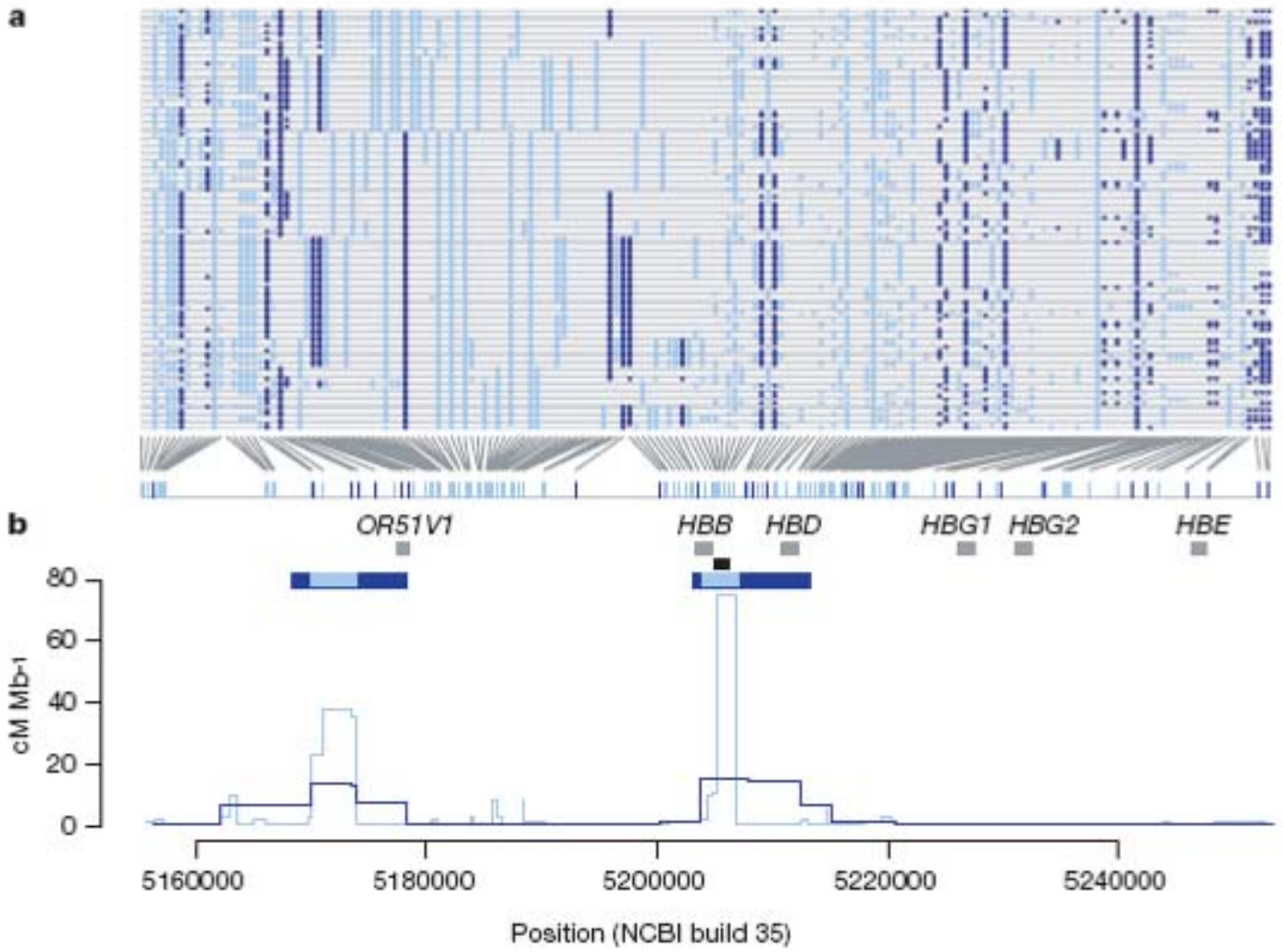
Example of the fine-scale structure of SNP density for a 100-kb region on chromosome 17 showing Perlegen amplicons (black bars). Polymorphic Phase I SNPs in the consensus data set (red triangles) and polymorphic Phase II SNPs in the consensus data set (blue triangles).

SNP density in the Phase II HapMap



The distribution of polymorphic SNPs in the consensus Phase II HapMap data (blue line and left-hand axis) around coding regions.

Haplotype structure and recombination rate estimates from the Phase II HapMap



a. **Haplotypes from YRI in a 100kb region around the ß-globin gene.**
b. **Recombination rates and the location of hotspots.**

内 容

- Introduction
- Construction of the Phase II HapMap
- **The use of the Phase II HapMap in association studies**
- New insights into linkage disequilibrium structure
- The distribution of recombination
- Natural selection
- 作物的单倍型作图

Phase II HapMap and genome-wide association studies

Using Phase II data, we estimated the coverage of several available products on which genome-wide association studies are already underway.

Similar to earlier estimates， these products typically perform well in CEU and CHB1JPT, and some also perform well in YRI

- Improved coverage of common variation
- Phase II HapMap and genome-wide association study

Table 4 | Estimated coverage of commercially available fixed marker arrays

Platform*	YRI		CEU		CHB+JPT	
	$r^2 \geq 0.8$ (%)	Mean maximum r^2	$r^2 \geq 0.8$ (%)	Mean maximum r^2	$r^2 \geq 0.8$ (%)	Mean maximum r^2
Affymetrix GeneChip 500K	46	0.66	68	0.81	67	0.80
Affymetrix SNP Array 6.0	66	0.80	82	0.90	81	0.89
Illumina HumanHap300	33	0.56	77	0.86	63	0.78
Illumina HumanHap550	55	0.73	88	0.92	83	0.89
Illumina HumanHap650Y	66	0.80	89	0.93	84	0.90
Perlegen 600K	47	0.68	92	0.94	84	0.90

*Assuming all SNPs on the product are informative and pass QC; in practice these numbers are overestimates.

Table 3 | Number of tag SNPs required to capture common (MAF ≥ 0.05) Phase II SNPs

Threshold	YRI	CEU	CHB+JPT
$r^2 \geq 0.5$	627,458	290,969	277,831
$r^2 \geq 0.8$	1,093,422	552,853	520,111
$r^2 = 1.0$	1,616,739	1,024,665	1,078,959

MA：Minor allele frequency

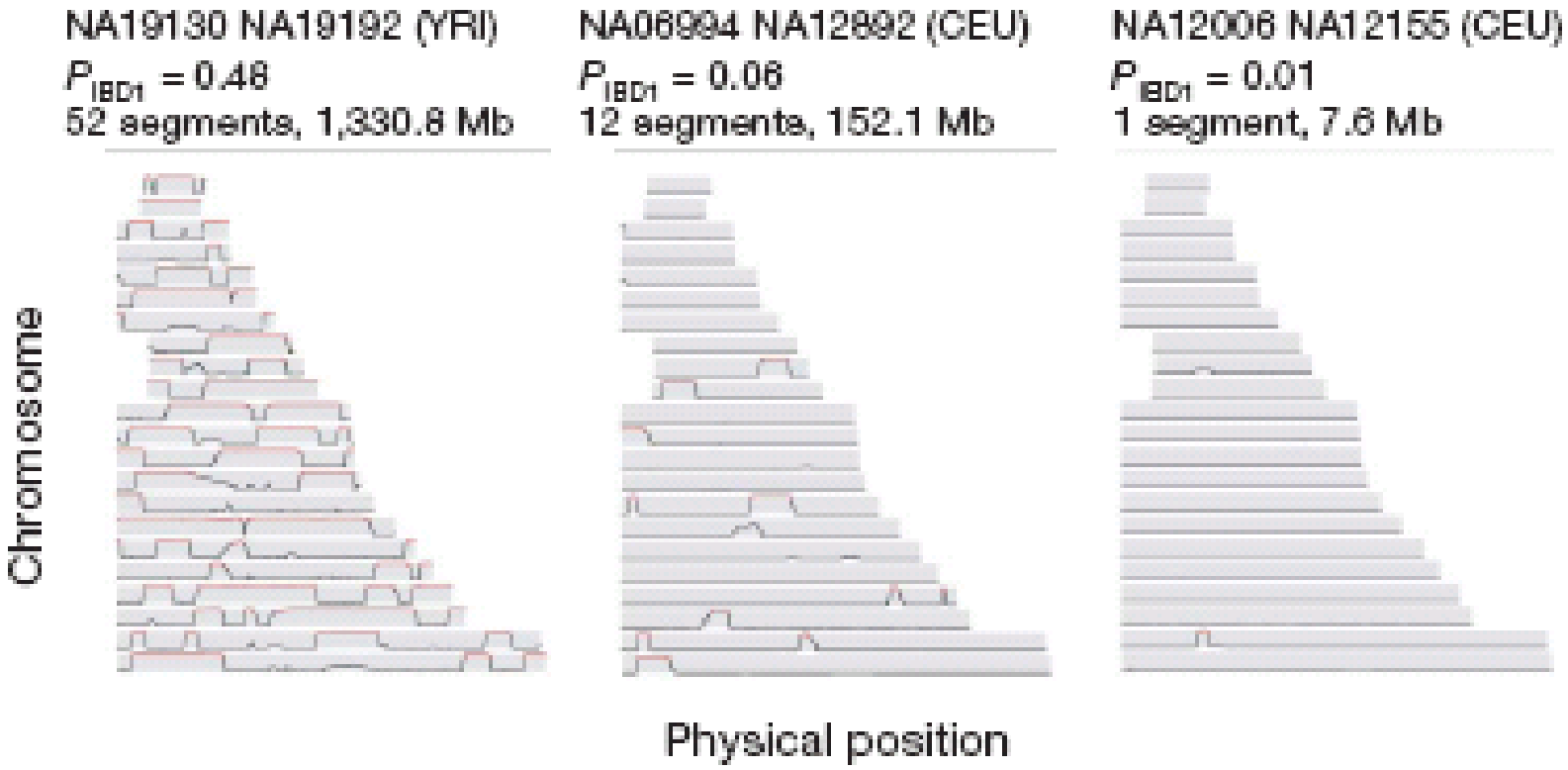
内 容

- Introduction
- Construction of the Phase II HapMap
- The use of the Phase II HapMap in association studies
- **New insights into linkage disequilibrium structure**
- The distribution of recombination
- Natural selection
- 作物的单倍型作图

