LECTURE 12: INSIGHTS FROM GENOME SEQUENCING

Read 337-345; p327 (ortholog)
Table 10.1
Fig. 10.14; 10. 15; 10.18, 10.19

Genome sequencing changed the practice of biology, genetics and genomics

1. High density molecular markers
   - facilitate gene mapping and cloning of disease genes
   - disease diagnosis, prevention, and cure
   - forensic, identity, defense etc.

2. Global insights into genome organization and structure
   - how much repeats/transposons

3. Comparative genomics/evolutionary insights
   ortholog vs. paralog

4. Facilitate understanding related genomes

5. Facilitate gene expression and functional analyses
   - discover noncoding RNA/RNA splicing/protein coding

Insights from genome sequencing

Comparison of total gene numbers in sequenced genomes:

Near constant number of genes in all genomes irrespective of genome sizes

25,000 Arabidopsis
20-25,000 human
19,099 in C. elegans
13,600 in Drosophila

Smaller than originally expected

Human genome thought to have 100,000 genes
Now thought to be closer to 20,000–25,000 genes

Many new functions arise in gene expression

- -Use old genes in new ways
- - Alternative splicing

Selective expansion of genes (paralogs)

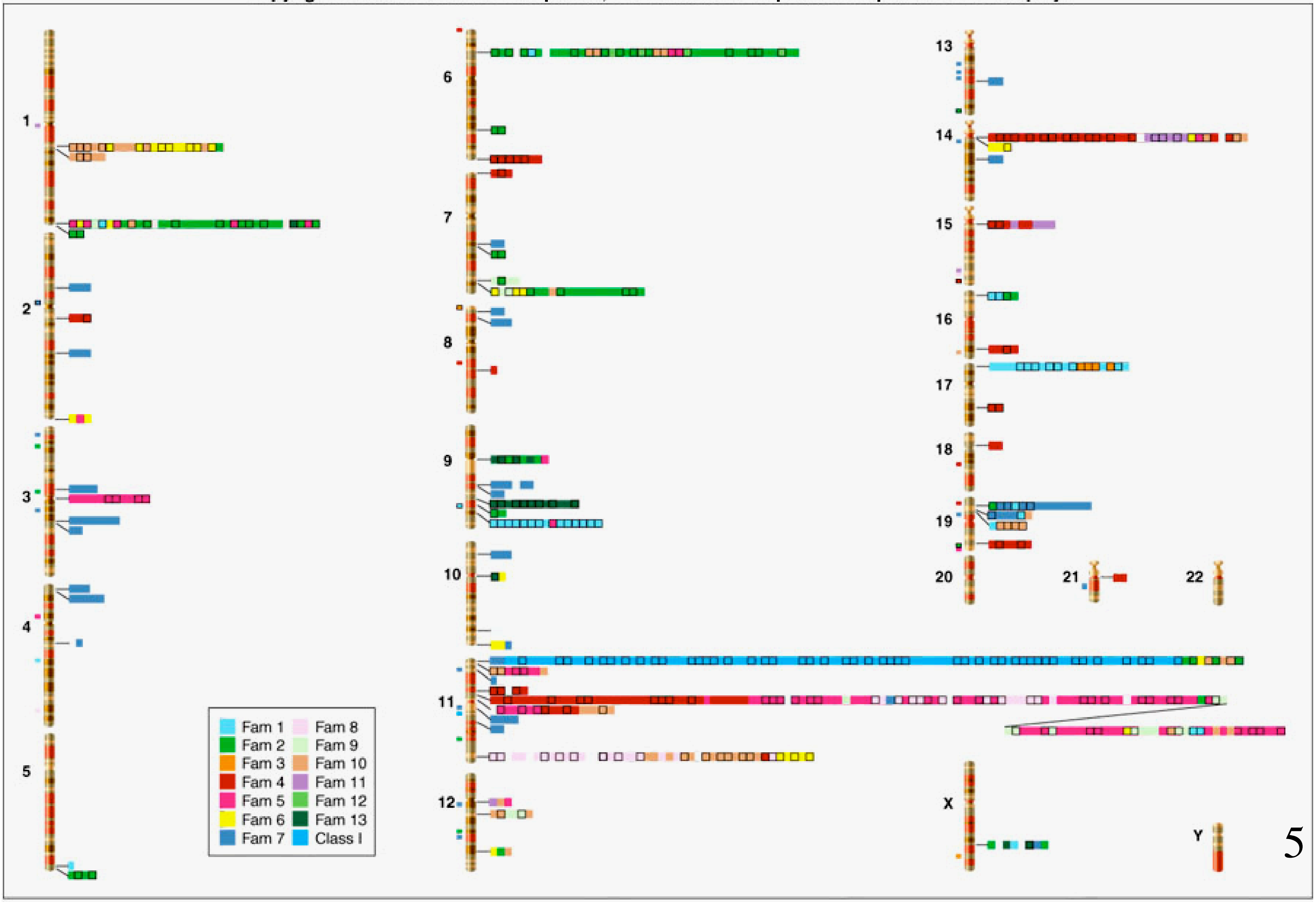
- –Roundworm, *C. elegans*, has a large number of nuclear receptor genes
- –*Drosophila* has a large number of zinc-finger transcription factors
- –Plants have no *G*-protein-coupled receptors

