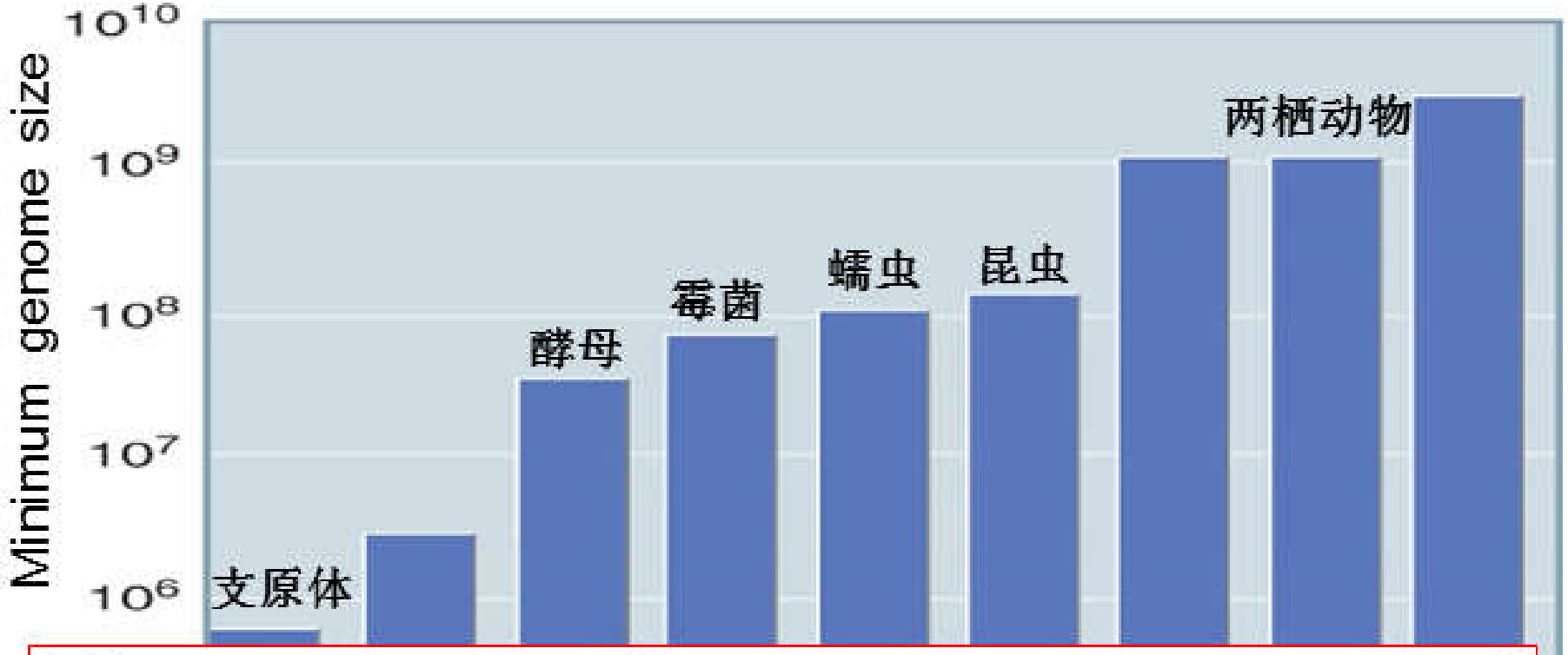
4. Genome complexity

4.1 Genome and genome size

- **Genome** is the complete set of sequences in the genetic material of an organism (haploid cell).
  一特定生物体(单倍体细胞)中的整套遗传信息(全部DNA序列)的总和称为基因组。

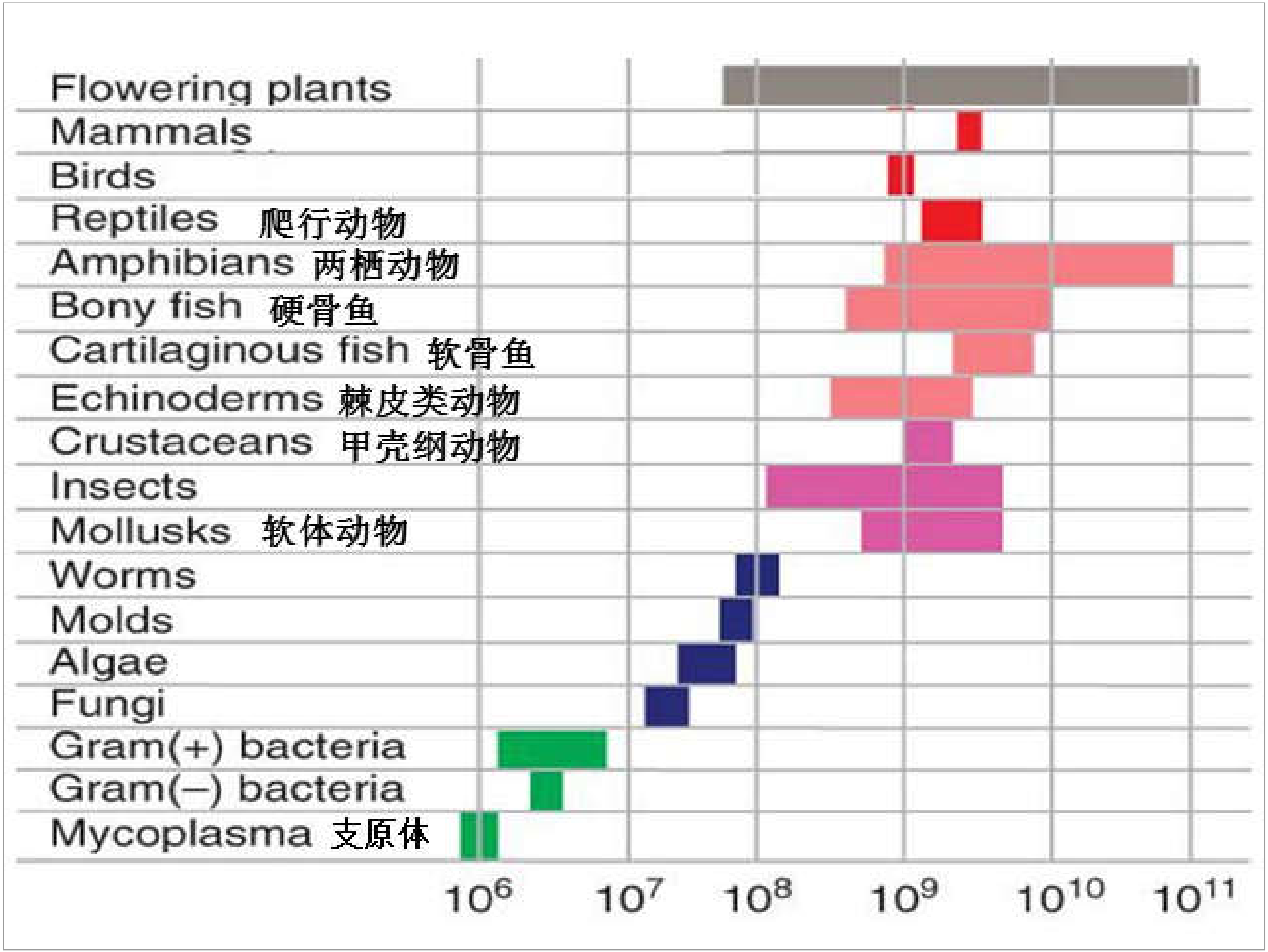
  - Genome includes the sequence of each chromosome plus any DNA in organelles (细胞器).

- **Genome size** is the total amount of DNA in the genome (**C-value, C值**).



There is an increase in the minimum genome size associated with organisms of increasing complexity.

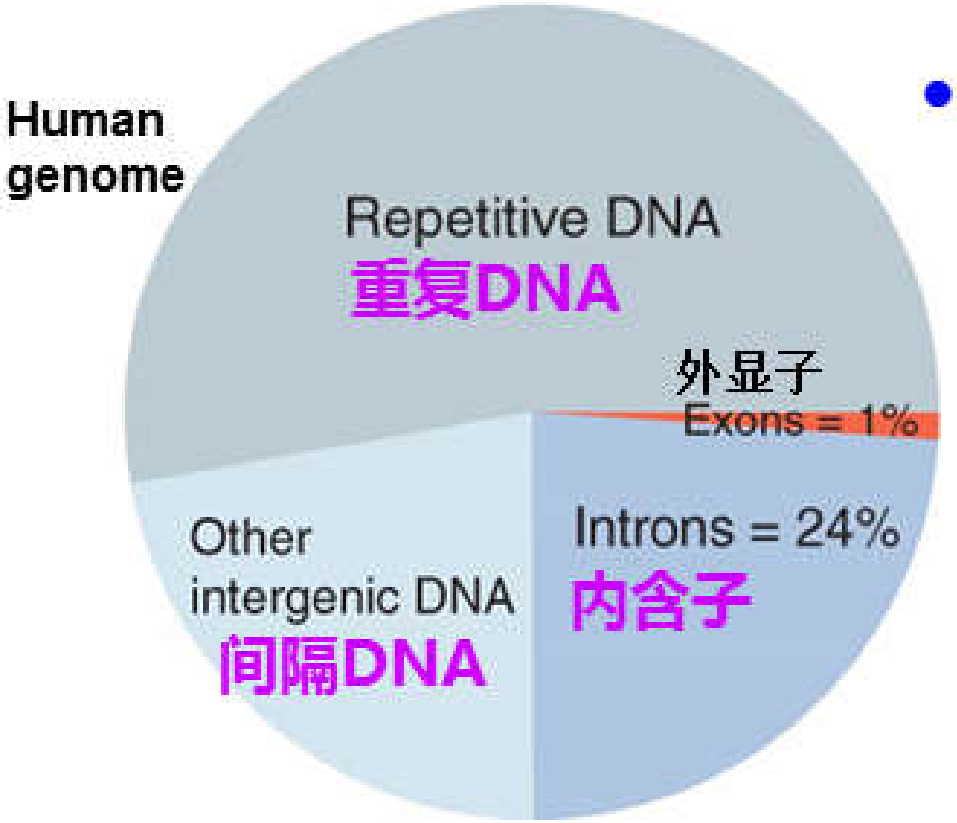
随着生物的复杂程度增高，其(最小)基因组越大。


Flowering plants
Mammals
Birds
Reptiles 爬行动物
Amphibians 两栖动物
Bony fish 硬骨鱼
Cartilaginous fish 软骨鱼
Echinoderms 棘皮类动物
Crustaceans 甲壳纲动物
Insects
Mollusks 软体动物
Worms
Molds
Algae
Fungi
Gram(+) bacteria
Gram(−) bacteria
Mycoplasma 支原体
10^6
10^7
10^8
10^9
10^{10}
10^{11}

