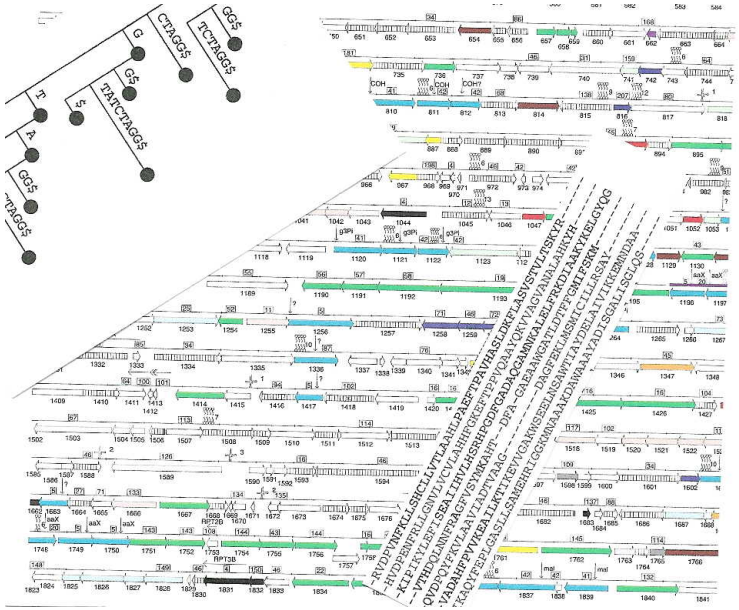
Lecture 10 : Whole genome sequencing and analysis

Introduction to Computational Biology

Teresa Przytycka, PhD

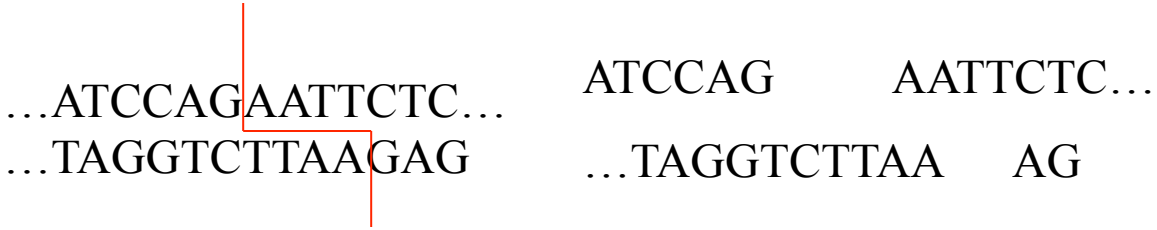


Sequencing DNA

- Goal – obtain the string of bases that make a given DNA strand.
- Problem –Typically one cans sequence directly only DNA of short length (400-700 bp – Sanger; <200 - Illumina).
- Sequence assembly – the process of putting together the fragments.

Cutting and breaking DNA

- Restriction enzymes – proteins that catalyze hydrolysis (breaking the molecule by adding water) of DNA at certain points called restriction sides.
- Example: EcoRI restriction side GAATTC. Note that the complement of GAATTC is GAATTC (a sequence equal to its reverse is called a palindrome)



