CSE206B-L2

Genome rearrangements: WGD

Yeast Whole Genome Duplication?

Molecular evidence for an ancient duplication of the entire yeast genome

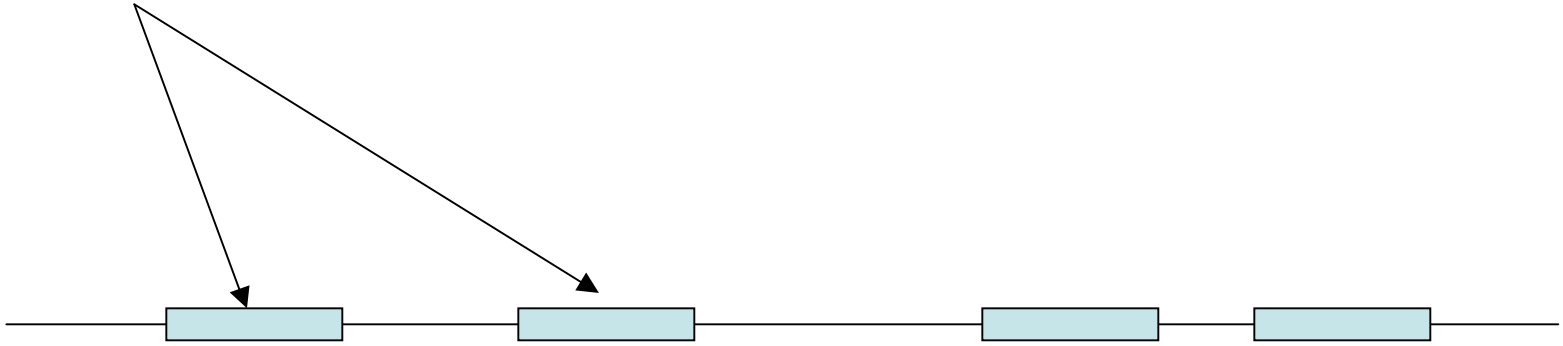
Kenneth H. Wolfe & Denis C. Shields

Department of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland

be phenotypically redundant[3–7]. Here we show that the arrangement of duplicated genes in the *S. cerevisiae* genome is consistent with Ohno's hypothesis. We propose a model in which this species is a degenerate tetraploid resulting from a whole-genome duplication that occurred after the divergence of *Saccharomyces* from *Kluyveromyces*. Only a small fraction of the genes were subsequently retained in duplicate (most were deleted), and gene order

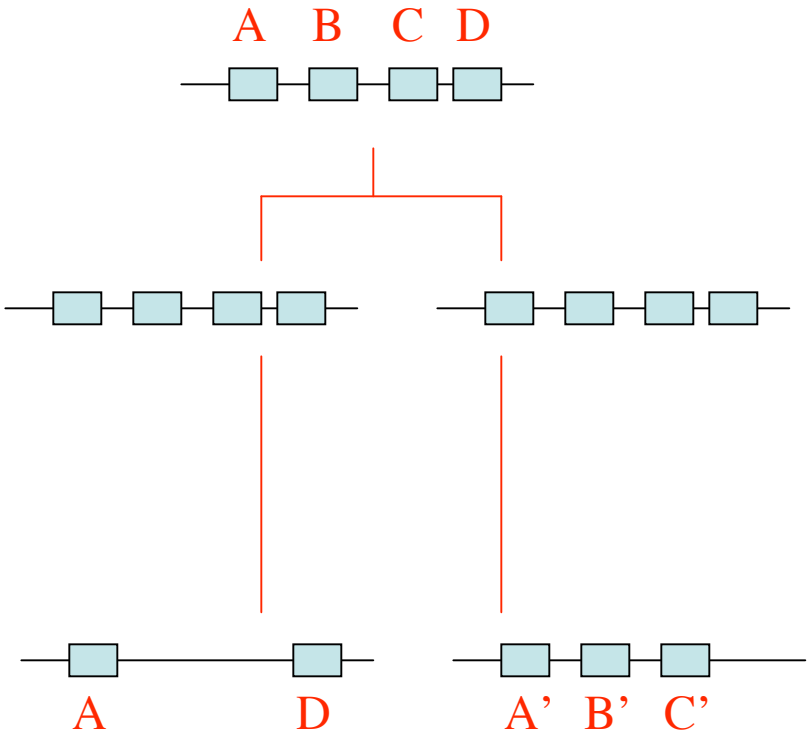
Prove or Disprove the WGD hypothesis

- First, some terms:
  - DNA
  - Proteins
  - Genes



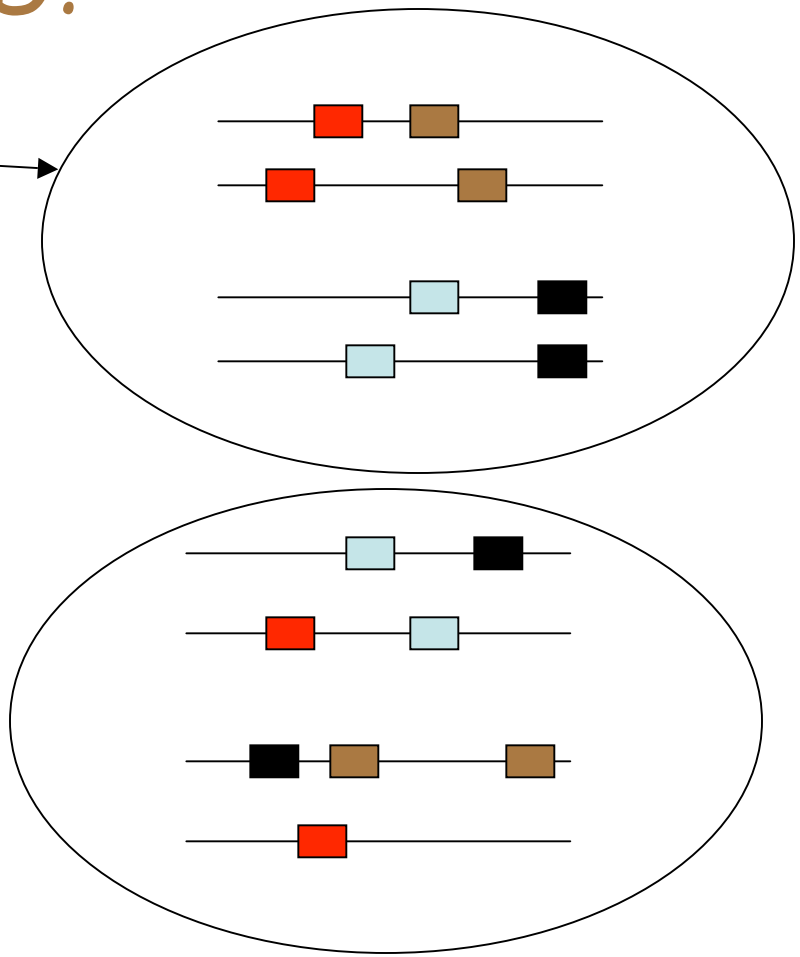
WGD

- Each ancestral chromosome split into two.
- Genes were rapidly deleted.
- A few duplicates remain



