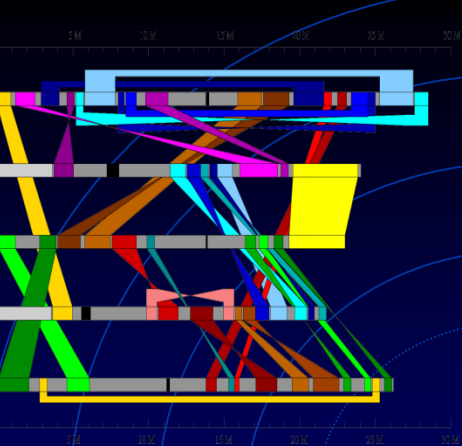




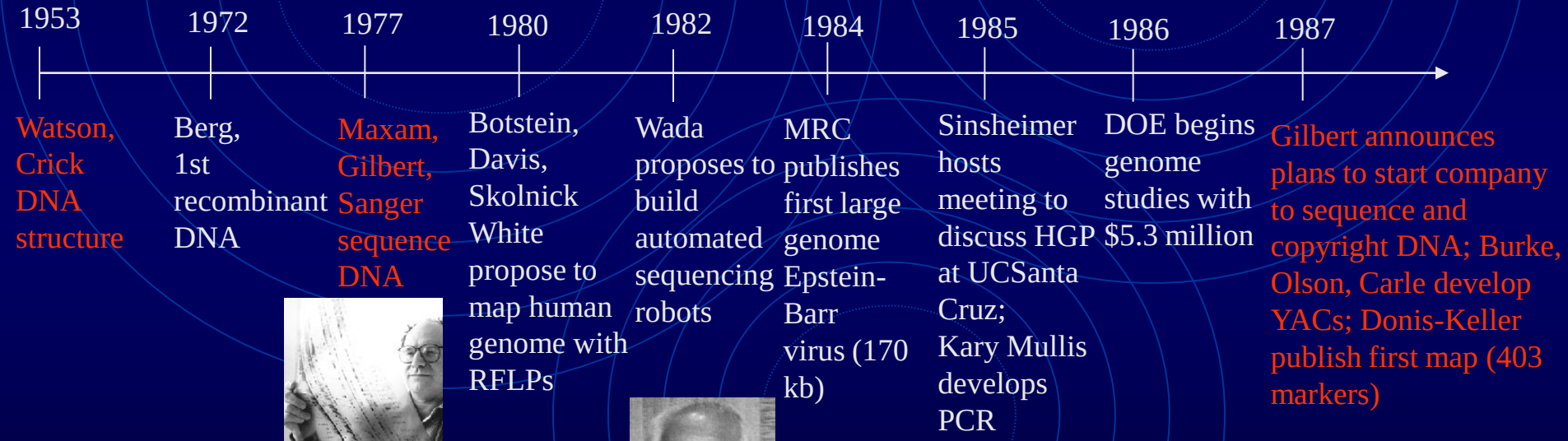
4. GENOME ANALYSIS

- *4.1 Genome sequencing*
- *4.2 Composition and evolution of genome*
- *4.3 Gene finding*
- *4.4 Non-coding gene finding*
- *4.5 Repeat and other elements finding*
- *4.6 Genome-wide expression profiling*

4.1 Genome Sequencing

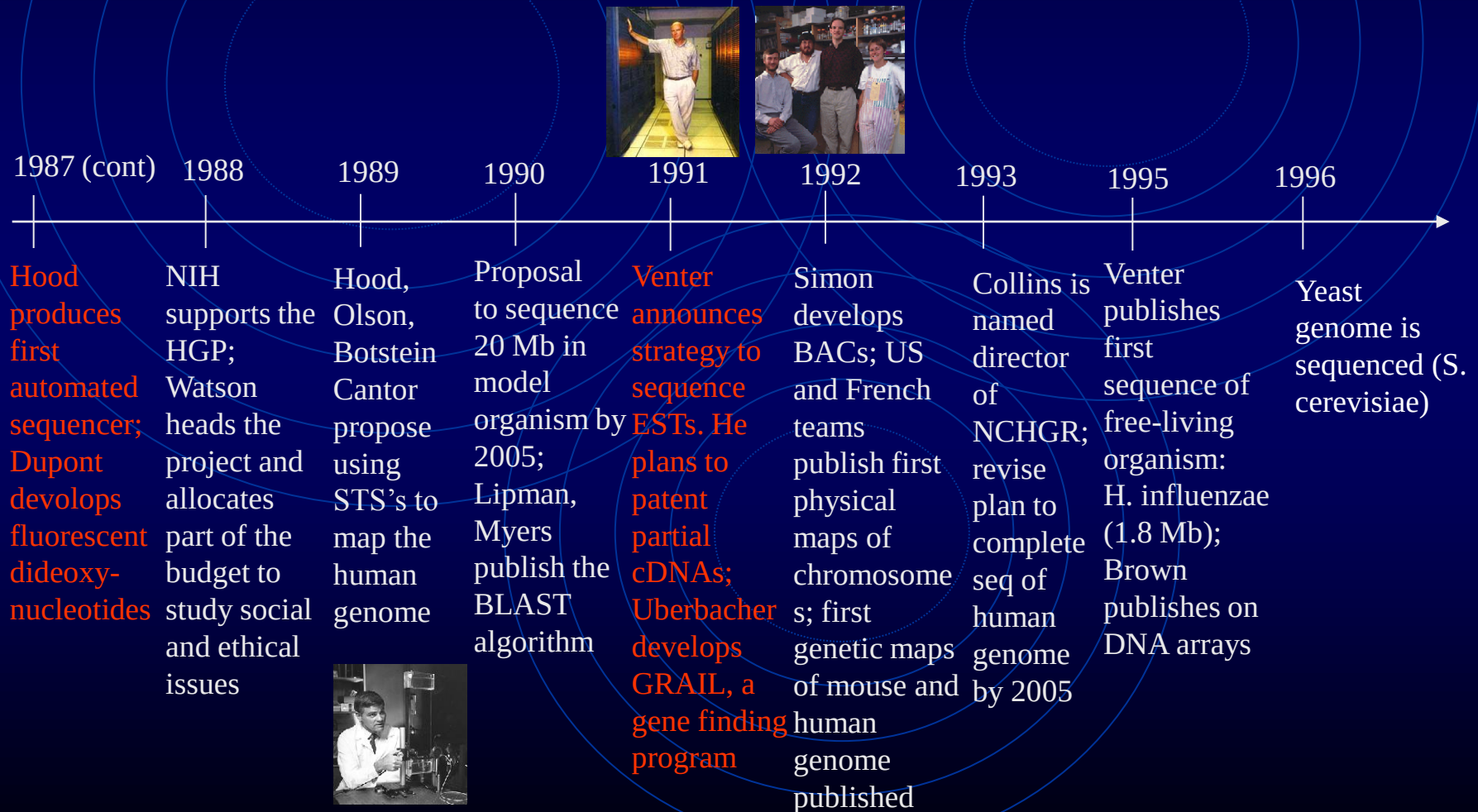
- *4.1.1 History of genome investigation: the human genome project*
- *4.1.2 Strategic issues for genome sequencing*

History of the Human Genome Project



From *SCIENCE* the human genome issue

History of the Human Genome Project (continued)



History of the Human Genome Project (continued)



1997
Blattner, Plunket complete *E. coli* sequence; a capillary sequencing machine is introduced.

1998
SNP project is initiated; rice genome project is started; Venter creates new company called Celera and proposes to sequence HG within 3 years; *C. elegans* genome completed

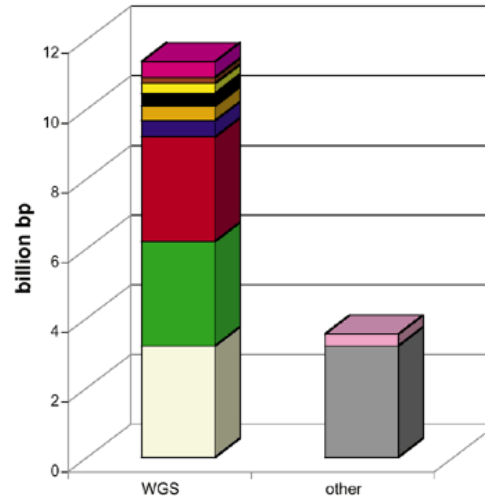
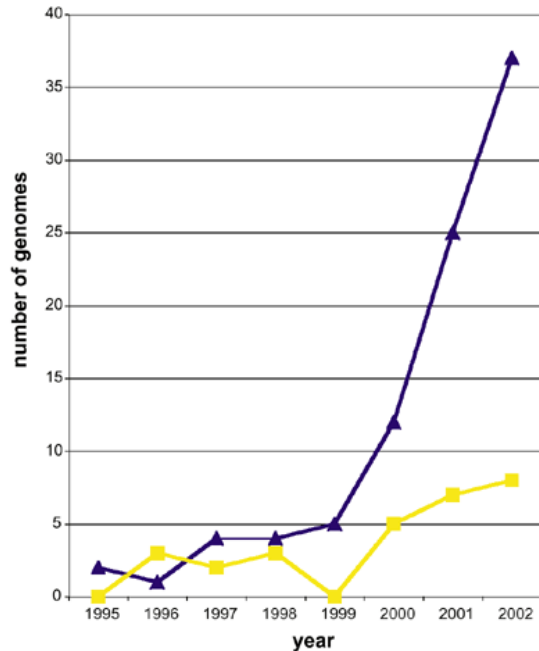
1999
NIH proposes to sequence mouse genome in 3 years; first sequence of chromosome 22 is announced

2000
Celera and others publish *Drosophila* sequence (180 Mb); human chromosome 21 is completely sequenced; proposal to sequence puffer fish; Arabidopsis sequence is completed

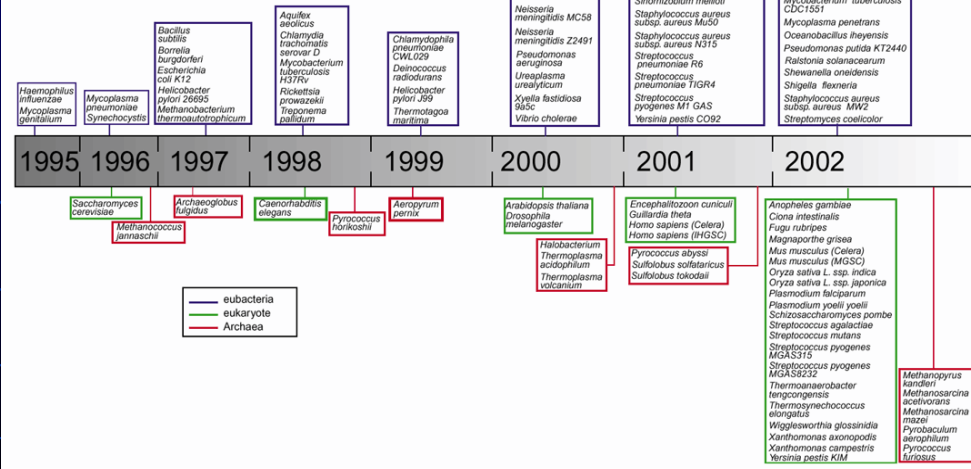
2001
Celera publishes human sequence in *Science*; the HGP consortium publishes the human sequence in *Nature*



The number of genomes sequenced each year since 1995



completed genomes



Citation tree for the WGS sequencing of *H. influenzae*

WGS

H. influenzae
Fleischmann et al. 1995 (ref. 37) (2294)

genome sequencing

Methanococcus jannaschii - Bult et al. 1996 (ref. 94) (1468W)
Streptomyces coelicolor Chr A3 - Redenbach et al. 1996 (ref. 95) (229C)
S. cerevisiae - Goffeau et al. 1996 (ref. 8) (662C)
E. coli K12 - Blattner et al. 1997 (ref. 9) (1747C)
Helicobacter pylori - Tomb et al. 1997 (ref. 96) (1172W)
Baillus subtilis - Kunst et al. 1997 (ref. 97) (1111)
Archaeoglobus fulgidus - Klenk et al. 1997 (ref. 98) (713W)
Borrelia burgdorferi - Fraser et al. 1997 (ref. 99) (550W)
Rickettsia prowazekii - Andersson et al. 1998 (ref. 100) (385)
Aquifex aeolicus - Deckert et al. 1998 (ref. 101) (453)
Treponema pallidum - Fraser et al. 1998 (ref. 102) (304W)
Plasmodium falciparum (Chr2) - Gardner et al. 1998 (ref. 103) (243)
Drosophila melanogaster - Adams et al. 2000 (ref. 67) (1073)
Neisseria meningitidis - Tettelin et al. 2000 (ref. 104) (194)
Homo sapiens - Venter et al. 2001 (ref. 18) (1258W), Lander et al. 2001 (ref. 19) (1014C)

comparative genomics

Mycoplasma species - Himmelreich et al. 1996 (ref. 38) (472)
Mycobacteria species - Philipp et al. 1996 (ref. 40) (177)
Archae and Eubacterial species - Nelson et al. 1999 (ref. 39) (347); Brown et al. 1997 (ref. 41) (166)
D. melanogaster, *C. elegans* and yeast - Rubin et al. 2000 (ref. 42) (383)
H. influenzae and *E. coli* - Tatusov et al. 1997 (ref. 45) (346)

gene discovery/functional assignment

proteosome - Coux et al. 1996 (ref. 105) (860R)
DNA topoisomerases - Wang 1996 (ref. 106) (786R)
multidrug efflux systems - Paulsen et al. 1996 (ref. 107) (300R); Nikaido et al. 1996 (ref. 108) (226R)
DNA glycosylases - Krokan et al. 1997 (ref. 109) (230)
establishing gene function - Miklos et al. 1996 (ref. 110) (219R)
metal resistance in Bacteria - Silver et al. 1996 (ref. 111) (166R)

host pathogen interactions

cystic fibrosis - Govan et al. 1996 (ref. 50) (269)
Rhizobium/Legumes - Freiberg et al. 1997 (ref. 51) (259); Zumft 1997 (ref. 52) (237R)
nitrogen regulation - Merrick et al. 1995 (159R)

minimal genome

Fraser et al. 1995 (ref. 85) (1088)
Mushegian et al. 1996 (ref. 86) (161)

proteomics

Yates 1998 (ref. 49) (212)
Blackstock 1999 (ref. 48) (183)

tRNA detection

computational - Lowe et al. 1997 (ref. 112) (182)

yeast reverse genetics

Dujon 1996 (ref. 46) (223)

bacterial identification

DNA typing - Gurtler et al. 1996 (ref. 113) (193)

protein classification

COGs - Tatusov et al. 1997 (ref. 45) (344)

DNA microarray

yeast ORFs - Lashkari et al. 1997 (ref. 47) (193)

Genomes in numbers

- Sizes:

- virus: 10^3 to 10^5 nt
- bacteria: 10^5 to 10^7 nt
- yeast: 1.35×10^7 nt
- mammals: 10^8 to 10^{10} nt
- plants: 10^{10} to 10^{11} nt

- Gene number:

- virus: 3 to 100
- bacteria: ~ 1000
- yeast: ~ 7000
- mammals: ~ 30'000
- Plants: 30'000-50'000?