- C-value paradox: the amount of DNA in a haploid genome does not seem to correspond strongly to the complexity of an organism.
  C值悖论/C值矛盾: 基因组大小往往与生物的复杂程度不一致。
- 具体表现：(1)某些高等生物的C值小于某些低等生物；(2)同类生物不同种属之间C值相差很大。

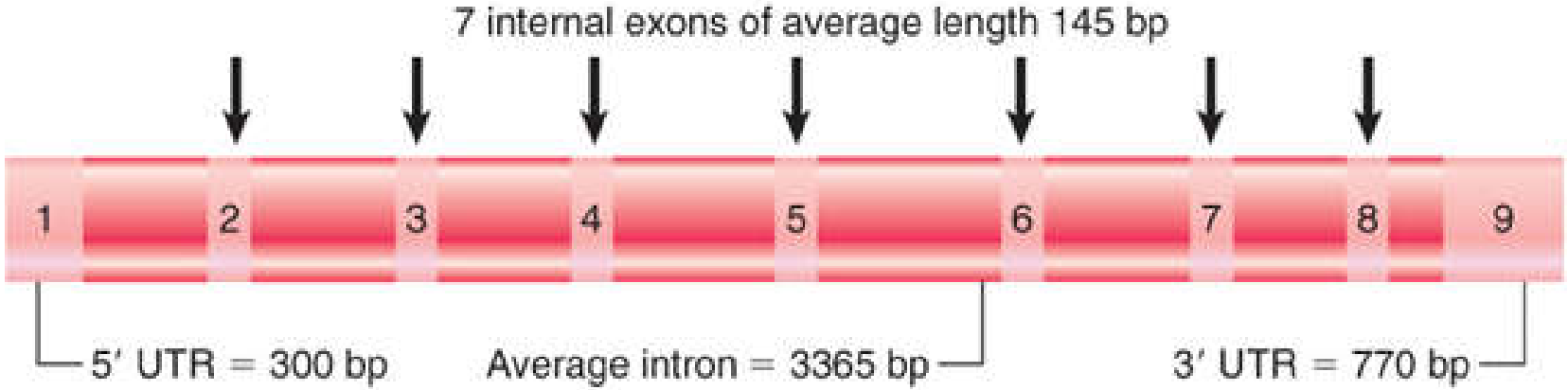
4.2 Noncoding DNA

- **Noncoding DNA** is component of an organism's DNA that does not code for protein.
  生物体内不编码蛋白质的DNA称为非编码DNA。



- Complex eukaryotic organisms contain much noncoding DNA.
  - Much of noncoding DNA consists of multiple repeats (repetitive DNA).

- The coding regions of genes are interrupted by intron sequences (noncoding DNA in the gene).

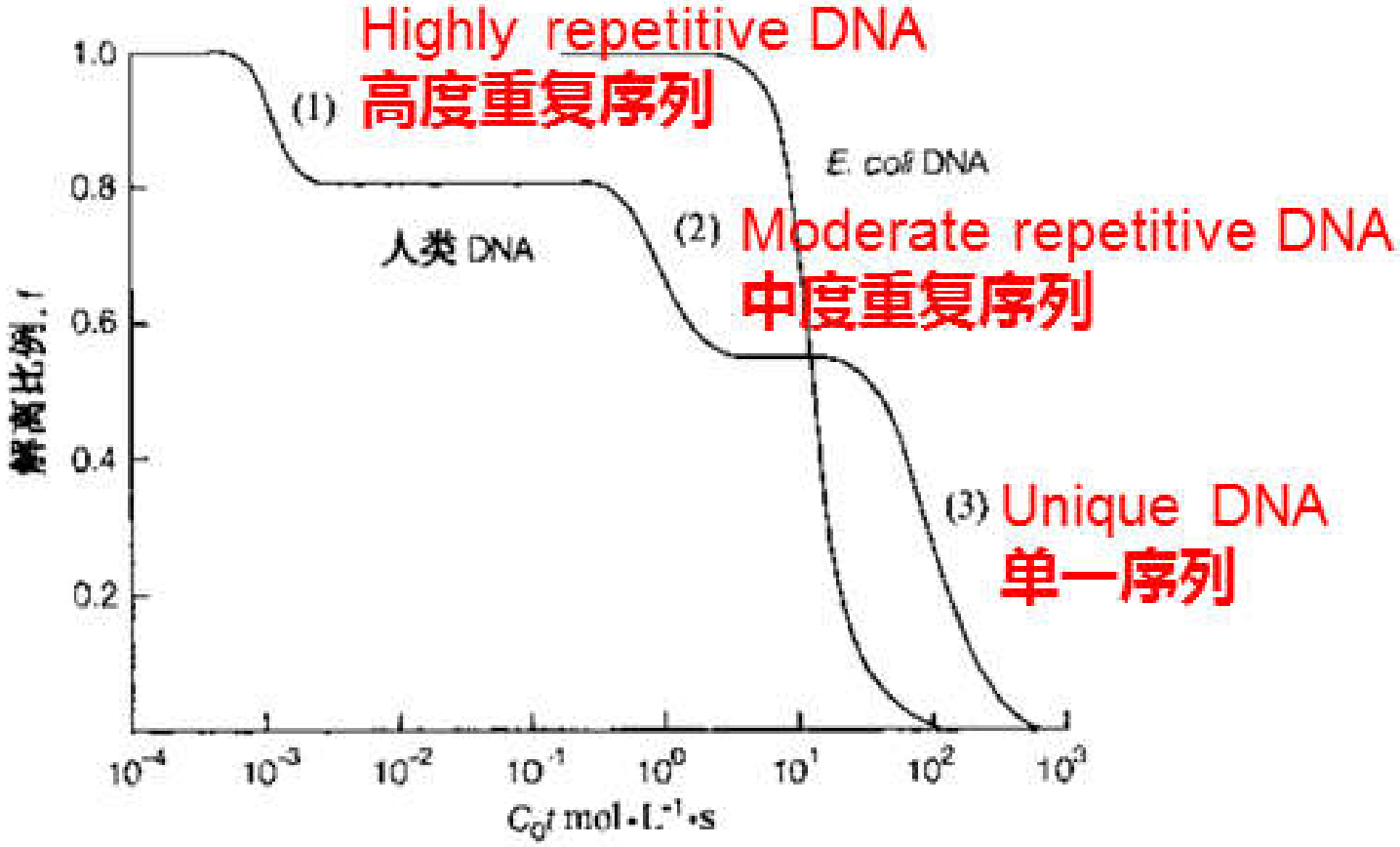


- Intergenic DNA – noncoding DNA between genes.

  - There are two kinds of intergenic DNAs: unique (单一的) and repeated. About one-quarter of the intergenic DNA is unique.

4.3 Reassociation kinetics (复性动力学)

- **Reassociation kinetics is a technique that measures the rate of reassociation of complementary strands of DNA derived from a single source.**
  - Step 1: Broken by sonication (超声波) into same size
  - Step 2: DNA fragments are thermally denatured.
  - Step 3: Reanneal at a low concentration
  - Step 4: Spectroscopic (光谱) or hydroxyapatite chromatography (羟基磷灰石色谱) analysis



C_0 - the initial concentration of ssDNA
t - the reannealing time (seconds)

- **Increasing numbers of DNA copies – decreasing C_0t**

Decreasing numbers of DNA copies – increasing C_0t

4.4 Unique sequence DNA

- One or a few copies per haploid genome
- The slowest fraction to reassociate on a C_0t curve
- Most protein genes in eukaryotes are unique sequence DNA.
- Only a small fraction of unique DNA is used to encode proteins.