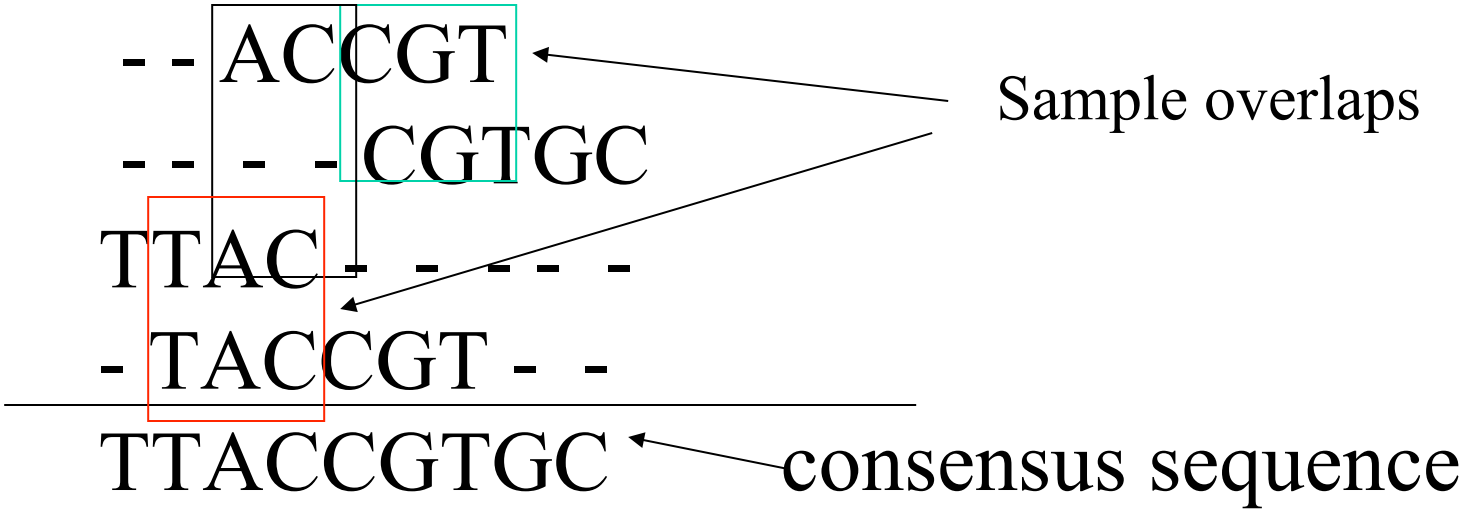
Fragment assembly

- After DNA fragments (reads) are sequenced we want to assemble then together to reconstruct the entire target sequence.
- If the overlaps were unique and error free, this would be relatively easy task… but they are not.
- In addition : fragments can come from any of the two DNA strands and we do not know which

The "ideal" example

Input: ACCGT
CGTGC
TTAC
TACCGT

Assume target sequence of about 10bp.



Fragment assembly

- After DNA fragments (reads) are sequenced we want to assemble then together to reconstruct the entire target sequence.
- Most fragment assembly algorithms include the following 3 steps:
  - Overlap - finding potentially overlapping fragments
  - **Layout** – finding the order of the fragments
  - Consensus – deriving DNA sequence from the layout.
- Usually we know with some approximation the length of the target sequence.

Finding overlaps

- In theory we should test for overlaps all pairs of fragments. For every pair we will consider all relative orientations.
- One possible method: perform alignment without charging for flanking gaps

```
    - - TAATG
    TGTAA - -
```

Representing overlaps

F - fragments. Overlap graph :

vertices = elements of F

weighted edges: if a, b ∈ F then the weight of edge from a to b is equal t where maximum integer such that

suffix(a,t) = prefix(b,t)

suffix(a,t) = last t symbols of a

prefix(b,t) = first t symbols of b

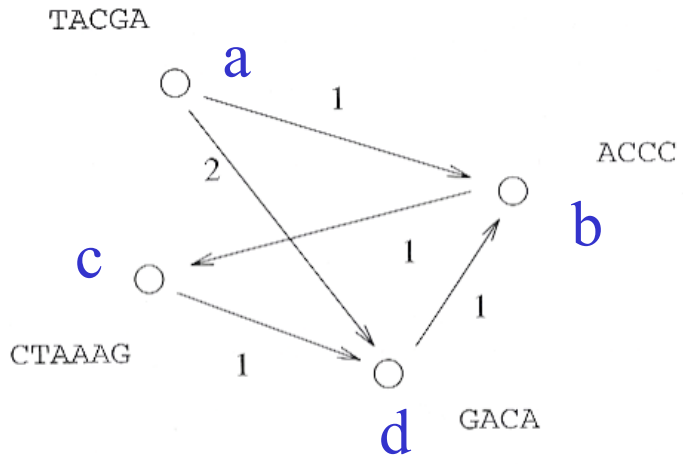


FIGURE 4.15

Overlap multigraph with zero-weight edges omitted.

Each simple path (simple = not using the same vertex more than once) in overlap graph defines an alignment.

Two assumptions:

-no fragment completely included in another

-Direction of fragments is known

Path dbc leads to alignment

```
GACA--------
---ACCC-----
------CTAAAG
```

Path abcd leads to alignment

```
TACGA----------
----ACCC--------
-------CTAAAG---
-----------GACA
```

Finding Layout

Definition: **Hamiltonian path** – a path that visits each vertex exactly once.

Let P – path, A the set of fragments involved in A

$|S(P)| = ||A|| - w(P)$

Where ||A|| sum of lengths of fragments in A

w(P) the sum of weight of path P (sum of the edge weights on this paths).

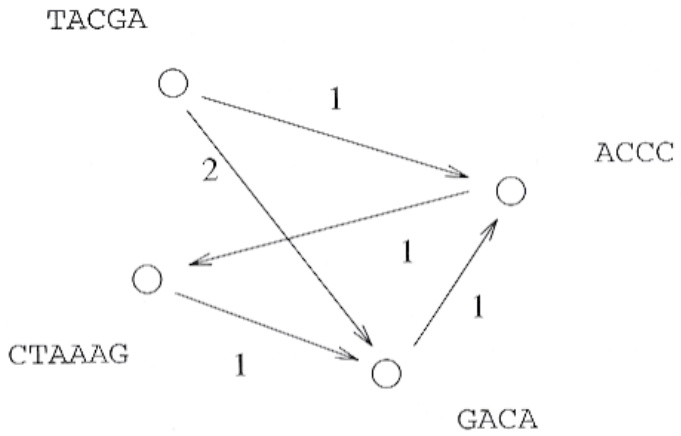


FIGURE 4.15

Overlap multigraph with zero-weight edges omitted.