Sequencing projects

- « small » genomes ($<10^7$): bacteria, virus
 - Many already sequenced (industry excluded)
 - More than 90 microbial genomes already in the public domain
 - More to come! (one new every two weeks...)
- « large » genomes (10^7 - 10^{10}) eucaryotes
 - 12 finished (*S.cerevisiae*, *S. Pombe*, *E. cuniculi*, *C.elegans*, *D.melanogaster*, *A. gambiae*, *D. rerio*, *F. rubripes*, *A.thaliana*, *O. sativa*, *M. musculus*, *Homo sapiens*)
 - Many more to come: rat, pig, cow, maize (and other plants), insects, fishes, many pathogenic parasites (*Plasmodium*...)
- EST sequencing
 - Partial mRNA sequences
 - $\sim 12 \times 10^6$ sequences in the public domain



“基因组地理”探险：四张图

- 哥伦布几百年前发现了美洲、澳洲等新大陆，最终使人类认识了地球上的每一块处女地。于是有人形象地把人类目前的基因组研究形象地比喻为“地球探险”，并把基因组研究称为基因组地理（genomic geography）。
- 我们当前的测序技术还不允许我们把整条染色体拿来测序，而只能一小段一小段（一般一个测序反应为500个碱基对长度）地进行。
- 这些小段长度（500bp）的DNA片段对于整条染色体而言，就仿佛是染色体长线上的一个个小点，那么如何把这些测序完成的小片段连接在一起而获得整条染色体的序列呢？这是人类基因组计划的最大难题之一。

A、遗传图谱(genetic map)

- 遗传学图距是以形成精子或卵子的减数分裂过程中，两个位点之间进行交换、重组百分率（cM,厘摩尔根）为单位的，反映基因遗传效应的基因图谱。
- 限制性片段长度多态性（RFLP）\微卫星标记（microsatellite marker）\SNP（single nucleotide polymorphysm）的遗传标记系统，即单核苷酸多态性

B、物理图谱(physical map)

- 物理图以Mb、kb、bp作为图距，以DNA探针的STS（sequence-tagged site，序列标签位点）序列为路标。
- 构建物理图谱的一个主要内容是把含有STS对应序列的DNA的克隆片段连接成相互重叠的“片段重叠群（conting）”。
- 在YAC载体的基础上，现有BAC（细菌人工染色体库）、PI（一种源于λ噬菌体的载体）、PAC（来源于PI的人工染色体）、MAC（一种类似于YAC，但以哺乳细胞作为宿主细胞的新型载体）、Fosmid（一种类似于BAC，来源于大肠杆菌的F基因的载体）。

C、序列图谱(sequence map)

- 由全部核苷酸组成的基因组序列图。前面所谈的遗传图与物理图的构建都是为了绘制序列图而建的。
- 1977年，人类完成对自然界第一个基因组（全长5.3kb的 ϕ x174噬菌体）全序列测序，整个测序历时将近一年时间。1987年第一台自动测序仪投入使用后，1990年完成了长230kb的人巨细胞病毒（HCMV）的全序列测序工作。后来又得到了第一个细菌基因组全序列（1995，1.9Mb）、模式生物酵母基因组全序列（1996，12Mb）、线虫基因组全序列（1998，97Mb）、果蝇基因组全序列（1999，136Mb），直到人类自身全序列（2000，3286Mb）。水稻（2001，450M）。

D、基因图谱(gene map)

- 基因图谱就是基因组中全部基因的位置、结构与功能的明细图。
- 基因图谱的意义在于它能有效地反映在正常或受控条件中表达的全基因的时空图。

4.1.2 基因组测序的策略

How to sequence a genome?

两个基因组测序策略:

逐步克隆方法

(clone by clone)

全基因组鸟枪方法

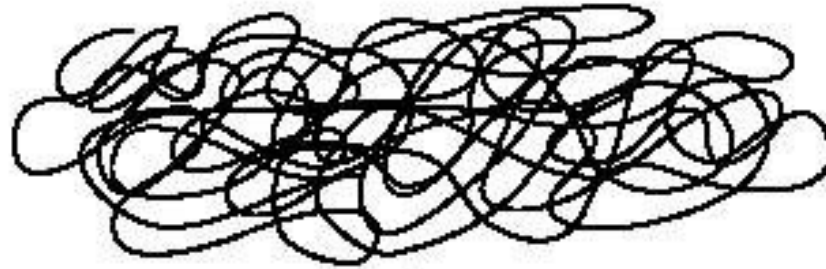
(whole genome shotgun,
WGS)



Clone by Clone Approach

Hierarchical shotgun sequencing

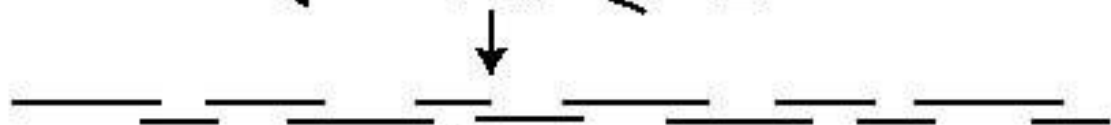
Genomic DNA



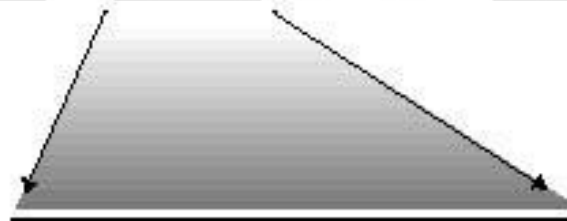
BAC library



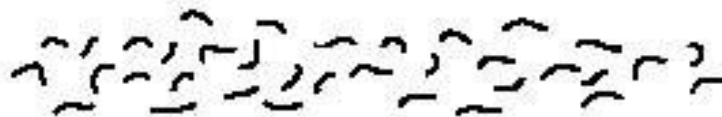
Organized mapped large clone contigs



BAC to be sequenced



Shotgun clones



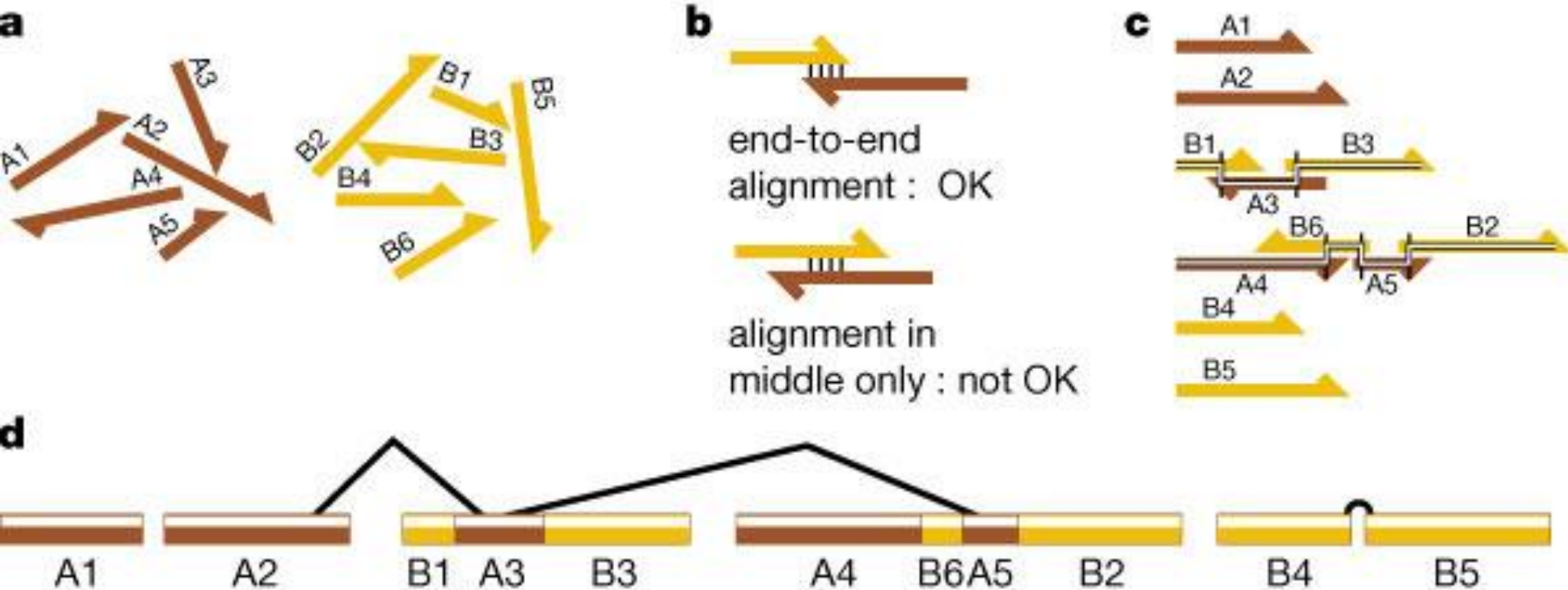
Shotgun sequence

...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...

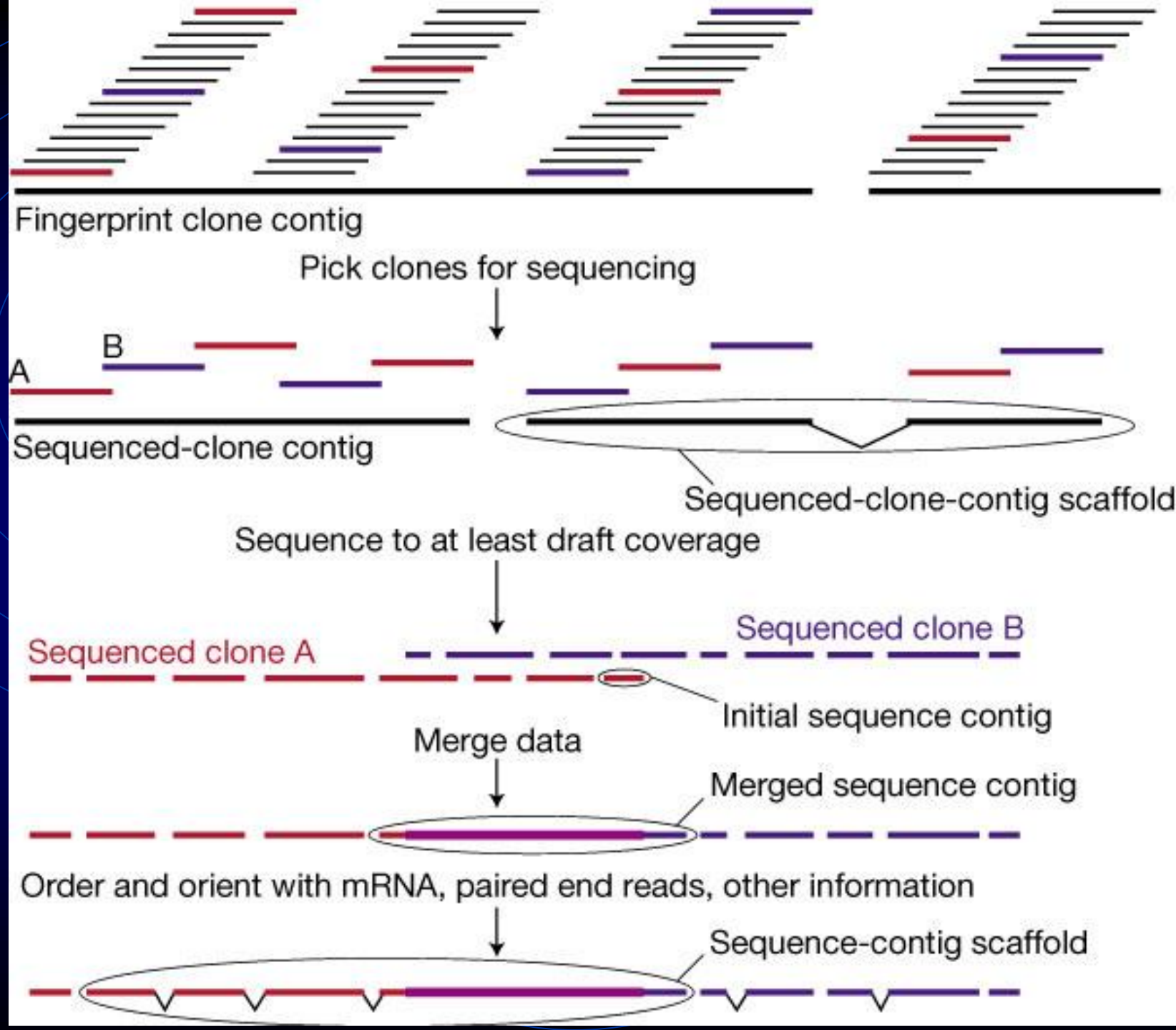
Assembly

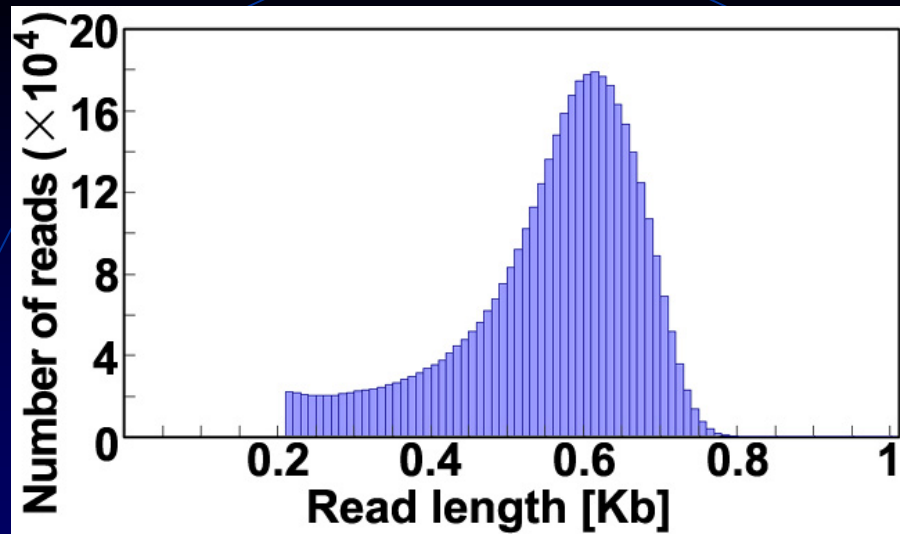
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...