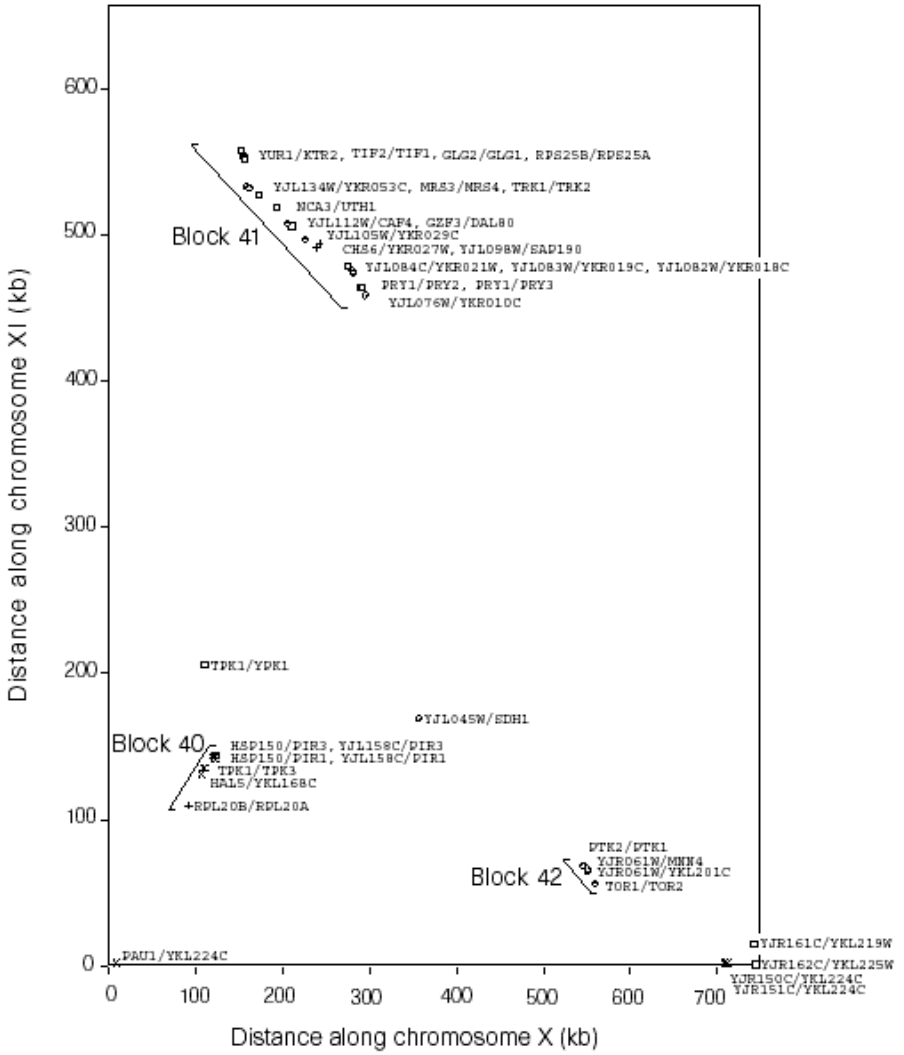
Are few duplicates enough to show WGD?

- Yes!
- Unfortunately, the genomes are also undergoing rearrangements.
- This complicates the picture



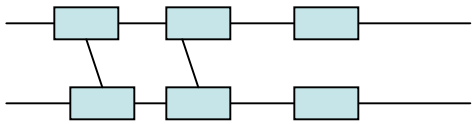
Evidence

- All S. cerevisae proteins were compared against each other using BLASTP.
- 55 genomic regions were duplicated containing 376 duplicated genes



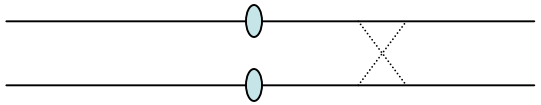
Duplicated regions

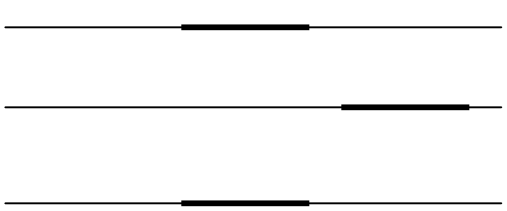
- What defines a duplicate region
  - BLAST score >= 200 for each pair
  - At least 3 pairs within 50kb each
  - Conservation of gene order and orientation (some local inversions allowed).
- 25% of the genes in duplicate regions were duplicated.
- 42% of all duplicate genes were in duplicated regions



How did the 55 regions arise?

- Independent duplications of each region.
- Alternatively, WGD followed by reciprocal translocations.
- 1. 50/55 regions have the same orientation w.r.t centromere.
- 2. Based on a Poisson distribution, 55 successive independent duplications would results in 7 triplicate regions. None are seen.
- Therefore, independent duplication is unlikely.





WGD or Not?

- The Wolfe paper concludes that yeast is an ancestral teraploid. They conjecture that this happened when two diploid yeast cells fused to form a tetraploid.
- Sequence decay/deletions led back to diploidy.
- About 15% of the gens are still in duplicate form
- This conclusion was not widely accepted!

Genomic exploration of related gene families

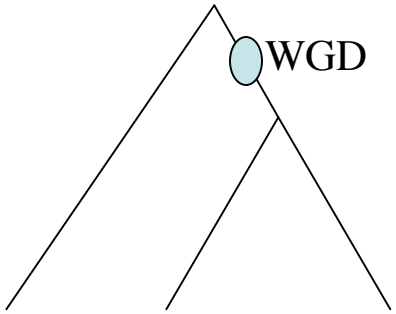
- Llorente at al., 2000 presented a different argument.
- Conducted survey sequencing of 13 yeast species.
- Looked at gene families, and found that the overall degree of gene redundancy seems conserved across all species.
- This includes species that predated the duplication, which should have a different distribution.
- Conclusion. No evidence for WGD.

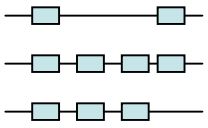
Wong et al. 2002

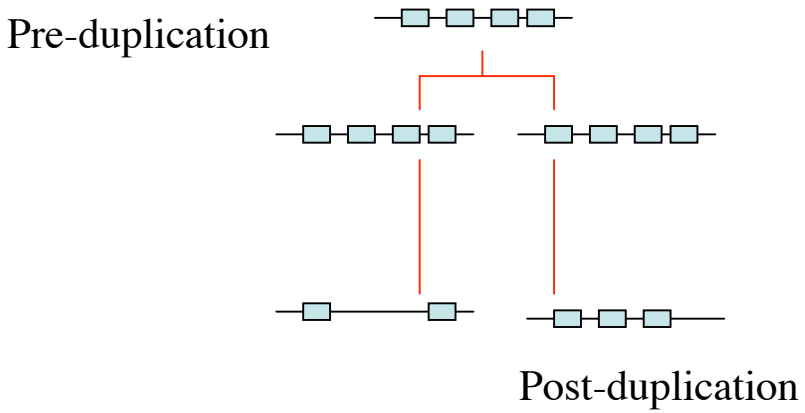
- They find even more compelling evidence of WGD from the 14 survey sequences!
- Can you suggest a comparative experiment?

Clues from pre-duplicated genomes

- Pairs of post-duplicated genomic regions should correspond to one pre-duplicated genomic region.