4.5 Moderate repetitive DNA

- 10^1~10^6 copies per haploid genome

Tandem gene clusters
（串联基因簇）

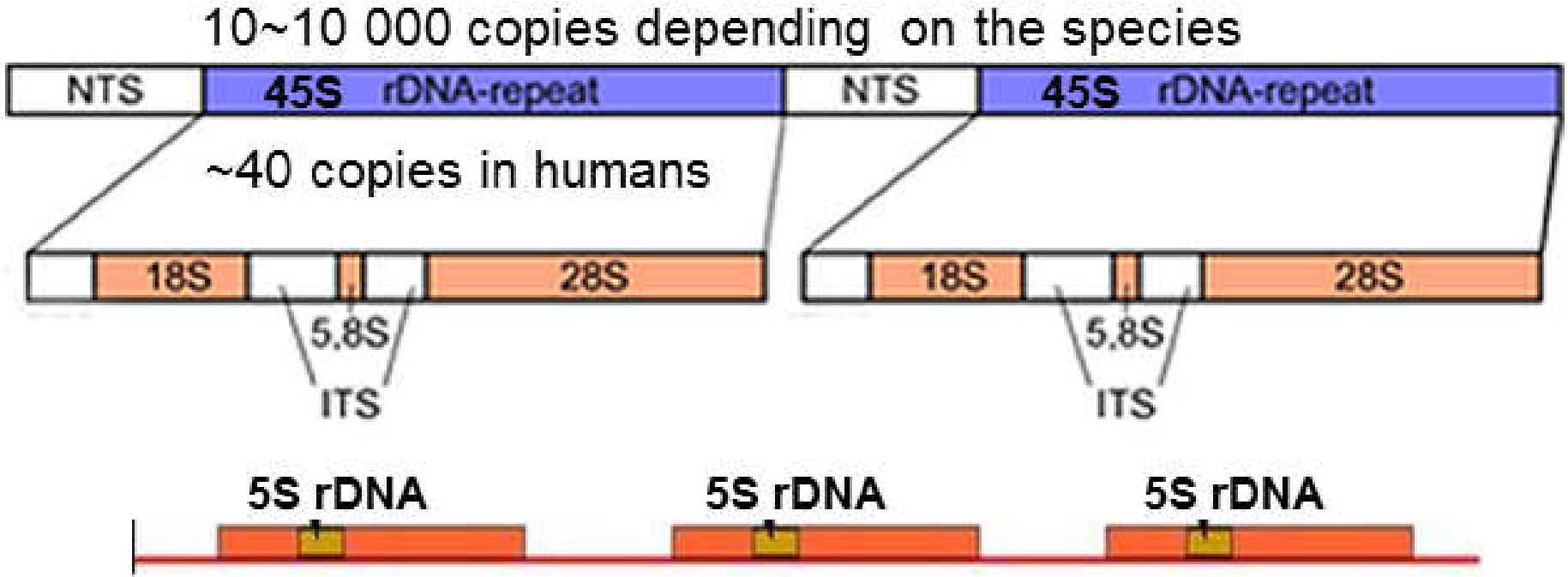
Dispersed repetitive DNA
（分散重复DNA）

4.5.1 Tandem gene clusters

- **Tandem gene clusters** is a group of identical or related genes that are arranged in tandem.
- One kind of gene family
- **Gene family** is a set of several similar genes, formed by duplication of a single original gene, and generally with similar functions.
  来源相同、结构相似、功能相关的一组基因构成**基因家族**。

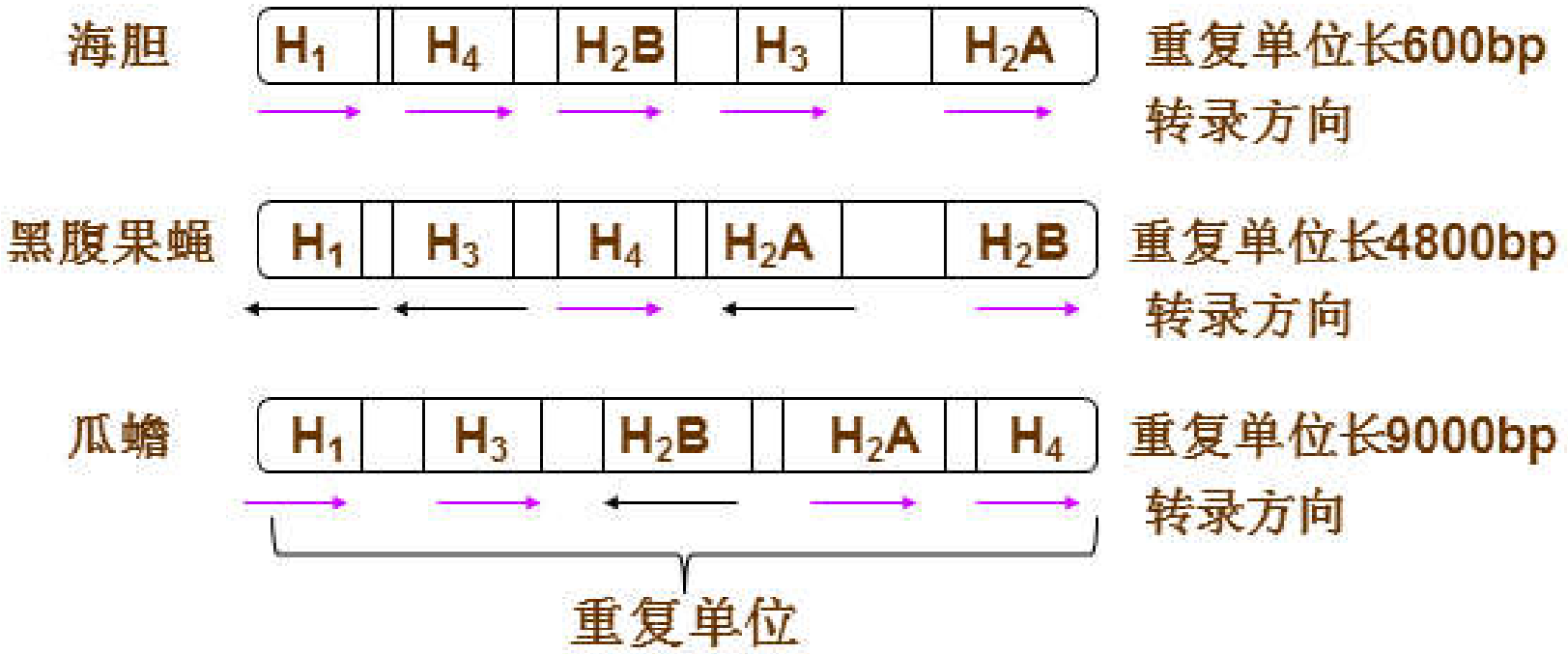
(1) rRNA-encoding gene (rDNA)

- The gene which encodes the 45S precursor of the 18S, 5.8S, 28S rRNA is repeated in arrays.
- In interphase, 45S precursor genes are spatially located together in the nucleolus (核仁).



(2) Histone gene clusters

- Transcribed in large quantities during S-phase
- The five histone genes occur together in a cluster.



- Directly repeated ten to several hundred times

4.5.2 Dispersed repetitive DNA

- Repeated 10^3~10^6 times and scattered throughout the whole genome

Short interspersed elements
（SINE, 短散布元件）

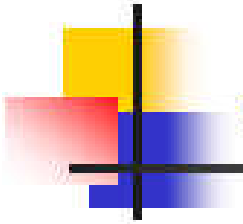
Long interspersed elements
（LINE, 长散布元件）

(1) Short interspersed elements (SINE)

- The average length of the repeat sequence is a few hundred base pairs.
- *Alu* element
  - The length of the repeat unit: 300 bp
  - Repeat times: 300 000 ~ 500 000
  - The copies are all 80~90% identical.
  - Most contain the *Alu* Ⅰ restriction site (限制位点).

(2) Long interspersed elements (LINE)

- The average length of the repeat sequence is several thousand base pairs.
- *L1* element
  - The length of the repeat unit: ~6 000 bp
  - Repeat times: hundreds of thousands
  - *L1* are found in all mammals and make up almost 10% of the human genome
  - Occurring between genes and in introns



Functions of dispersed repetitive DNA

(1) Origins of replication

(2) Gene regulation sequences

(3) Most are transposons (转座子).

4.6 Highly repetitive DNA

- >10^6 copies per haploid genome
- Noncoding DNA

Inverted repeat
（IR, 反向重复）

Satellite DNA
（卫星DNA）

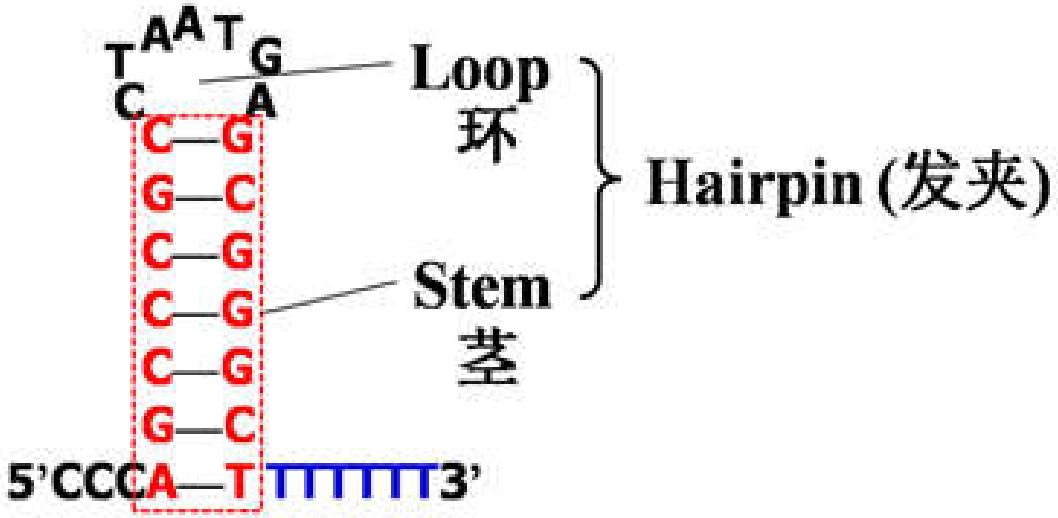
4.6.1 Inverted repeat

- **An inverted repeat** is that two identical copies of the sequence are arranged in reverse in the DNA strand.

