LECTURE 12: INSIGHTS FROM GENOME SEQUENCING

Read Chapter 12 (p500-523)

p724 (ortholog vs paralog)

DOE's genomics and its impact

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
20-30,000 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 20,000–30,000 genes

How is the diversity generated with limited number of genes?

Many new functions arise in gene expression

- Alternative splicing
-Chemical modifications to the proteins
-Noncoding RNAs

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

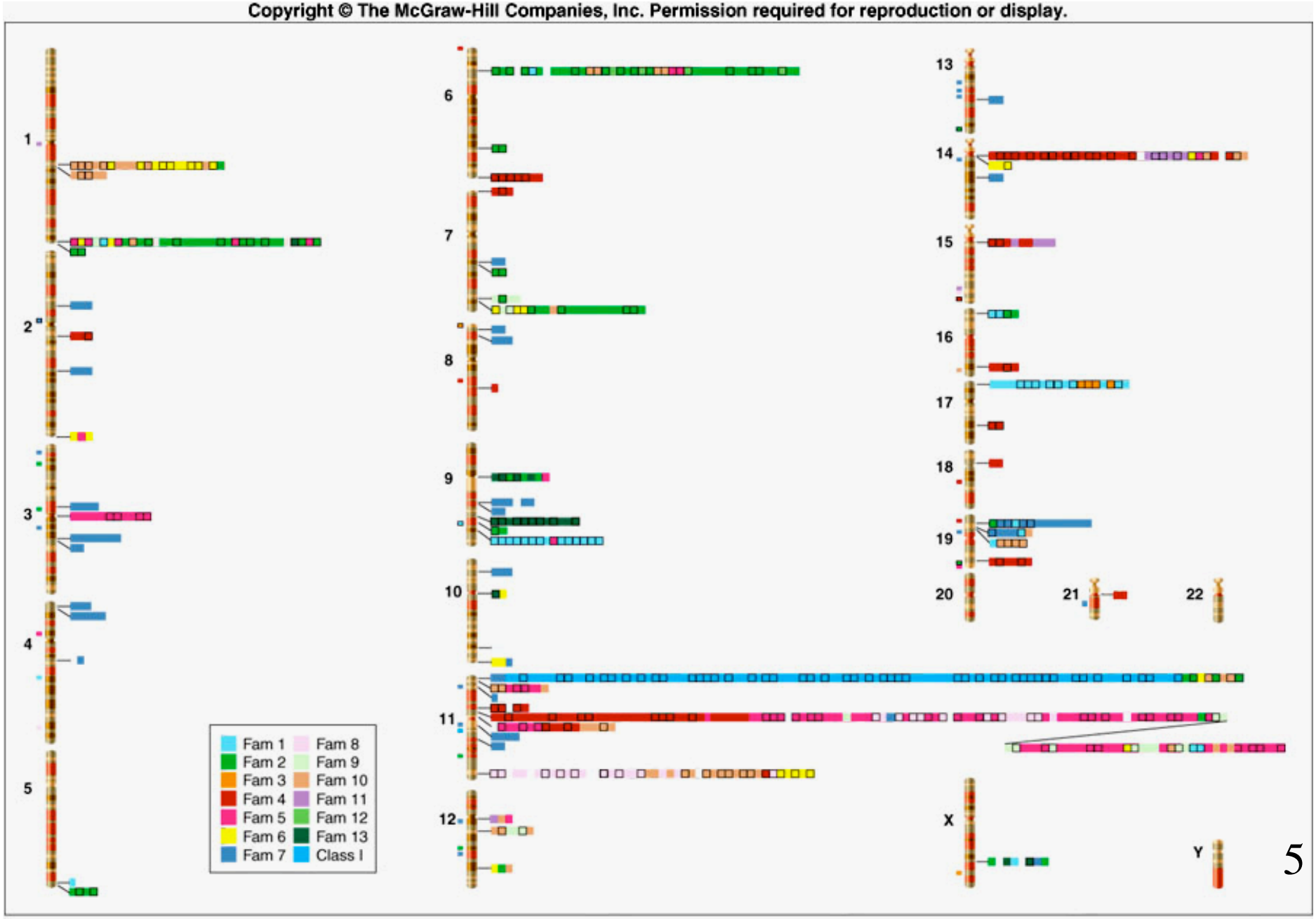
–Plants have no *G*-protein-coupled receptors