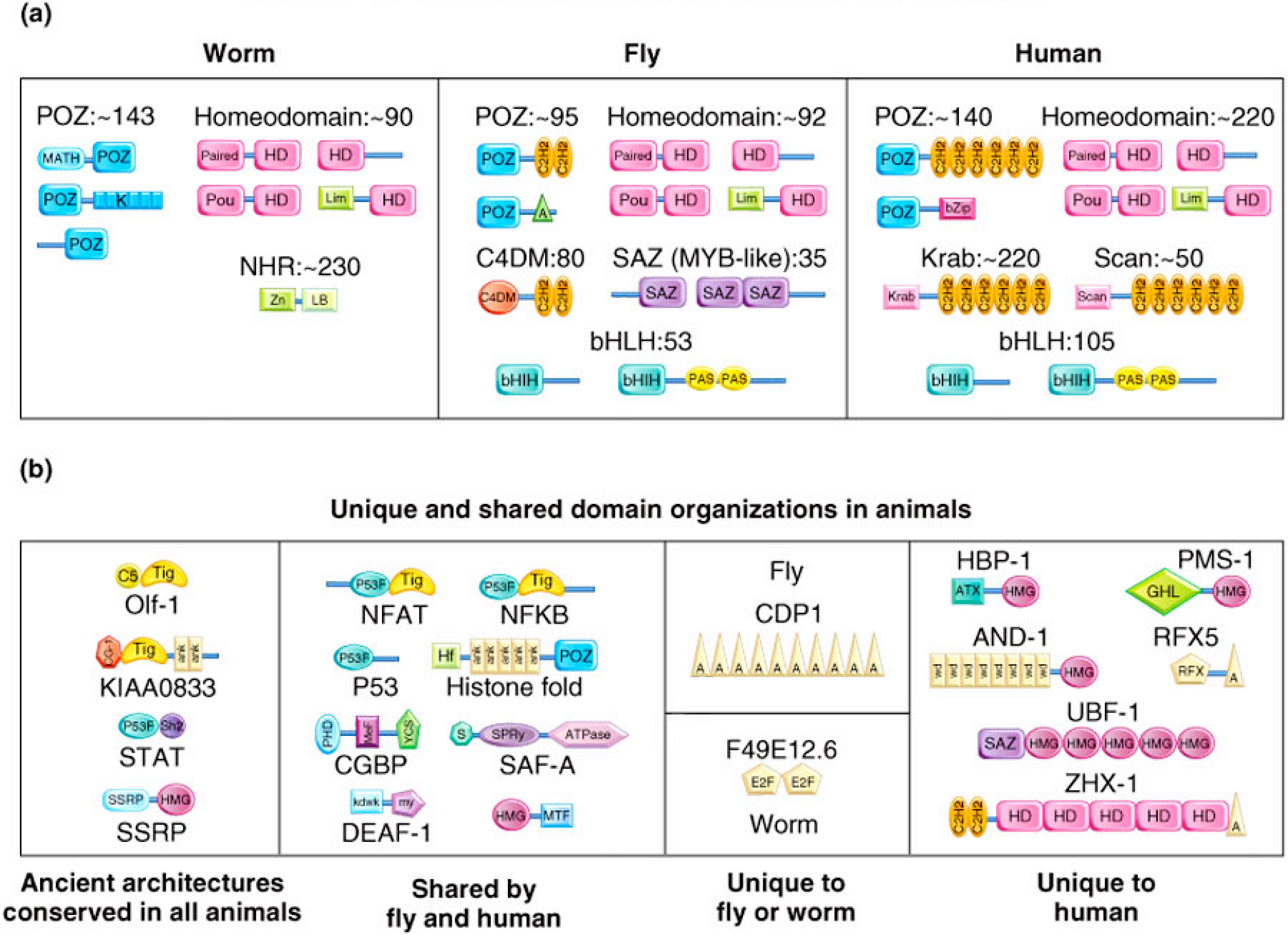
Different shuffling of discrete functional units (ie. protein domains)