The greedy algorithm

- Goal: find a Hamiltonian path with large w(P).
- Heuristic: iteratively find the heavies edge and try to add it to the path:
- Acceptance test: An edge can be added to the path, if it will not create brunching point on the path.

From Setubal/Meidanis book

Algorithm Greedy:

sort edges by weight

for each edge (f,g) in decreasing order

perform acceptance test for (f,g)

if accepted add it to the path

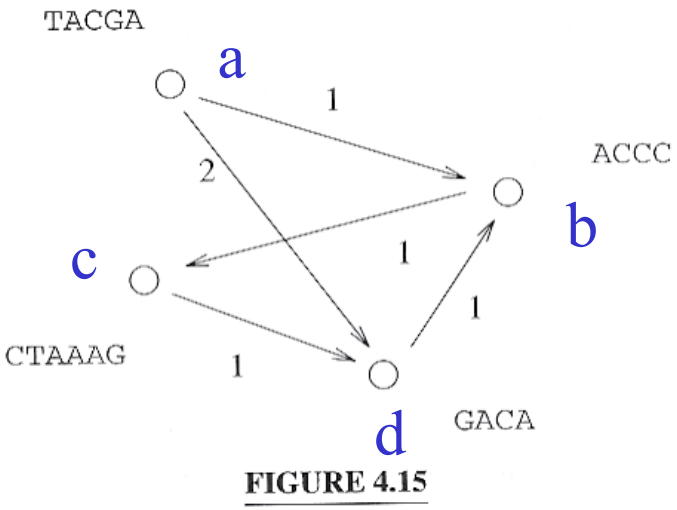


FIGURE 4.15

Overlap multigraph with zero-weight edges omitted.

Example:

greedy choice

Try: (a,d) – ok, selected

Try: (d,b) – ok, selected

Try: (a,b) – acceptance test false

Try: (b,c) – ok, selected

From Setubal/Meidanis book

Complication - repeated regions

Repeated regions: sequences that appears more than once in the molecule. The copies of repeats do not need to be exactly the same. Problems are illustrated below:

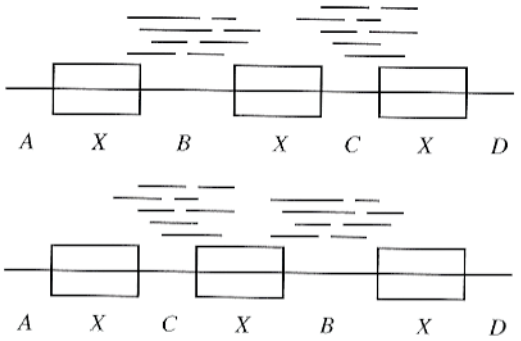


FIGURE 4.8

Target sequence leading to ambiguous assembly because of repeats of the form XXX.

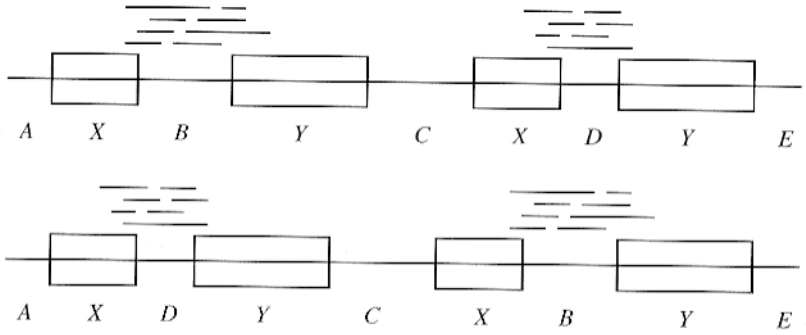


FIGURE 4.9

Target sequence leading to ambiguous assembly because of repeats of the form XYXY.

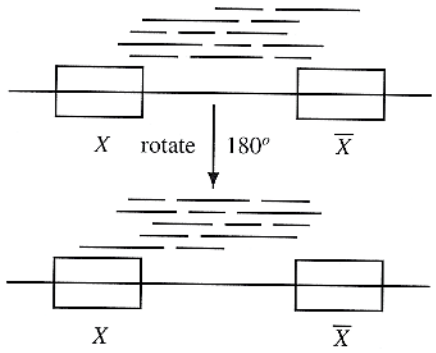


FIGURE 4.10

Target sequence with inverted repeat. The region marked $\overline{X}$ is the reverse complement of the region marked X.