反向重复序列是指两个序列相同的拷贝在DNA链上呈反向排列。

- When the intervening length is zero, the inverted repeat is a palindromic sequence (回文序列).

- When the inverted repeat is denatured and then renatured, the complementary sequences in the same strand can form a hairpin structure.



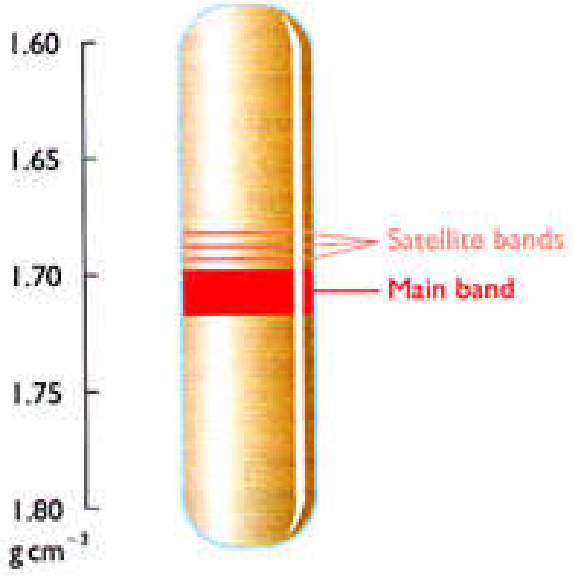
- Mainly found in transposons and regulatory sequences of gene expression.

4.6.2 Satellite DNA

- **Satellite DNA** – DNA that consists of **many** identical or related **tandem repeats** (>10^6 copies) of a **short** basic repeating unit.

  卫星DNA: 由**许多**相同或相关的**短**的基本重复单元组成的**串联重复**DNA。

- Identified as satellite bands in CsCl density gradients of chromosomal DNA (have nonaverage G+C content)
- The highest component of the higher eukaryotic repetitive DNA



6 Satellite DNA from the human genome.

(1) Minisatellite DNA (小卫星DNA)

- The length of the repeat unit: tens of base pairs
- Also called variable number tandem repeats (VNTR, 可变数目串联重复)

(2) Microsatellite DNA (微卫星DNA)

- The length of the repeat unit: 1~10 bp
- Also called simple tandem repeats (STR, 简单串联重复)

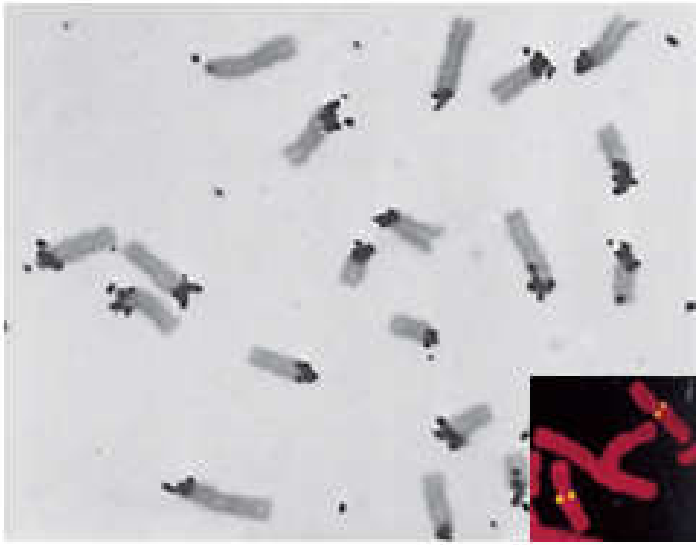
Usually, minisatellites and microsatellites are together classified as VNTR.

(3) Other satellite DNA

- The length of the repeat unit: 100~1 000 bp

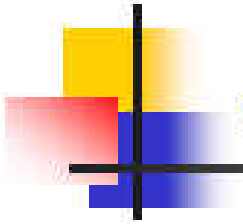
- **Satellite DNA mainly located in the centromeres and telomeres of chromosomes, which form a large part of heterochromatin.**

➢ *in situ* hybridization – hybridization of a probe to tissue or cell to locate its complementary strand (target sequence) by autoradiography or fluorescence.



原位杂交：将探针与组织或细胞杂交，通过放射自显影或荧光定位其互补链（靶序列）所在位置。

Fluorescent *in situ* hybridization (FISH, 荧光原位杂交)



Applications of VNTR

DNA fingerprinting technique is a method used to identify individuals from DNA samples based on VNTR.

DNA指纹技术是一种基于VNTR从DNA样本中识别个体的方法。

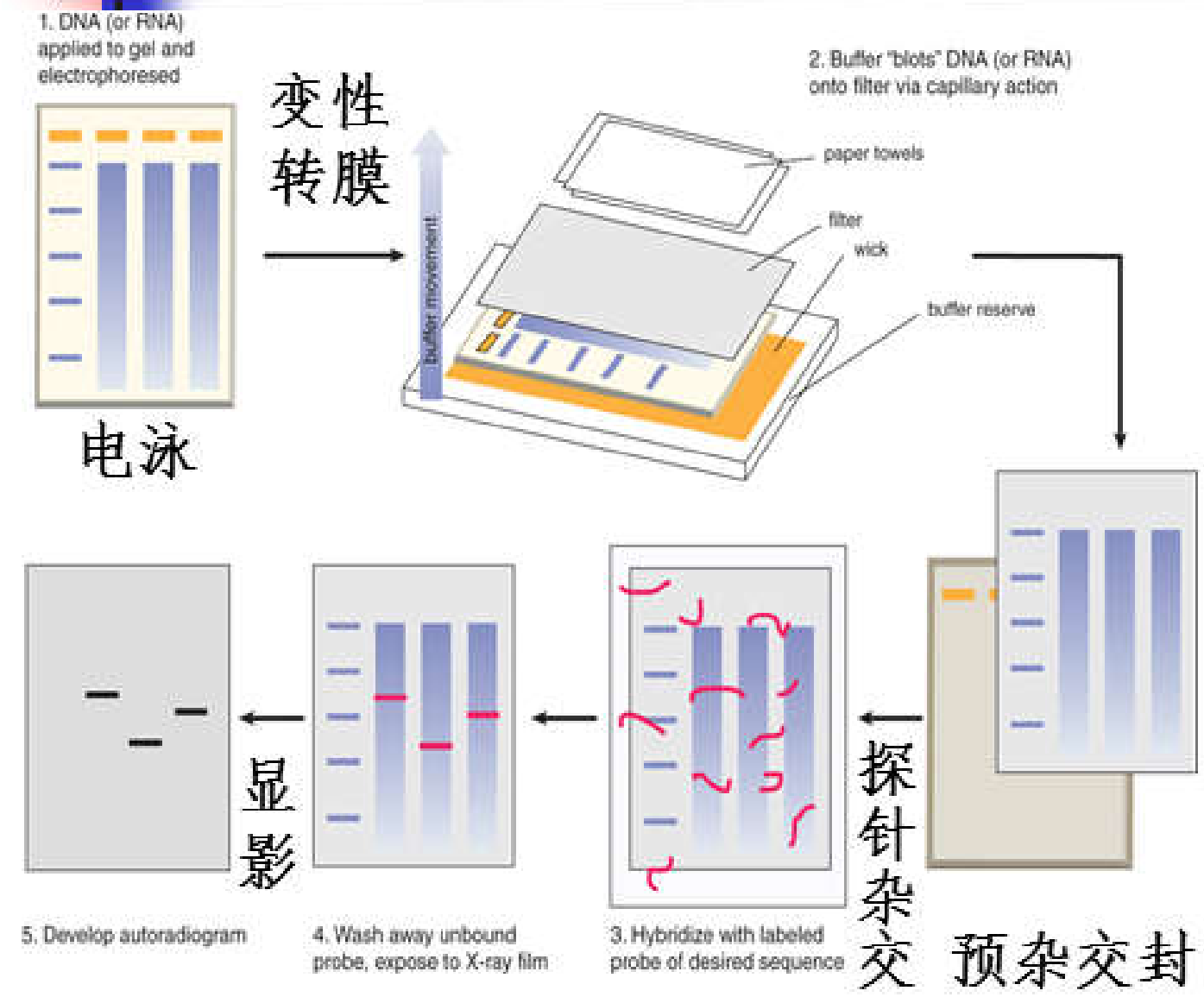
➢ Paternity test (亲子鉴定)

(1) The numbers of repeats in the arrays of some satellite sequences are hypervariable (高度可变) – examples of genetic polymorphism (遗传多态性).

(2) VNTR exhibits codominant inheritance (共显性遗传), and children can simultaneously display parental VNTR bands.