New insights into linkage disequilibrium structure

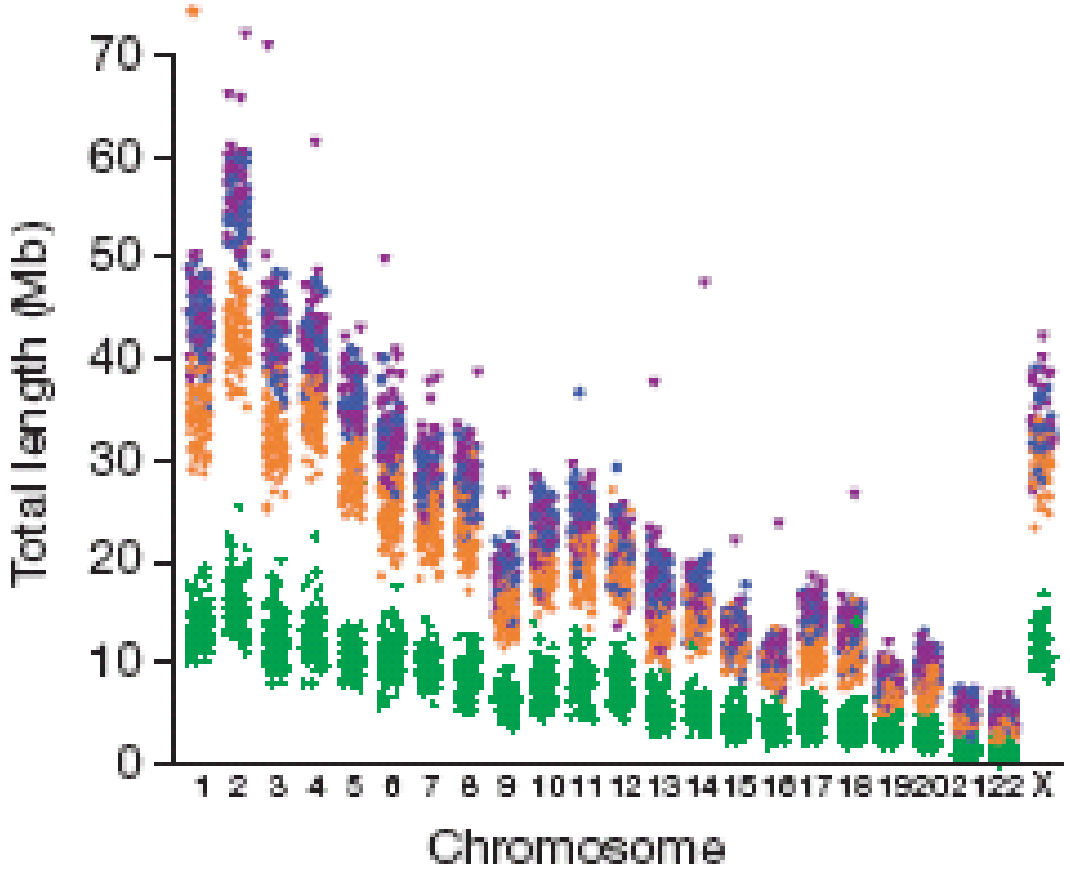
- The extent of recent common ancestry and segmental sharing
- Any two individuals from the same population share approximately 0.5% of their genome through recent IBD ,shared segments over 1-megabase (Mb) long and containing at least 50 SNPs,
- 10–30% of pairs in each analysis panel share regions of extended identity resulting from sharing a common ancestor within 10–100 generations. These regions typically span hundreds of SNPs and can extend over tens of megabases

The extent of recent co-ancestry among HapMap individuals



Three pairs of individuals with varying levels of IBD sharing illustrate the continuum between very close and very distant relatedness and its relation to segmental sharing

The extent of recent co-ancestry among HapMap individuals

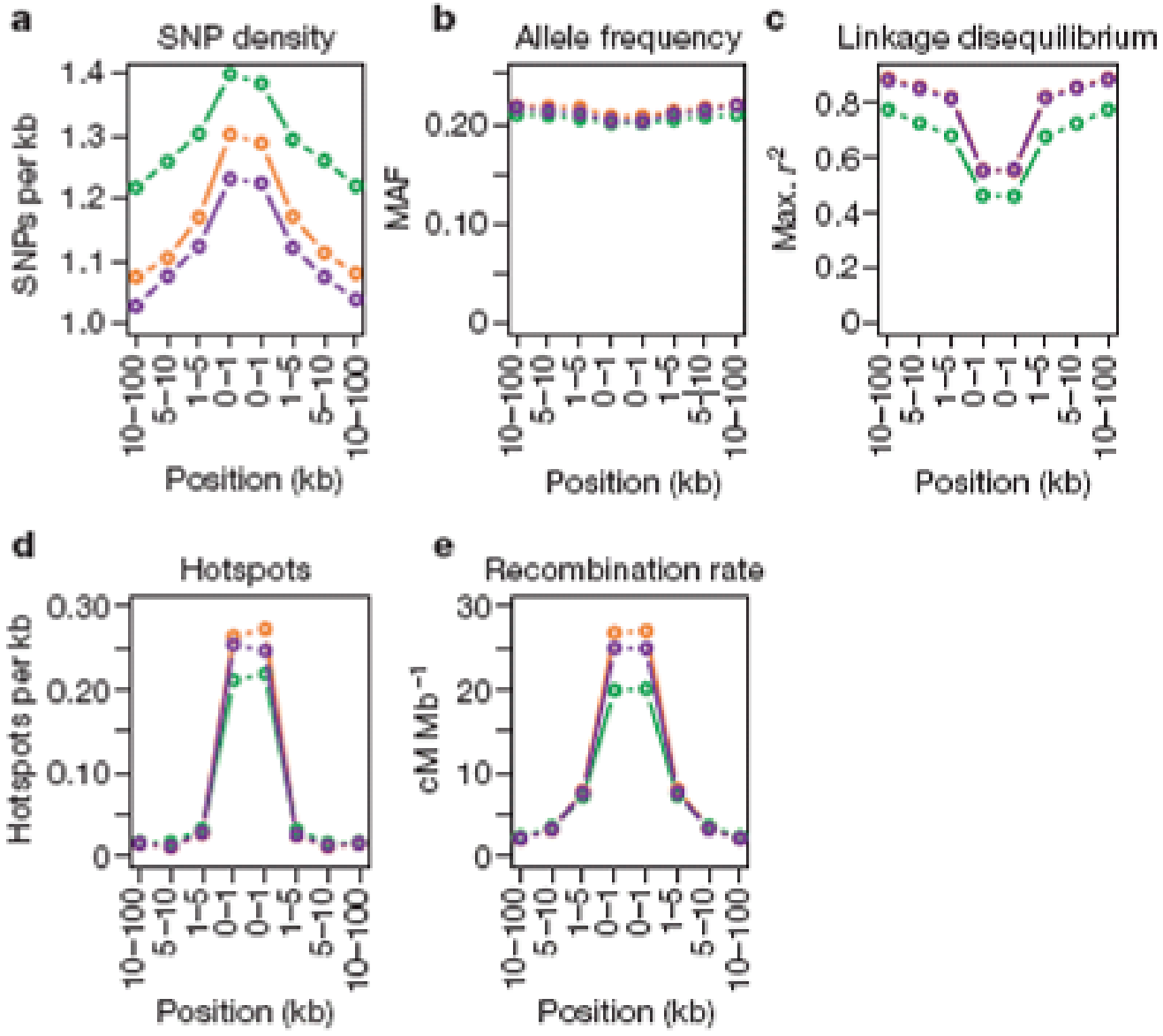


The extent of homozygosity on each chromosome for each individual in each analysis pane. YRI, green; CEU, orange; CHB blue; JPT, magenta.

The distribution and causes of untaggable SNPs

- We marked as untaggable SNPs to which no other SNP within 100 kb has an r^2 value of at least 0.2
- In Phase II, approximately 0.5–1.0% of all high-frequency SNPs are untaggable and the proportion in YRI is approximately twice as high as in the other panels. Similar proportions are observed across the ten HapMap ENCODE regions

Properties of untaggable SNPs



Properties of untaggable SNPs

- **The proximity of untaggable SNPs to the centre of hotspots suggests that they may lie within gene conversion tracts associated with the repair of double-strand breaks.**
- **Double-strand breaks are thought to resolve as crossover events only 5–25% of the time.**
- **Consequently, SNPs lying near the centre of a hotspot are liable to be included within gene conversion tracts and will experience much higher effective recombination rates than predicted from crossover rates alone.**

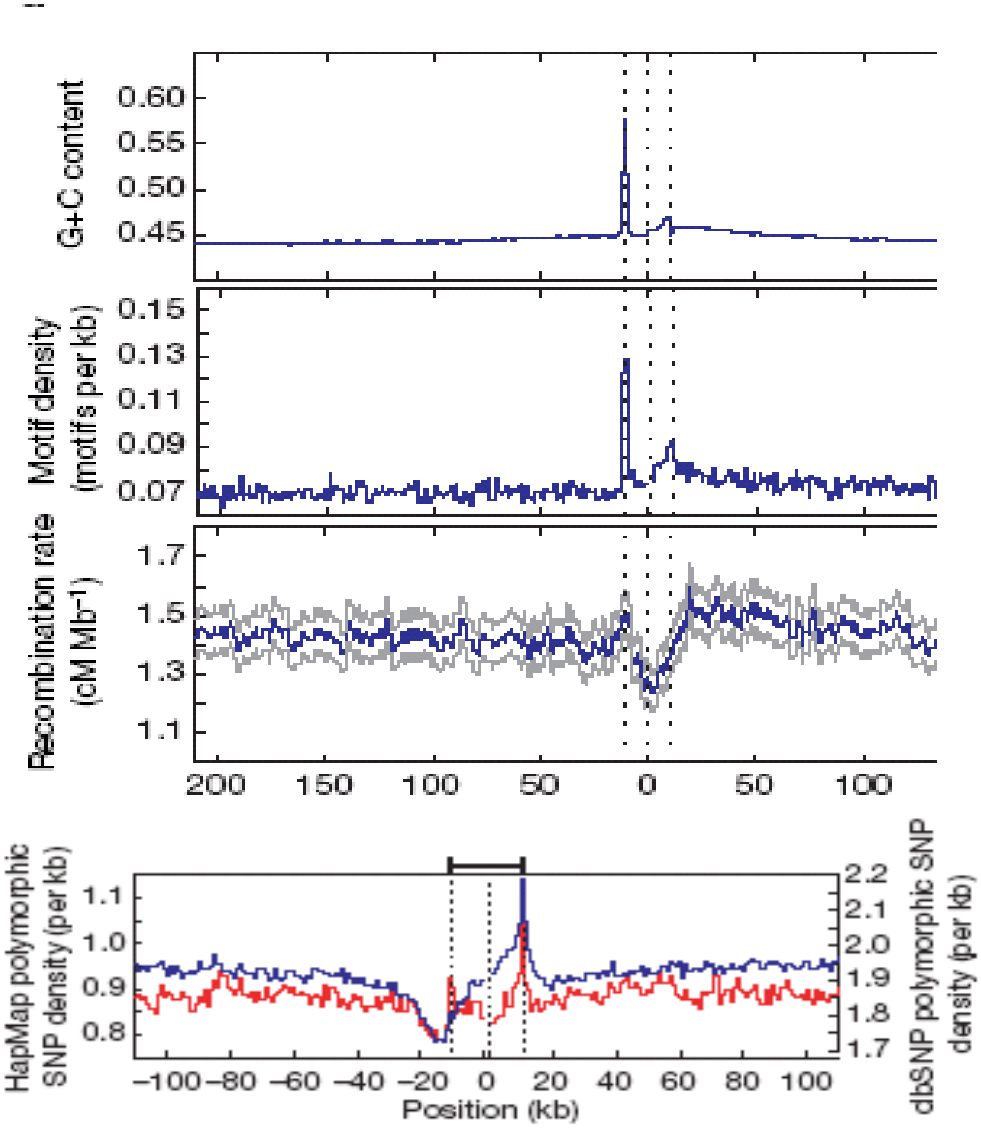
内 容

- Introduction
- Construction of the Phase II HapMap
- The use of the Phase II HapMap in association studies
- New insights into linkage disequilibrium structure
- **The distribution of recombination**
- Natural selection
- 作物的单倍型作图