From Setubal/Meidanis book

Coverage and linkage

- coverage = number of times given position is included in a an aligned fragment.
- if a coverage equals 0 at some column – we do not have continuous layout.
- linkage amount of overlaps between fragments:

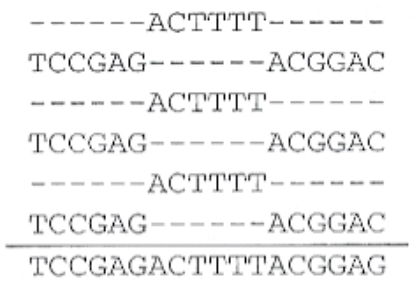
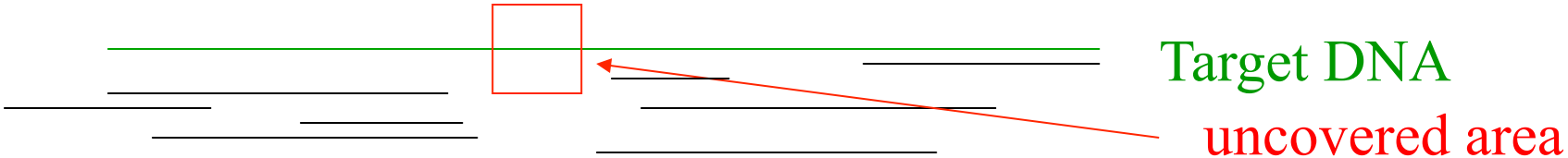


FIGURE 4.22

Good coverage but bad linkage.

Complication – lack of coverage



- Coverage at position i of the target is the number of fragments that cover this position.
- A conting – continuously covered region.

Closing gaps

- sequence walking (direct sequencing)
  - derive a primer from a sequence near the end of a conting
  - replicate the sequence starting at the primer
  - sequence this the replicated sequence
  - if the replicated sequence did not cover the gap, repeat the above steps.
  -Problems: tedious for larger gap, region of interest must be unique in the genome
- dual end sequencing. Recall that the inserts are much longer than the sequenced fragments. If we sequence both ends of the insert, we obtain mate pairs which can be used as follows:
  if two ends of a mate pair are in two different contigs, we can deduce the orientation and distance between two contings.

Scaffold – sequence of contigs where the order and distances between the contigs are approximately known.,

What do we learn form whole genome sequence

- Using gene finding algorithm we can discover significant portion of genes
- Understand the structure of a genome
- Understand genome evolution
- Searching for genes associated with diseases

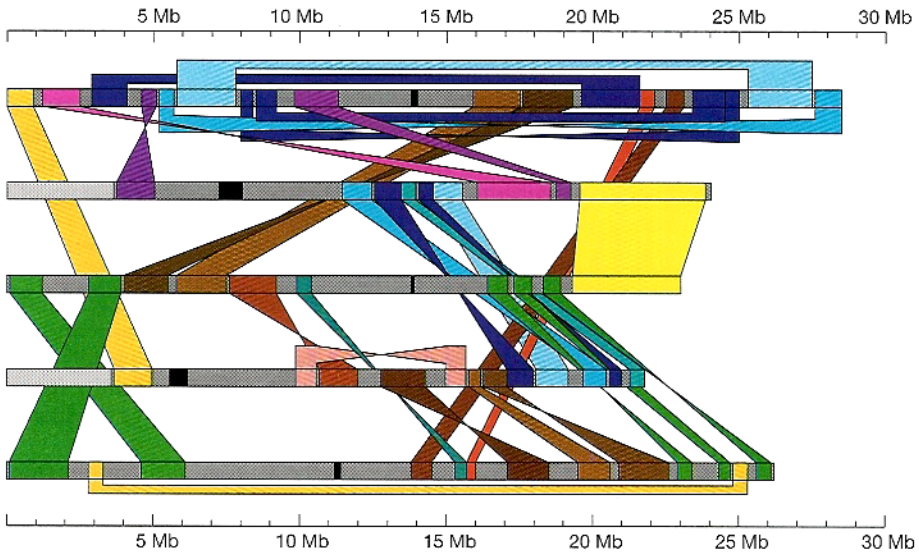
Genome duplication

- Gene duplication – widely accepted method for creation of new genes
- Ohno proposes that whole genome duplication (polyploidization) provides material for new genomes (1970)
- 2R Hypothesis: two rounds of polyploidization followed by gene loss and functional divergence occurred early in vertebrate lineage.

Syntheny blocks

In comparative genome analysis synteny blocks = regions containing the homologous genes

Below: Segmental duplications in the Arabidopsis genome fund using program MUMer.

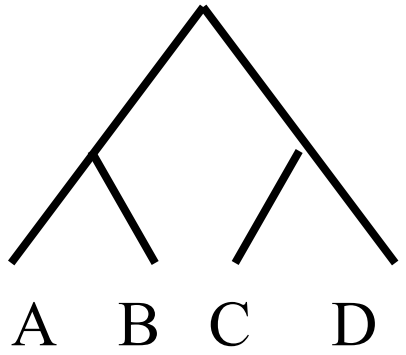


Results filtered to report segments at least 1000bp, at lest 59% identity

NATURE 1 VOL 40S 114 DECEMBER 20001 www.nature.com 801

How many rounds of genome duplication?