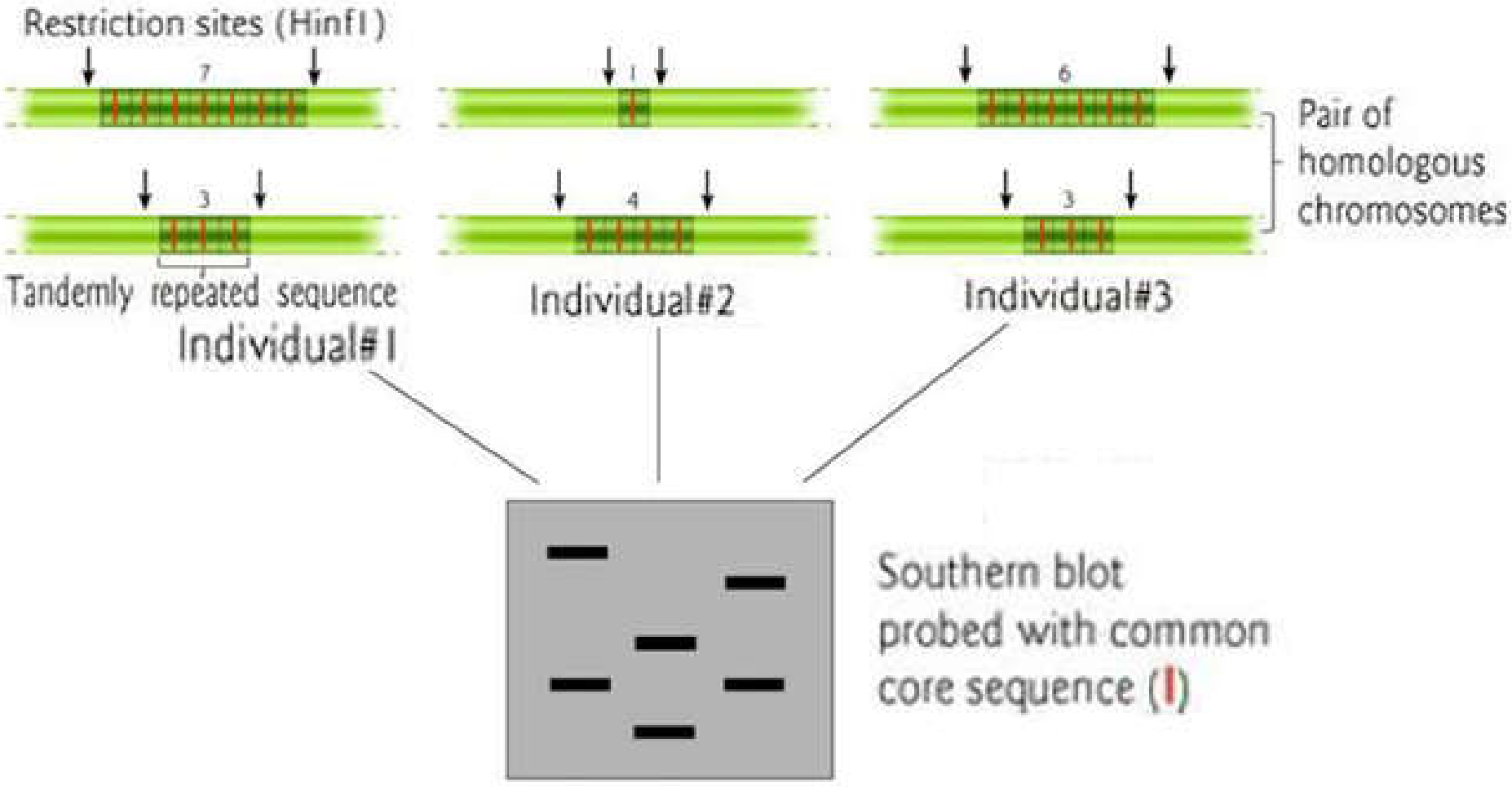
(3) Southern blot and hybridization (杂交)

Southern blot杂交



根据毛细管作用的原理，使在电泳凝胶中分离的DNA片段转移而结合在滤膜上，然后通过同DNA探针（荧光标记或放射同位素标记）的杂交作用检测这些被印迹的DNA片段。

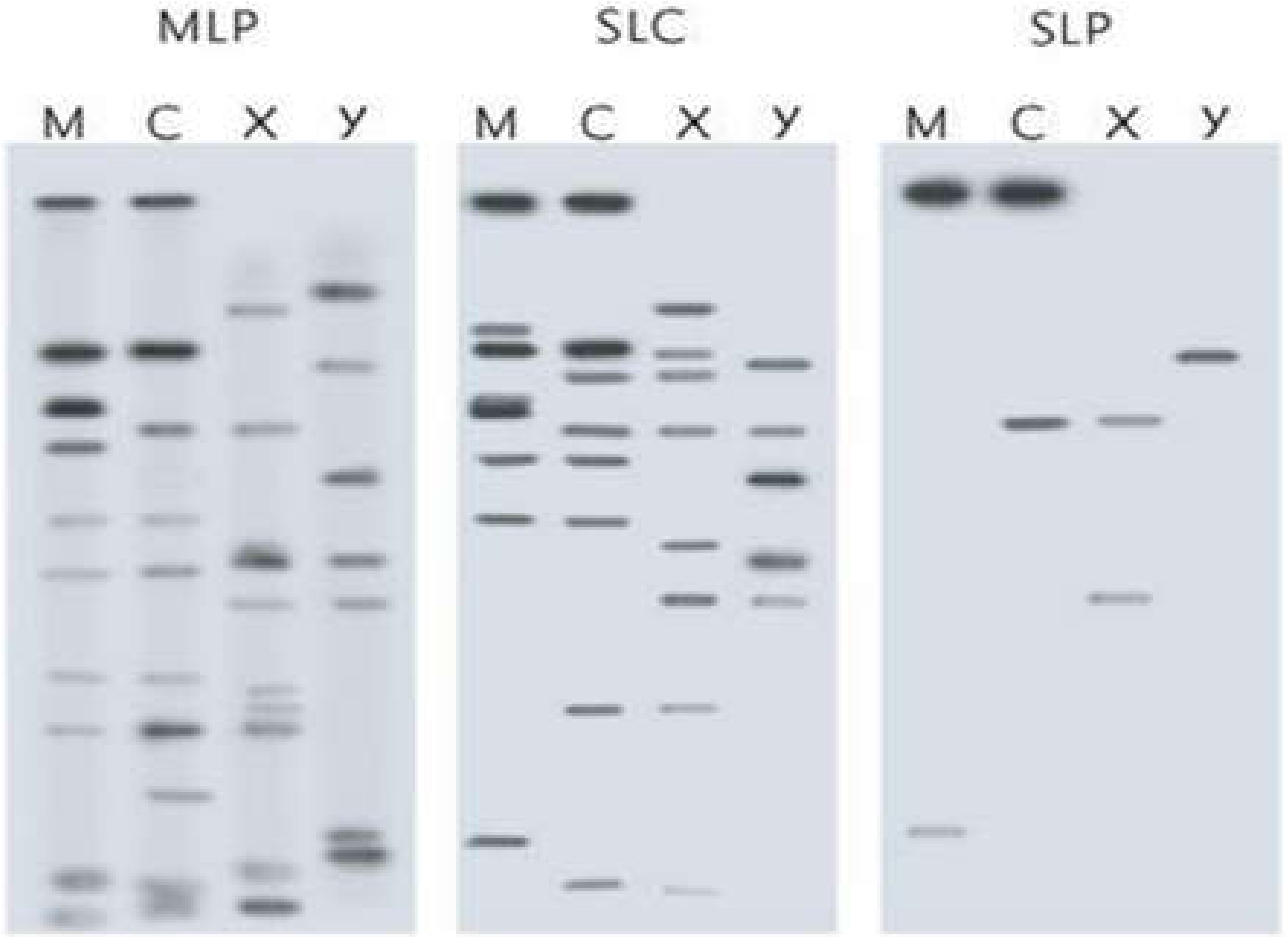
VNTR核心序列制备探针，酶切后进行Southern blot



多位点探针

单位点混合探针

单位点探针



M:母亲，C:孩子　X和Y谁是父亲?

4.7 Genetic polymorphism (遗传多态性)

- **Genetic polymorphism** – Base changes in a gene or a chromosomal locus create **multiple forms of that locus**.
  遗传多态性：基因或染色体基因座上的碱基变化使该**基因座产生多种形式**。
  - Genetic polymorphism can describe different alleles of a single copy gene in a single individual as well as the different sequences present in different individuals in a population.
- Genetic polymorphism is caused by **non-lethal mutation** (非致死性突变).