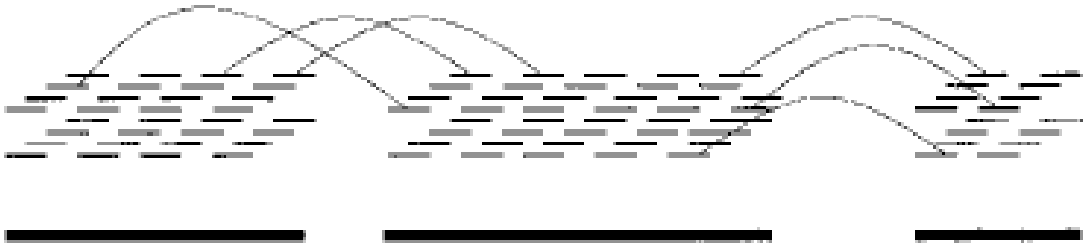






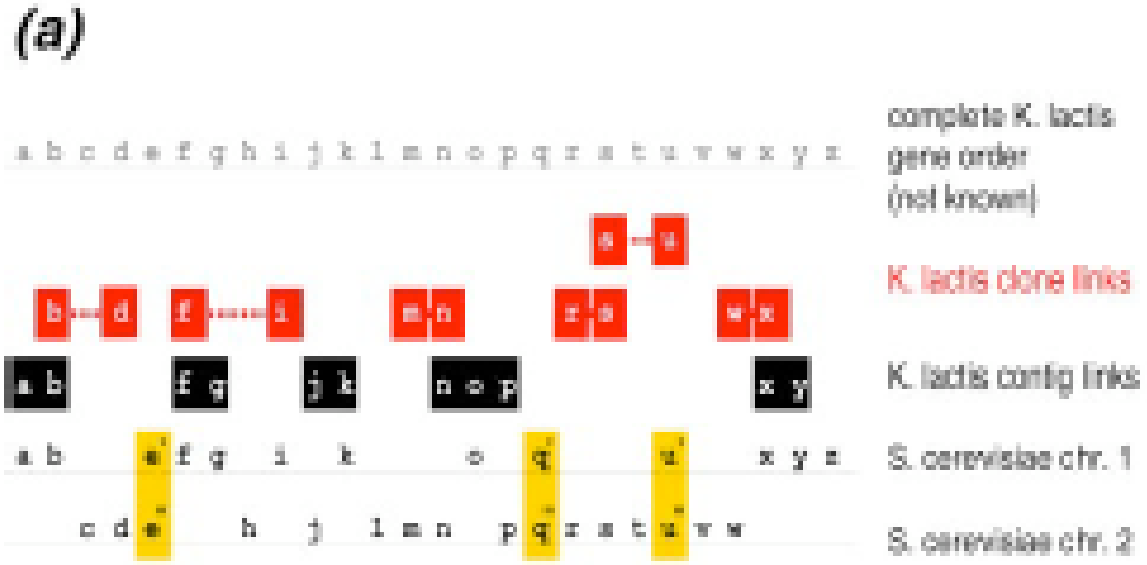
- In reality, the picture is convoluted by local rearrangements. However, there is still considerable evidence remaining.

Genome sequencing & assembly



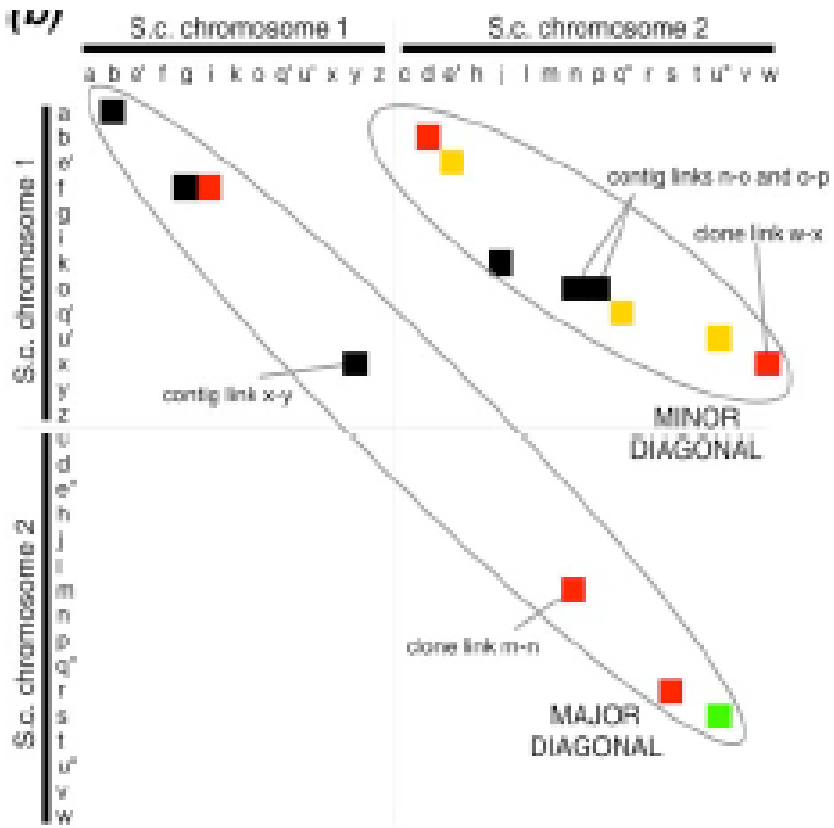
- Fragments are sequenced from the ends of large clones.
- Overlapping fragments are assembled (alignment and consensus generation) to form contigs.
- Clone mate-pairs are used to connect contigs.

Pre-duplicated genome

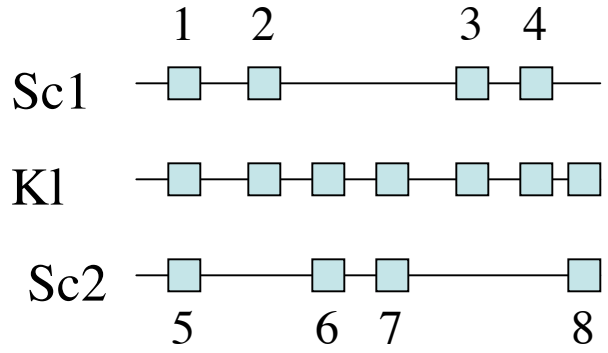


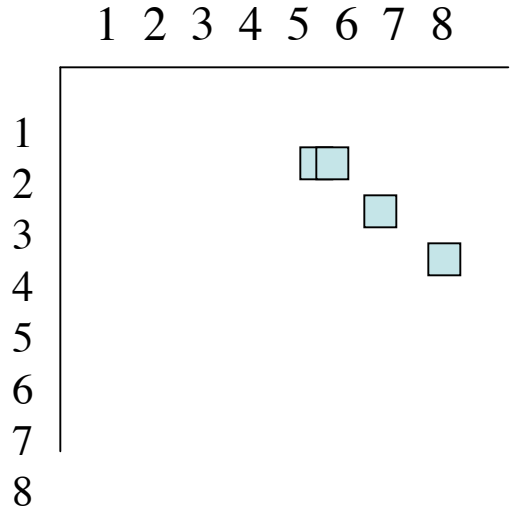
Comparing pre, and post WGD genomes

- Both axes represent S. cerevisae genes
  - Black dots represent contig links in K.lactis
  - Red dot represent clone links
  - Yellow dots represent duplication in S.cere.
- What does this mean?



- Major diagonal represents synsteny between Sc and Kl
- Minor diaginal band represents similarity between 2 Sc chromosomes inferred by Kl gene order





- 88 minor diagonals were were identified.
- These diagonals covered 176 sections of Sc chromosomes
- 3900/5583 genes were included (70%)
- Almost no overlap between regions.
- Superimposing information about duplicated genes, the sister regions covered 82% of the genome.
- No overlap between sister regions rules out independent duplication