- Two round of genome duplication should lead to occurrences of groups of four synteny blocks
- Such tree should be then observed in the current genome
- They should be consistent
- For vertebrates evolution there is evidence for full genome duplication



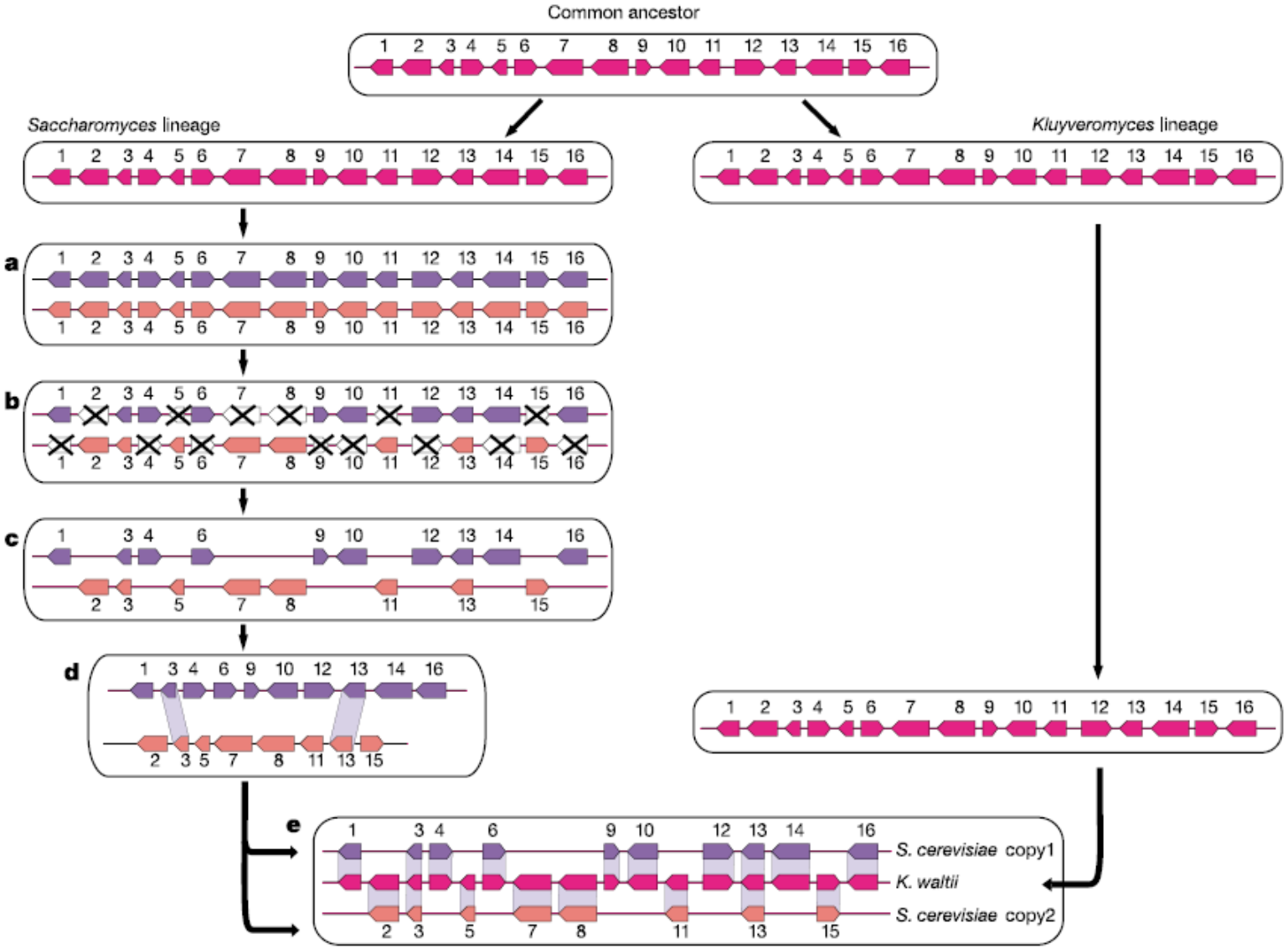
Whole genome duplications in yeast

articles

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]