The distribution of recombination

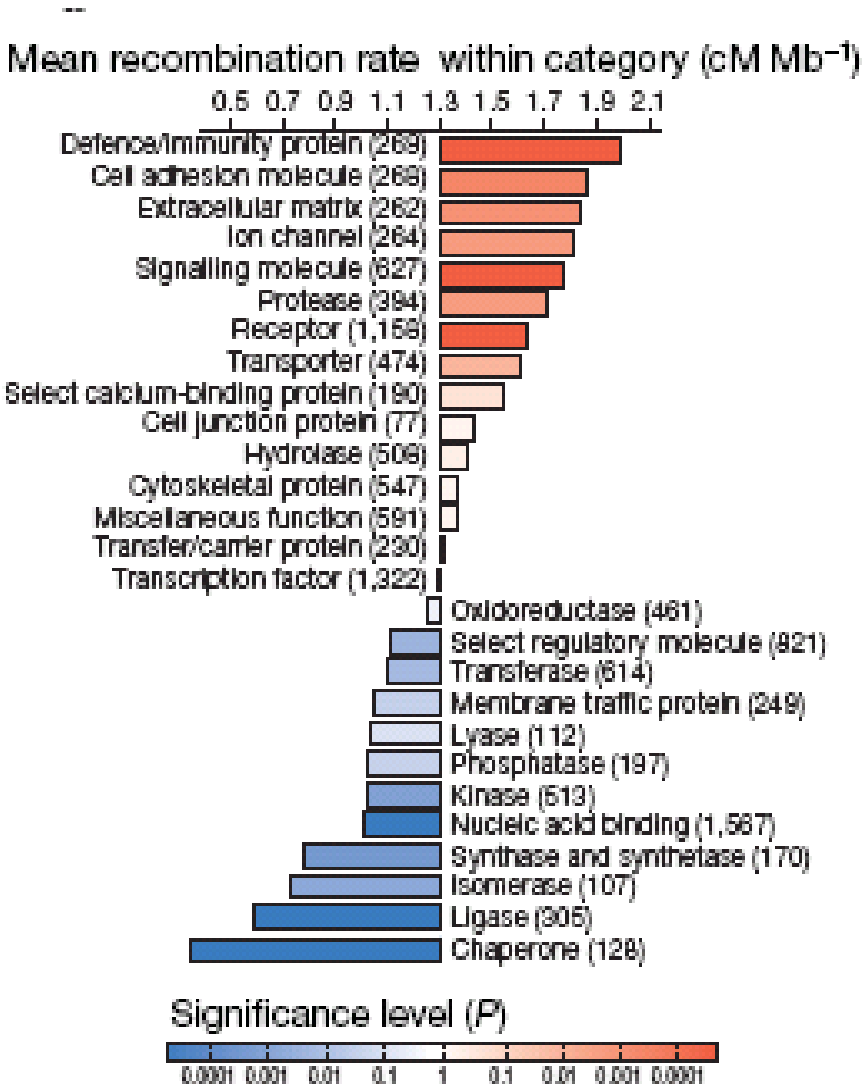
- 32,996 recombination hotspots
- 68% localized to a region of <5 kb
- The median map distance induced by a hotspot is 0.043 cM, the hottest identified is 1.2 cM
- Hotspots account for approximately 60% of recombination in the human genome and about 6% of sequence

Recombination rate around genes



- Within the transcribed region of genes there is a marked decrease in recombination rate
- 5'of the transcription start site is a peak in recombination rate with a corresponding local increase in the density of hotspot motifs,
- G+C content, reflecting the presence of CpG islands in promoter regions
- An asymmetry in recombination rate across genes

Systematic differences in recombination rate by gene class



- Gene functions associated with cell surfaces and external functions tend to show higher recombination rates (immunity, cell adhesion, extracellular matrix, ion channels, signalling)
- Those with internal to cells (chaperones, ligase, isomerase, synthase) lower recombination rates

内 容

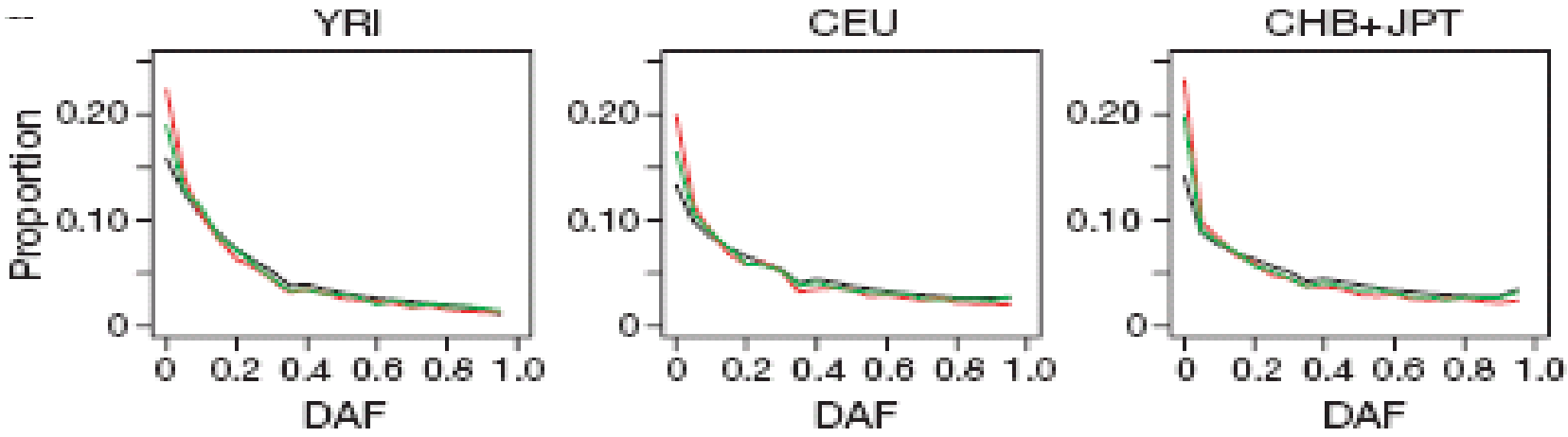
- Introduction
- Construction of the Phase II HapMap
- The use of the Phase II HapMap in association studies
- New insights into linkage disequilibrium structure
- The distribution of recombination
- **Natural selection**
- 作物的单倍型作图

Natural selection

The Phase I HapMap data have been used to identify genomic regions that show evidence for the influence of adaptive evolution, primarily through extended haplotype structure indicative of recent positive selection

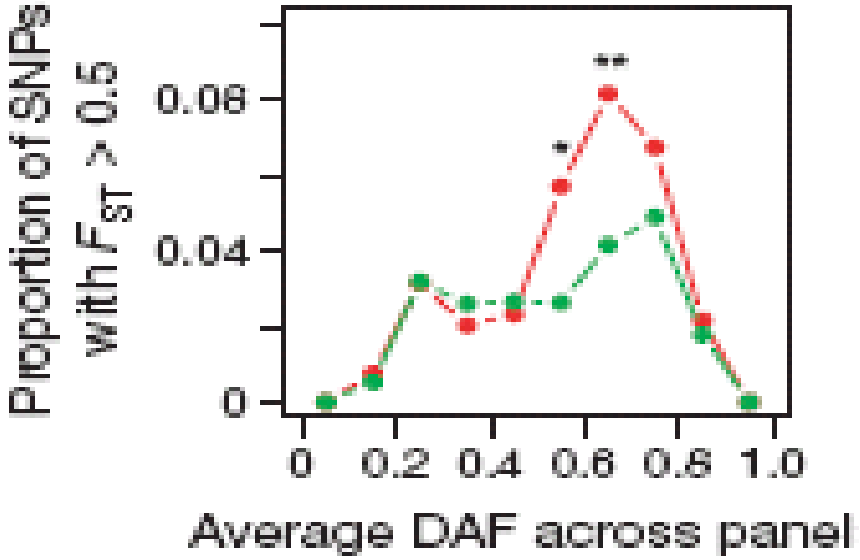
Approximately 200 regions with evidence of recent positive selection

Properties of non-synonymous and synonymous SNPs



The derived allele frequency (DAF) spectrum in each analysis panel for all SNPs (black), synonymous SNPs (green) and non-synonymous SNPs (red)

Properties of non-synonymous and synonymous SNPs



Enrichment of non-synonymous SNPs among genic SNPs showing high differentiation

意义

- The initial HapMap Project data had a central role in the development of methods for the design and analysis of genome-wide association studies
- Performing economically viable genome-wide genotyping
- Lead to a new phase in human medical genetics
- Have led to novel insights into the distribution and causes of recombination hotspots
- The prevalence of structural variation and the identity of genes that have experienced recent adaptive evolution

意义

- Because the HapMap cell lines are publicly available, many groups have been able to integrate their own experimental data to combine genome-wide data on such diverse aspects of genetic variation with molecular phenotypes collected in the same samples
- Provide a powerful framework to study the connection of DNA sequence to function
- Association studies is shifting from candidate gene approaches towards genome-wide analyses

内 容

- Introduction
- Construction of the Phase II HapMap
- The use of the Phase II HapMap in association studies
- New insights into linkage disequilibrium structure
- The distribution of recombination
- Natural selection
- 作物的单倍型作图