4.7.1 Single nucleotide polymorphisms

- **Single nucleotide polymorphism (SNP, 单核苷酸多态性) is a variation in a single nucleotide that occurs at a specific position in the genome.**
- **SNPs are caused by point mutation (点突变).**
- **SNPs are responsible for most of the genetic variation between individuals.**

- **SNPs in DNA fragments of the same length may be detected by gel electrophoresis (凝胶电泳) of restriction fragments (限制性片段) or PCR products.**

  - DNA fragments must be denatured.
  - The two separated strands can adopt specific conformations that depend on their nucleotide sequence. — single stranded conformational polymorphism (SSCP, 单链构象多态性)

4.7.2 Restriction fragment length polymorphism

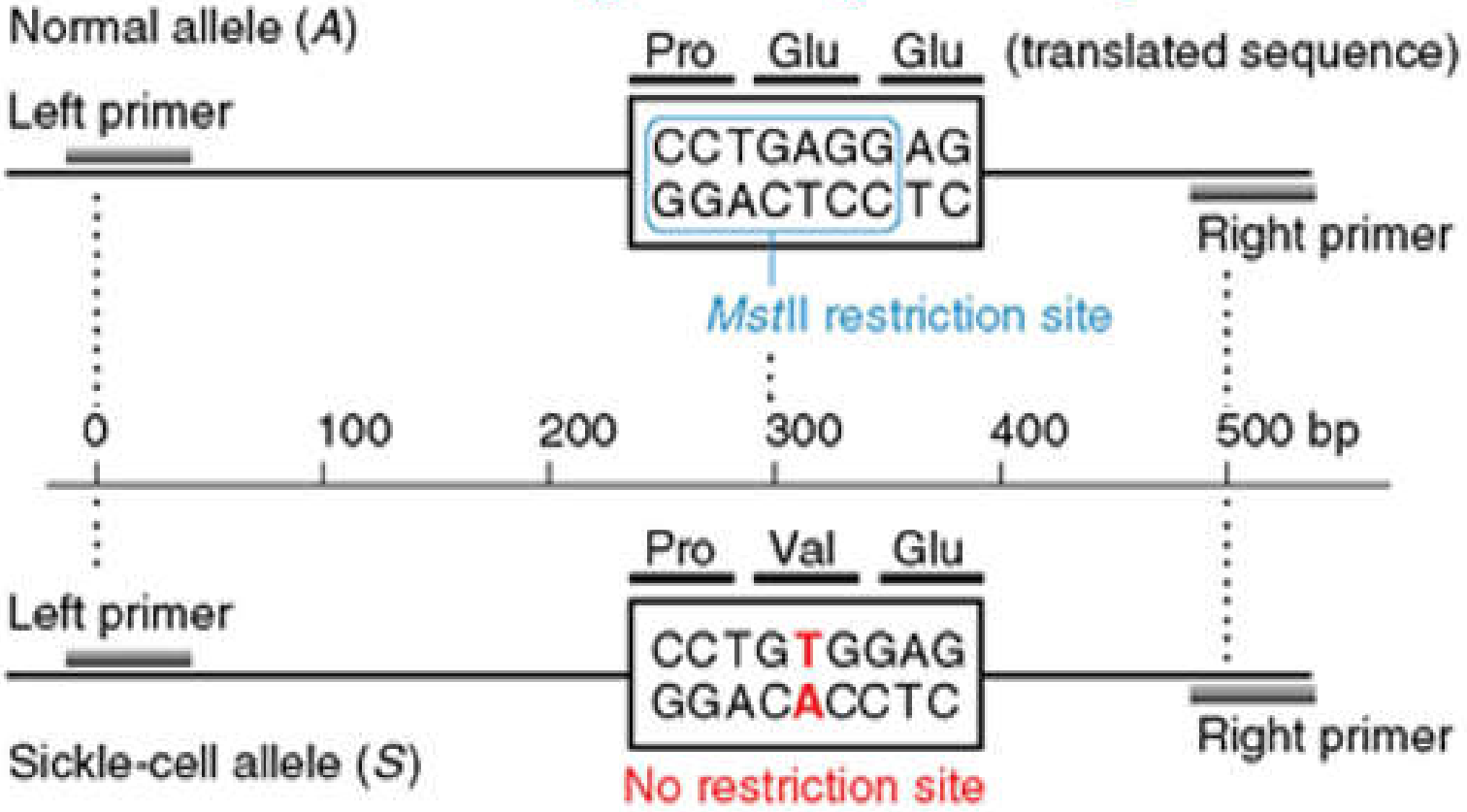
- **Restriction fragment length polymorphism (RFLP, 限制性片段长度多态性) – Inherited differences in sites for restriction endonuclease that result in differences in the lengths of the fragments produced by cleavage with the relevant restriction endonuclease.**

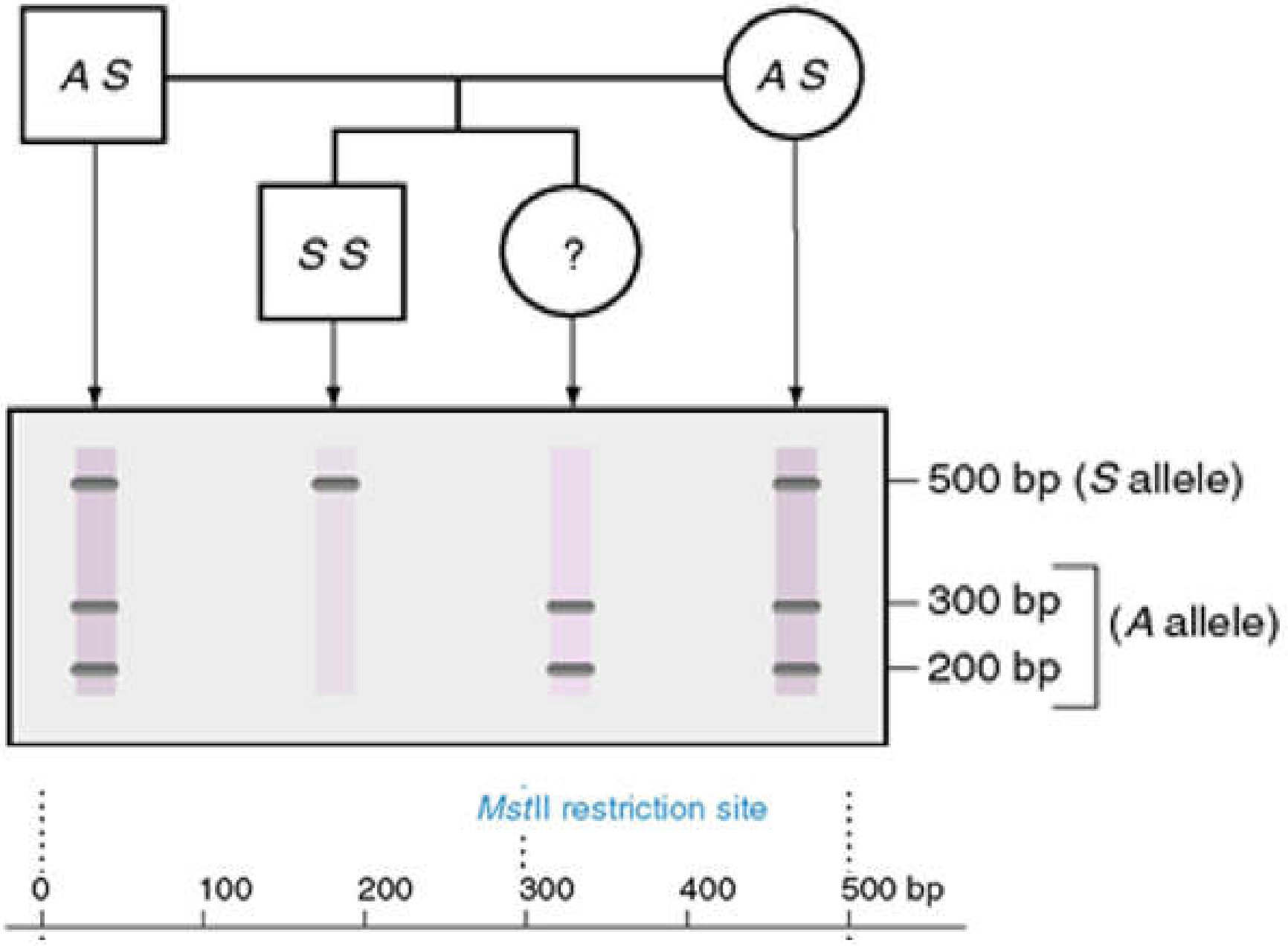
Applications of RFLP

(1) Disease diagnosis

➢ Sickle cell anemia（镰刀形红细胞贫血症）

PCR+restriction digestion+gel electrophoresis

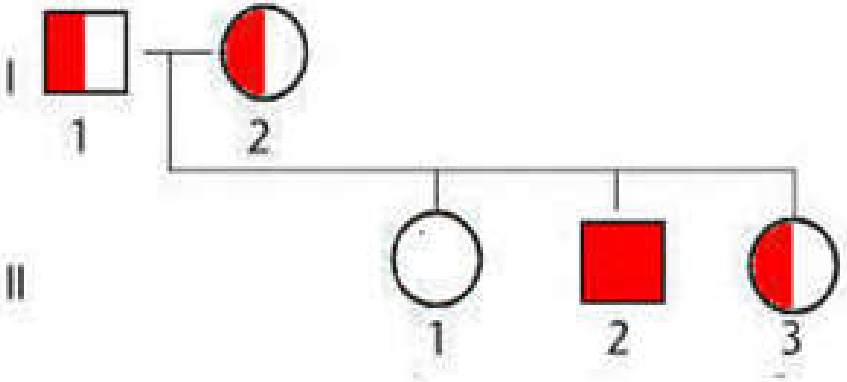


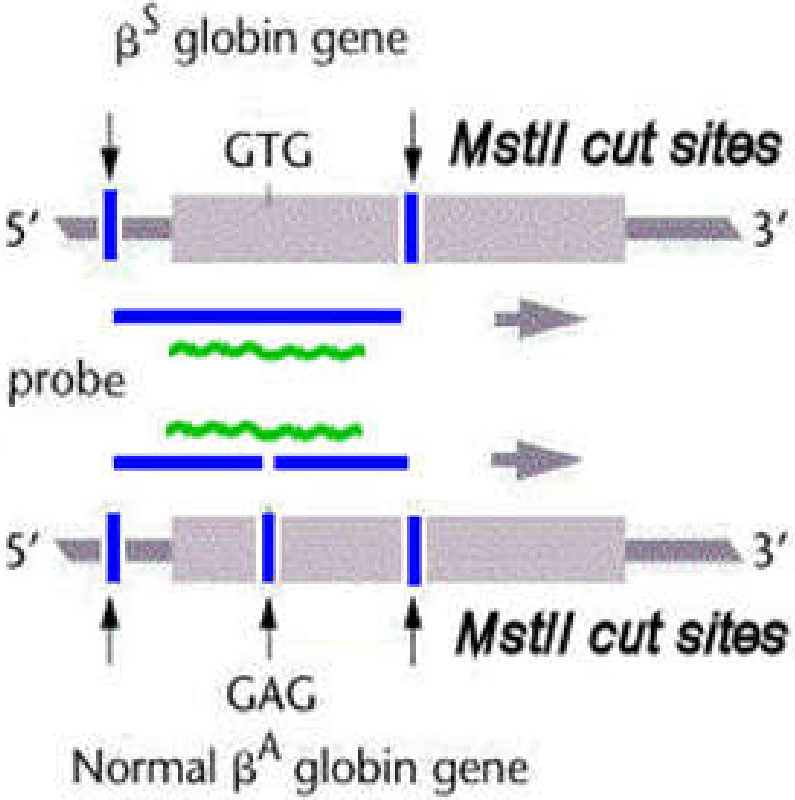

A S
A S
S S
?
500 bp (S allele)
300 bp
200 bp
(A allele)
MstII restriction site
0
100
200
300
400
500 bp