The key steps in assembling individual sequenced clones into the draft genome sequence

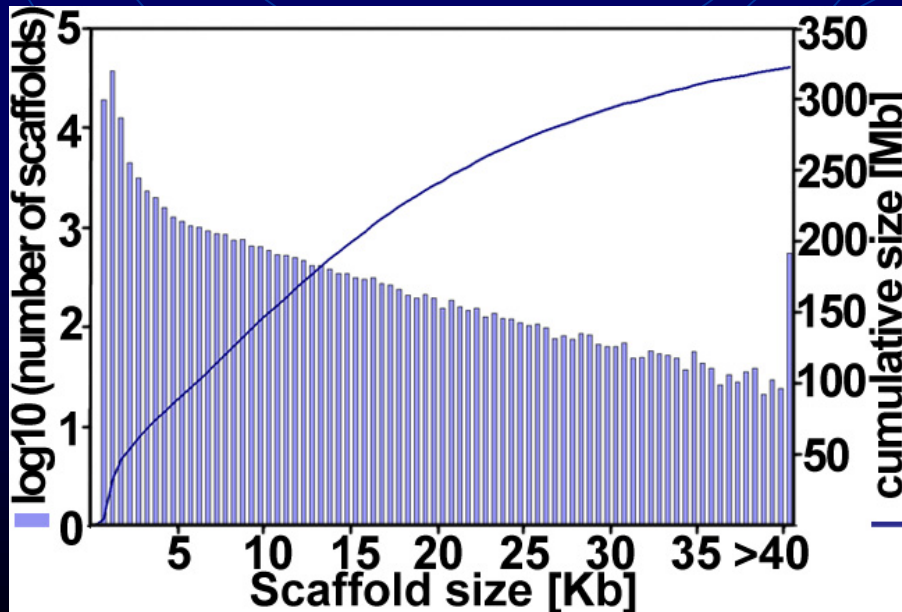


Contig and Scaffold

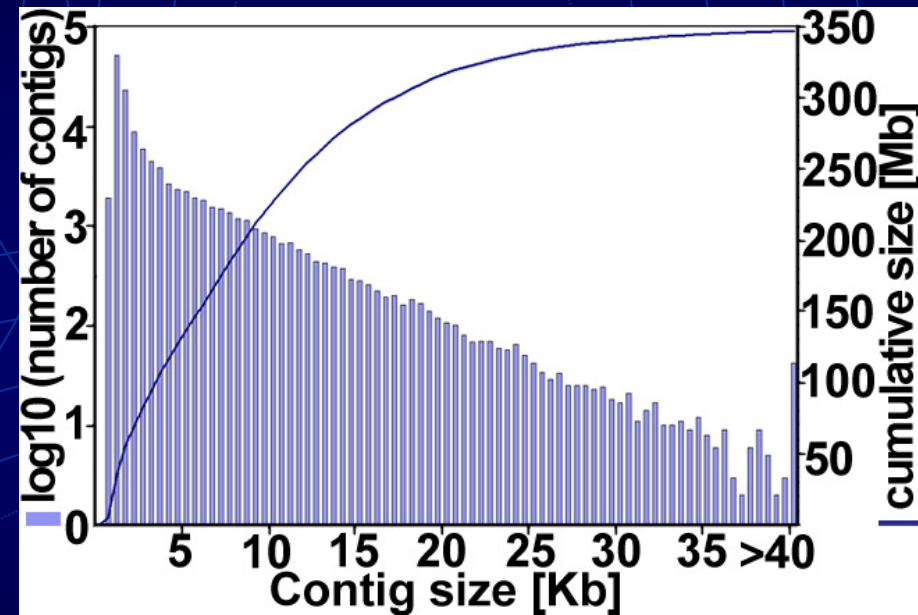




Contigs: 127,550
(N50=6,688 bp)



Quality:
546 bp at Q20
(*indica* rice genome)



Scaffolds: 102,444
(N50=11,764 bp)

4.2 Composition and evolution of genome



- *Genome Anatomy: Structure and Composition of Genomes*
- *Genomic Evolution*

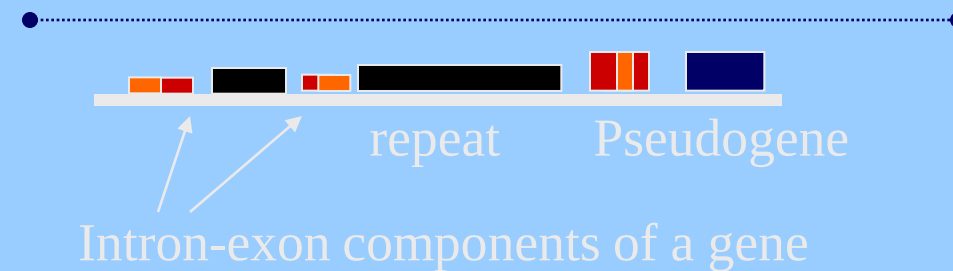
4.2.1 Appearance of Genome

🐎 One to many chromosomes

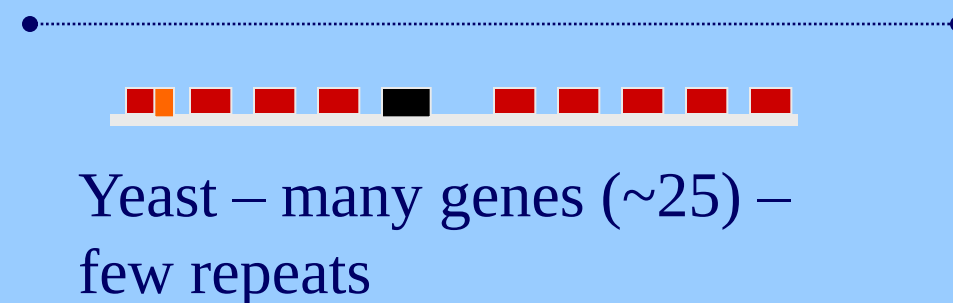
🐎 Repeat sequences common in some genomes e.g. 35% of human are transposable elements -10% *Alu*, 14.6% LINE1 sequences

🐎 Gene structure varies – no. and length of introns

What does 50 kb of sequence look like?



Human – very few genes - repeats



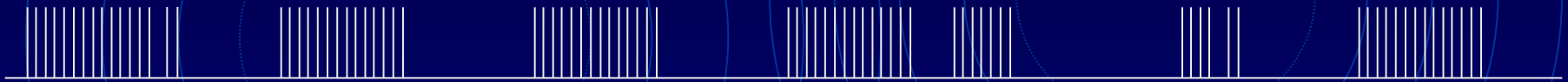
Yeast – many genes (~25) – few repeats



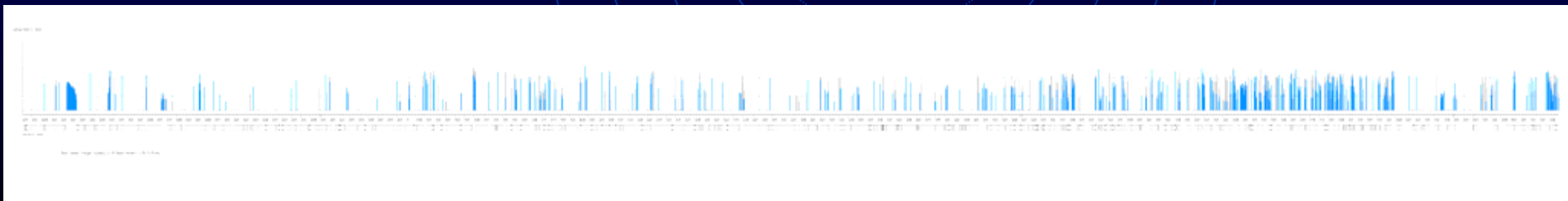
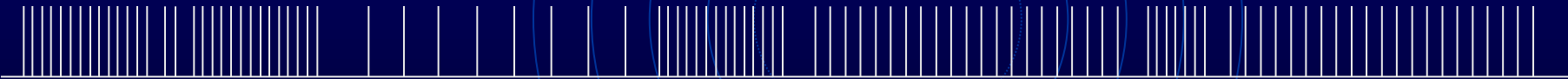
Maize – mostly repeats

The Reverse-phase of “Gene-jungles”:

“Repeat-jungles”, But No “Deserts” !
MDRs in rice genome

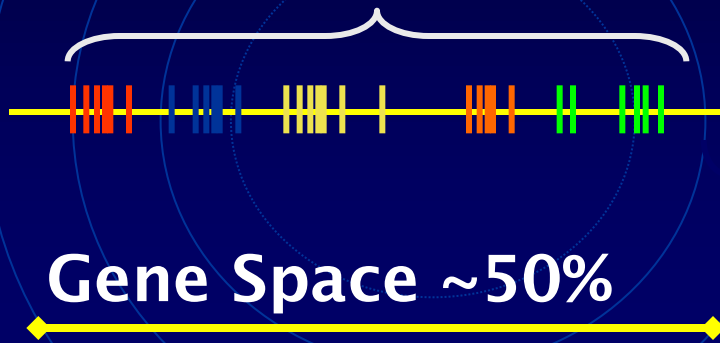


MDRs in human genome

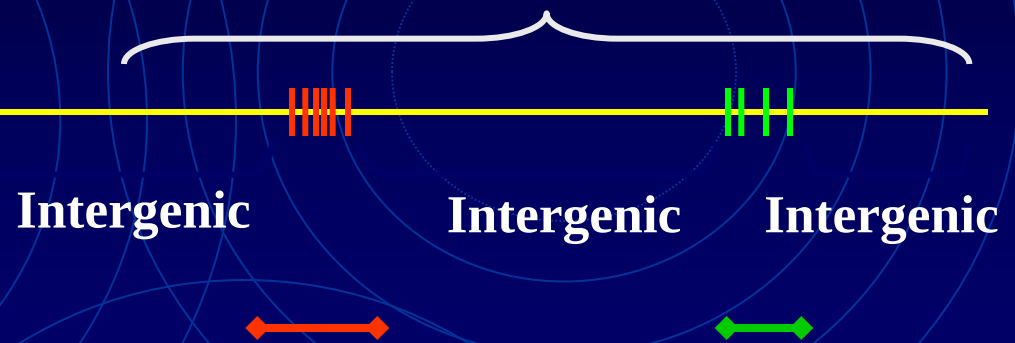


“Jungle-Desert” Or “Big –Small”?