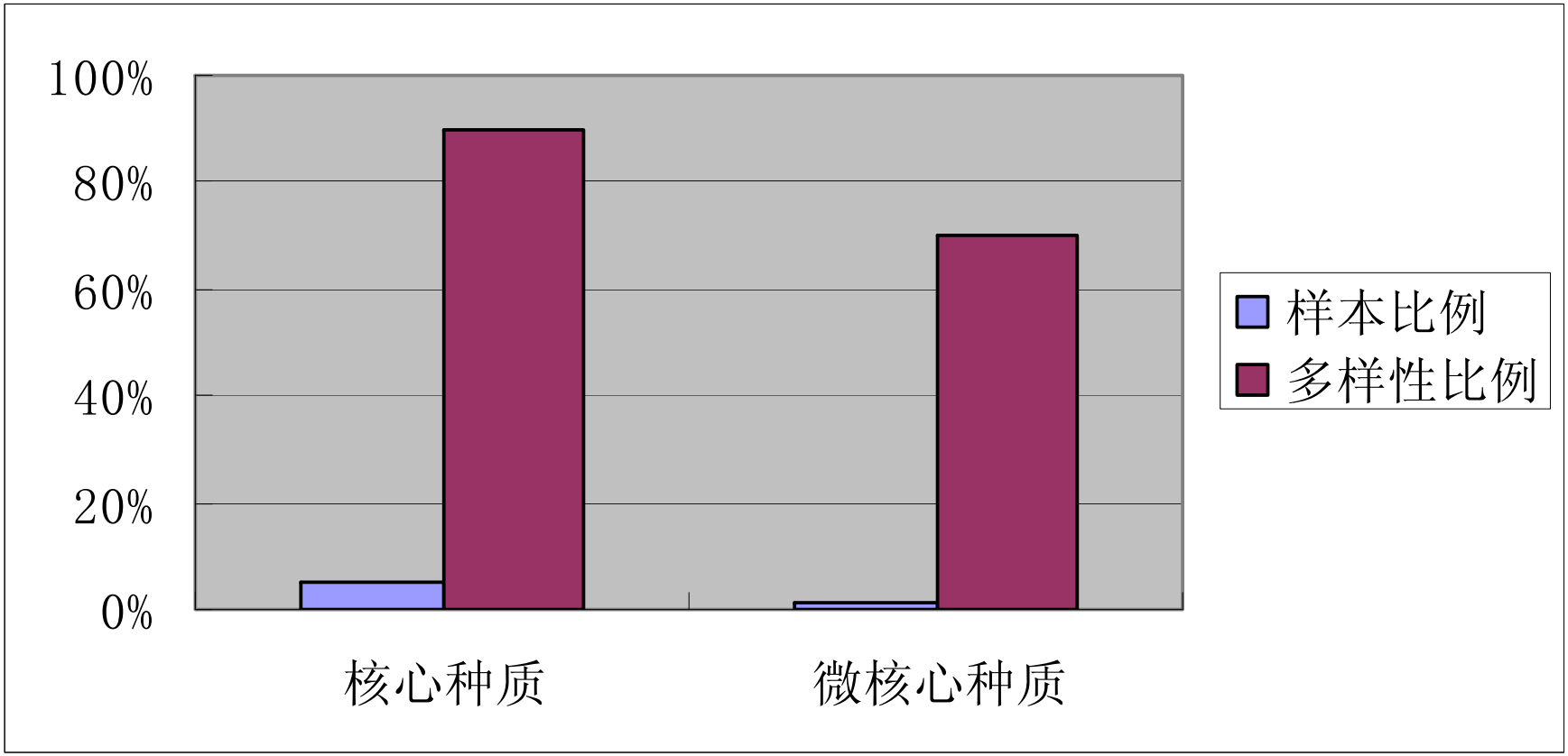
挑　战

- 作物种质资源数量多
- 作物的基因组大，结构复杂

  小麦：**$16x10^9bp$**

水稻、小麦与大豆核心种质构建

	基础种质	核心种质		微核心种质	
	No.	No.	%	No.	%
水稻	61479	1560	2.5	311	0.5
小麦	23135	1160	5	231	1
大豆	28809	1439	5	220	1


100%
80%
60%
40%
20%
0%
核心种质
微核心种质
样本比例
多样性比例

单倍型（haplotype）

- 组成生物的基因组碱基数量虽然巨大，但单倍型区段的数量却有限

- 单倍型的研究大大简化了作物基因资源研究的复杂性，是进行资源研究的有效手段

单倍型研究在作物资源研究上的应用

- 由于单倍型与单倍型区段的研究就是分区段进行资源的多样性研究
- 单倍型作图就是对资源的多样性进行作图
- 核心种质的单倍型作图是在基因组水平上进行资源研究的最有效途径

我国作物种质资源单倍型研究展望

- 材料：建立代表性更广泛的核心种质（共用）
- 方法

  高通量分子标记： SNP 384， 1536

  重测序：利用新一代测序仪进行直接测序

我国作物种质资源单倍型研究展望

新一代测序仪

Table 1 Second-generation DNA sequencing technologies

	Feature generation	Sequencing by synthesis	Cost per megabase	Cost per instrument	Paired ends?	1° error modality	Read-length	References
454	Emulsion PCR	Polymerase (pyrosequencing)	~$60	$500,000	Yes	Indel	250 bp	14,20
Solexa	Bridge PCR	Polymerase (reversible terminators)	~$2	$430,000	Yes	Subst.	36 bp	17,22
SOLiD	Emulsion PCR	Ligase (octamers with two-base encoding)	~$2	$591,000	Yes	Subst.	35 bp	13,26
Polonator	Emulsion PCR	Ligase (nonamers)	~$1	$155,000	Yes	Subst.	13 bp	13,20
HeliScope	Single molecule	Polymerase (asynchronous extensions)	~$1	$1,350,000	Yes	Del	30 bp	18,30

The pace with which the field is moving makes it likely that estimates for costs and read-lengths will be quickly outdated. Vendors including Roche Applied Science, Illumina, and Applied Biosystems have major upgrade releases currently in progress. Estimated costs-per-megabase are approximate and inclusive only of reagents. Read-lengths are for single tags. Subst., substitutions; indel, insertions or deletions; del, deletions.

我国作物种质资源单倍型研究展望

- 新一代测序仪的发展与使用将使单倍型研究进入一个全新的发展阶段

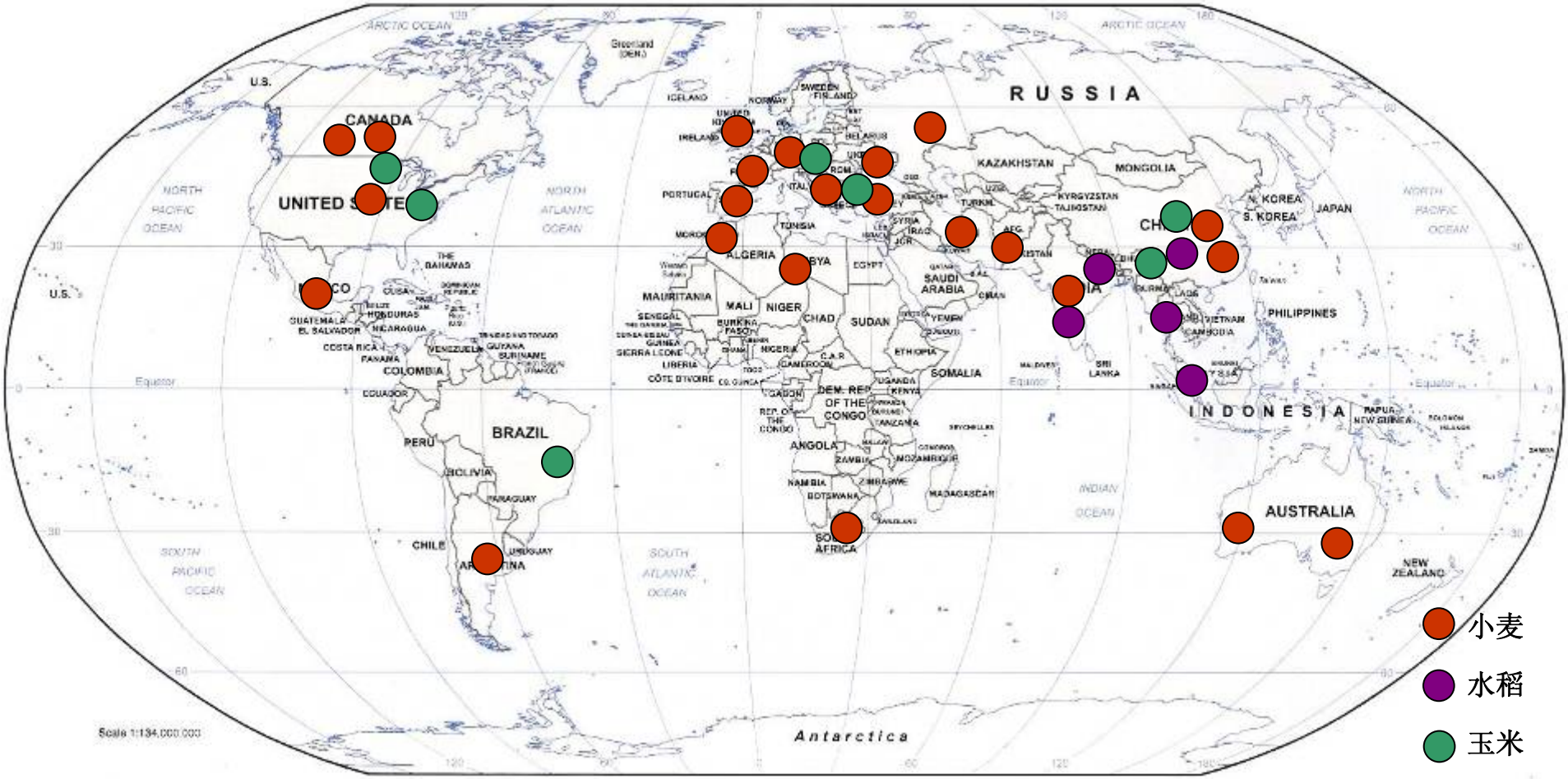
- 核心种质及其单倍型信息应该也能够成为未来资源研究、育种研究与基因组研究的材料平台与信息平台，使上述研究进入一个全新的发展时代

小麦光周期基因*Ppd-D1*单倍型的发现、评价与分布

引 言

- 植物通过感受昼夜长短变化而控制开花的现象称为光周期反应
- 小麦是长日照作物，光周期反应严重影响着小麦生育期、产量和适应性
- 小麦也是禾谷类作物中适应性最广泛的作物