-Each protein contains different combinations of protein domains. Protein composition may change with evolution

Fig. 10.19

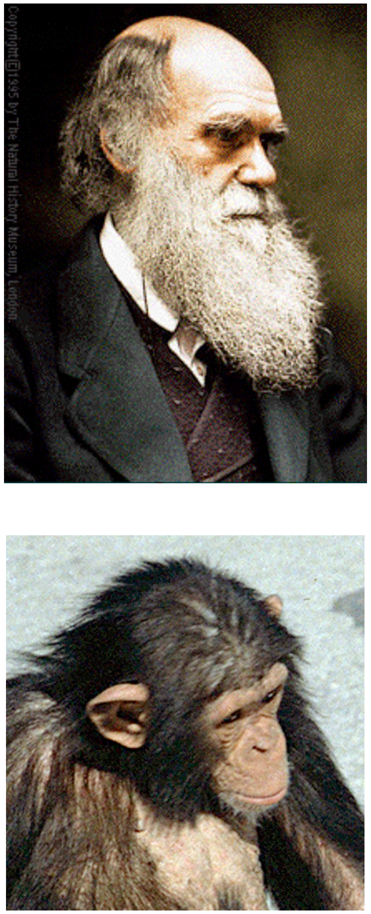
Olfactory gene families




(a)
Worm
Fly
Human
POZ:~143
MATH
POZ
K
Homeodomain:~90
Paired
HD
Pou
Lim
NHR:~230
Zn
LB
POZ:~95
C2H2
A
Homeodomain:~92
C4DM:80
C4DM
SAZ (MYB-like):35
SAZ
bHLH:53
bHlH
PAS
POZ:~140
bZip
Homeodomain:~220
Krab:~220
Krab
Scan:~50
Scan
bHLH:105
(b)
Unique and shared domain organizations in animals
C5
Tig
Olf-1
ank
KIAA0833
P53F
Sh2
STAT
SSRP
HMG
SSRP
NFAT
NFKB
P53
Hf
Histone fold
PHD
CGBP
S
SPRy
ATPase
SAF-A
kdwk
my
DEAF-1
MTF
Fly
CDP1
A
F49E12.6
E2F
Worm
HBP-1
ATX
PMS-1
GHL
AND-1
RFX5
RFX
UBF-1
ZHX-1
Ancient architectures conserved in all animals
Shared by fly and human
Unique to fly or worm
Unique to human

What is the difference between man and ape?