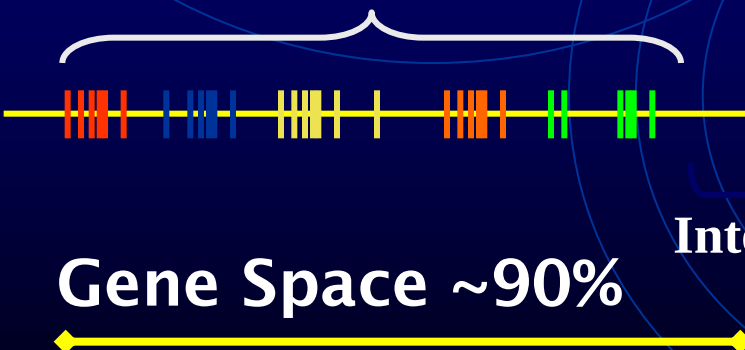
The Gene “Jungle”



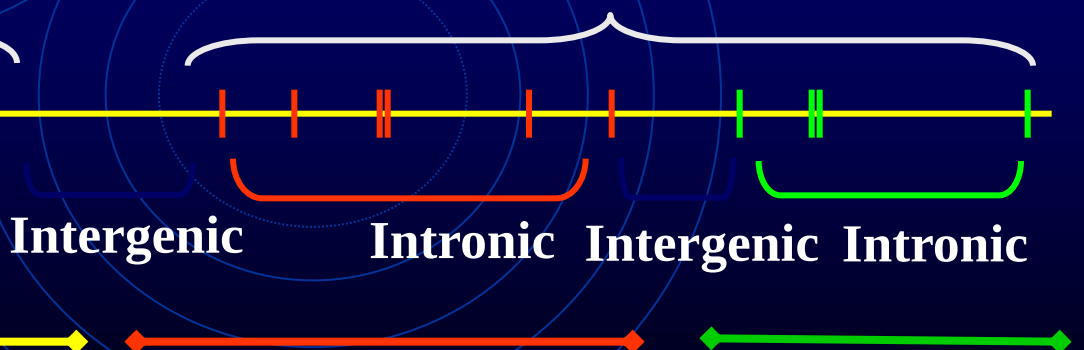
The Gene “Desert”



Many Small Genes

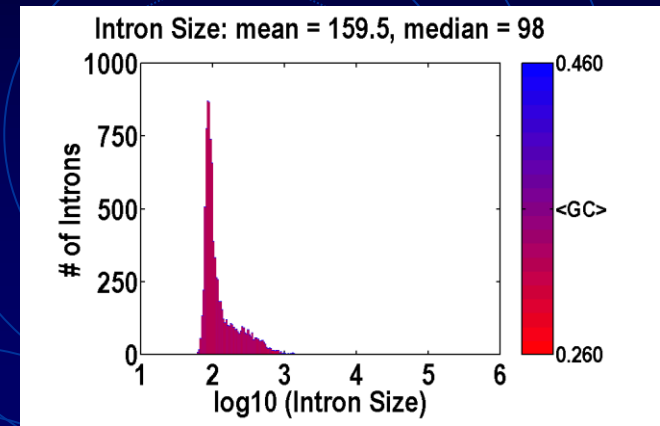
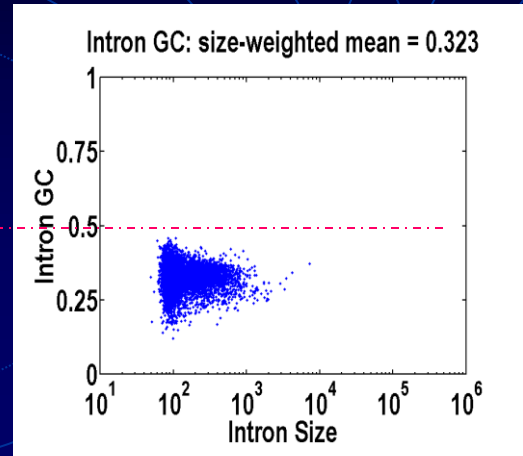
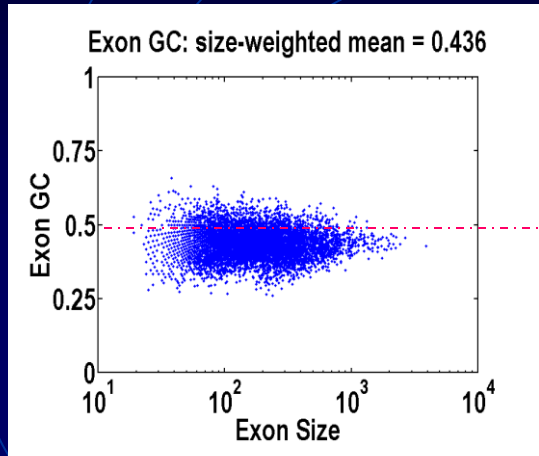


Two Big Genes

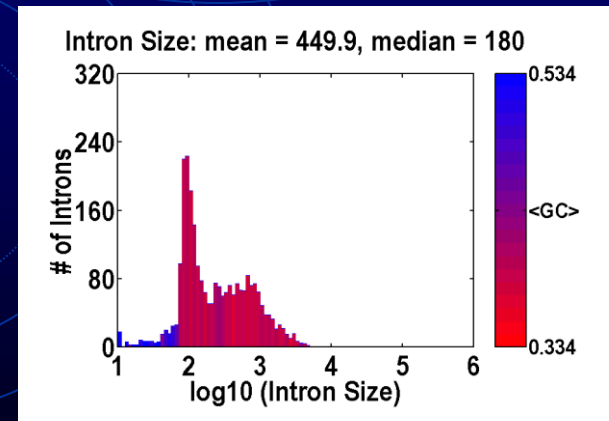
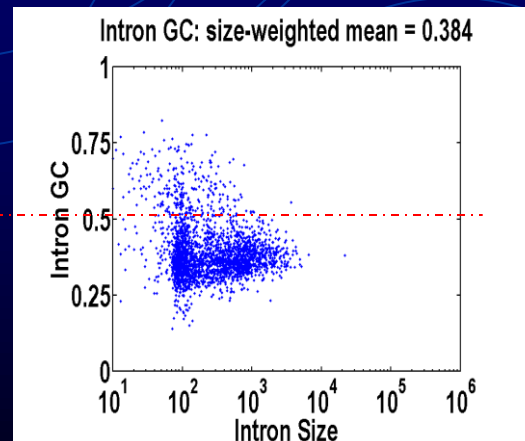
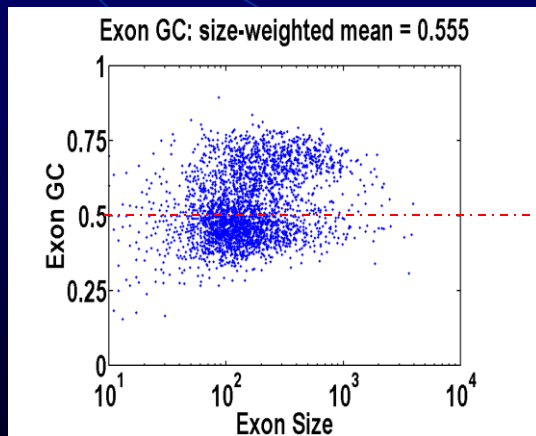


Exon-Intron Organization In Plant Genomes

Arabidopsis resembles nematode.

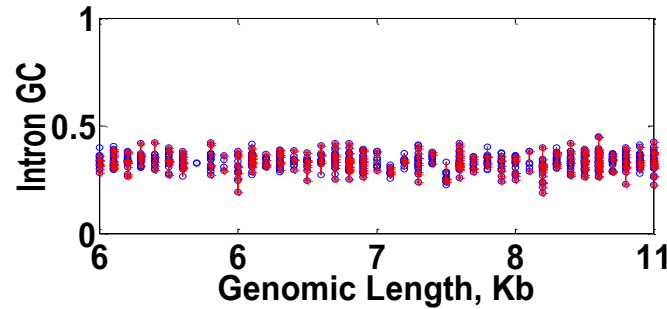
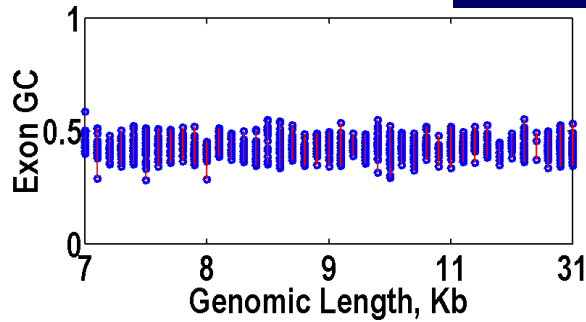


Rice resembles cold-blooded vertebrates.

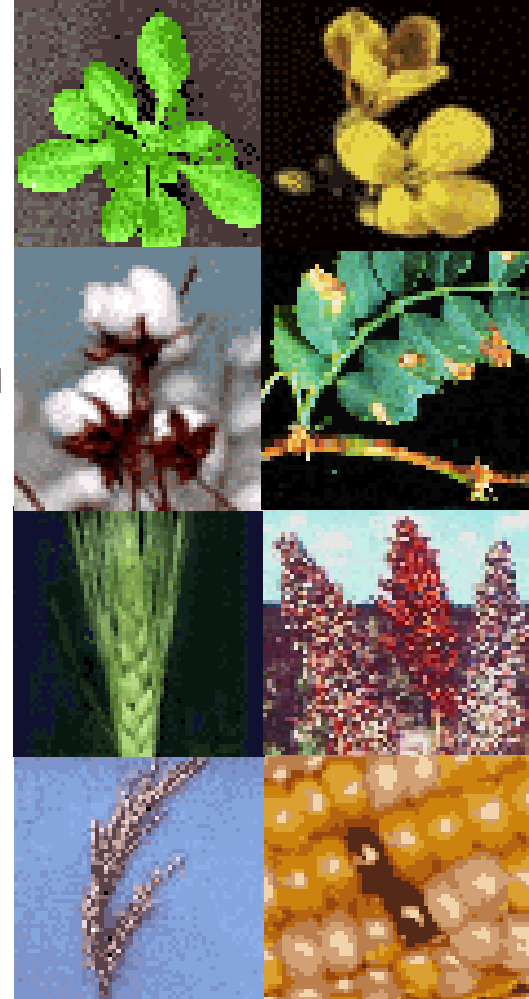
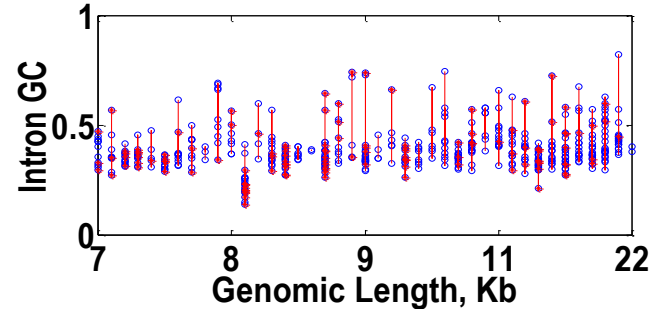
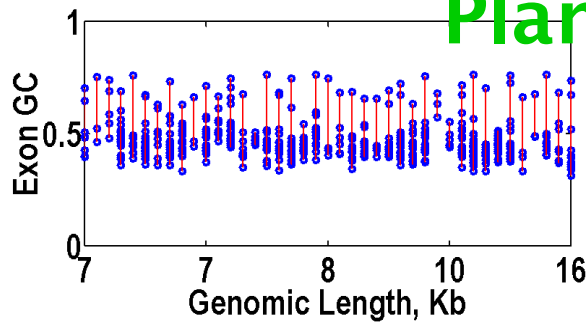