小麦在世界范围内的分布



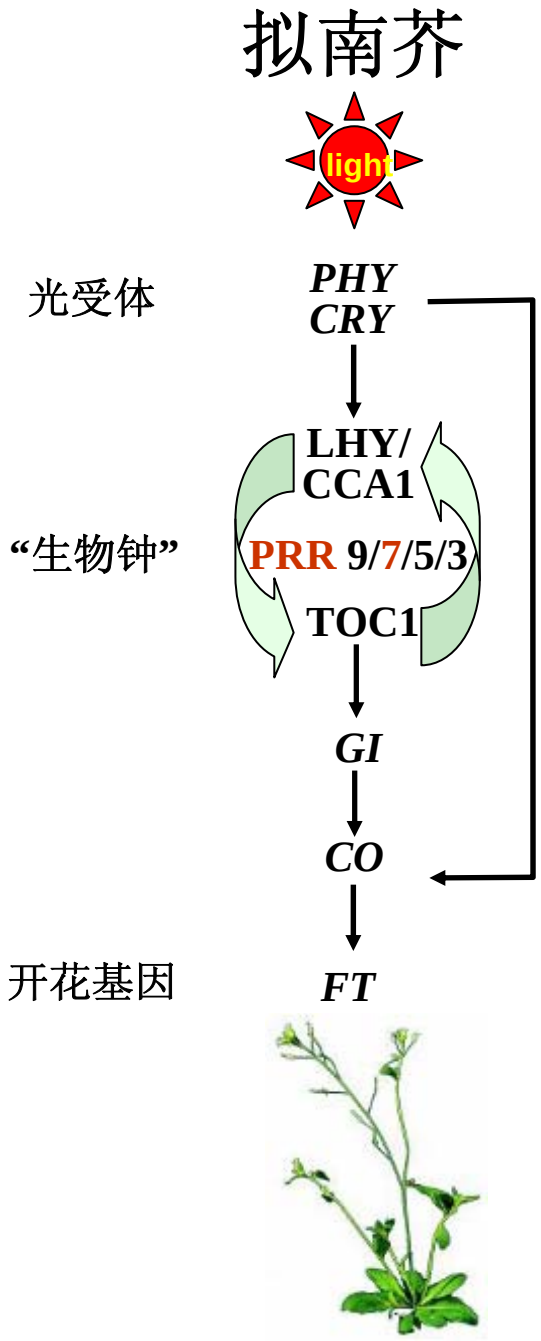
品种的光周期反应必须适应环境光周期大范围连续的变化

小麦光周期基因*Ppd-1*

- 光周期基因是控制小麦发育的主要基因之一，与小麦的产量、适应性、地理分布关系密切
- “绿色革命之父”Norman Borlaug曾将光周期不敏感基因*Ppd-D1*与矮秆基因*Rht1*并称为“绿色革命基因”
- 人们普遍认为光周期将植物对光周期的反应属于典型的质量性状，然而这一结果不能合理地解释小麦在全球范围的广泛适应性，即光周期的广泛变异。

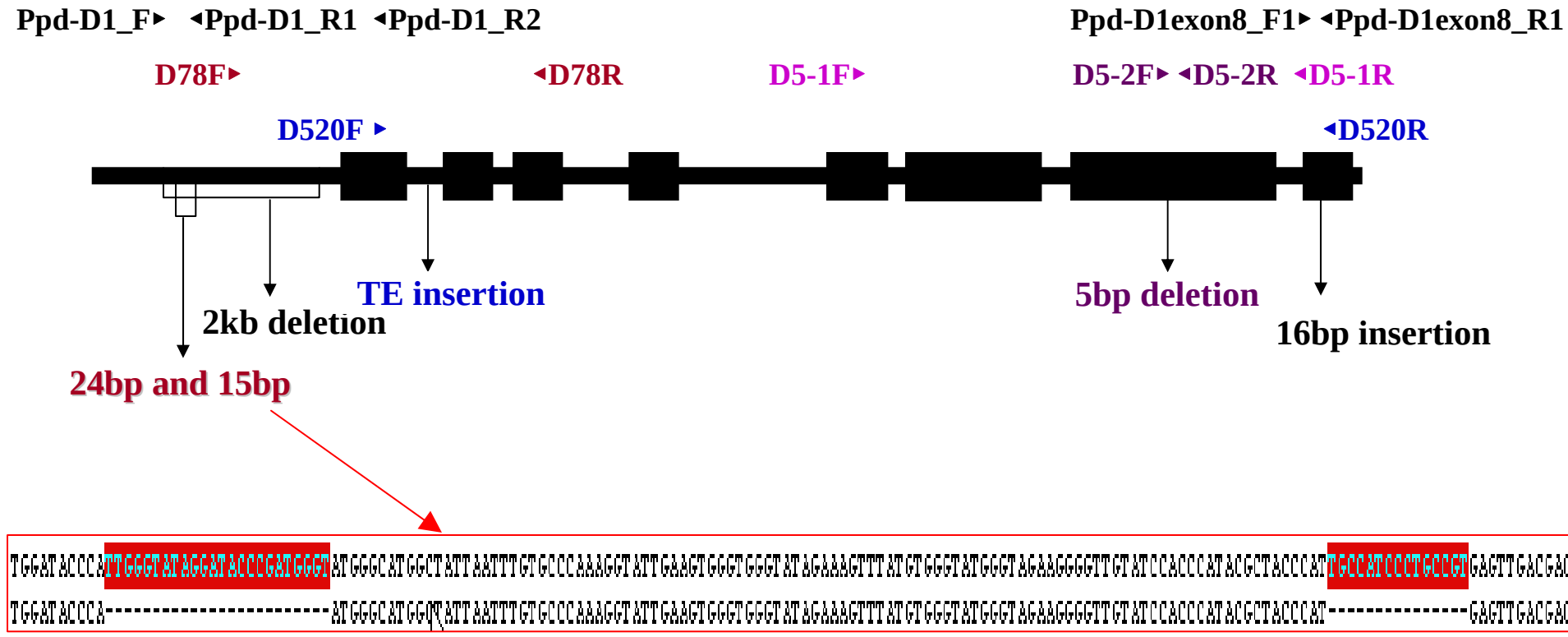
小麦光周期反应遗传学研究进展

- 图位克隆大麦***Ppd-H1***，发现是***PRRs***家族成员
- 同源克隆小麦***Ppd***基因



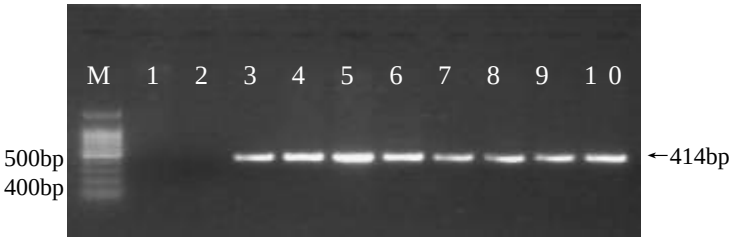
Turner et al. (2005) Science; Beales et al. (2007) TAG.

Ppd-D1 多态等位变异

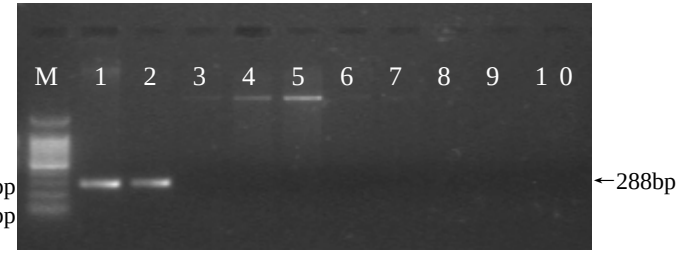


Ppd-D1 标记的开发

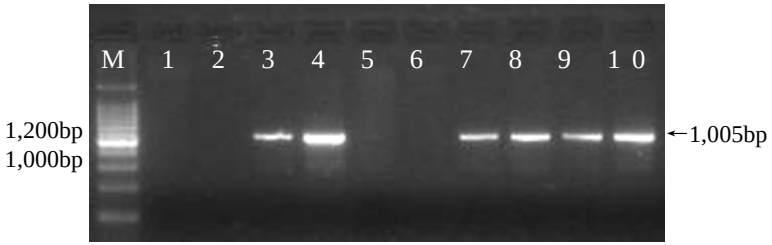
Ppd-D1_F* and Ppd-D1_R1*



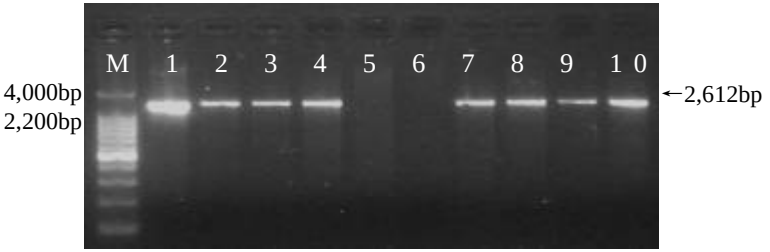
Ppd-D1_F* and Ppd-D1_R2*



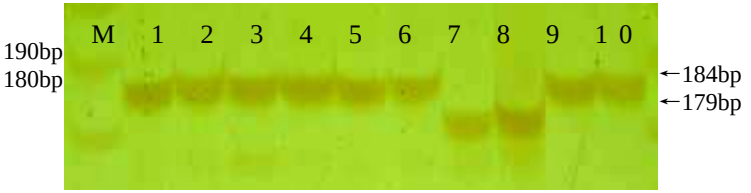
D78



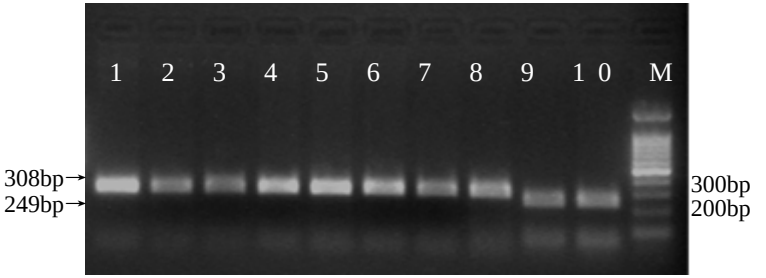
D520



D5



Ppd-D1exon8_F1* and Ppd-D1exon8_R1



M: DNA Markers100bp or 200bp, 1 Yanzhan 1, 2 Akagomughi, 3 Chinese Spring, 4 Ningchun 10, 5 Fr81-12, 6 Hussar, 7 Hezuo 2, 8 Early Premium, 9 96S-231 (synthetics), 10 96S-213(synthetics).