- Man and chimpanzee have a genomewide similarity of greater than 95%.
- What accounts for differences between species?
- Recent study suggests that differences between species are due to specific gene expression differences
  - Striking differences found only in brain



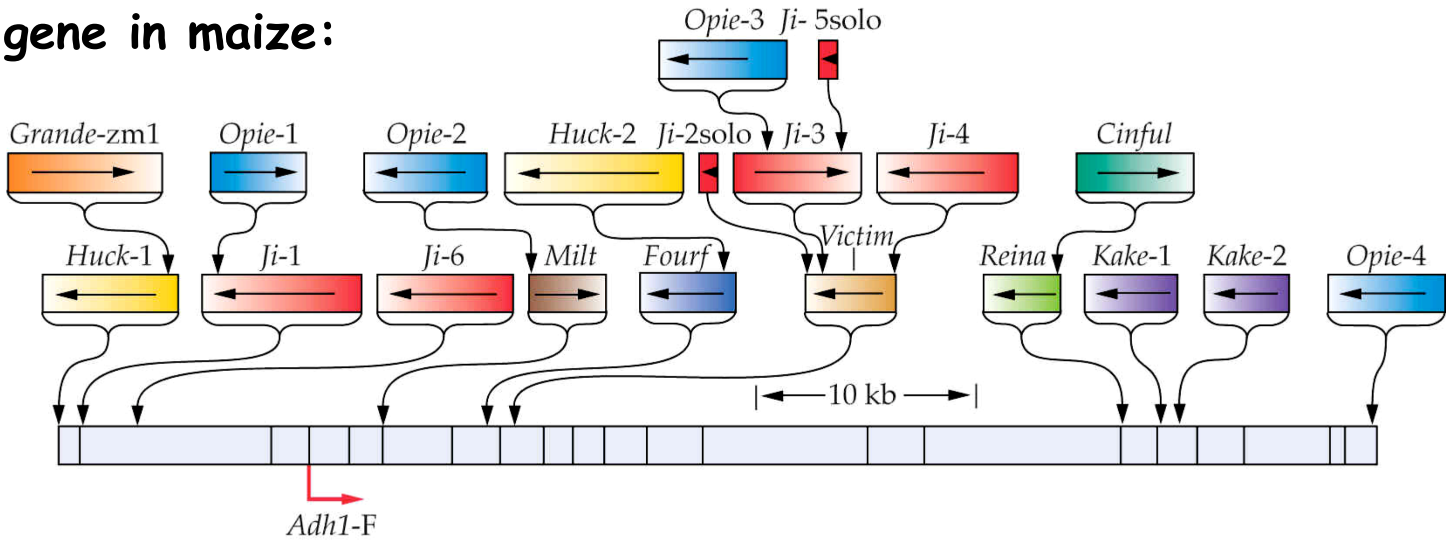
From Genomics by Benfey and Protopapas 2005

The C-value paradox

The bigger a genome, the more repetitive DNA

Arabidopsis:	1X 10^5 kb (14%)
Tomato:	1X 10^6 kb (15-20%);
Mung Bean:	4.5X10^5 kb (30%)
Pea:	4.1X 10^6 kb (70%)
Wheat, Corn	10^7 kb (60-80%)

-Adh1 gene in maize:

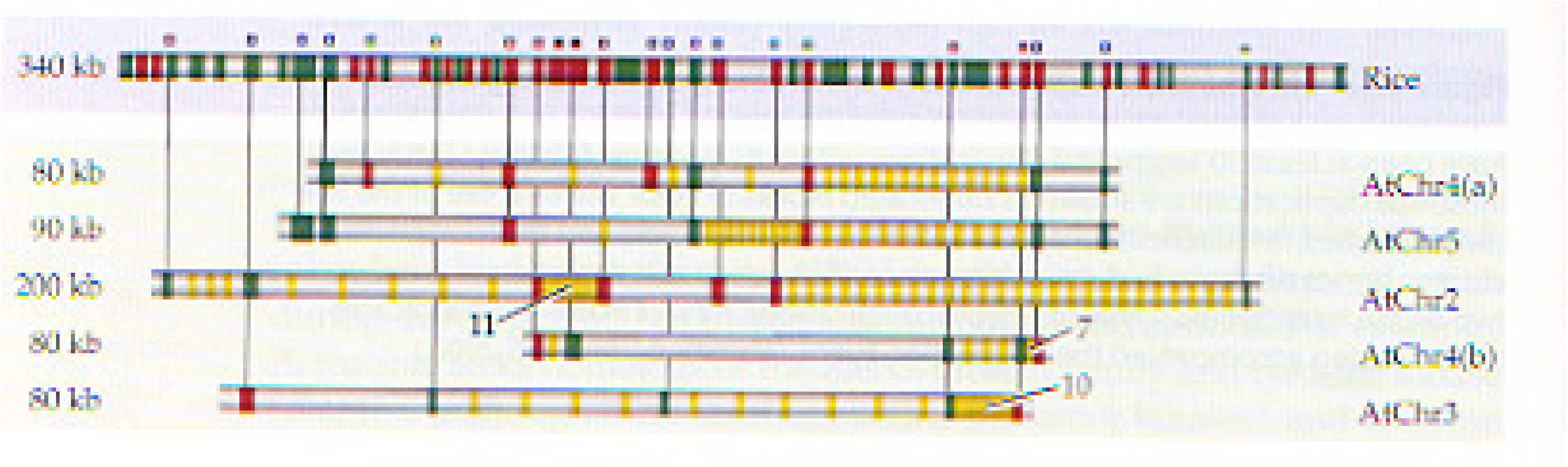


Comparative genomics

- **Synteny**: genes that are in the same relative position on two different chromosomes
- Genetic and physical maps compared between species
  - Or between chromosomes of the same species
- Closely related species generally have similar order of genes on chromosomes
- Synteny can be used to identify genes in one species based on map position in another

Synteny: Colinearity of loci (genes) among different plant species

i.e. Revolutionarily conserved organization and arrangement of single copy genes

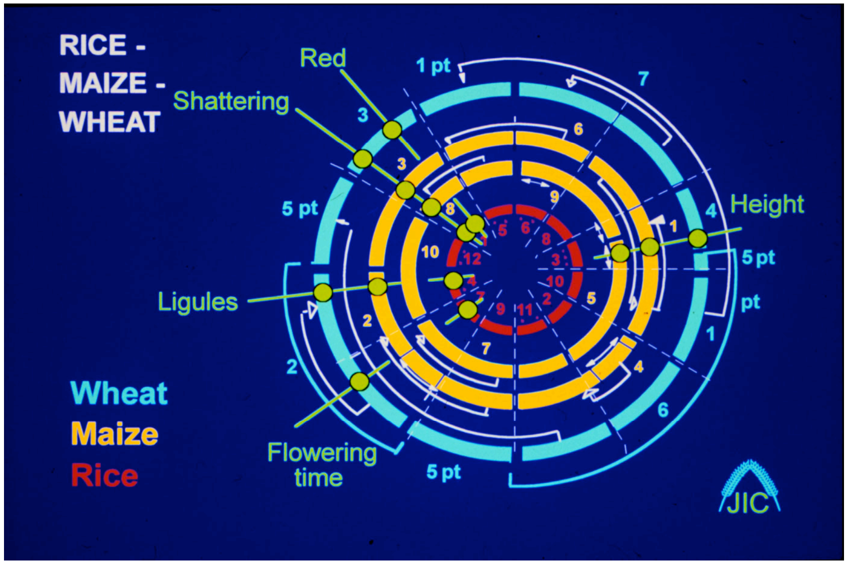


20 of the 54 genes in a 340 kb stretch of the rice genome (top) retain the same order in five different 80-200 kb regions of Arabidopsis genome