Not Between-gene, But Within-gene

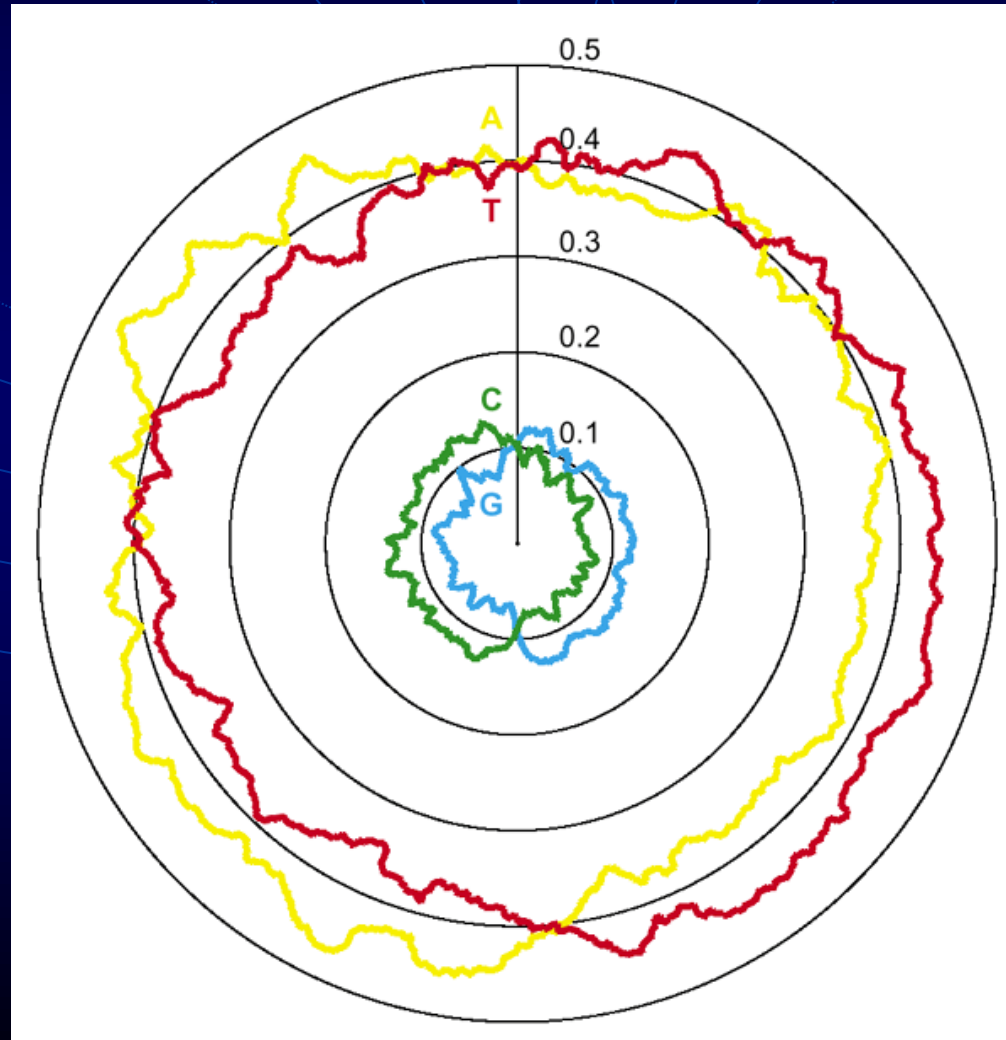
Dicot Plant



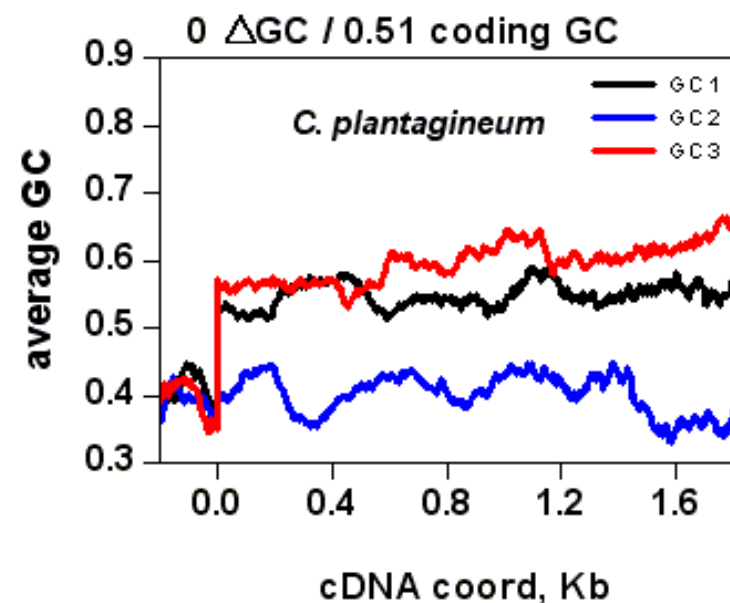
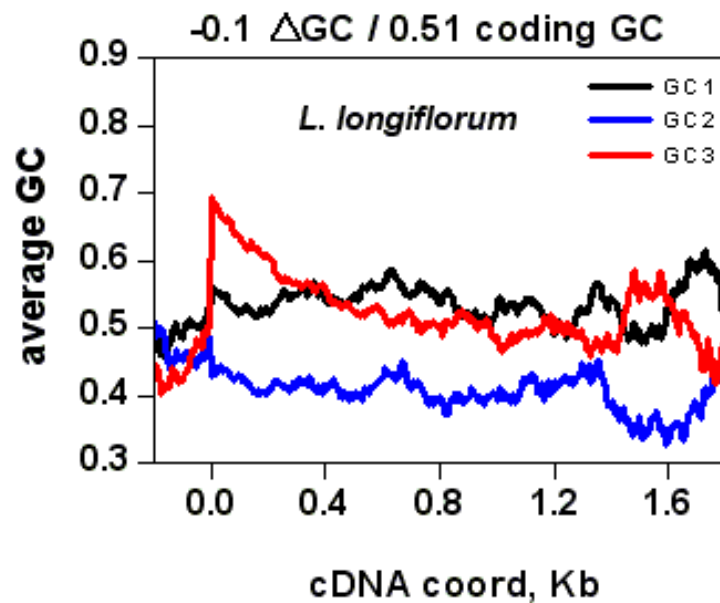
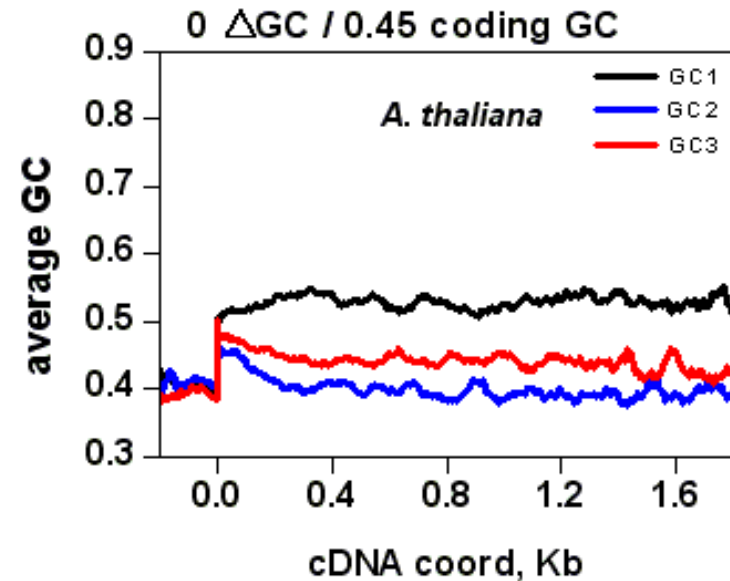
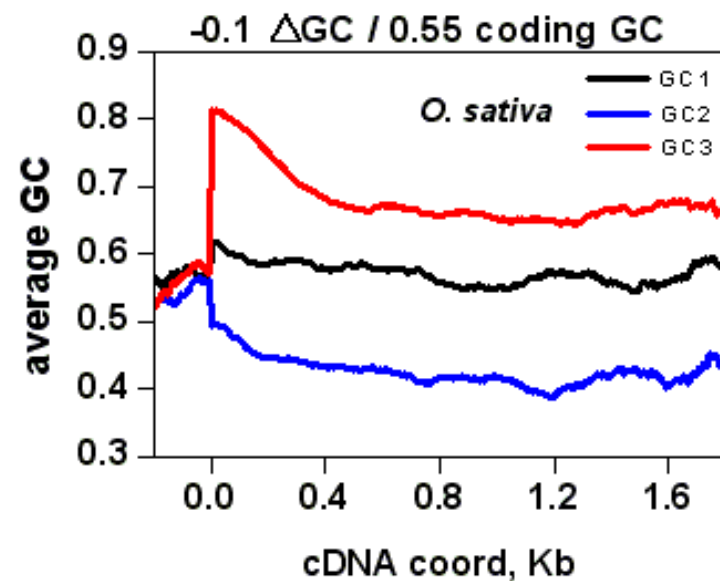
Monocot Plant



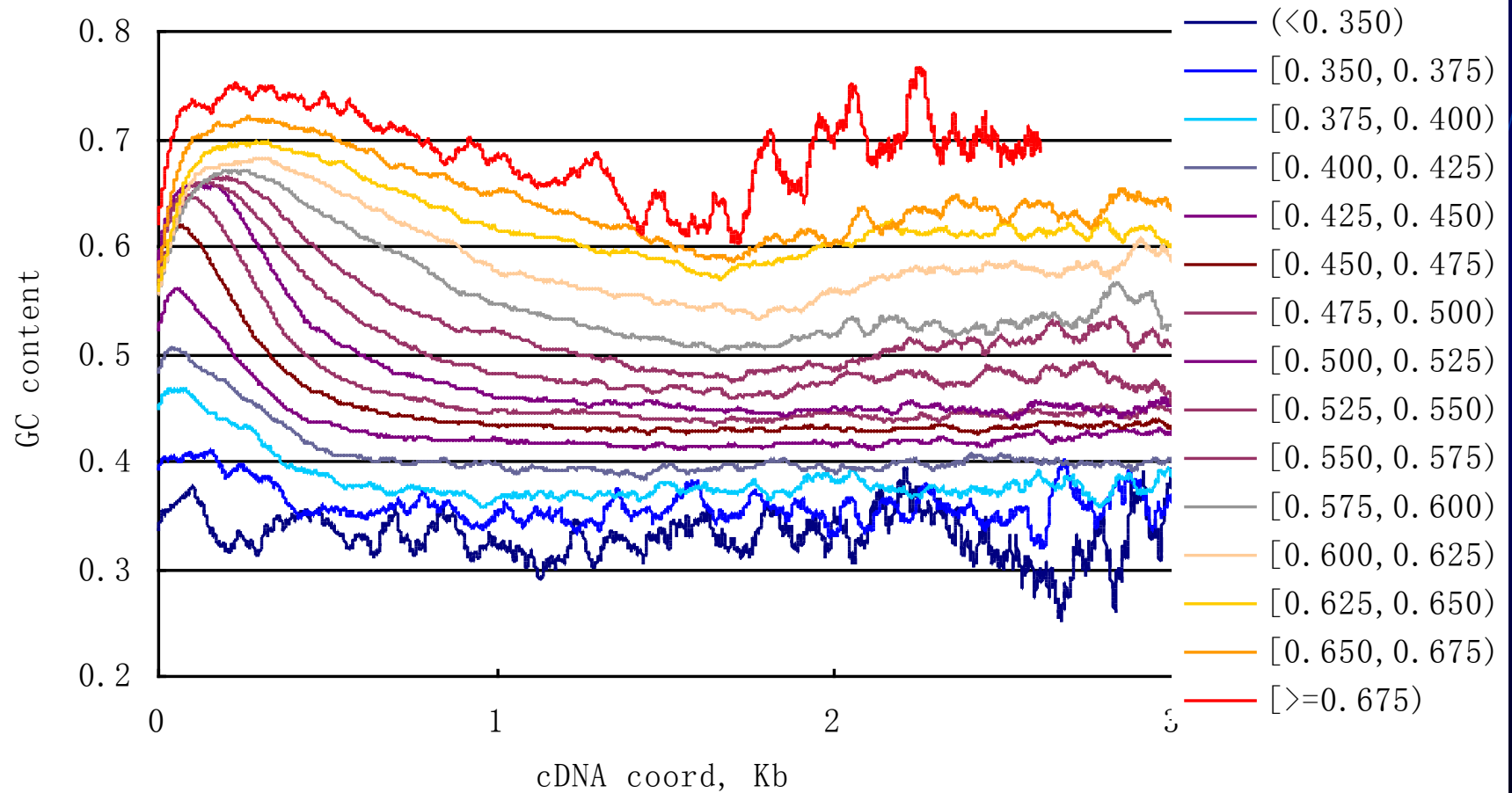
Variation in base composition around the genome of Camplobacter jejuni: Deamination Theory