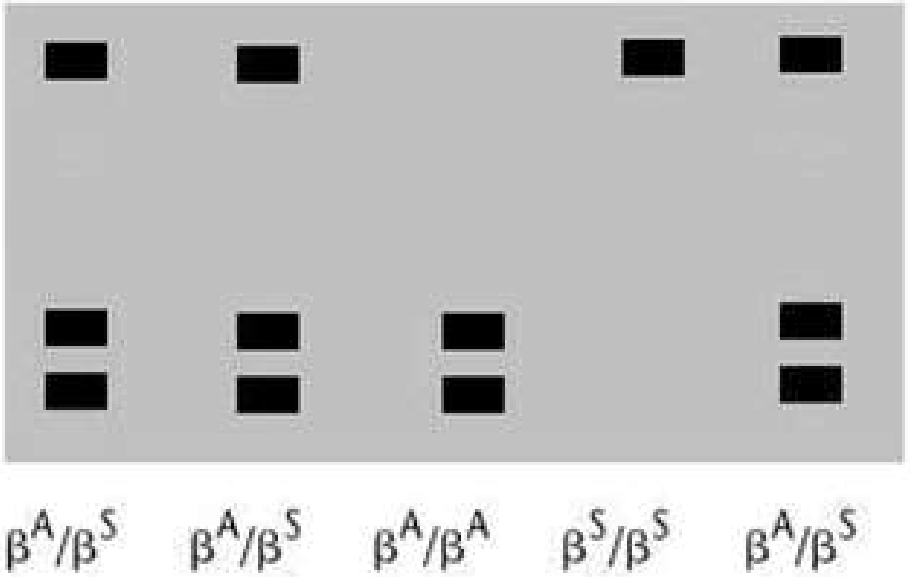
Restriction digestion+Southern blot+hybridization

β^A: CCTGAGGAG

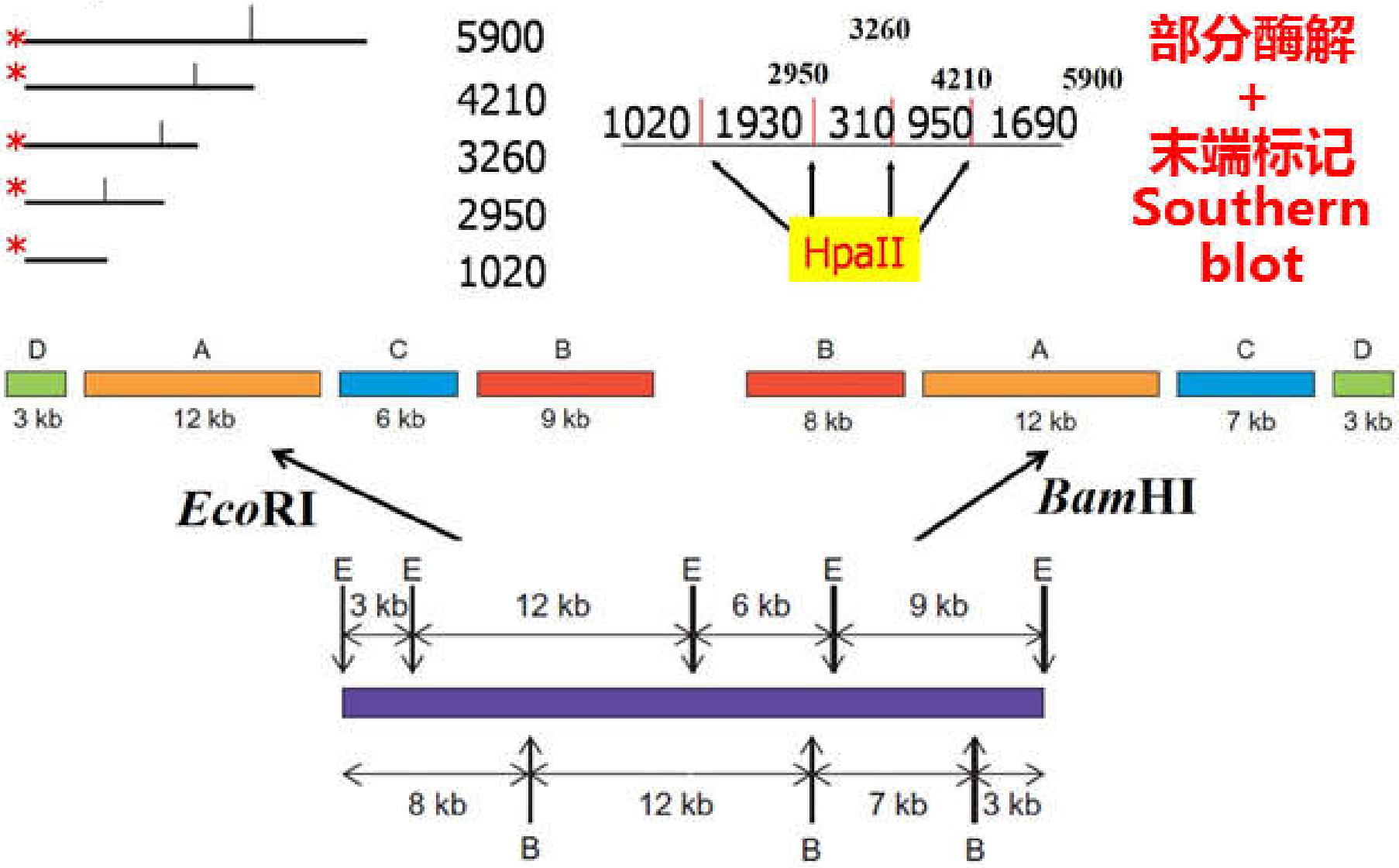
β^S: CCTGTGGAG







(2) Create genetic/physical maps of chromosomes



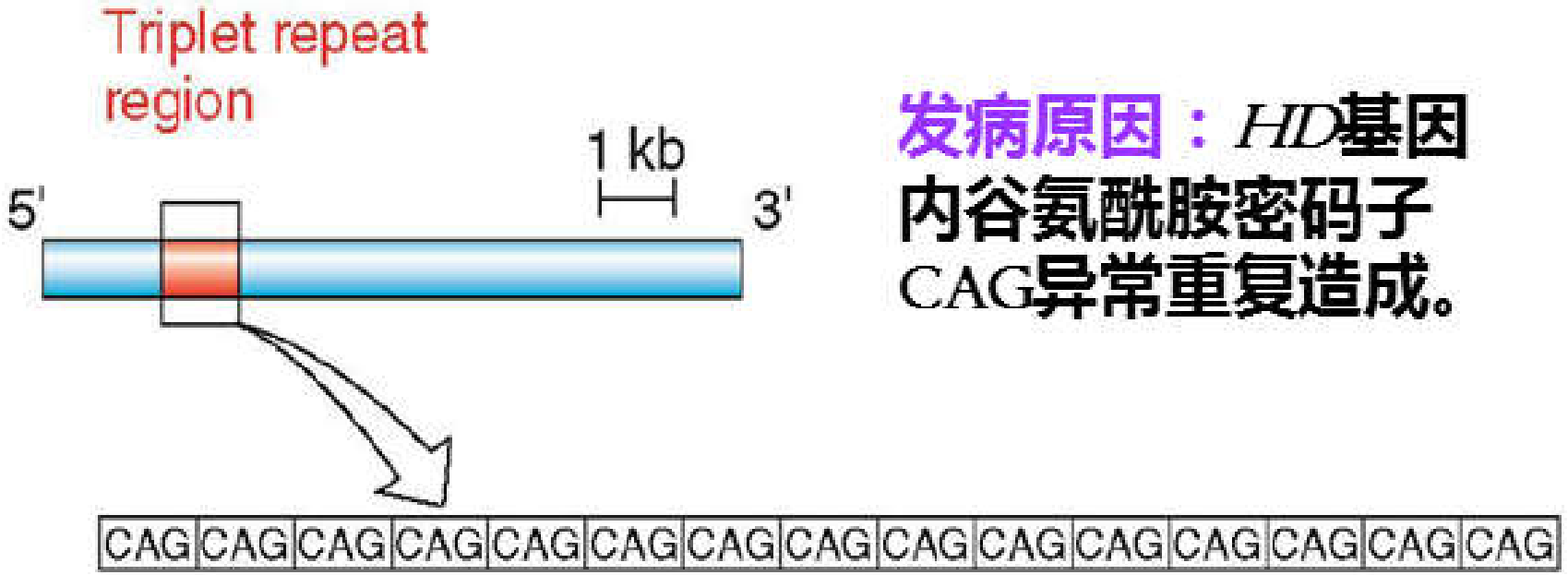
4.7.3 Simple sequence length polymorphism