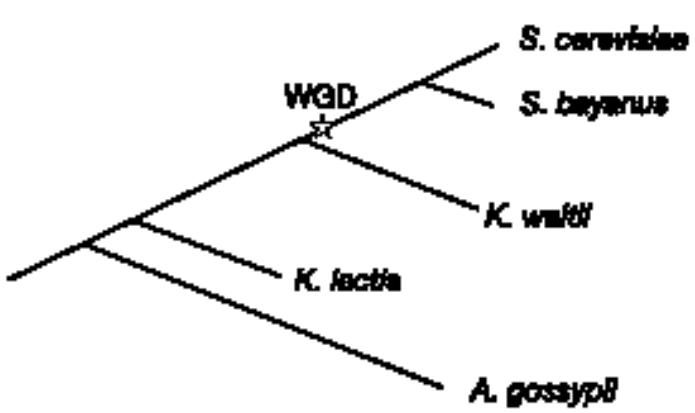
Final proof for WGD, 2004

Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*

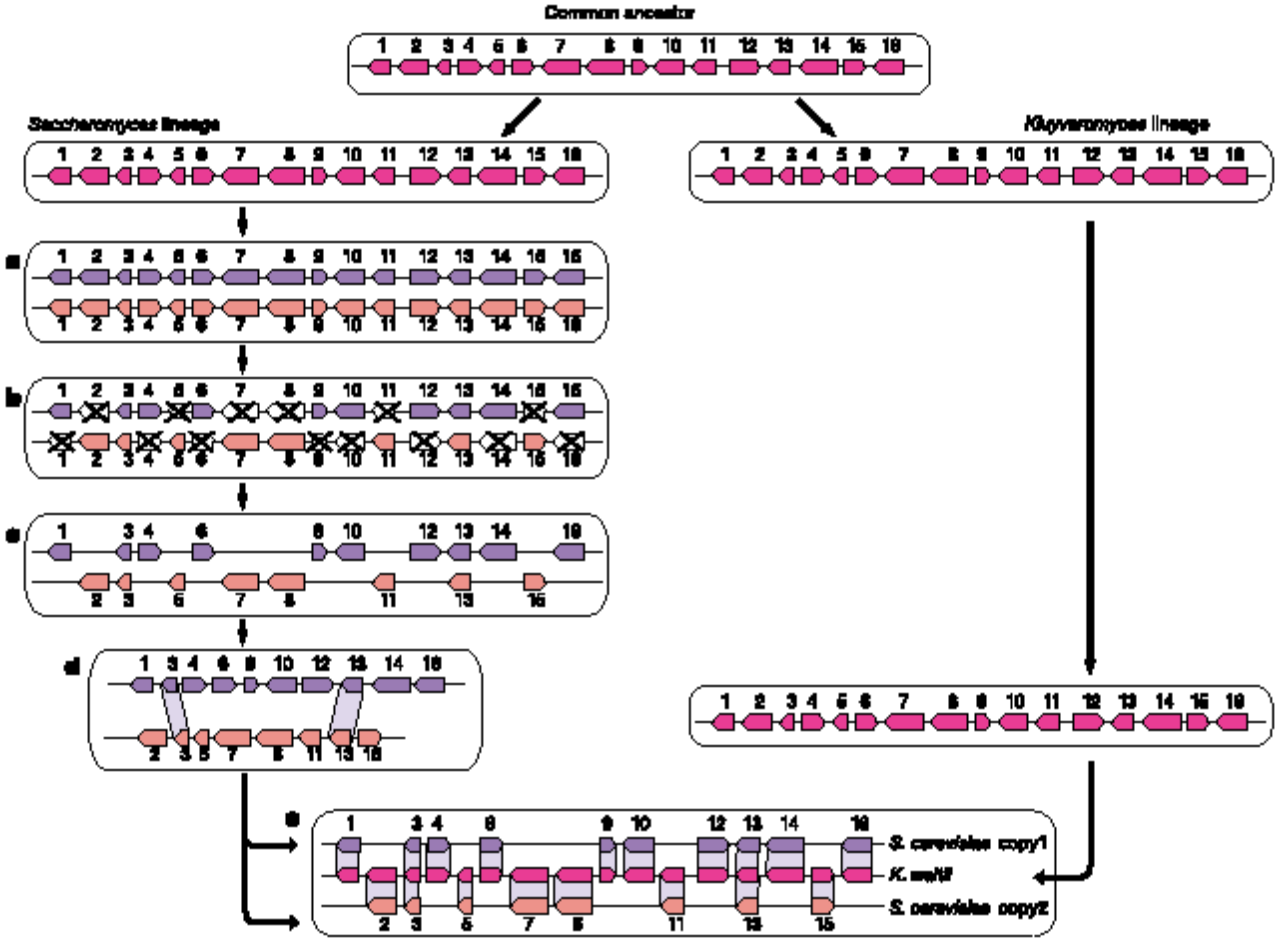
Manolis Kellis[1,2], **Bruce W. Birren**[1] **& Eric S. Lander**[1,3]

[1]*The Broad Institute, Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02138, USA*
[2]*MIT Computer Science and Artificial Intelligence Laboratory, and* [3]*Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02139, USA*

- Complete sequence of a pre duplicated genome was generated.



Evidence is similar to Wong et al.



Doubly Conserved Synteny Blocks (DCS)

- 253 DCS blocks were identified containing 75% of K. waltii genes and 81% of S. cerevisae genes
- A typical DCS block has 27 genes (largest block has 81 genes).
- DCS blocks are separated by ~3 genes on the average.
- ~1% of the K. waltii genome lies in segments that match 3 ore more Sc regions. Mainly due to local rearrangements.
- In a DCS block 90% of Kw genes have a match in at least 1 of the 2 Sc regions.
- 47 Sister regions have no duplicated gene.

Analysis of duplicated genes

- Gene loss can occur by large segmental deletions, or individual gene deletions.
- DCS blocks indicate that gene loss occurred by many small deletions, and is balanced between the 2 sister regions (Average: 0.57-0.43)