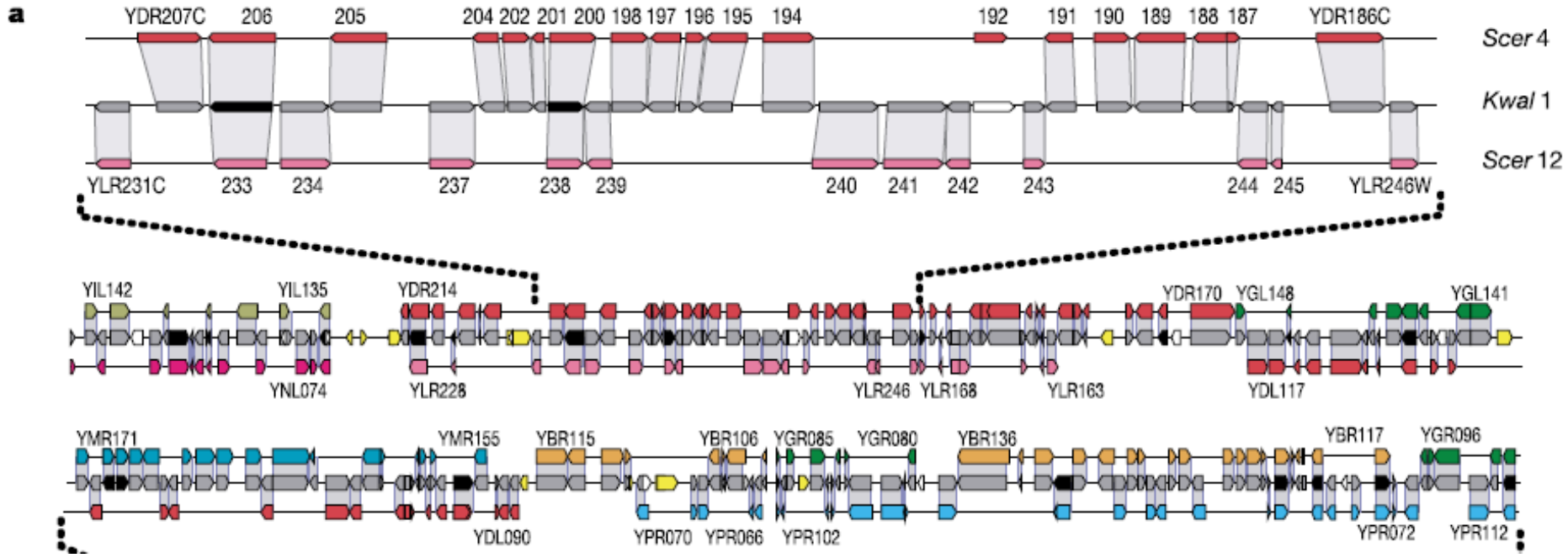
[1]*The Broad Institute, Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02138, USA*
[2]*MIT Computer Science and Artificial Intelligence Laboratory, and* [3]*Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02139, USA*



ure 1 Model of WGD followed by massive gene loss predicts gene interleaving in sister ons. **a**, After divergence from *K. waltii*, the *Saccharomyces* lineage underwent a ome duplication event, creating two copies of every gene and chromosome. **b**, The majority of duplicated genes underwent mutation and gene loss. **c**, Sister segments ined different subsets of the original gene set, keeping two copies for only a small ority of duplicated genes, which were retained for functional purposes. **d**, Within *S. cerevisiae*, the only evidence comes from the conserved order of duplicated genes (numbered 3 and 13) across different chromosomal segments; the intervening genes are unrelated. **e**, Comparison with *K. waltii* reveals the duplicated nature of the *S. cerevisiae* genome, interleaving genes from sister segments on the basis of the ancestral gene order.

Computational Approach

- Find syntheny blocks
- Find overlaps in syntheny blocks
- Use duplicate synteny blocks do define “sister” regions in S. cerevisiae (145 sister regions covering 88% of the genome)



Some lessons from whole genome alignment of closely related species

Neutral evolution/natural selection

- **natural selection:** a process by which biological populations are altered over time, as a result of the propagation of heritable traits that affect the capacity of individual organisms to survive.
  - responsible for organisms being adapted to their environment.
  - The theory of natural selection was proposed by Charles Darwin and Alfred Russel Wallace in 1858, though vaguer and more obscure formulations had been arrived at by earlier researchers.
- **neutral theory of evolution** (Kimura 1960):
  - vast majority of molecular differences are selectively neutral.
  - these genome features are neither subject to, nor explicable by, natural selection.
  - most evolutionary change is the result of genetic drift acting on neutral alleles. Through drift, these new alleles may become more common within the population. They may subsequently decline and disappear, or in rare cases they may become fixed--meaning that the substitution they carry becomes a universal feature of the population or species
- **The neutralist-selectionist debate – which is the prevalent evolutionary force?**

Comparative Genome analysis tools

K_A / K_S ratio

Assume two closely related organisms (closely for this purpose is that probability of a back substitutions A→X→A are unlikely: example muse/rat; human chimpanzee)

K_A - #of coding base substitutions that results in amino-acid change

K_S - of coding base substitutions that do not results in amino-acid change (synonymous substitution rate)

K_A / K_S – measure of evolutionary constraints

$K_A / K_S << 1$; strong purifying selection

$K_A / K_S > 1$; possible adaptive or positive selection

Comparison mouse/rat human/chimpanzee

Initial sequence of the chimpanzee genome and comparison with Human genome, The Chimpanzee Genome Sequencing and Analysis Consortium, Nature, August 2005

K_A / K_S human-chimpanzee = 0.20
K_A / K_S mouse – rat = 0.13
Difference attributed to relaxed evolutionary constrains
4.4% human-chimpanzee orthologs have $K_A / K_S > 1$
and are hypothesized to be under positive selection (e.g.. genes involving reproduction)

Same species comparison

- HapMap project: a multi-country effort to identify and catalog genetic similarities and differences in human beings.
- In the initial phase of the Project, genetic data are being gathered from four populations with African, Asian, and European ancestry.
- First version 2005; Second version 2007