🟥 🟩 genes on different strands
🟨 interspersed, unrelated genes

Synteny of Grass genomes

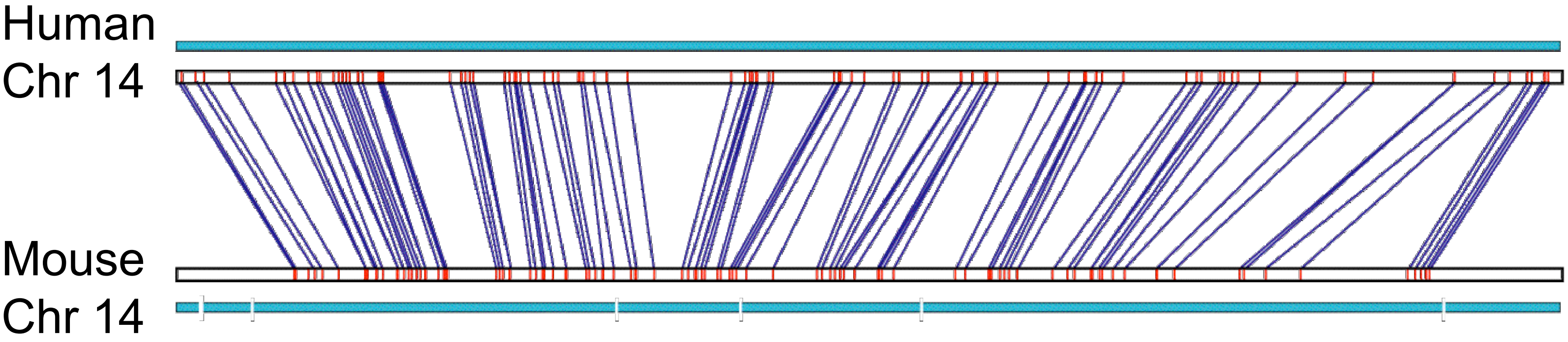
- Synteny among crop genomes: rice, maize, and wheat
- Rice is smallest genome–in center
- Wheat is largest genome–outer circle
- Genes found in similar places on chromosomes are indicated



From Genomics by Benfey and Protopapas 2005

Synteny of sequenced genomes

- When sequences from mouse and human genomes are compared, we find regions of remarkable synteny
- Genes are in almost identical order for long stretches along the chromosome