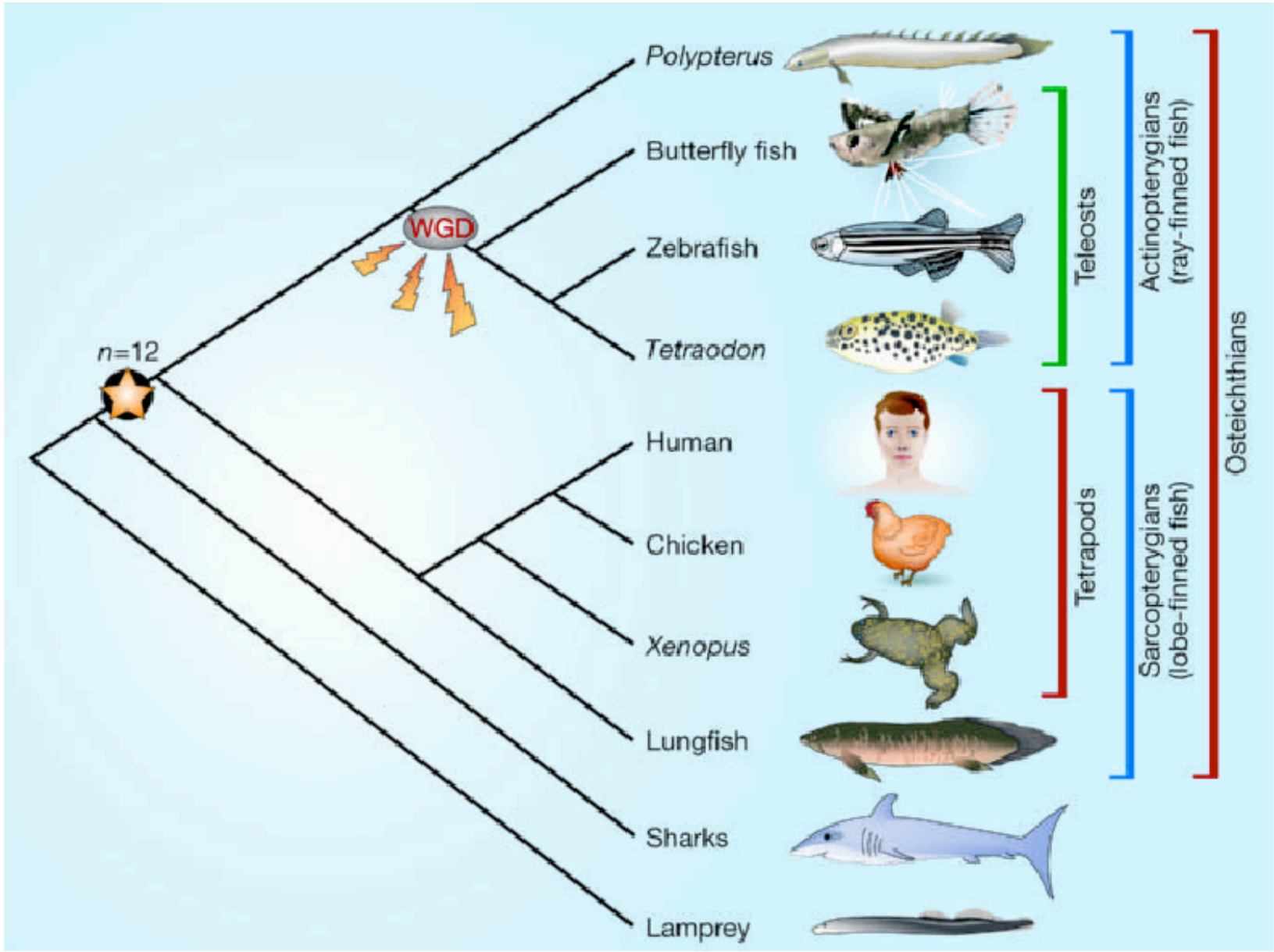
Accelerated divergence

- Ohno hypothesized that after duplication, one copy would preserve the original function, and the other copy would be free to diverge. Others argued that both copies would diverge.
- Kellis et al. Compared genes from K. waltii to corresponding pairs in Sc.
- 76 of 457 pairs show accelerated evolution. In 95% of the cases, acceleration was limited to one of the 2 paralogs.
- Deletion of the ancestral paralog is lethal in 18% of the cases.
- Deletion of a derived paralog is never lethal.

WGD in Pufferfish: 2004

Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype

Olivier Jaillon[1], Jean-Marc Aury[1], Frédéric Brunet[2], Jean-Louis Petit[1], Nicole Stange-Thomann[3], Evan Mauceli[3], Laurence Bouneau[1], Cécile Fischer[1], Catherine Ozouf-Costaz[4], Alain Bernot[1], Sophie Nicaud[1], David Jaffe[3], Sheila Fisher[3], Georges Lutfalla[5], Carole Dossat[1], Béatrice Segurens[1], Corinne Dasilva[1], Marcel Salanoubat[1], Michael Levy[1], Nathalie Boudet[1], Sergi Castellano[6], Véronique Anthouard[1], Claire Jubin[1], Vanina Castelli[1], Michael Katinka[1], Benoît Vacherie[1], Christian Biémont[7], Zineb Skalli[1], Laurence Cattolico[1], Julie Poulain[1], Véronique de Berardinis[1], Corinne Cruaud[1], Simone Duprat[1], Philippe Brottier[1], Jean-Pierre Coutanceau[4], Jérôme Gouzy[8], Genis Parra[6], Guillaume Lardier[1], Charles Chapple[6], Kevin J. McKernan[9], Paul McEwan[9], Stephanie Bosak[9], Manolis Kellis[3], Jean-Nicolas Volff[10], Roderic Guigó[6], Michael C. Zody[3], Jill Mesirov[3], Kerstin Lindblad-Toh[3], Bruce Birren[3], Chad Nusbaum[3], Daniel Kahn[8], Marc Robinson-Rechavi[2], Vincent Laudet[2], Vincent Schachter[1], Francis Quétier[1], William Saurin[1], Claude Scarpelli[1], Patrick Wincker[1], Eric S. Lander[3,11], Jean Weissenbach[1] & Hugues Roest Crollius[1]*


Polypterus
Butterfly fish
Zebrafish
Tetraodon
Human
Chicken
Xenopus
Lungfish
Sharks
Lamprey
WGD
n=12
Teleosts
Tetrapods
Actinopterygians
(ray-finned fish)
Sarcopterygians
(lobe-finned fish)
Osteichthians

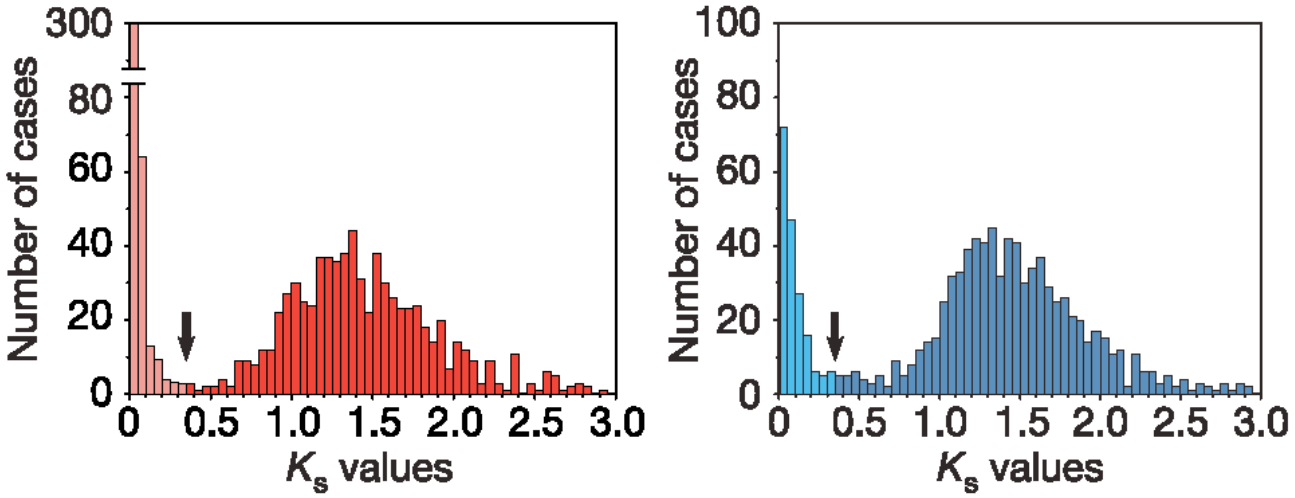
2 Main arguments for WGD

- Evidence for a number of 'paired' duplicate regions. The regions do not overlap and must not show higher redundancy
- Evidence for DCS when compared against a pre-duplicated 'nearby' genome.

Selecting ancient duplications

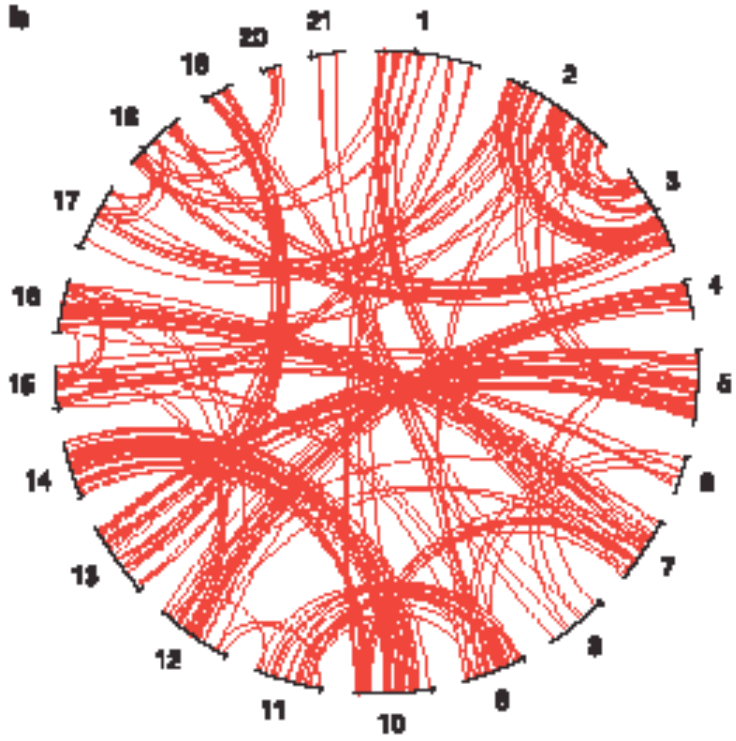
Duplicated genes

- 1078 pairs in tetra, and 995 pairs in fugu
- based on Ks, 75% are ancient duplications before tetra-fugu separation