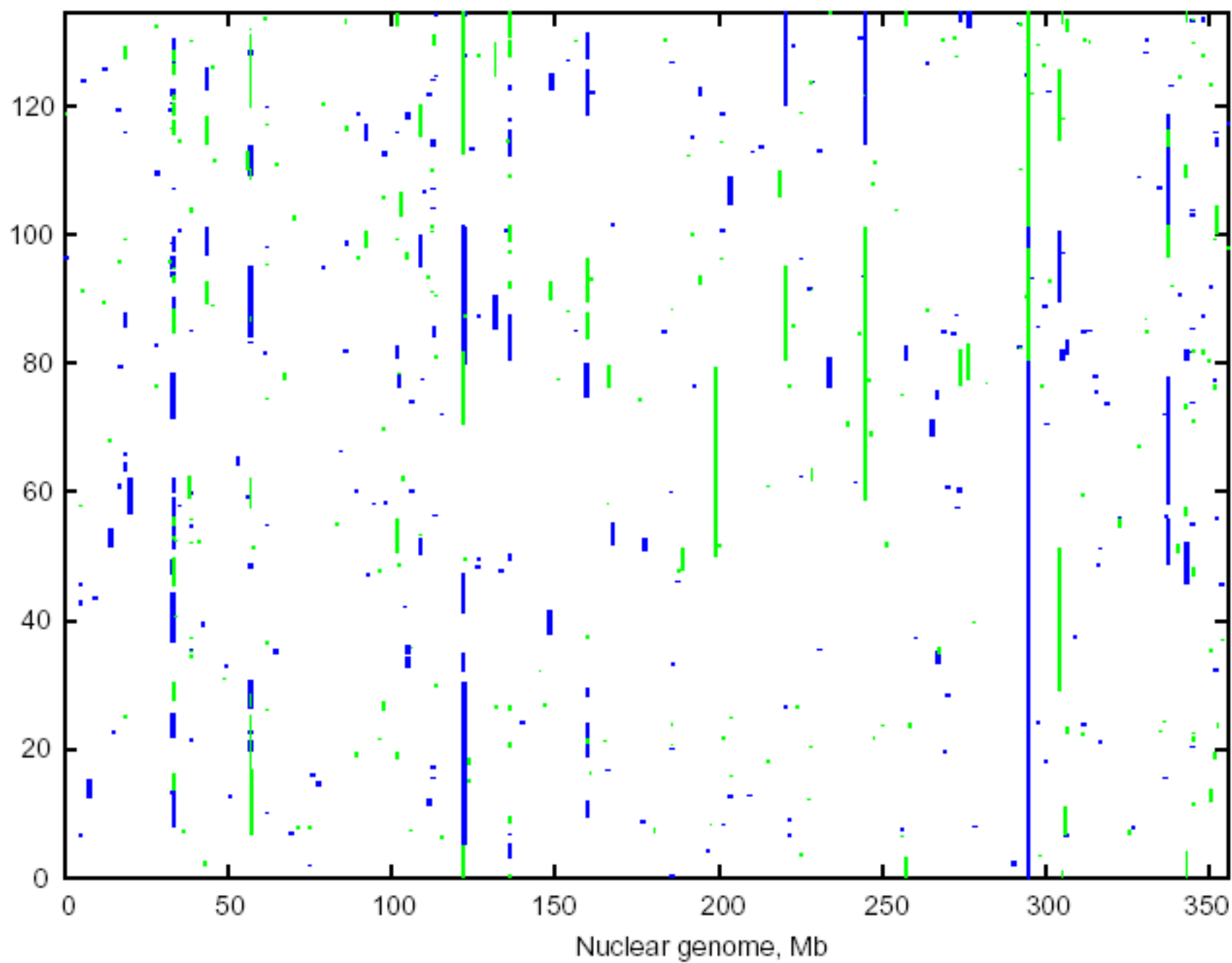
Nature Genetics,
Supplement, 33,
2003

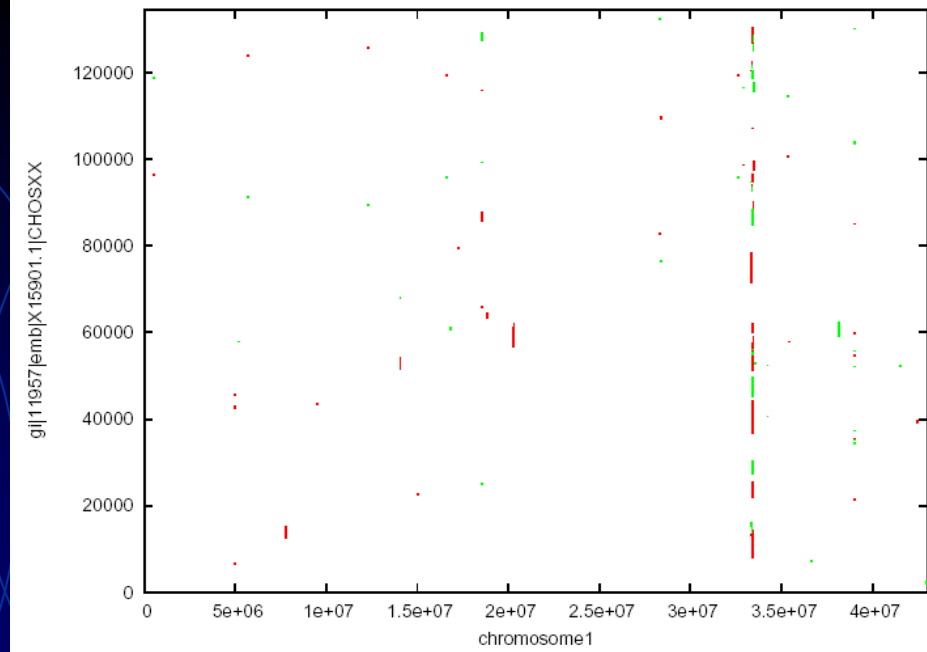
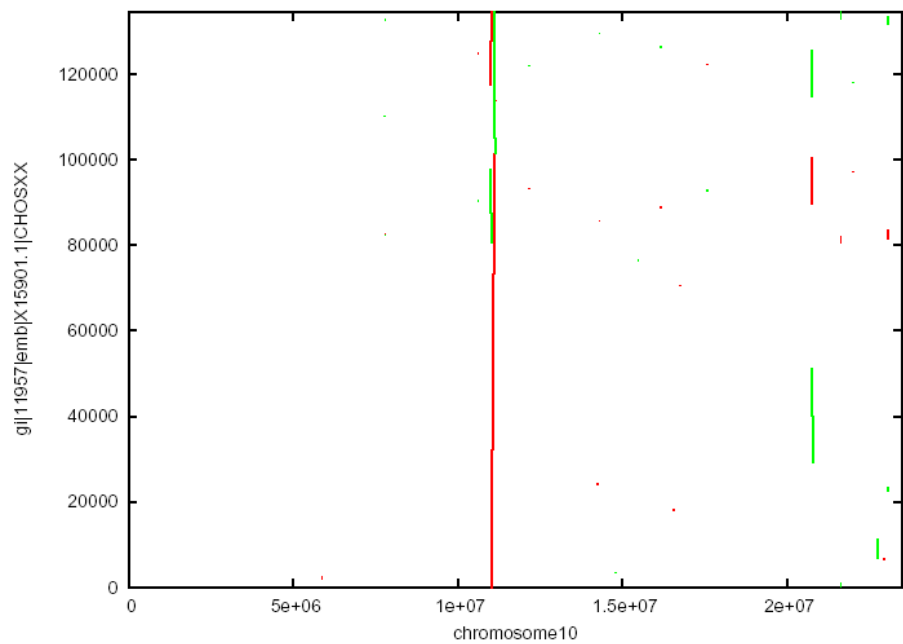


GC content and gradient of rice genes

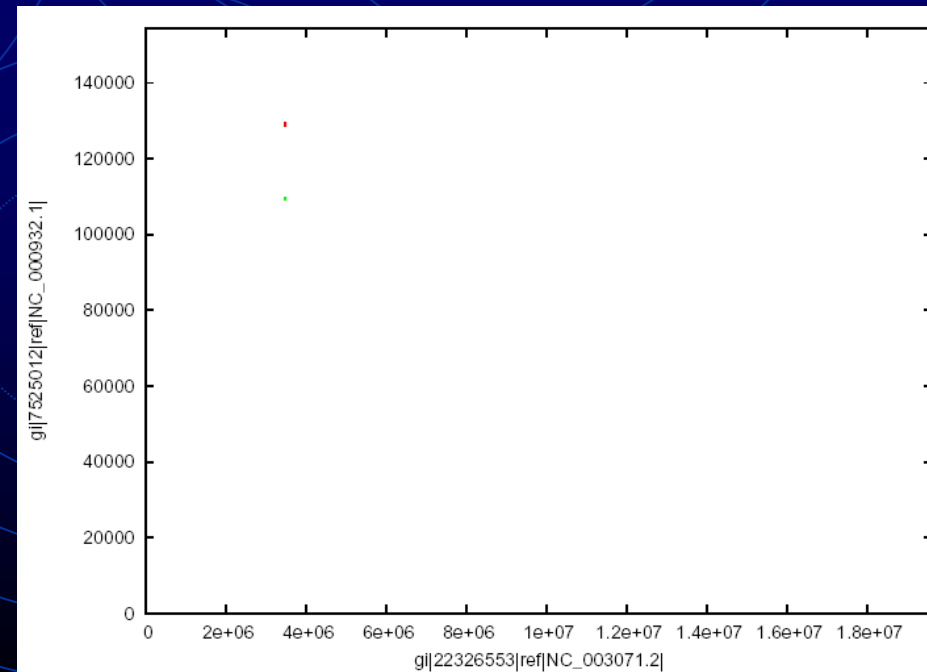
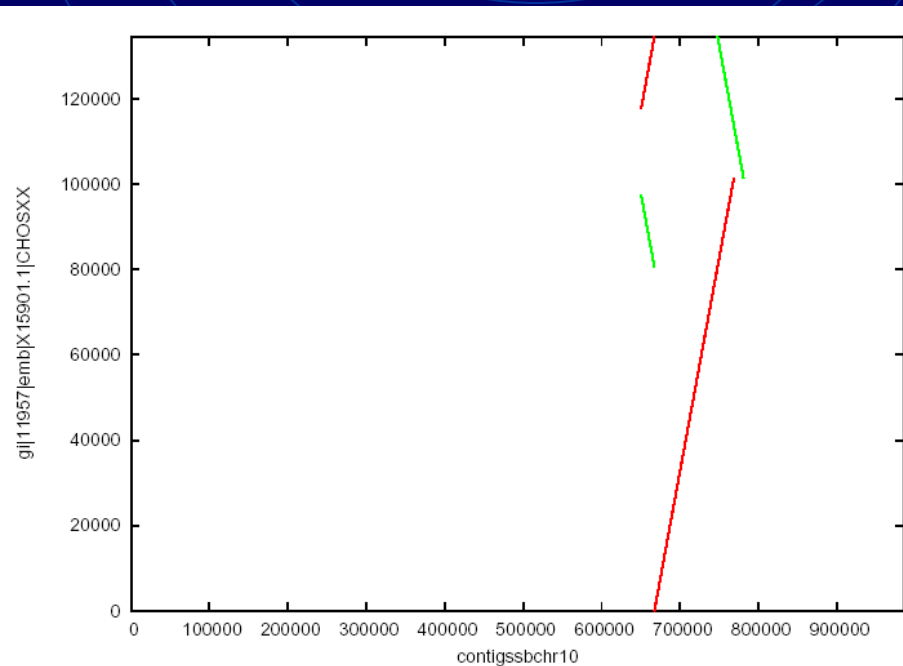


Chloroplast genome, kb





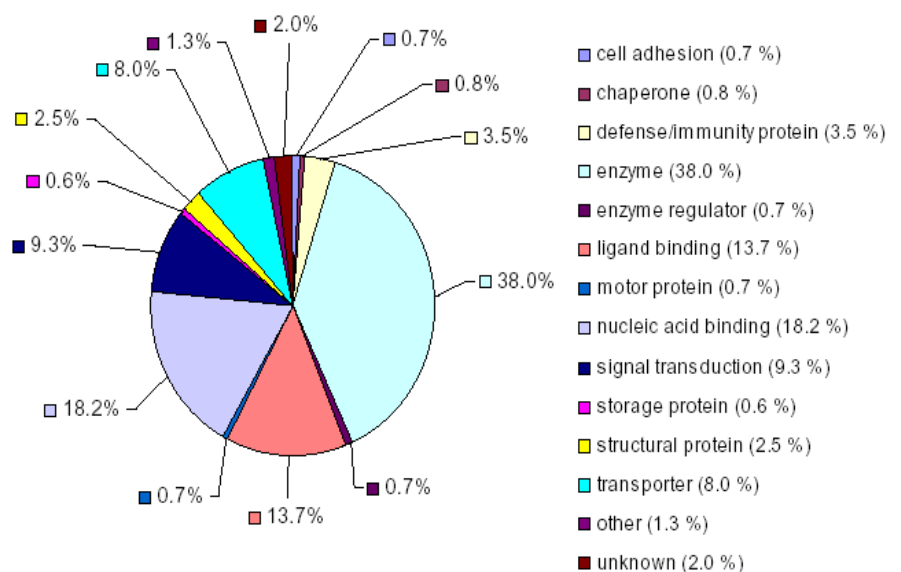
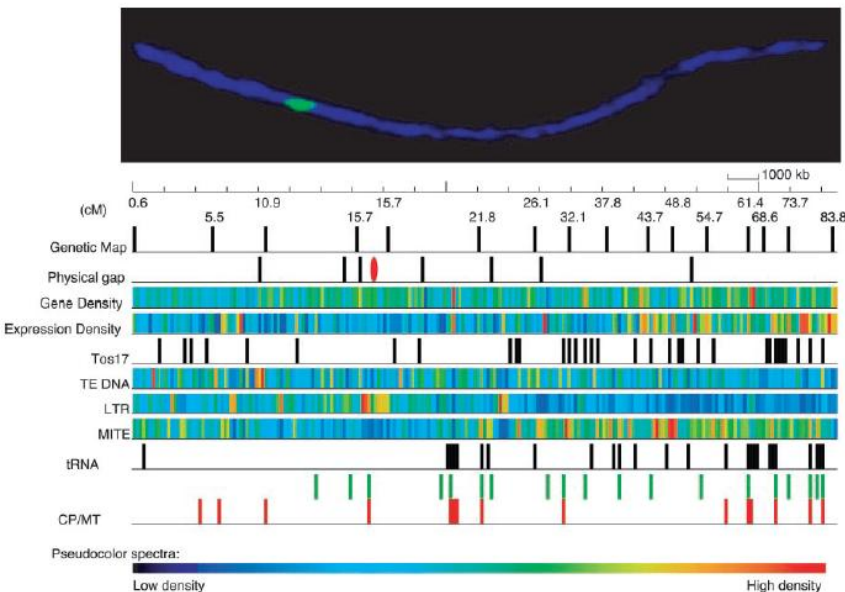
Chloroplast Insertion in Rice and Arabidopsis Genome



Rice: Chromosome 10

(SCIENCE, 2003)

Features	Rice chromosome 10	Arabidopsis whole genome
Total number of BACs/PACs	202	1,578 (5 chromosomes)
Total BAC length (Mb)	28.1	132.9
Total non-overlapping sequence (Mb)	22.4	117.3
Short arm (Mb)	7.6	
Long arm (Mb)	14.8	
G + C Content		
Overall	43.5%	35.9%
Exons	53.6%	43.8%
Introns	39.4%	32.6%
Intergenic regions	41.2%	31.7%
Protein-coding DNA	53.6%	44.0%
Total number of genes	3,471	29,084
Integrated genetic markers	140 (loci)	
Average gene size (bp)	2,556	1,975
Average gene density (bp per gene)	6,464	4,008
Average number exons per gene	4	4.9
Average exon size (bp)	344	275
Average intron size (bp)	389	163
Known genes	1,785 (51.4%)	19,301 (66.4%)
Average known gene size (bp)	3209	2253
Unknown genes	288 (8.3%)	4,725 (16.2%)
Average unknown gene size	2322	1967
Hypothetical genes	1,398 (40.3%)	5,058 (17.4%)
Average hypothetical gene size	1771	1784
tRNA genes	67	611



4.2.2 Genome Evolution

To answer why the genome looks like current that

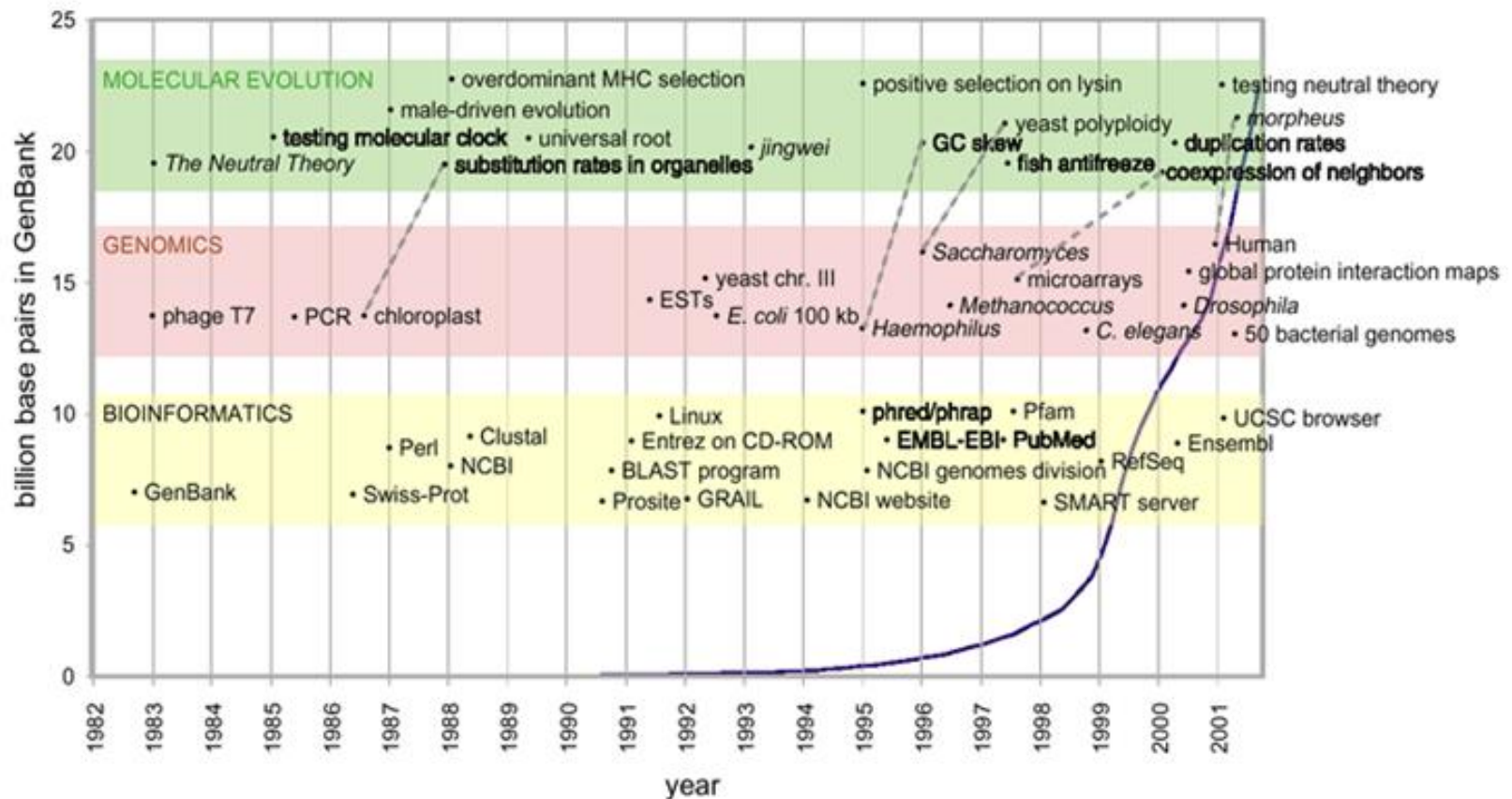


Figure 1: Timeline of developments in bioinformatics, genomics and molecular evolution, charted against the accumulation of DNA sequence information in GenBank, which was established in 1982.