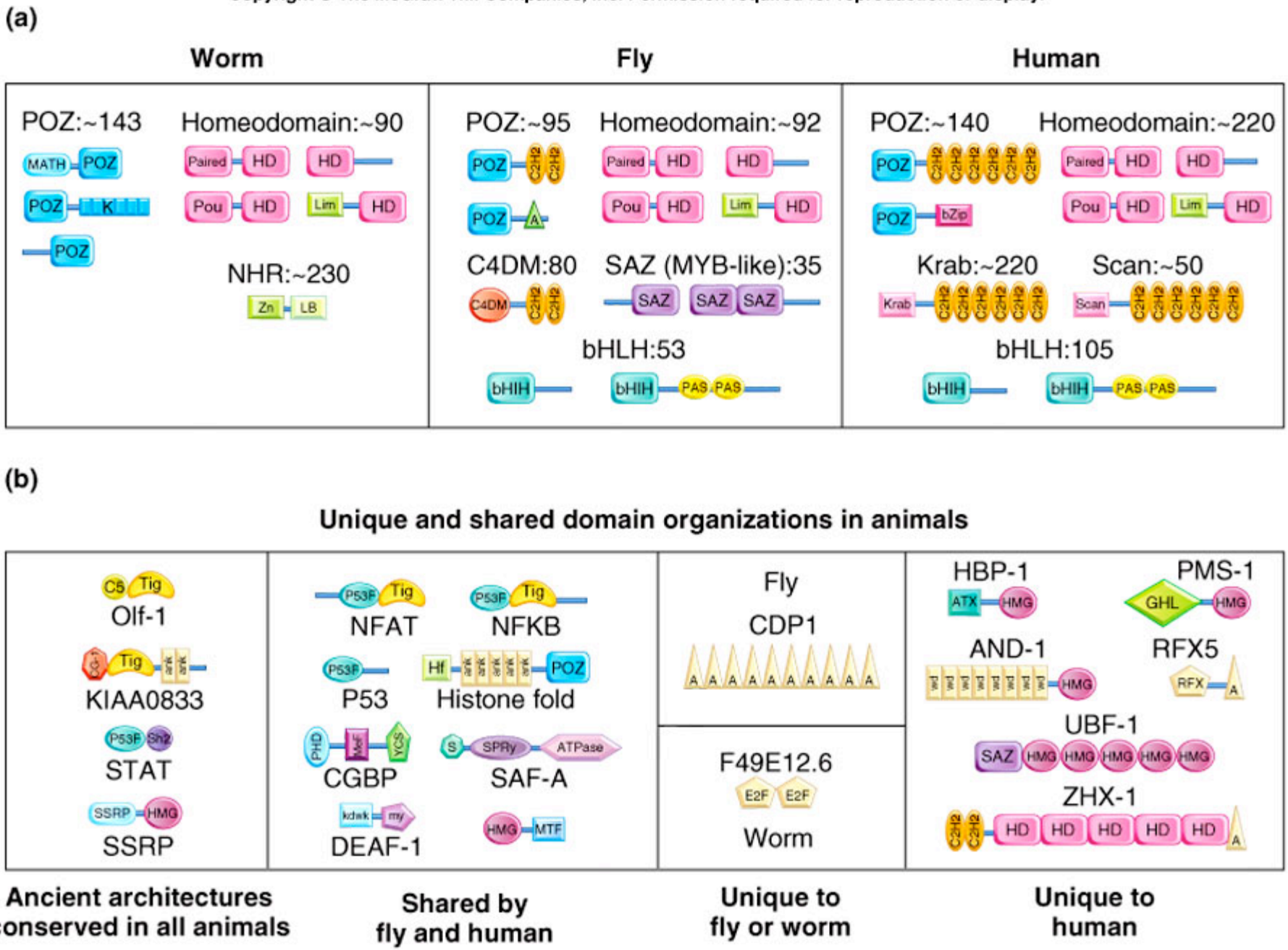
–Olfactory gene family

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

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

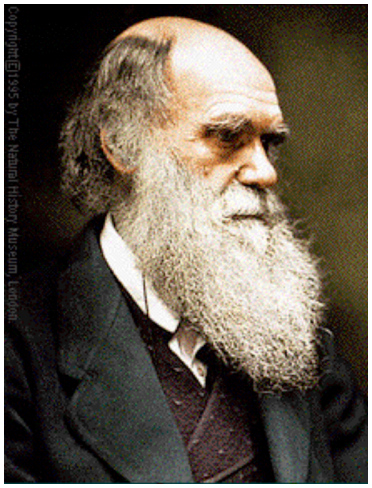