实验材料

- **492** 份普通小麦材料（六大洲**41**个国家）

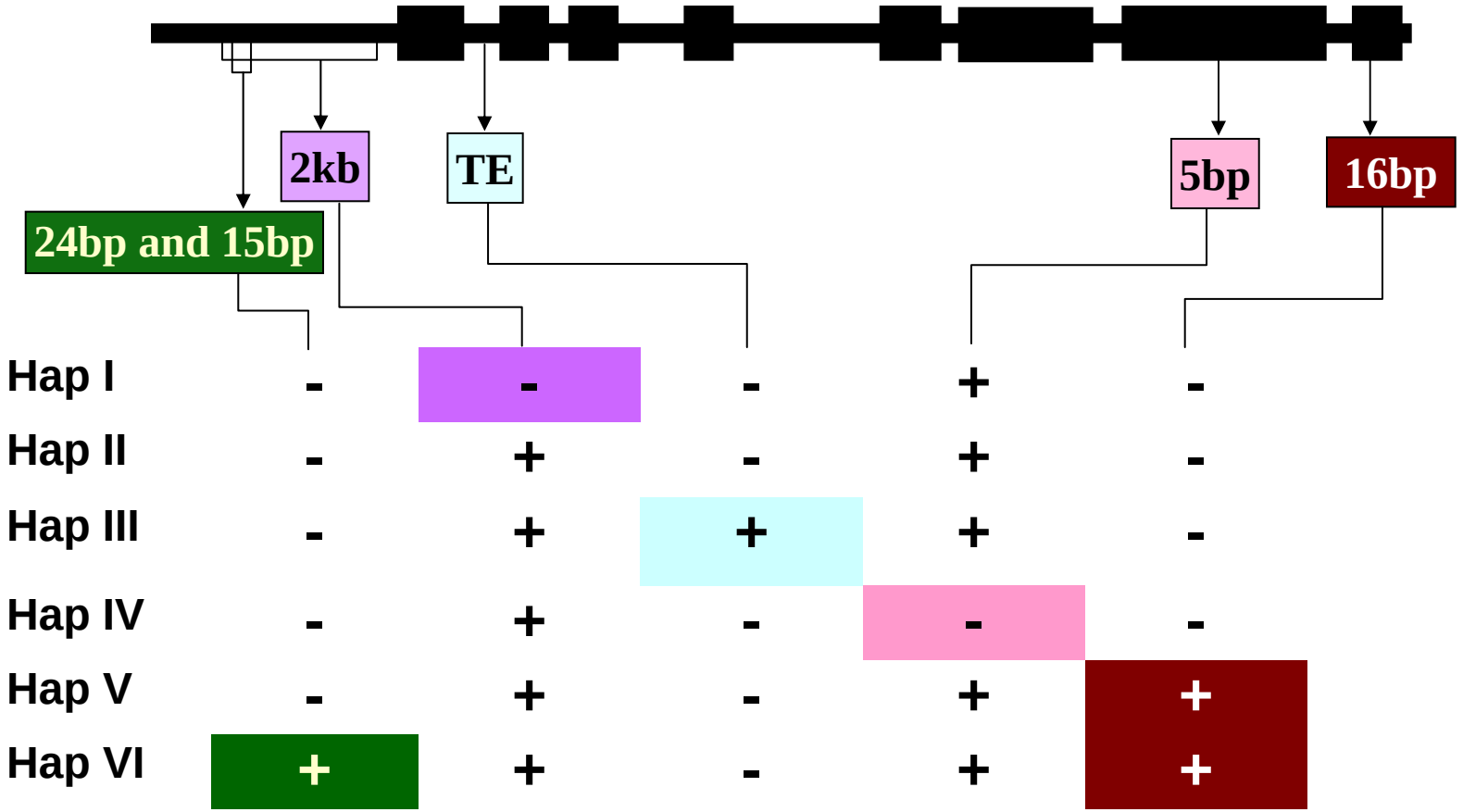
  我国材料**(**近一半**)**：地方品种和育成品种

  国外引进材料：骨干亲本材料和世界各地种质资源

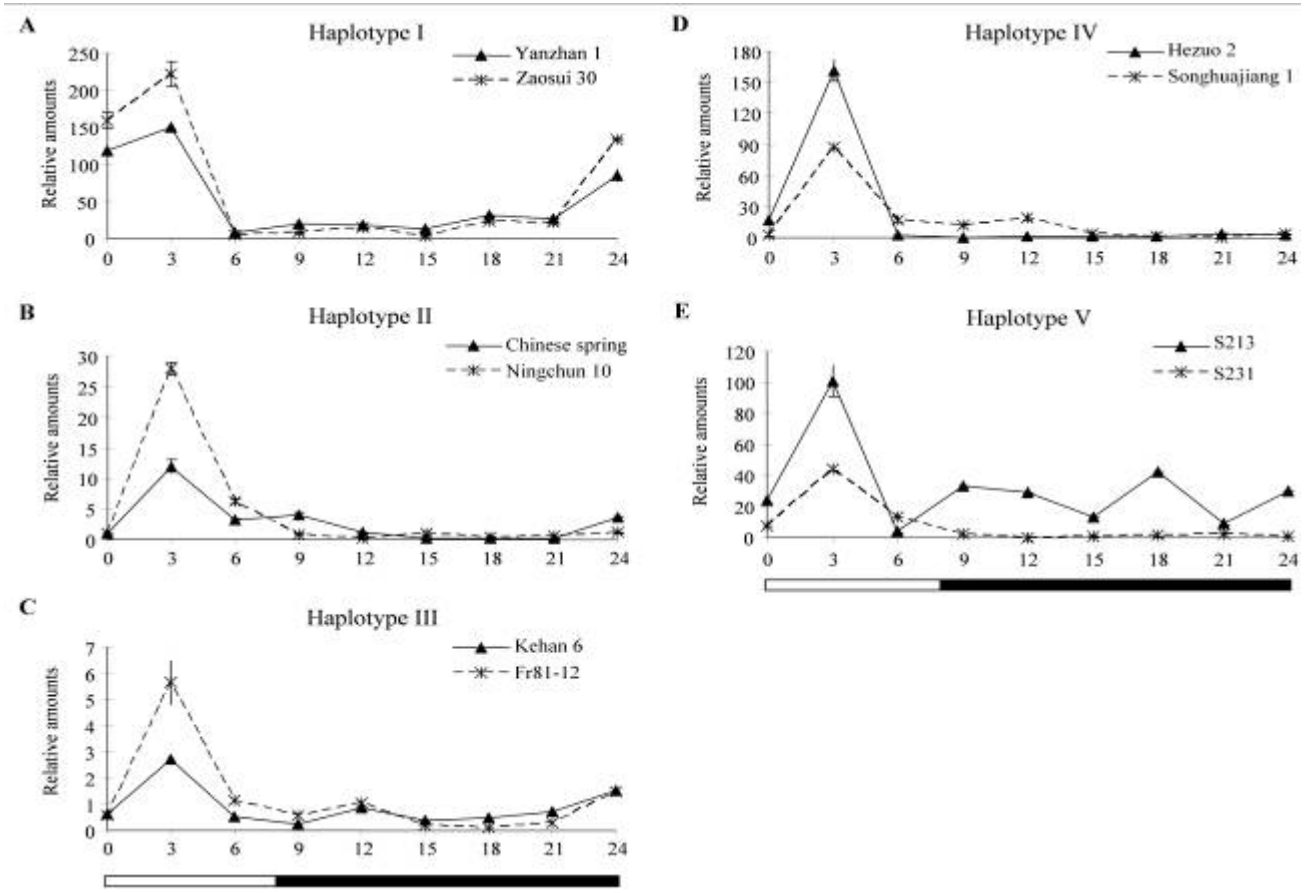
- **55**份近缘种

  25 份合成种和**30** 份粗山羊草**(*Ae. tauschii*)**

Ppd-D1 单倍型的分类



Ppd-D1 五种单倍型的表达模式



每个单倍型选用两份春性材料，在短日照条件下（**8h**光照和**16h**黑暗）研究表达。

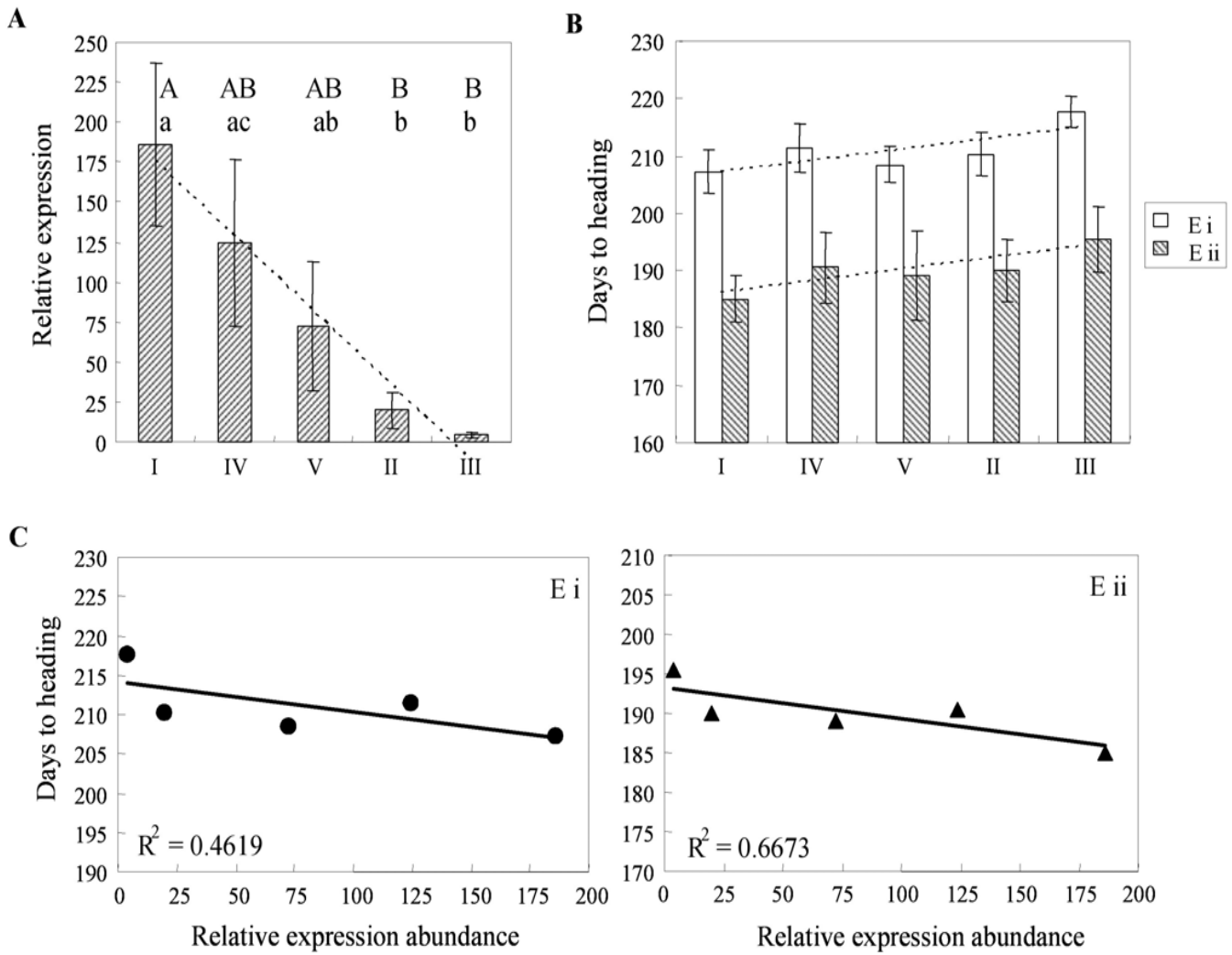
*Ppd-D1*五种单倍型与主要农艺性状关联分析

性状	环境	I		II		III		IV		V	
		N	Means	N	Means	N	Means	N	Means	N	Means
抽穗期	i	103	207.21a(A)	37	210.27b(B)	4	217.75c(C)	14	211.36b(B)	6	208.50ab(AB)
	ii	116	185.08a(A)	43	190.09b(B)	9	195.44c(C)	15	190.53b(BC)	10	189.10b(B)
开花期	i	76	245.36a	29	245.55a	1	251	8	247.88b	5	246.40ab
	ii	117	222.89a(A)	42	224.07ac(AC)	8	230.00b(B)	15	226.47bc(BC)	10	227.90b(BC)
株高	i	81	81.99a(AC)	35	97.97b(B)	5	80.48ac(ABC)	11	90.28ab(AB)	3	61.85c(C)
	ii	105	103.73a(A)	41	123.17b(B)	6	111.50ab(AB)	12	117.00bc(AB)	9	103.19ac(A)
	iii	110	115.70a(A)	42	139.51b(B)	9	116.28ac(A)	14	127.65bc(AB)	8	111.70ac(A)
穗下节	i	81	28.10a(AB)	35	30.82b(A)	5	26.92abc(AB)	11	30.22ab(AB)	3	20.13c(B)
	ii	108	31.11a	41	32.34ab	6	32.45ab	12	35.74b	9	30.44a
	iii	103	36.02a(A)	42	39.67bc(BC)	9	34.36a(AB)	14	41.66c(C)	8	35.64ab(ABC)

不同的字母代表显著差异，大写代表*P*<0.05水平差异显著，小写代表*P*<0.01差异极显著。环境（E） i 、ii 和iii分别代表北京昌平（116.2^{o}E, 40.2^{o}N）、洛阳（111.6^{o}E, 33.8^{o}N）旱地和水地。

性状	环境	I		II		III		IV		V	
		N	Means	N	Means	N	Means	N	Means	N	Means
穗长	i	81	9.39a(A)	35	10.48b(B)	5	12.54c(B)	11	11.15bc(B)	3	10.77abc(AB)