GRASS GENOMES

Oats

Triticeae

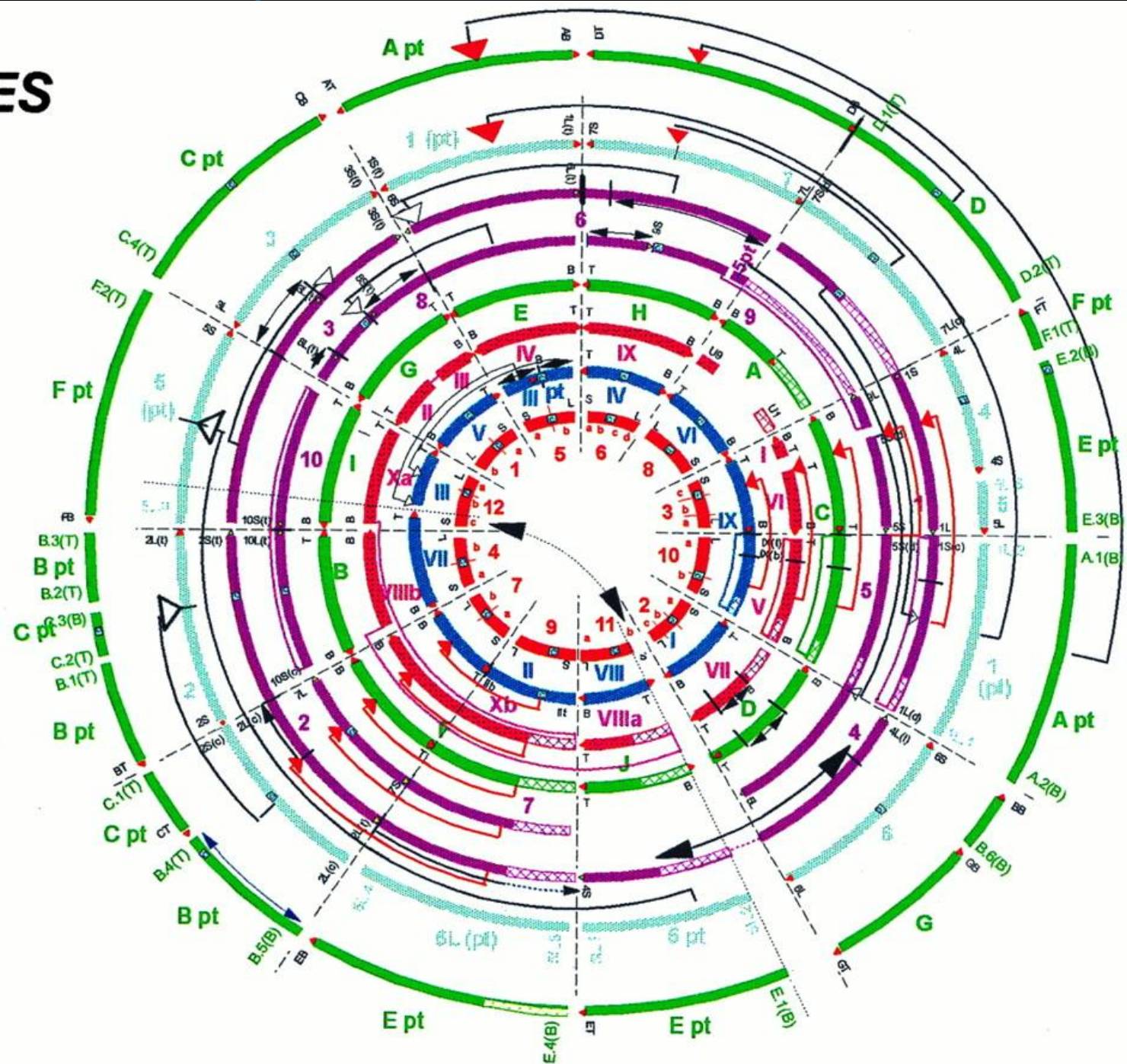
Maize

Sorghum

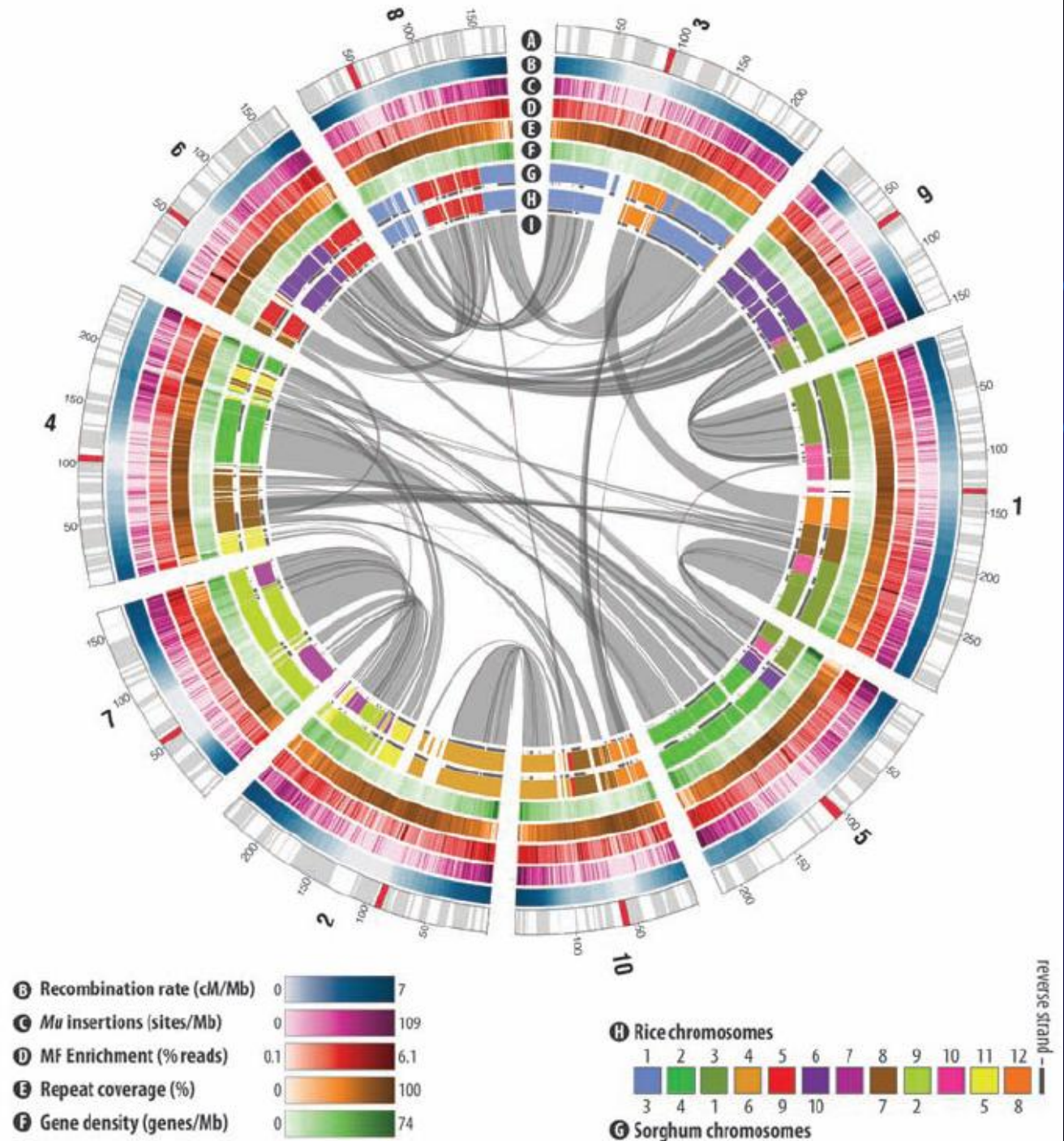
Sugar cane

Foxtail millet

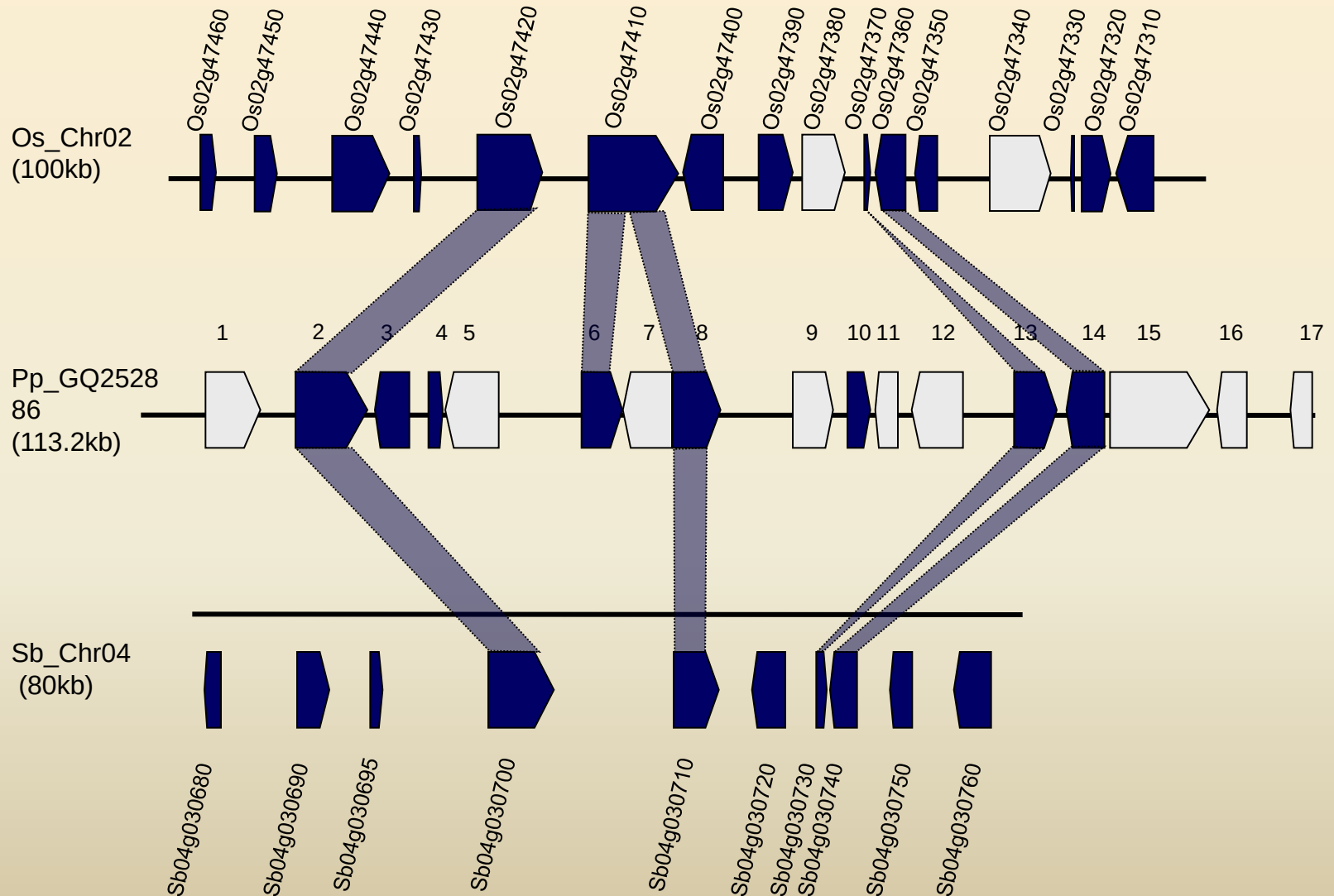
Rice

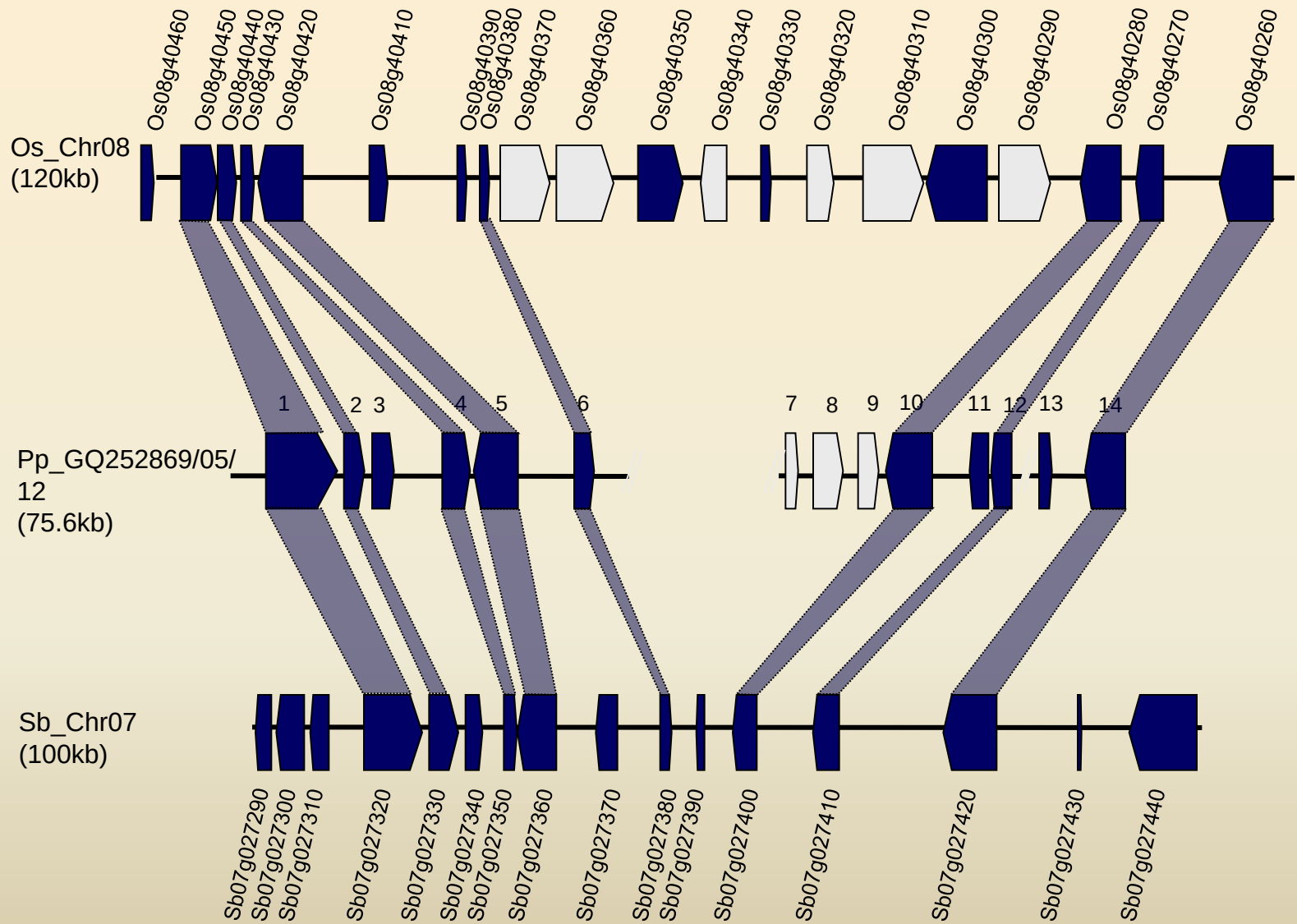


Schnable et al.
2009. The B73
maize genome:
Complexity,
diversity, and
dynamics.
Science.
326:1112



Genomic synteny: rice/sorghum/bamboo



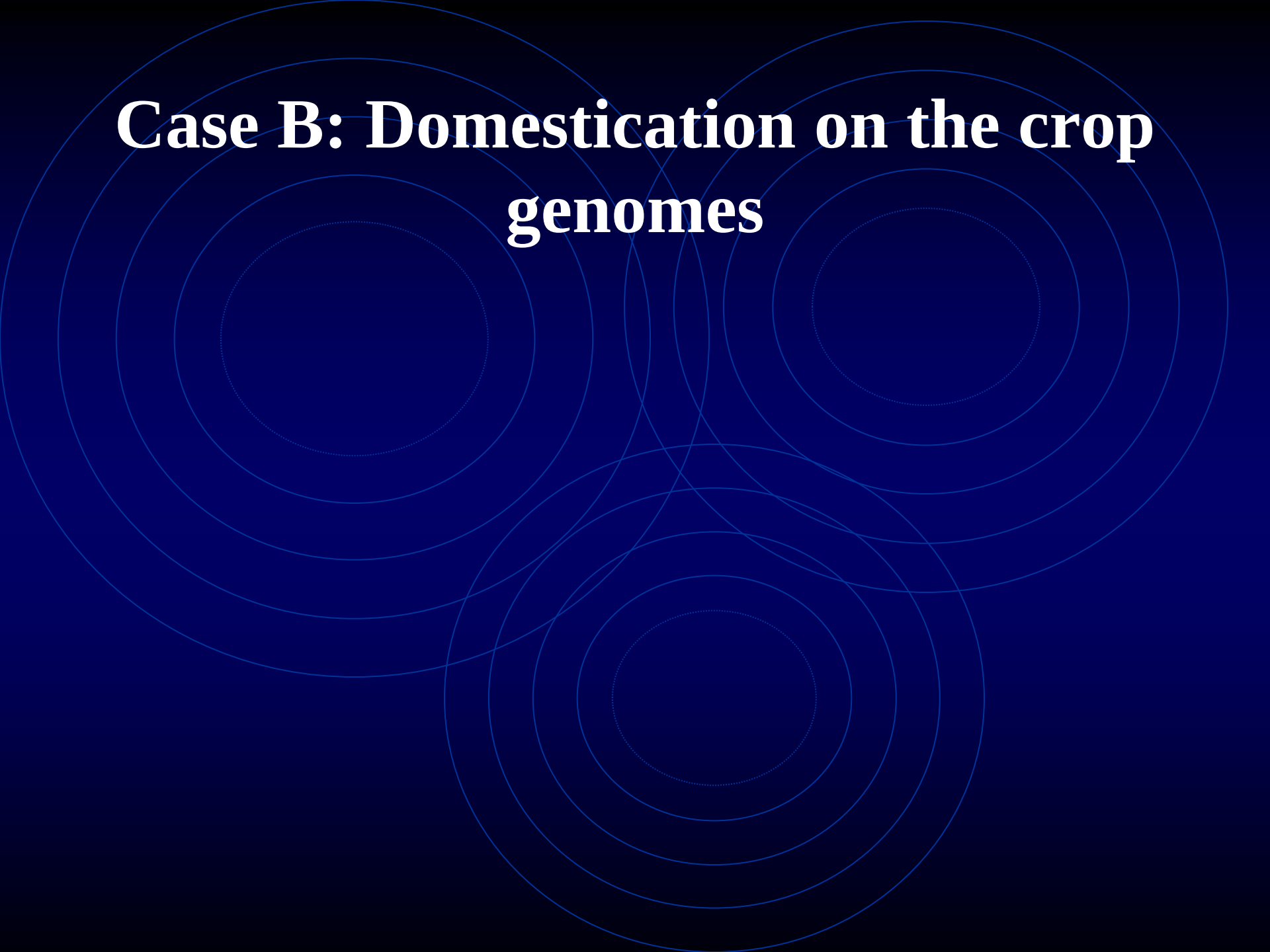


Case A: Rice Genome Duplication

Structural Dynamics of Eukaryotic Chromosome Evolution

Evan E. Eichler and David Sankoff
Science 2003 August 8; 301: 793-797

Case B: Domestication on the crop genomes

The background of the slide is a solid dark blue. Overlaid on this background are three sets of concentric circles, each consisting of four thin, light blue lines. These circles are arranged in a triangular pattern: one set is on the left, one is on the right, and one is centered at the bottom. The circles overlap with each other and with the text area.

4.3 Gene finding

- ***4.3.1 Introduction***
- ***4.3.2 Gene finding strategies***

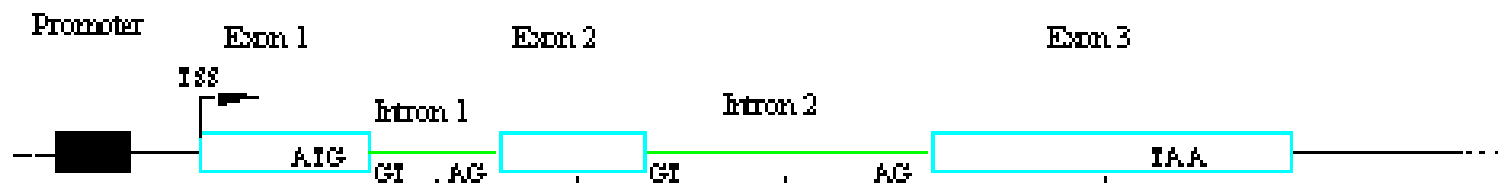
4.3.1 Introduction

- Given the sequence of a genome, we would like to be able to identify:
 - Genes
 - Exon boundaries & splice sites
 - Beginning and end of translation
 - Alternative splicings
 - Regulatory elements (e.g. promoters)
- Only certain way to do this is experimentally, but computational methods can achieve reasonable accuracy quickly, and help direct experimental approaches.



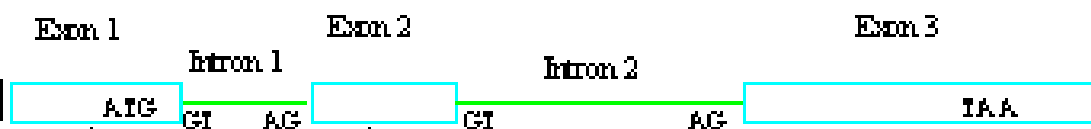
Schematic gene structure

DNA:



transcription

preRNA:



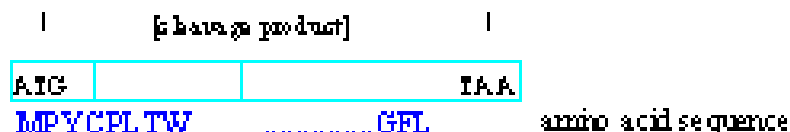
splicing

mRNA:



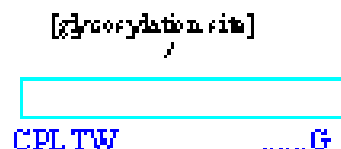
translation

primary translation



modification

active protein:



ORFs

- ORF = Open Reading Frame. Region of sequence that has the potential to be translated into a protein.
 - Six possible “reading frames” (3 starts x 2 strands)
 - Starts with a atg (methionine)
 - Ends with a stop codon (taa, tag or tga).
 - No intervening stop codons
- Not all ORFs are exons, but all exons must be in an ORF.

How to identify ORFs that are exons?

- Observed length distribution is non-random
 - Long ORFs more likely to be exons
 - Proposed 12 part mixture model of length distributions
- Signatures of exons
 - CpG islands
 - intron splice junctions
 - (hexa)nucleotide frequencies
- Signatures of non-exons (repeats, ALUs, etc.)
- Compatibility with surrounding sequence & other exons (e.g. consistent reading frame)

General challenges: gene finding is difficult

- DNA sequence signals have low information content (degenerated and highly unspecific) and It is difficult to discriminate real signals
- Short exons are hard to find, but not uncommon. Shortest human exon is 3bp!
- First and last exons are particularly difficult