Olfactory gene families





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



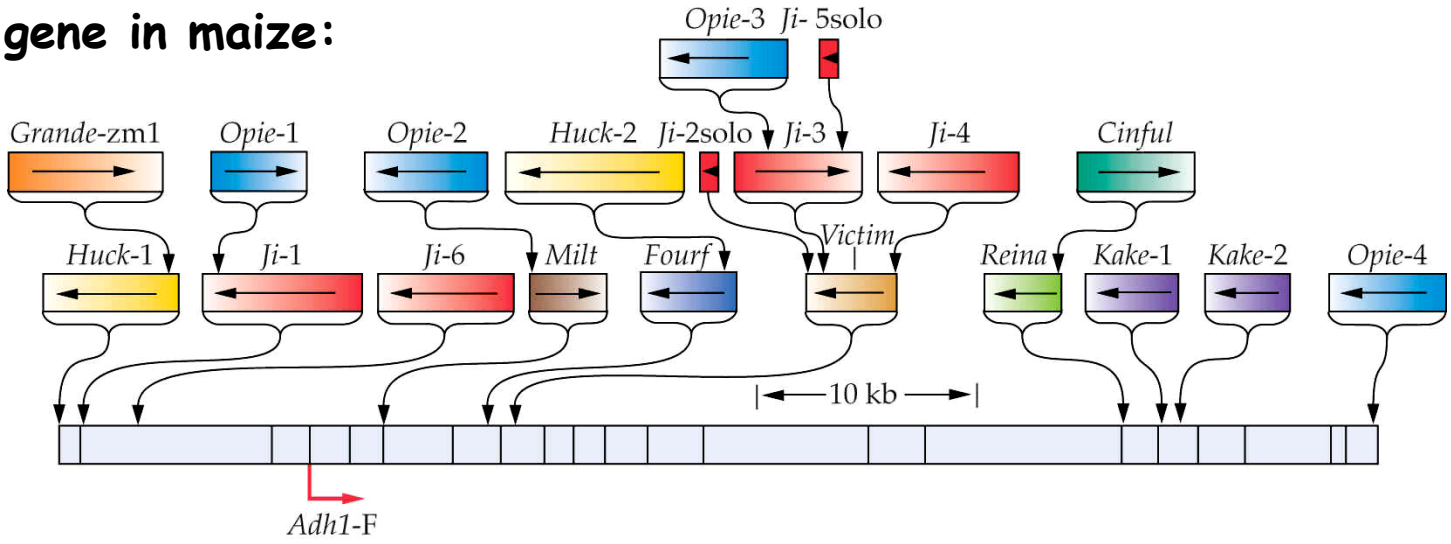
From Genomics by Benfey and Protopapas 2005

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:

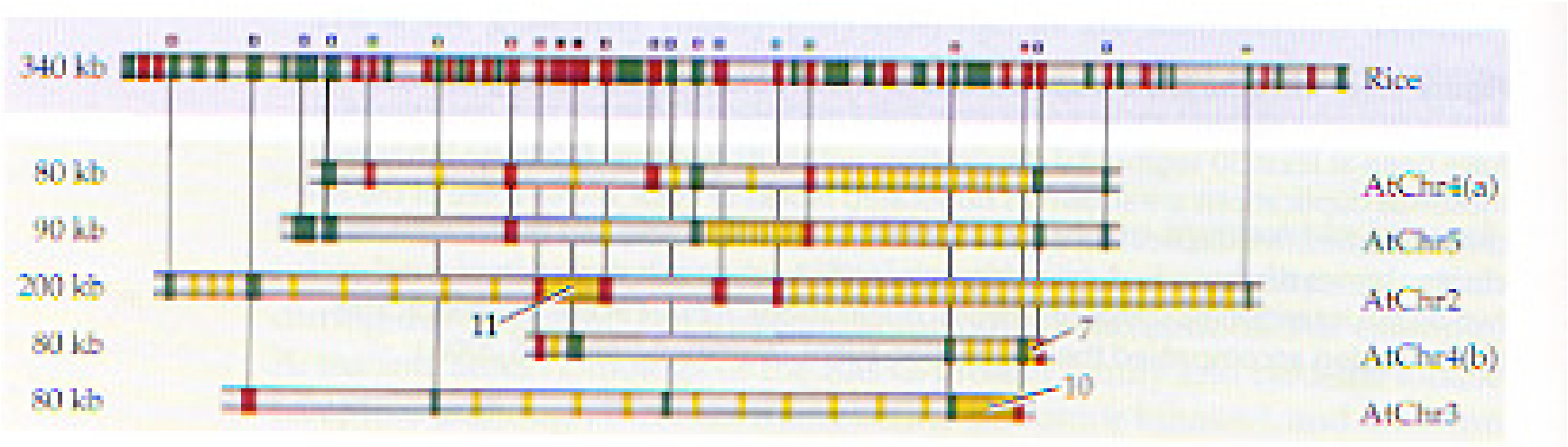


Comparative genomics

- **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

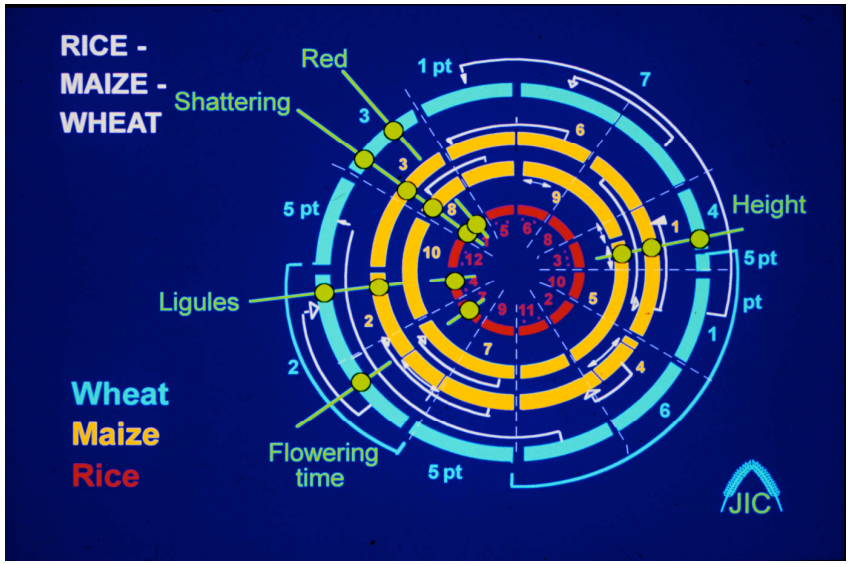


20 of the 54 genes in a 340 kb stretch of the rice genome (top) retain the same order in five different 80-200 kb regions of Arabidopsis genome