SNPs

- Single Nucleotide Polymorphism (SNP): a variation is a single nucleotide in a genome
- Typically we have two say *alleles* (here C and T) minor (less common) and major.
- minor allele frequency - the ratio of chromosomes in the population carrying the less common variant to those with the more common variant
- A second generation human haplotype map has over 3.1 million **SNPs**

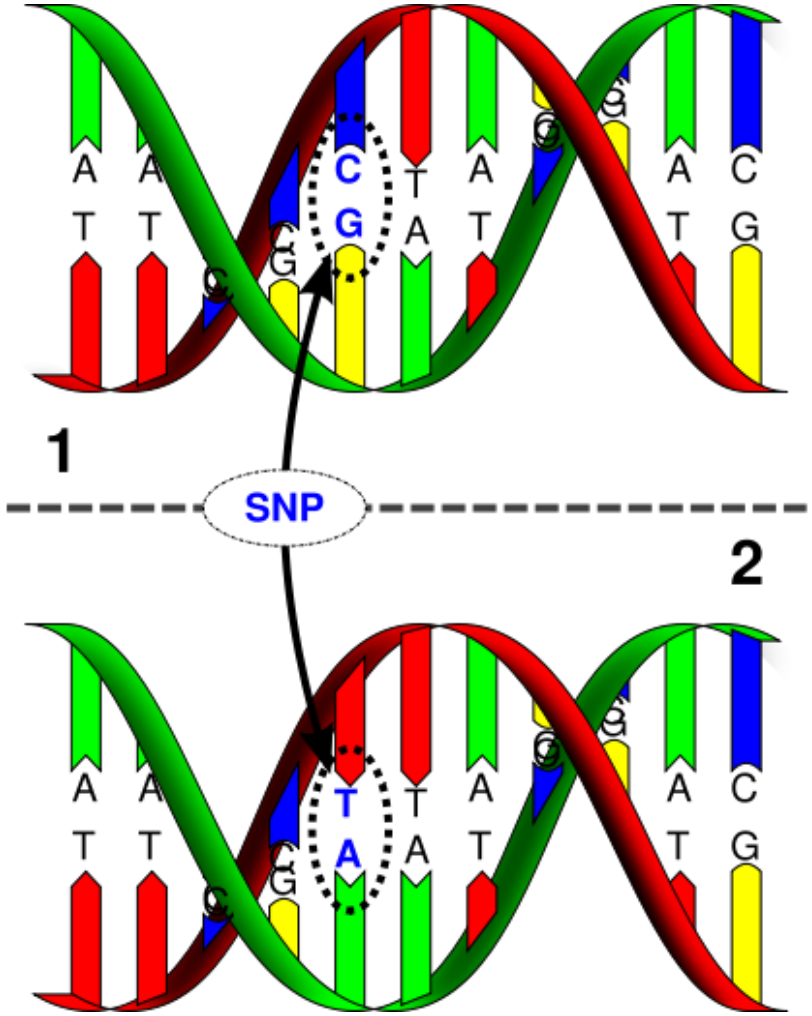


Figure from Wikipedia

Infinite site mutation model
(Kimura 1969)

- Under the **infinite sites mutation model** (Kimura,. 1969), mutations never occur twice at the same position
- This is condition is equivalent to the prefect phylogeny

Recombination

- In eukaryotes recombination commonly occurs during meiosis as chromosomal crossover between paired chromosomes
- It has been demonstrated that the points of recombination crossover are not uniformly distributed but instead most of recombinations occur in the so called recombination hotspots
- Recombination hotspots are not well preserved between human and chimpanzees

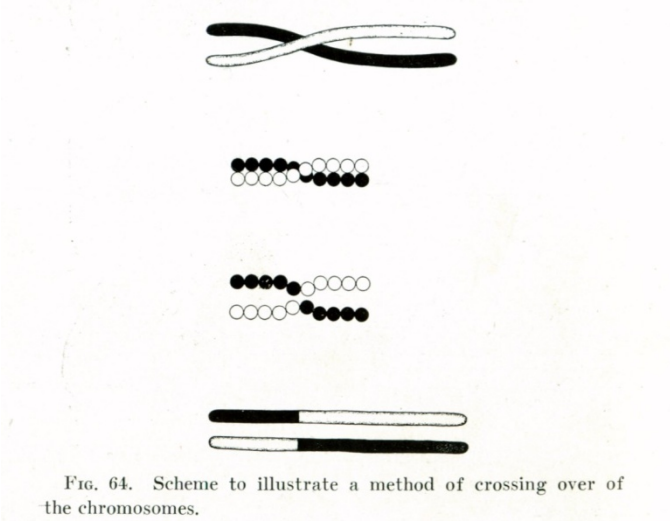


FIG. 64. Scheme to illustrate a method of crossing over of the chromosomes.