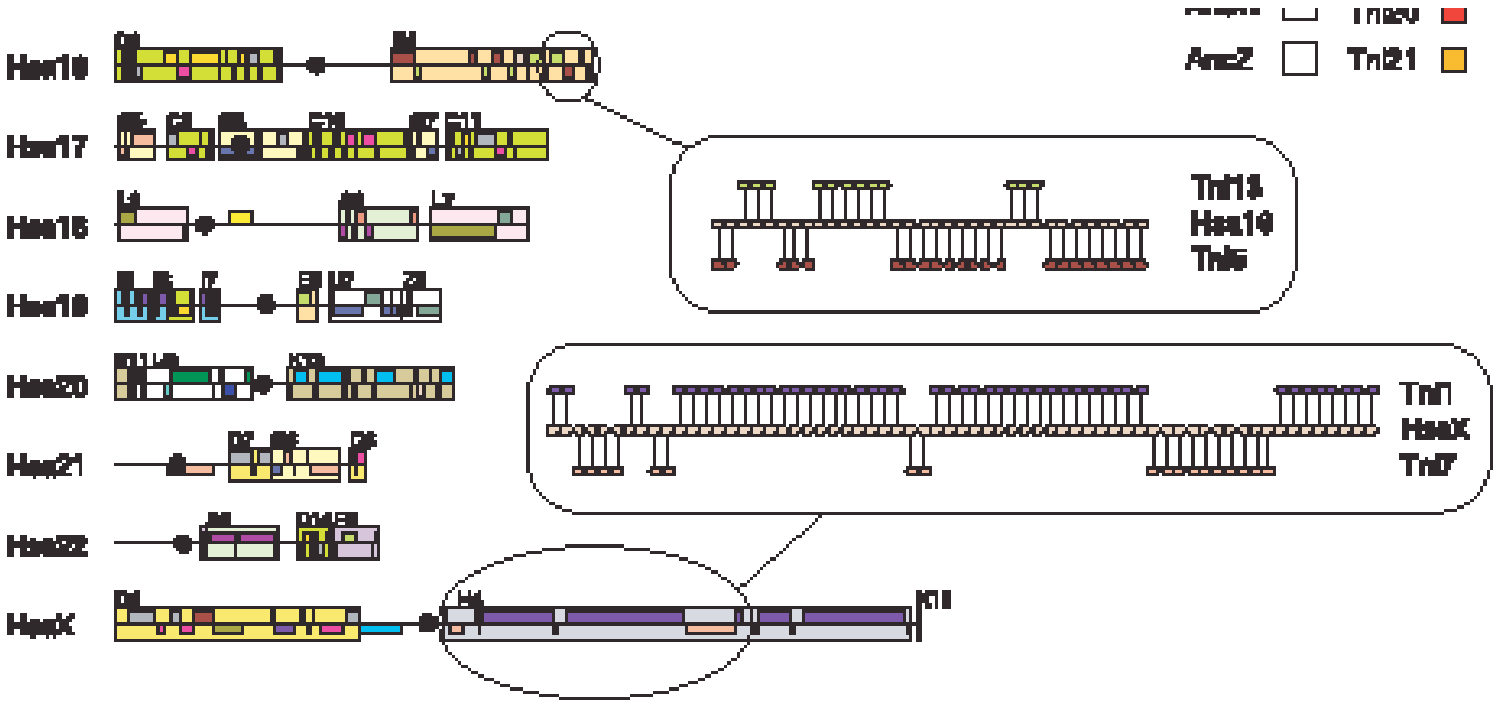


Evidence 1

- Duplicated sections of the chromosome.