	ii	108	9.55a	41	9.65ab	6	10.47ab	12	10.39ab	9	10.91b
	iii	110	9.80a	42	9.59a	9	11.42b	14	10.76ab	8	11.56b
小穗数	i	81	18.04a	35	18.97ab	5	20.61b	11	19.62ab	3	18.33ab
	ii	108	19.09a	41	19.50a	6	20.00a	12	19.15a	9	18.19a
	iii	109	19.33a	42	19.23a	9	19.64a	14	19.46a	8	19.54a
穗粒数	i	81	43.19ab	35	39.76a	5	49.07b	11	44.16ab	3	51.93b
	ii	108	43.95a(A)	41	42.36a(AB)	6	39.42ab(AB)	12	41.20ab(AB)	9	37.32b(B)
	iii	109	49.08a	42	45.42b	9	42.28b	14	47.48ab	8	47.71ab
穗数	i	81	8.30a	34	8.03a	5	7.16a	11	9.61a	3	8.60a
	ii	108	10.61a	41	10.70ab	6	10.69ab	11	12.12b	9	10.48ab
	iii	109	11.98a	42	12.17a	7	10.89a	14	12.07a	8	12.42a
千粒重	i	81	29.52a(A)	35	26.83bd(A)	5	22.64d(A)	11	30.56ab(AB)	3	38.84c(B)
	ii	108	29.75a(A)	41	26.13b(B)	6	28.88ab(ABC)	12	30.59a(ABC)	9	36.14c(C)
	iii	110	33.47a(A)	40	29.24bc(B)	9	26.63b(B)	14	32.25ac(AB)	8	38.03d(A)

*Ppd-D1*单倍型表达量与品种抽穗期相关分析

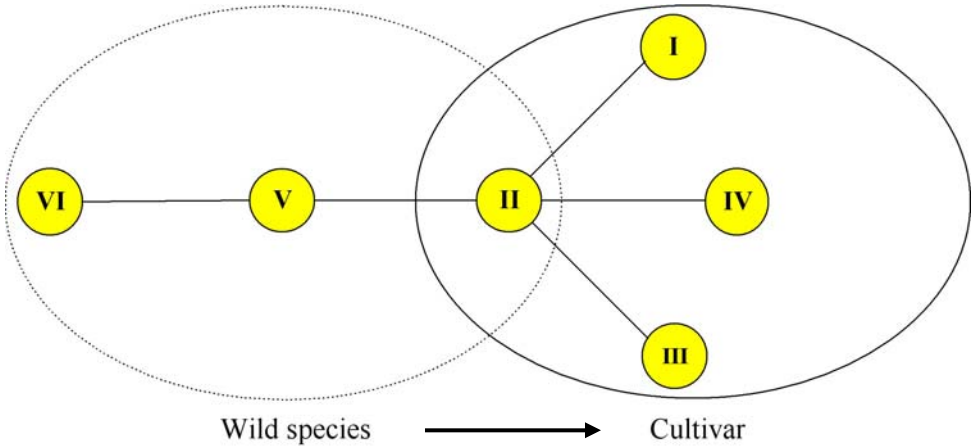


(A) 短日照条件下早晨（表达峰时间点）***Ppd-D1***单倍型的相对表达。大写表示在**1%**水平上的显著，小写表示**5%**水平上的显著。（**B**）不同单倍型的品种在不同环境下的抽穗期。环境**i**：北京（**116.2ºE, 40.2ºN**），环境**ii**： 洛阳（**111.6ºE, 33.8ºN**）。（**C**）***Ppd-D1***单倍型的相对表达量与品种在不同环境下的抽穗期相关分析。环境**i**（**R2=0.4619, *P*<0.207**），环境**ii**（**R2=0.6673, *P*<0.092**）。

Ppd-D1 单倍型不同材料中的分布

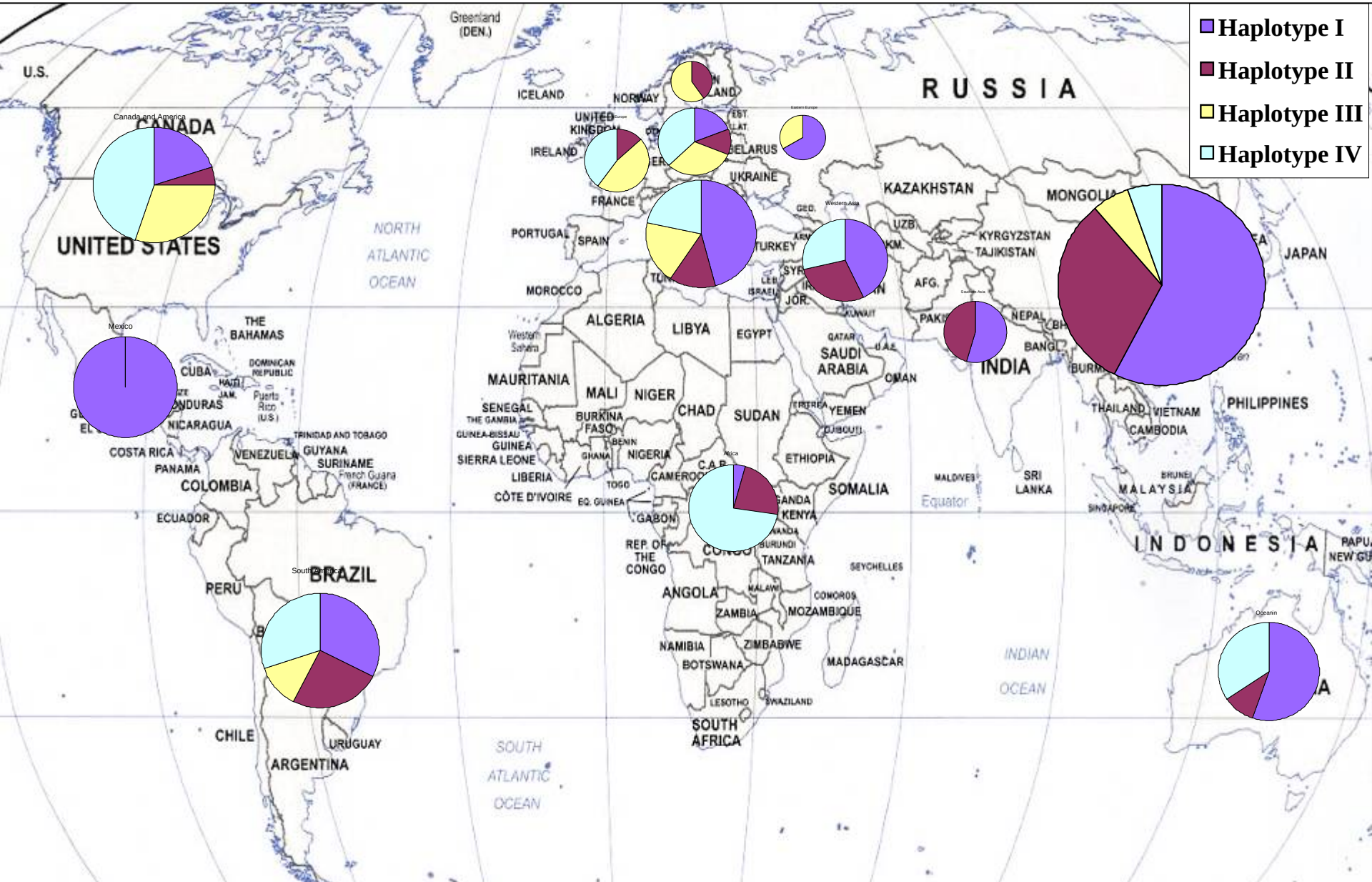
来源	材料数	Hap					
		I	II	III	IV	V	VI
普通小麦	492	47.47%	22.63%	10.71%	19.19%	0.00%	0.00%
近缘种	55	0.00%	3.64%	0.00%	0.00%	52.73%	41.82%
共计	547	42.73%	20.73%	9.64%	17.24%	5.27%	4.18%