Copy of the original figure by Morgan (1916)

Discovering recombination events

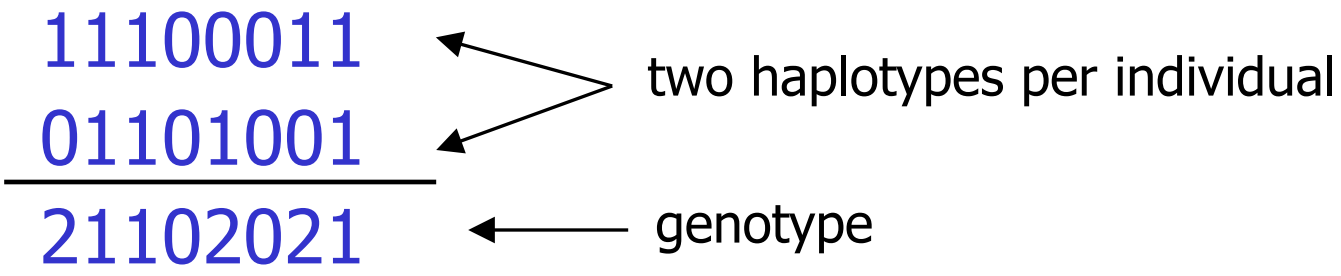
- Note infinite mutation site assumption is equivalent to perfect phylogeny.
- Four gamete test (Hudson,Kaplan 1985):

SNP1	SNP2
0	0
1	0
1	1
0	1

Under infinite site mutation model (perfect phylogeny) there must be a recombination event between these SNPs

Haplotype, genotype, phasing problem

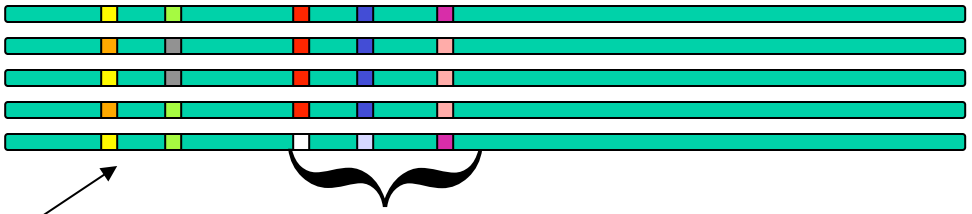
- **Haplotype:** description of SNP alleles on a chromosome
  - 0/1 vector: 0 for major allele, 1 for minor
- **Genotype:** description of alleles on both chromosomes
  - 0/1/2 vector: 0 (1) - both chromosomes contain the major (minor) allele; 2 - the chromosomes contain different alleles



- **Phasing:** the problem of assigning haplotypes given genotypes
  - Popular program: PHASE

Linkage disequlibrium

- the non-random association of alleles at two or more loci

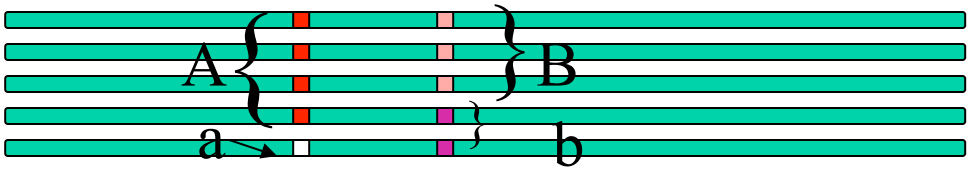


random

haplotype frequencies deviate from the values they would have if the genes at each locus were combined at random.

Strong disequlibrium – low recombination level

Measuring linkage disequilibrium



p_{AB}=3/5	p_{aB} =0/5	q_B =3/5
p_{Ab} =1/5	p_{ab} =1/5	q_b =2/5
q_A =4/5	q_a =1/5	1

Haplotype frequency

Allele frequency

$$D = p_{AB} - q_A q_B = 3/5 - (3/5)(4/5) = 15/25 - 12/25 = 3/25$$

observed expected

This deviation of the observed frequencies from the expected is referred to as the linkage disequilibrium parameter, D, introduced by Robbins (1918) and named by Lewontin and Kojima (1960)

Other measures

$D' = D/D_{max}$

where

$D_{max} = -\min\{q_A q_B, q_a q_b\}$ if $D' < 0$.

$D_{max} = \min\{q_a q_B, q_A q_b\}$ if otherwise

In our example: (3/25)/(3/25)=1 (1 implies at least one of the possible haplotypes was not observed)

Other measure

$r^2 = D^2 / (q_A q_B q_a q_b)$.

Haplotype blocks

- human genome has a haplotype block structure, such that it can be divided into discrete blocks of limited haplotype diversity (high LD).
- To carry out a genome-wide association study, researchers use two groups of participants: people with the disease being studied and similar people without the disease.
- If certain genetic variations (usually represented by SNPs) are found to be significantly more frequent in people with the disease compared to people without disease, the variations are said to be "associated" with the disease.
- Assoction of SNP with a disease does not necessarily means that this SNP is causative, but rather points to haplotype block that may contain gene of interest.