Evidence 2

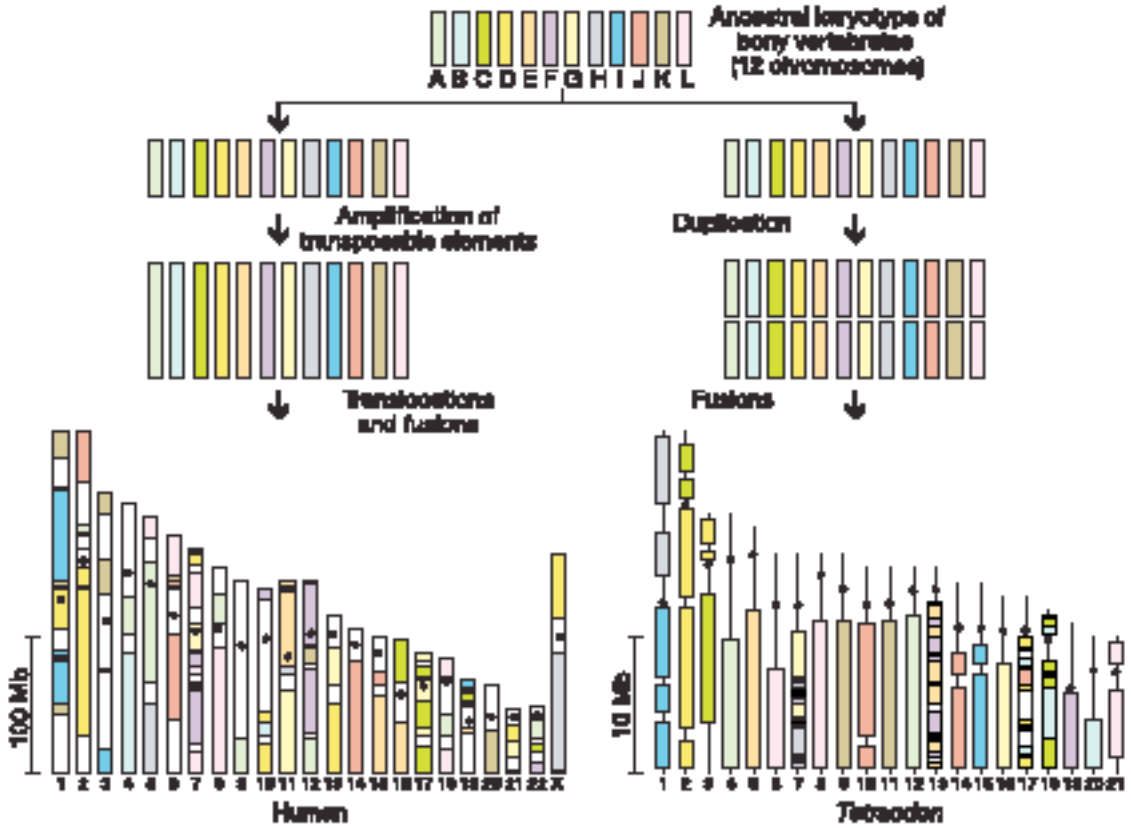


- Compared against human and mouse

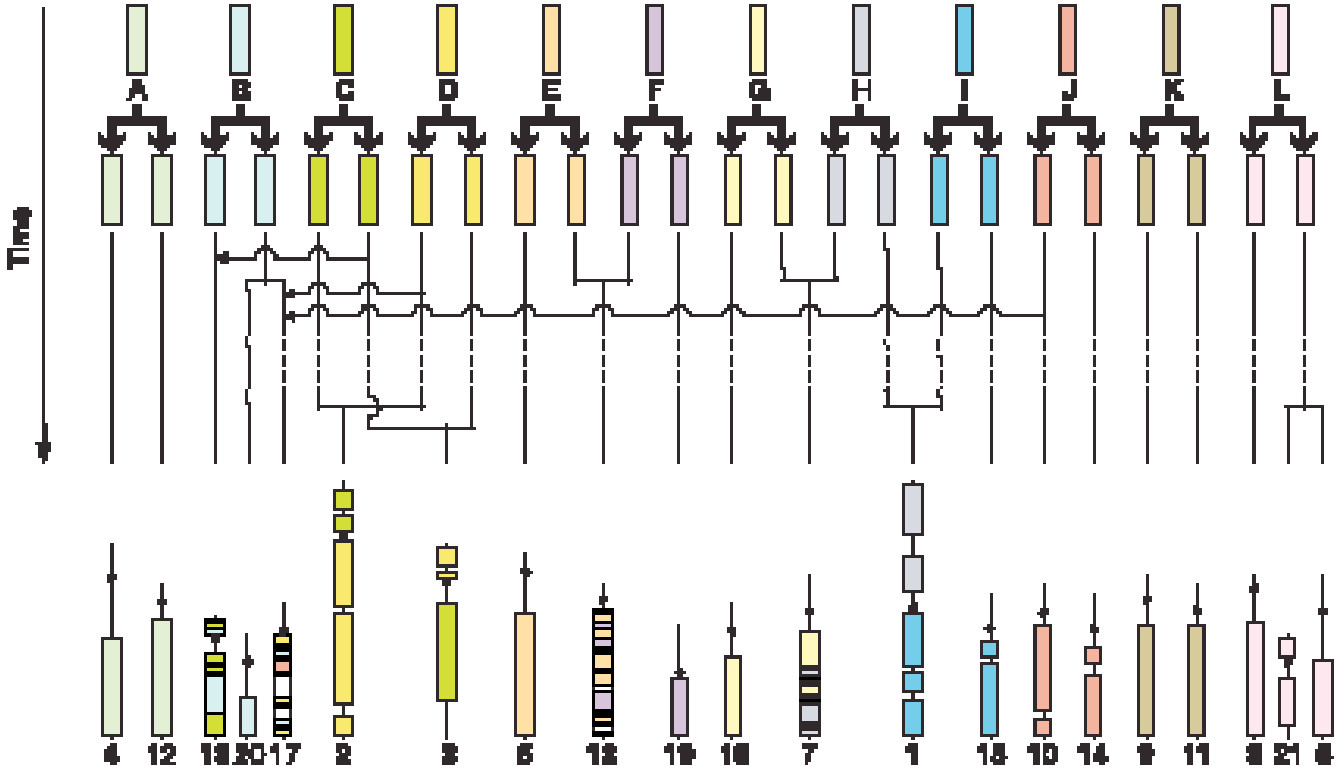
Composition of the ancestral genome?

- What did the pre-duplicated ancestor look like?
- Can you postulate the architecture of the current (human and tetraodon genomes) genomes in terms of the common ancestor?
- What was the sequence of events?

Model for distribution of ancestral segments in current Hs and Tn genomes



An evolutionary scenario for Tn



Genome rearrangement scenarios

- In subsequent classes, we will cover some of the theoretical approaches to studying these genome rearrangements.

Other topics

How many genes do we have?

● There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.

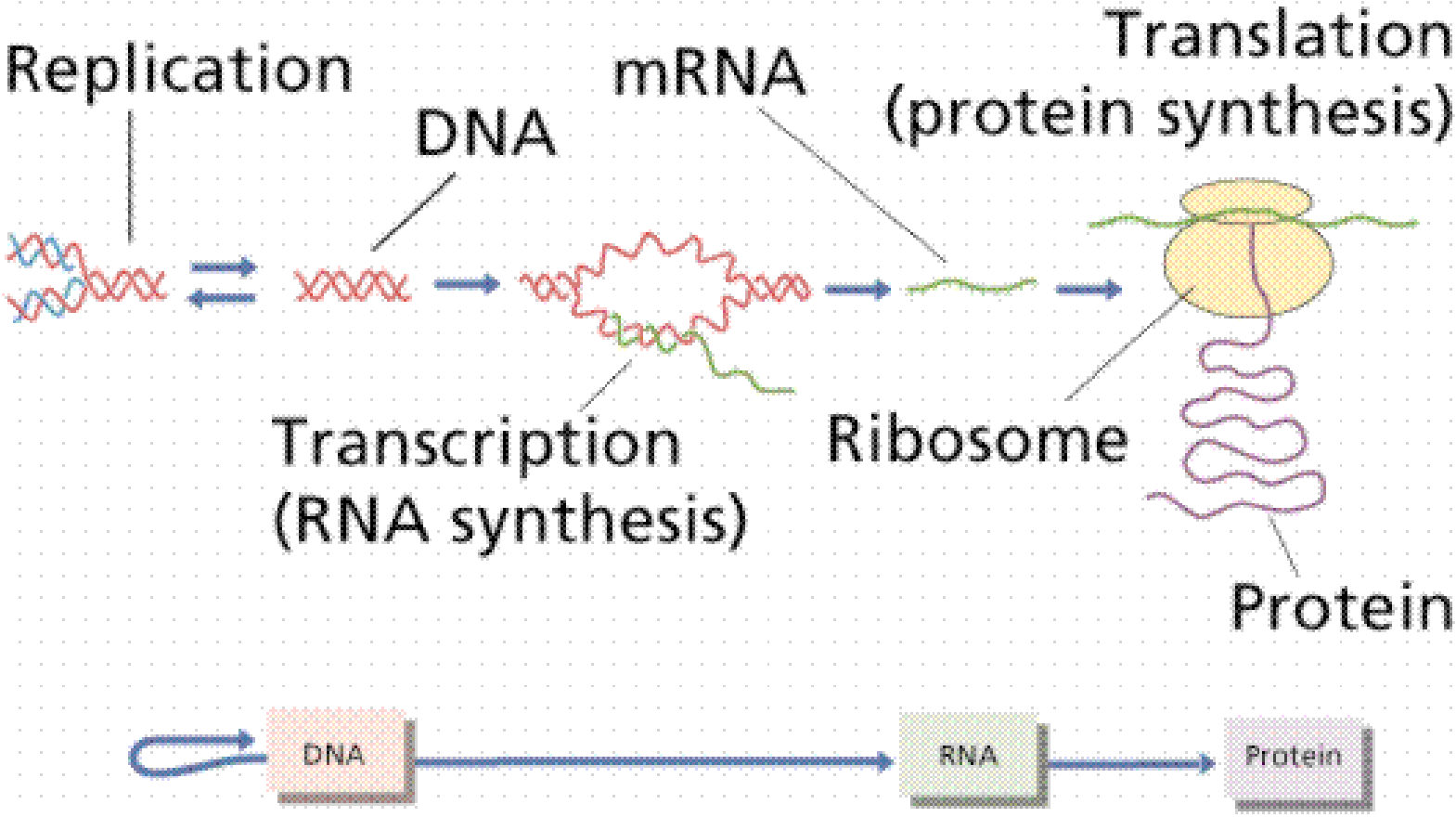
Nature 2001

Science

pressed. Conservative criteria, requiring at least two lines of evidence, were used to define a set of 26,383 genes with good confidence that were used for more detailed analysis presented in the subsequent sections.