- **Simple sequence length polymorphism (SSLP, 简单序列长度多态性) is a variation in the length of the repeated sequences.**
- **SSLPs are caused by insertion/deletion mutation.**

Huntington's disease（亨廷顿氏疾病）

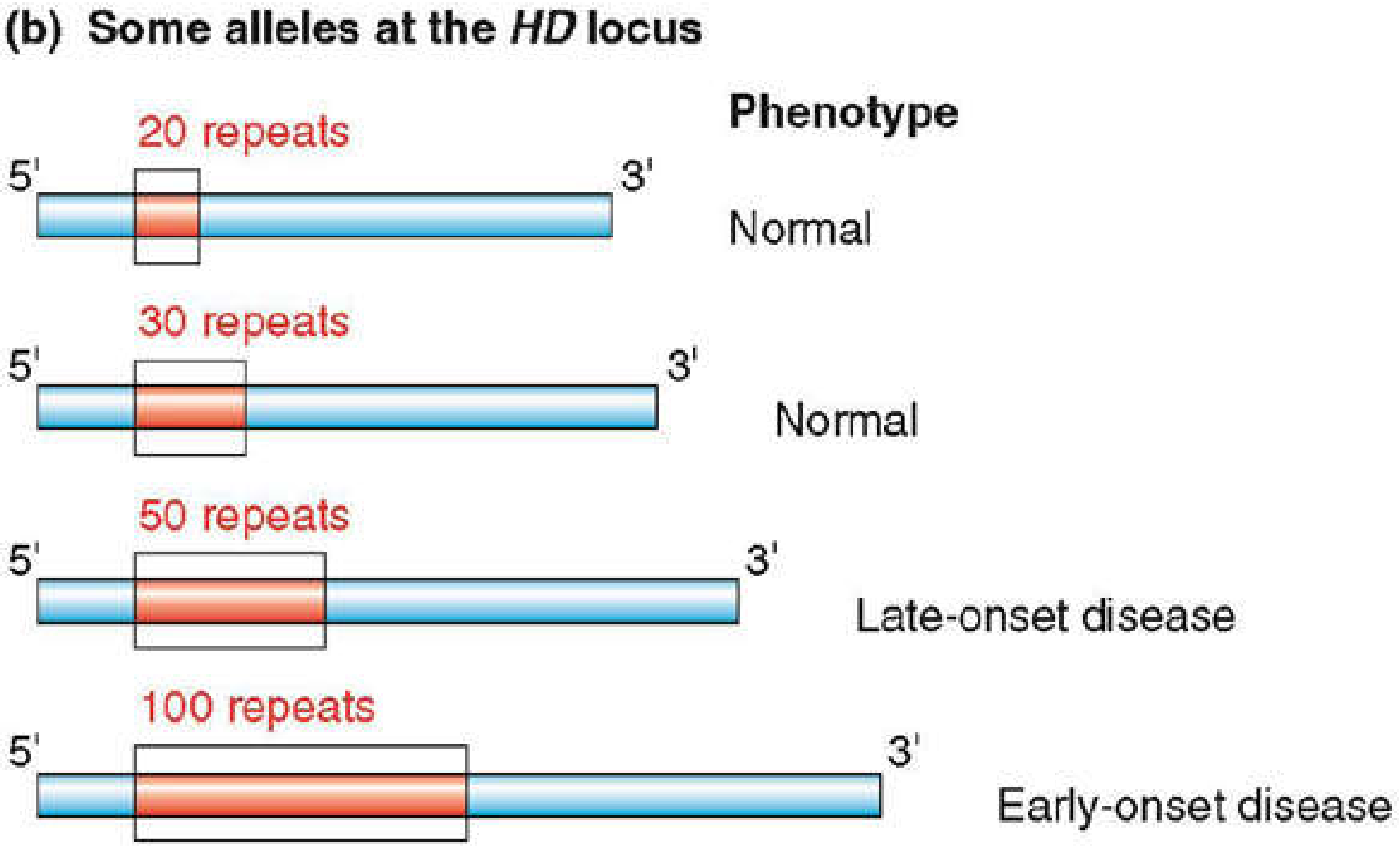
致病基因（*HD*）位于4号染色体上。

(a) Basic structure of the *HD* gene's coding region



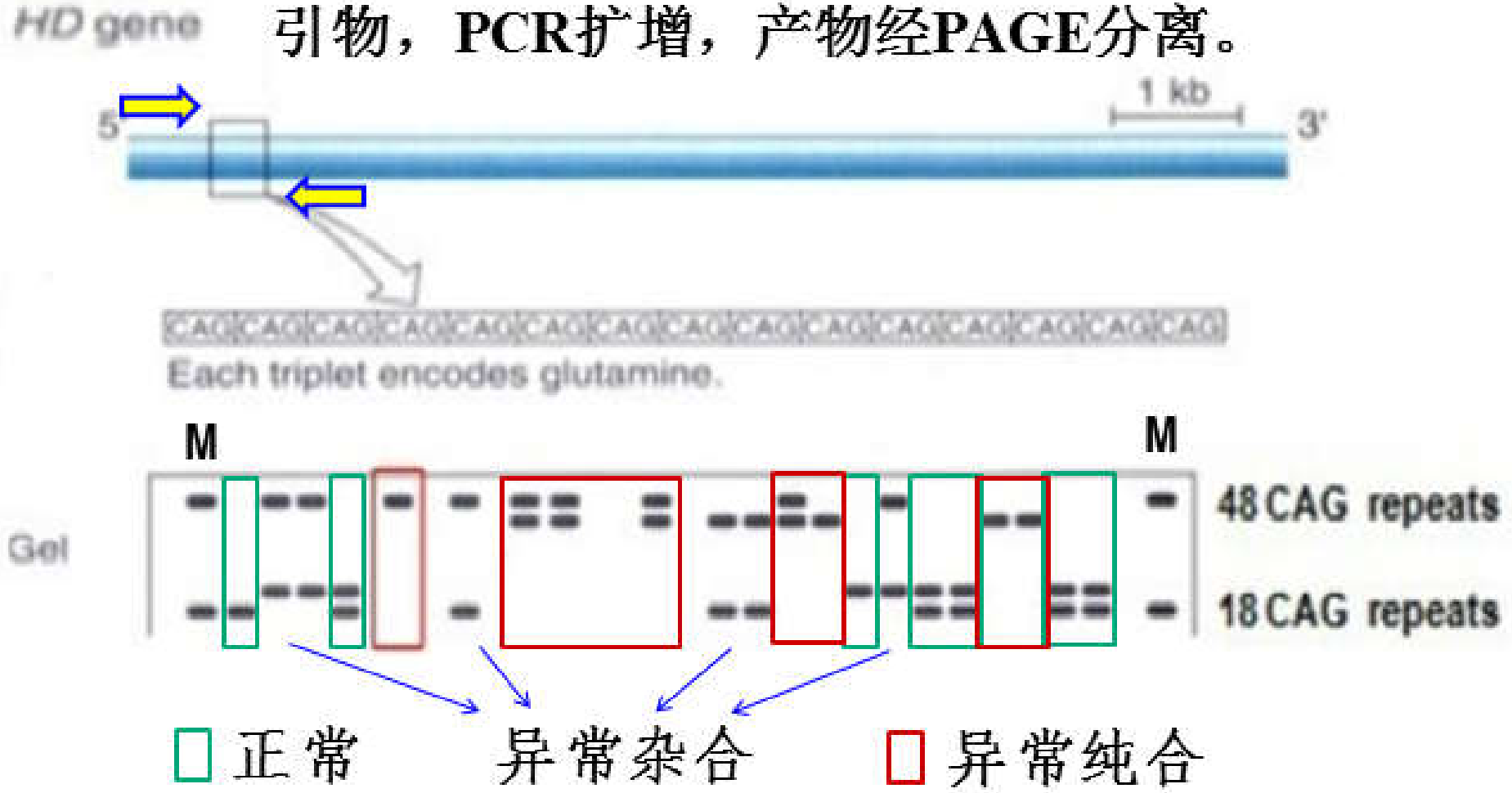
发病原因：*HD*基因内谷氨酰胺密码子CAG异常重复造成。

正常情况下，CAG重复11~34次。 >40会患亨廷顿氏病。重复次数越多，发病越早，症状越严重。



PCR+Polyacrylamide gel electrophoresis

根据HD基因CAG重复两侧的碱基序列设计引物，PCR扩增，产物经PAGE分离。



Summary

1. Characteristics of prokaryotic and eukaryotic genome
2. Chromosome, polycistronic mRNA, overlapping gene, chromatin, centromere, telomere, genome, C-value paradox, Genetic polymorphism, SNP, RFLP
3. The composition of nucleosomes
4. Structure and function of centromere and telomere
5. Differences between heterochromatin and euchromatin
6. Effects of CpG methylation and histone modification to the chromosomal structure and gene expression
7. Types and characteristics of eukaryotic DNA sequences