小麦*Ppd-D1* 单倍型进化模型

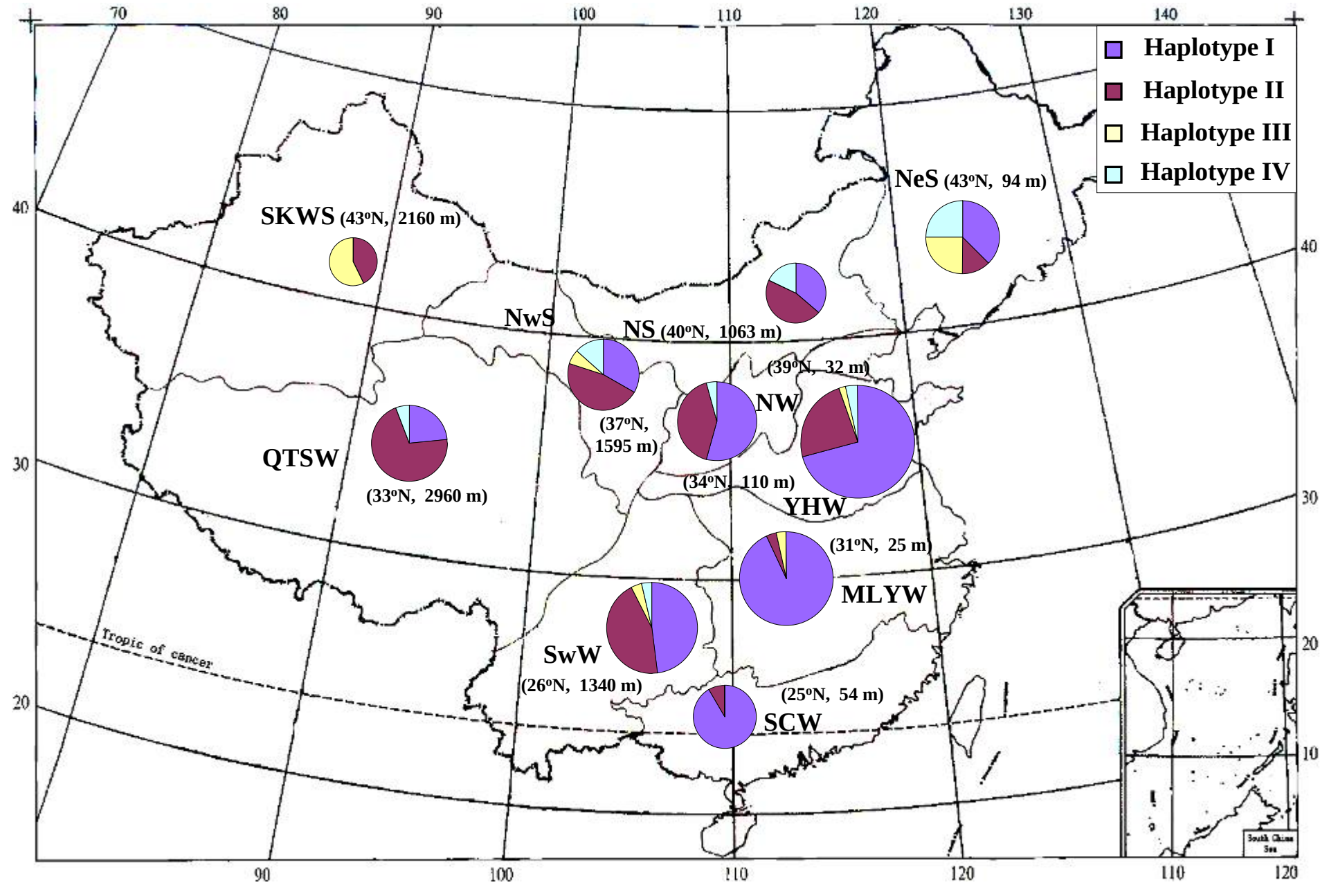


Haplotype	24bp and 15bp	2kb upstream region	TE intron 1	5bp exon 7	16bp exon8
I	-	-	-	+	-
II	-	+	-	+	-
III	-	+	+	+	-
IV	-	+	-	-	-
V	-	+	-	+	+
VI	+	+	-	+	+

*Ppd-D1*单倍型在世界范围内的分布



*Ppd-D1*单倍型在我国麦区的分布

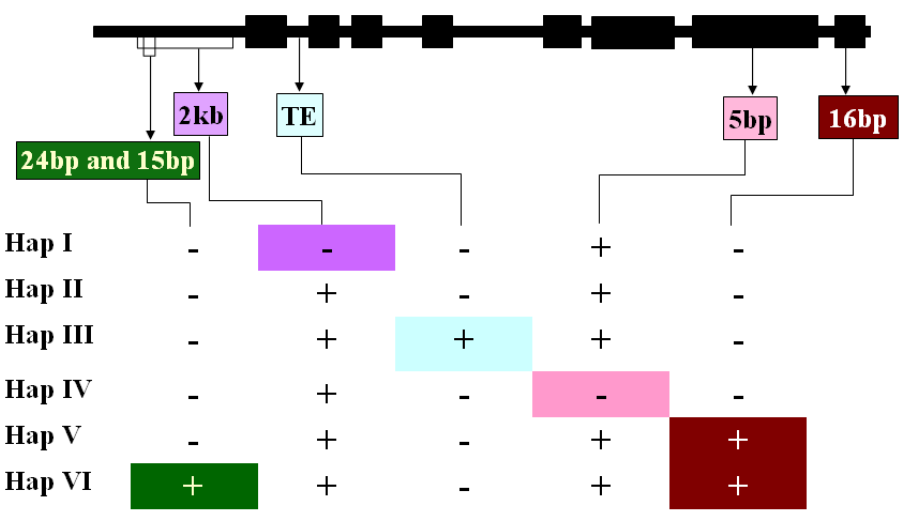


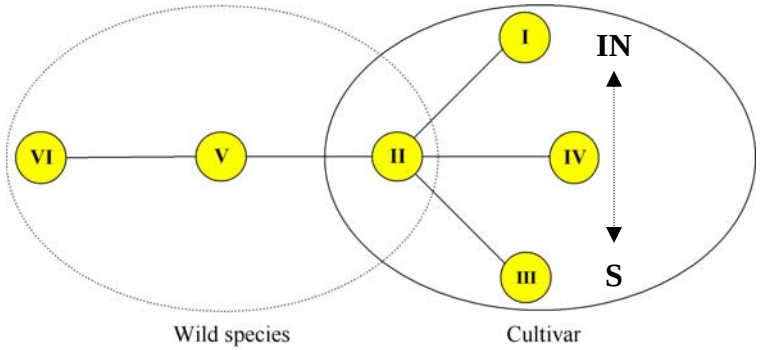
*Ppd-D1*单倍型分布频率与地理纬度、海拔的相关分析

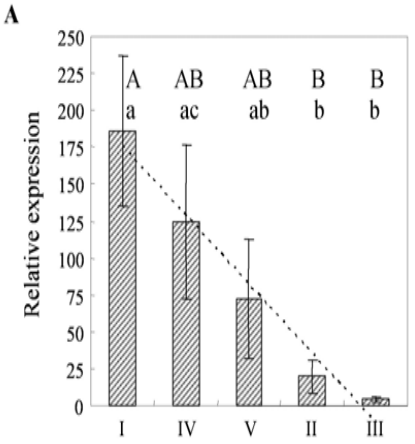
	变量		
	Hap I	**Hap II+Hap III**	**Hap IV**
纬度	**-0.8872****	**0.7666***	**0.5577**
海拔	**-0.9190****	**0.9038****	**-0.1208**

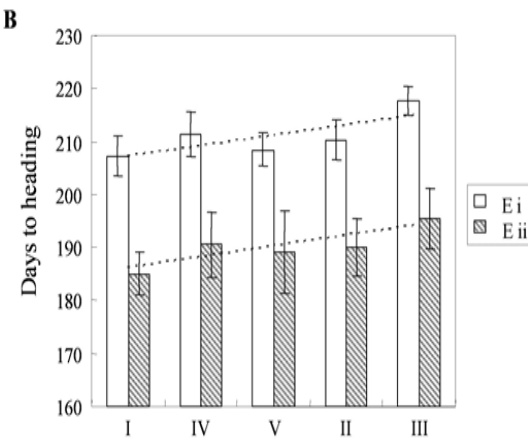
** 和* 分别表示*P*<0.01和 *P*<0.05水平上的显著性。

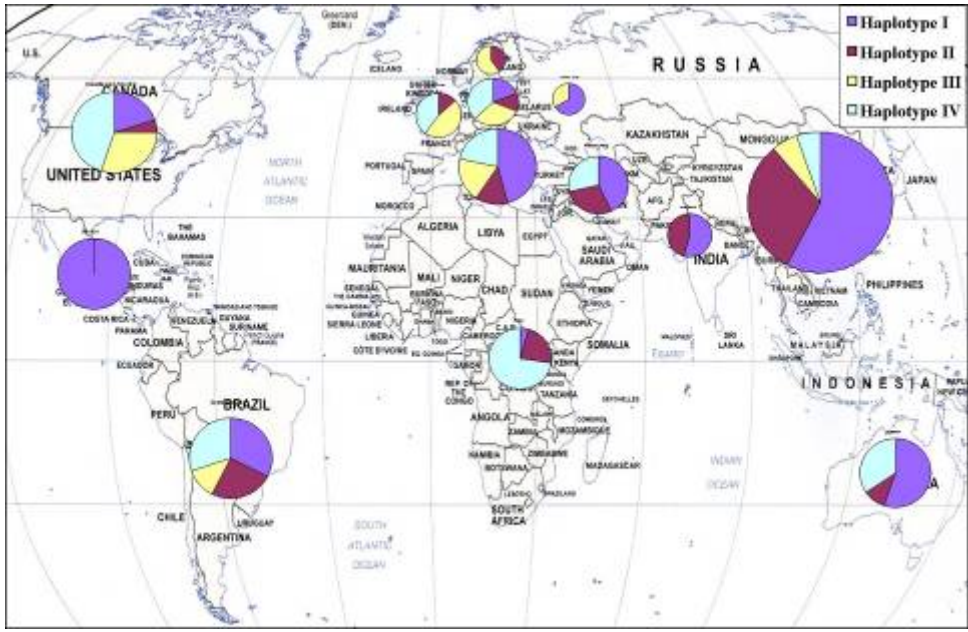
小　结


2kb
TE
5bp
16bp
24bp and 15bp
Hap I - - - + -
Hap II - + - + -
Hap III - + + + -
Hap IV - + - - -
Hap V - + - + +
Hap VI + + - + +


I
IN
VI
V
II
IV
III
S
Wild species
Cultivar


A
Relative expression
250
225
200
175
150
125
100
75
50
25
0
A AB AB B B
a ac ab b b
I IV V II III


B
Days to heading
230
220
210
200
190
180
170
160
I IV V II III
Ei
Eii


Haplotype I
Haplotype II
Haplotype III
Haplotype IV
RUSSIA
CANADA
UNITED STATES
INDIA
INDONESIA
BRAZIL

质量性状≌数量性状

- **QTL**的克隆使“数量性状”转变为“质量性状”
- 本研究发现原来认为的质量性状基因***Ppd-D1***有**6**种单倍型，其表达量及其对农艺性状的影响呈连续变异，具有明显的数量性状特征
- 随着越来越多等位基因的发现，原来认为的质量性状实际上可能是数量性状。


谢 谢