From Genomics by Benfey and Protopapas 2005

Mouse–human synteny

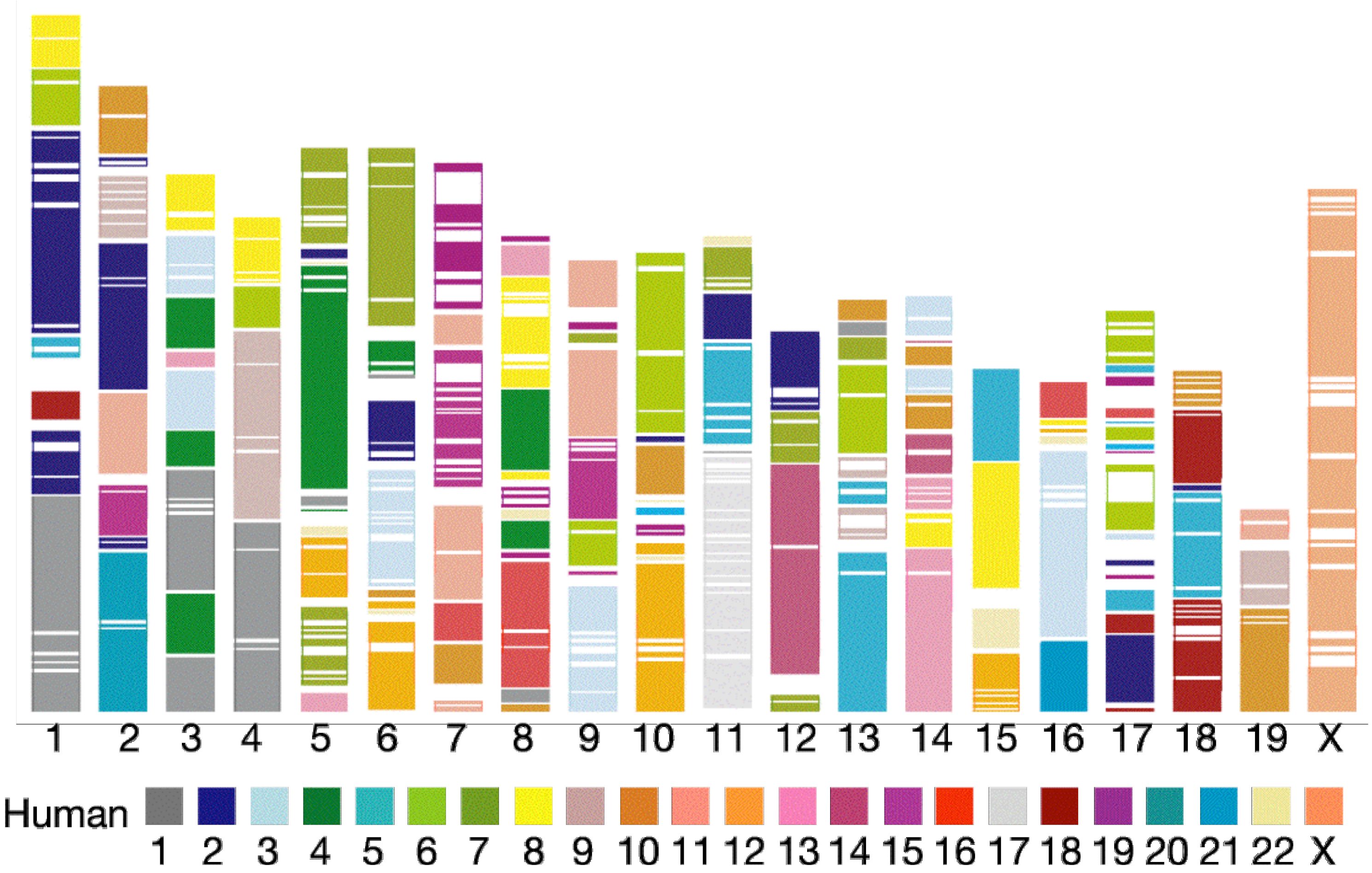


Fig. 10.15

Orthologs and Paralogs

- When comparing sequence from different genomes, must distinguish between two types of closely related sequences
  - Orthologs are genes found in two species that had a common ancestor
  - Paralogs are genes found in the same species that were created through gene duplication events

Arabidopsis thaliana (www.arabidopsis.org)

Genome sequence completed in 2000, published in 5 installment

See "Arabidopsis Genome Intiative, 2000 (pdf)"

-115 Mb, 25,500 predicted genes,

-Whole genome duplication 2X followed by extensive shuffling of chromosomal regions and gene loss

-The majority of the genes can be assigned to just 11,000 families, which might represent the minimal complexity or "toolkit" to support complex multicellularity. Animal and plant genomes might evolve from this toolkit

-Distinctive features of plant genome:

~ 800 genes are of plastid decent

~10% genome are transposable elements

~ plant specific genes:

Enzymes for cell wall biosynthesis, photosynthesis, secondary metabolites

Photptrophic, gravitrophic

Transport proteins for nutrient, ion, toxic compound, metabolites between cells

Pathogen resistant genes

http://www.ncbi.nlm.nih.gov/genome/guide/human/

Human genome: 3200 Megabases
20-40,000 genes

Proteome: The collective translation of the 30,000 predicted genes into proteins

Gene families: 1200
92 or 7% are vertebrate-specific
(involved in immunity, defense, nervous system)