- genes on different strands
- interspersed, unrelated genes

Synteny of Grass genomes

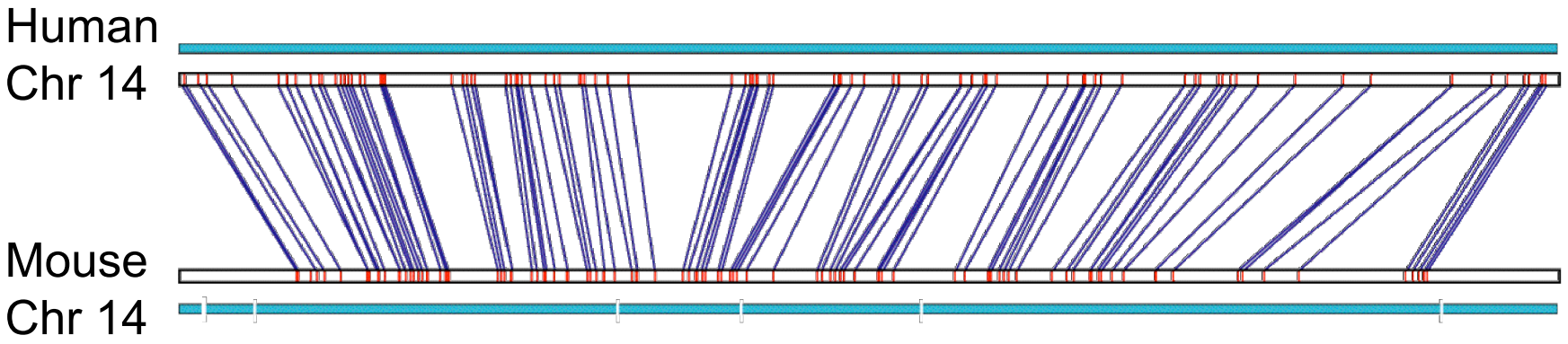
- Synteny among crop genomes: rice, maize, and wheat
- Rice is smallest genome–in center
- Wheat is largest genome–outer circle
- Genes found in similar places on chromosomes are indicated



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

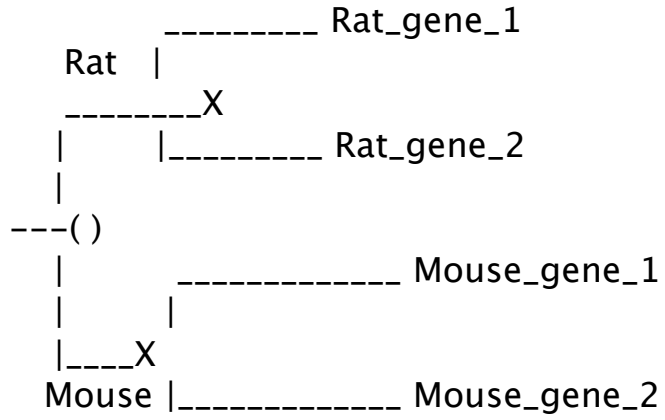


From Genomics by Benfey and Protopapas 2005

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

* **Two genes are to be orthologous if they diverged after a speciation event,**
* **Two genes are to be paralogous if they diverged after a duplication event.**



Mouse_gene_1 and Mouse_gene_2 are paralogous,
Rat_gene_1 and Rat_gene_2 are paralogous

Rat_gene_1 is orthologous to Mouse_gene_1 and to Mouse_gene_2
Rat_gene_2 is orthologous to Mouse_gene_1 and to Mouse_gene_2
Mouse_gene_1 is orthologous to Rat_gene_1 and to Rat_gene_2
Mouse_gene_2 is orthologous to Rat_gene_1 and to Rat_gene_2

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

Human Genome Project :1990-2003

Human genome: 3200 Megabases
20-30,000 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

What does the draft human genome sequence tell us?

• Less than 2% of the genome codes for proteins.

• Repeated sequences that do not code for proteins ("junk DNA") make up at least 50% of the human genome.

• Repetitive sequences are thought to have no direct functions, but they shed light on chromosome structure and dynamics. Over time, these repeats reshape the genome by rearranging it, creating entirely new genes, and modifying and reshuffling existing genes.

• The human genome has a much greater portion (50%) of repeat sequences than the mustard weed (11%), the worm (7%), and the fly (3%).

U.S. Department of Energy Genome Programs, Genomics and Its Impact on Science and Society, 2003

Anticipated Benefits of Genome Research

Molecular Medicine

- improve diagnosis of disease
- detect genetic predispositions to disease
- create drugs based on molecular information
- use gene therapy and control systems as drugs
- design “custom drugs” (pharmacogenomics) based on individual genetic profiles

Microbial Genomics

- rapidly detect and treat pathogens (disease-causing microbes) in clinical practice
- develop new energy sources (biofuels)
- monitor environments to detect pollutants
- protect citizenry from biological and chemical warfare
- clean up toxic waste safely and efficiently

U.S. Department of Energy Genome Programs, Genomics and Its Impact on Science and Society, 2003

Anticipated Benefits of Genome Research-cont.