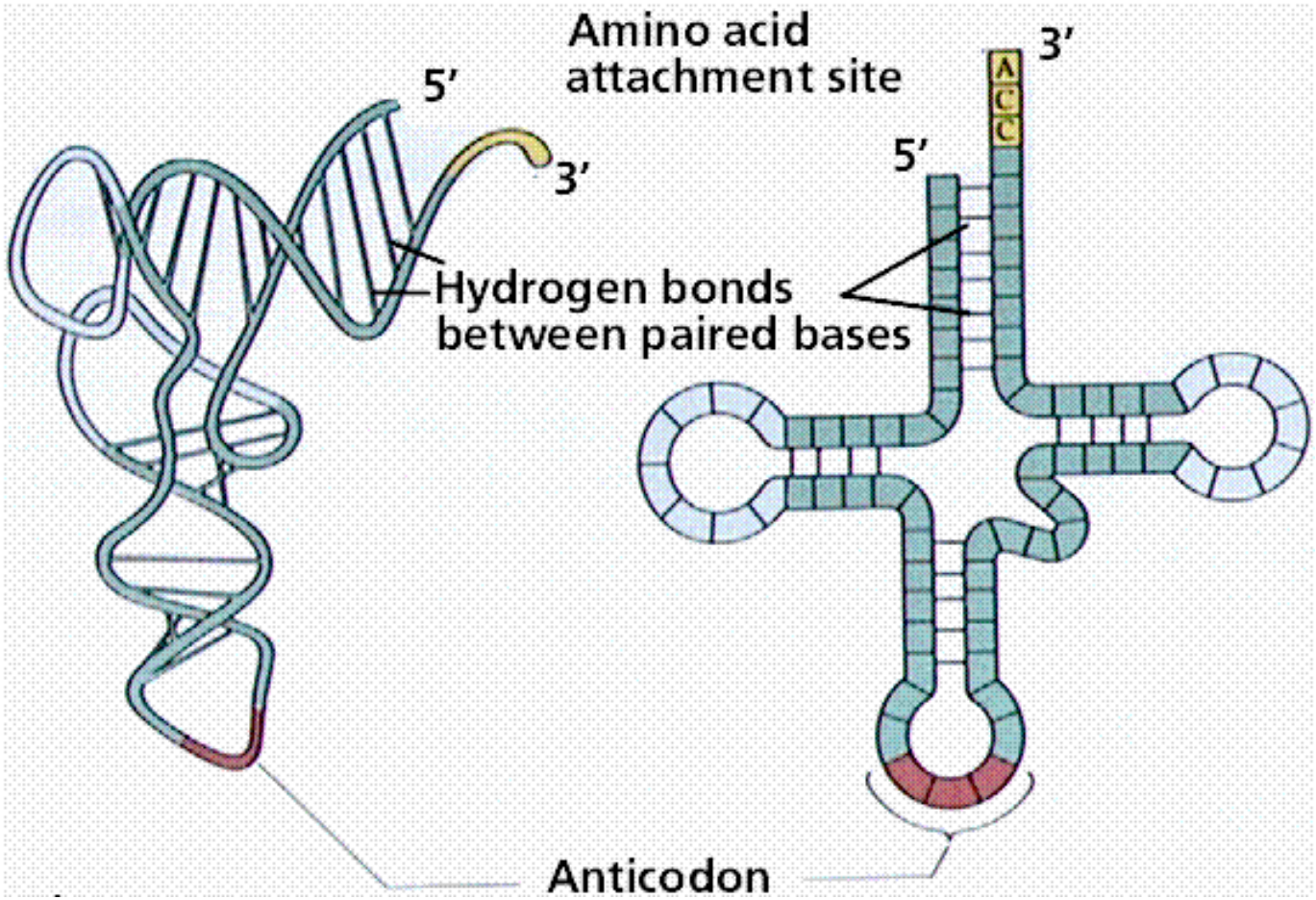
The Modern RNA World hypothesis

- The estimate of the number of genes is based largely on *computational* experiments. These methods works well for *large, highly expressed, protein coding* genes.
- "Could it be that a large class of genes has gone undetected because they do not make proteins?.."
  - Eddy: *Nature Reviews Genetics* 2001
- If true, what function can these genes have?


Replication
DNA
mRNA
Translation
(protein synthesis)
Transcription
(RNA synthesis)
Ribosome
Protein
DNA
RNA
Protein

Roles of ncRNA

- mRNA is translated. All untranslated RNA are lumped as ncRNAs
- However, the Ribosome (the translating machinery) is composed largely of RNA (rRNAs).
- tRNAs encode the template for translation.
- Splicing machinery in the nucleus (spliceosome) contains snRNA.
- Many other nc RNAs (snoRNA, siRNA, miRNA) are being found on a regular basis. Some of these are act as anti-sense regulators of gene expresssion (RNAi).
- The sequence specific binding makes RNA a more efficient regulator than proteins.

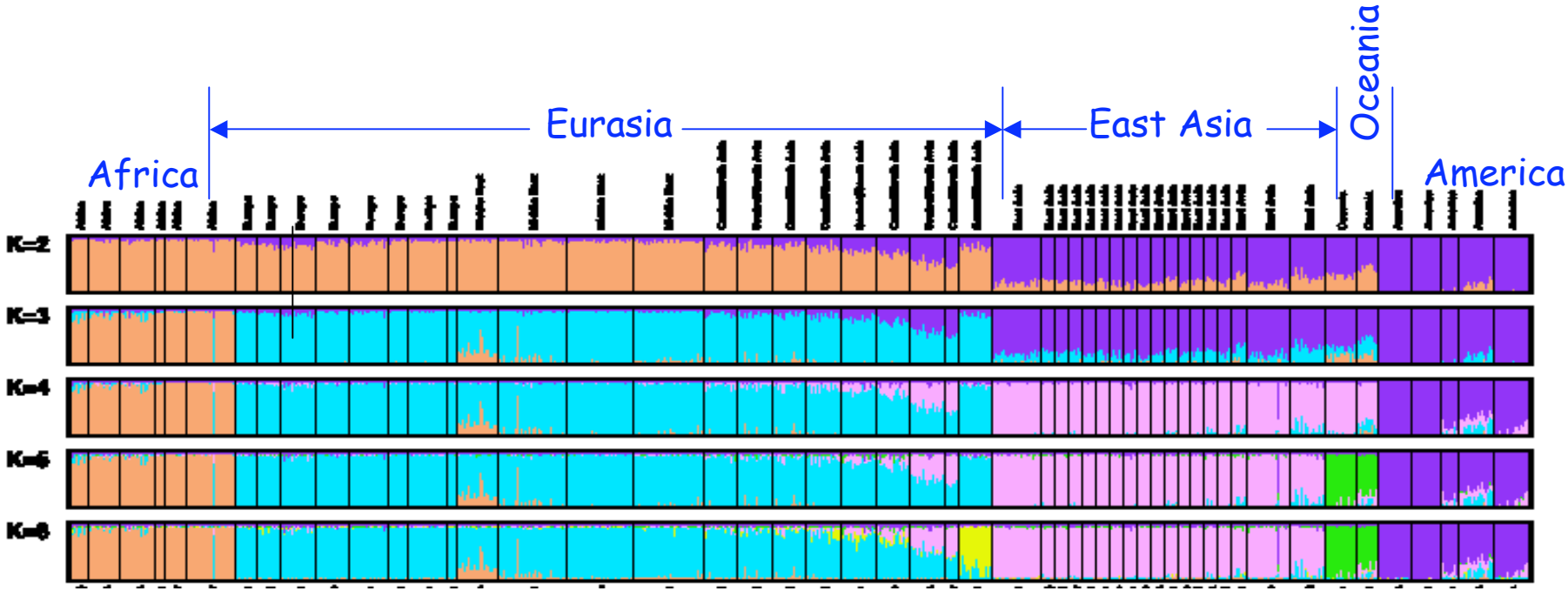

Amino acid
attachment site
5′
3′
5′
3′
A C C
Hydrogen bonds
between paired bases
Anticodon

Computational ncRNA genefinding

- Understanding & Computing RNA structure
- De novo: Predict signals for transcription initiation, termination, processing
  - Eukaryotic ncRNAs are transcribed by different polymerases.
  - Approach feasible in microbes
- Comparative genome analysis
  - Compensatory mutations preserve structure
  - Predict structure from alignment
  - Searching with homologs

Population Structure

- 377 locations (loci) were sampled in 1000 people from 52 populations.
- 6 genetic clusters were obtained, which corresponded to 5 geographic regions (Rosenberg et al. Science 2003)



Population Genetics

- What is it about our genetic makeup that makes us measurably different?
- These genetic differences are correlated with phenotypic differences
- With cost reduction in sequencing and genotyping technologies, we will know the sequence for entire populations of individuals.
- Here, we will study the basics of this polymorphism data, and tools that are being developed to analyze it.

Repeats

- Roughly 45-50% of the human genome contains transposable repeats.
- Like rearrangements, these are fossil records, and can offer unique insights into evolution.
- We will discuss tools for identifying, classifying, and analyzing repeats.