Repeats in the human genome: = >50%

Evidence of lateral gene transfer

Males have more than two fold mutation in meiosis over female

Different human races are genetically a single race

All living organisms evolve from a common ancestor

Repeats in the human genome = >50%

45% = transposon derived
- LINES (Long interspersed elements)
- SINES (Short interspersed elements)
- LTR-retrovirus
- DNA transposons

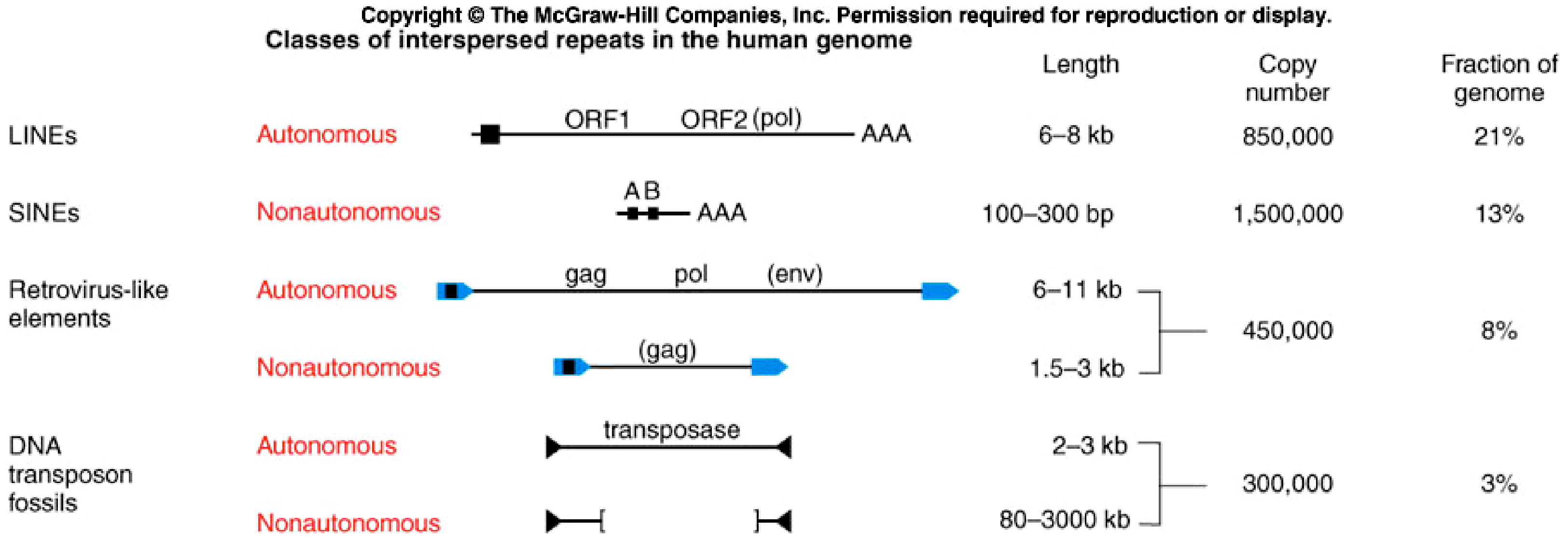
Pseudogenes

Simple sequence repeats

Segment duplication (10-300 kb) ~ >5%

Centromere and telomere repeats

Fig. 10.18



Classes of interspersed repeats in the human genome

		Length	Copy number	Fraction of genome
LINEs	Autonomous	6–8 kb	850,000	21%
SINEs	Nonautonomous	100–300 bp	1,500,000	13%
Retrovirus-like elements	Autonomous	6–11 kb	450,000	8%
	Nonautonomous	1.5–3 kb		
DNA transposon fossils	Autonomous	2–3 kb	300,000	3%
	Nonautonomous	80–3000 kb		