Risk Assessment

• evaluate the health risks faced by individuals who may be exposed to radiation (including low levels in industrial areas) and to cancer-causing chemicals and toxins

Bioarchaeology, Anthropology, Evolution, and Human Migration

• study evolution through germline mutations in lineages
• study migration of different population groups based on maternal inheritance
• study mutations on the Y chromosome to trace lineage and migration of males
• compare breakpoints in the evolution of mutations with ages of populations and historical events

Anticipated Benefits of Genome Research-cont.

DNA Identification (Forensics)

- identify potential suspects whose DNA may match evidence left at crime scenes
- exonerate persons wrongly accused of crimes
- identify crime and catastrophe victims
- establish paternity and other family relationships
- identify endangered and protected species as an aid to wildlife officials (could be used for prosecuting poachers)
- detect bacteria and other organisms that may pollute air, water, soil, and food
- match organ donors with recipients in transplant programs
- determine pedigree for seed or livestock breeds
- authenticate consumables such as caviar and wine

Anticipated Benefits of Genome Research-cont.

Agriculture, Livestock Breeding, and Bioprocessing

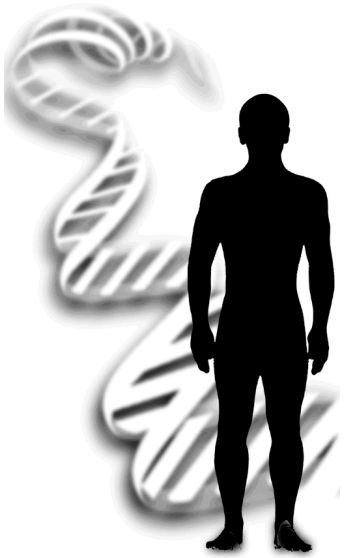
- grow disease-, insect-, and drought-resistant crops
- breed healthier, more productive, disease-resistant farm animals
- grow more nutritious produce
- develop biopesticides
- incorporate edible vaccines incorporated into food products
- develop new environmental cleanup uses for plants like tobacco

Medicine and the New Genetics

Gene Testing ! Pharmacogenomics ! Gene Therapy

Anticipated Benefits:

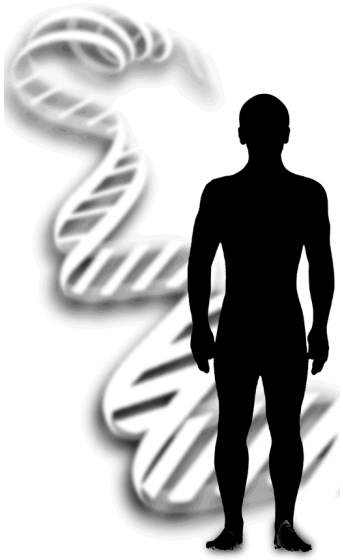
- improved diagnosis of disease
- earlier detection of genetic predispositions to disease
- rational drug design
- gene therapy and control systems for drugs
- personalized, custom drugs



ELSI: Ethical, Legal, and Social Issues

- **Privacy and confidentiality of genetic information**.
- **Fairness in the use of genetic information** by insurers, employers, courts, schools, adoption agencies, and the military, among others.
- **Psychological impact, stigmatization, and discrimination** due to an individual's genetic differences.
- **Reproductive issues** including adequate and informed consent and use of genetic information in reproductive decision making.
- **Clinical issues** including the education of doctors and other health-service providers, people identified with genetic conditions, and the general public about capabilities, limitations, and social risks; and implementation of standards and quality_control measures.

U.S. Department of Energy Genome Programs, Genomics and Its Impact on Science and Society, 2003



ELSI Issues (cont.)

• **Uncertainties associated with gene tests for susceptibilities and complex conditions** (e.g., heart disease, diabetes, and Alzheimer's disease).

• **Fairness in access to advanced genomic technologies**.

• **Conceptual and philosophical implications** regarding human responsibility, free will vs genetic determinism, and concepts of health and disease.

• **Health and environmental issues** concerning genetically modified (GM) foods and microbes.

• **Commercialization of products** including property rights (patents, copyrights, and trade secrets) and accessibility of data and materials.