 - No bounding intron, so no splice site signal (although start and end codons are related signals)
 - Generally contain non-coding sequence as well as coding sequence, so hexamer signals are diluted.
- Alternative splicing means that there are multiple true solutions for some genes.
- Sequencing errors

Prokaryotes

- High gene density and simple gene structure
- Short genes have little information
- Overlapping genes

Eukaryotes

- Low gene density and complex gene structure
- Alternative splicing
- Pseudo-genes

4.3.2 Gene finding strategies

There is no (yet known) perfect method for finding genes. All approaches rely on combining various “weak signals” together and assemble into a consistent gene model

Homology method

- Gene structure can be deduced by homology
- Requires a not too distant homologous sequence

Ab initio method

- Requires two types of information
 - . compositional information
 - . signal information

LOCUS AP003280 147201 bp DNA linear PLN 27-NOV-2003

DEFINITION *Oryza sativa* (japonica cultivar-group) genomic DNA,
chromosome 1, PAC clone:P0557A01.

COMMENT Genes were predicted from the integrated results of the following: GENSCAN1.0, BLASTN2.0, BLASTX2.0 as well as SplicePredictor (October 1998 version). The genomic sequence was searched against NCBI NonRedundant Protein database, nr (<ftp://ncbi.nlm.nih.gov/blast/db>) and the cDNA sequence database at RGP. Protein homologies of the coding regions were searched against NCBI NonRedundant Protein database with BLASTP2.0. ESTs represent the identified cDNA sequences using BLASTN 2.0 with the corresponding DDBJ accession no. and RGP clone ID.

A gene with identity or significant homology to a protein is classified based on the protein name to indicate the homology level such as same name, 'putative-' and '-like protein'. A gene without significant homology to any protein but with EST homology (covering almost the entire length of partial sequence) is classified as an 'unknown' protein. A gene predicted with a gene prediction program is classified as a 'hypothetical' protein.

LOCUS OSJN00244 151936 bp DNA linear PLN 14-NOV-2003
DEFINITION Oryza sativa genomic DNA, chromosome 4, BAC clone:
OSJNBa0053B21, complete sequence.

COMMENT

----- Summary Statistics -----

Assembly program: phrap

Genes were identified by a combination of several methods:

Gene prediction programs including Fgenesh

(<http://www.softberry.com/>), genscan (<http://CCR-081.mit.edu/GENSCAN.html>), GeneMarkHMM

(<http://genemark.biology.gatech.edu/GeneMark/>), tRNAscan-

SE (Sean Eddy, <http://genome.wustl.edu/eddy/tRNAscan-SE/>),

searches of the complete sequence against NCBI non-redundant protein database (nr) (<ftp://ncbi.nlm.nih.gov/blast/db>) and the EST database at NCGR.

LOCUS NT_030059 43027086 bp DNA linear CON 17-OCT-2003
DEFINITION Homo sapiens chromosome 10 genomic contig

COMMENT GENOME ANNOTATION REFSEQ:

Features on this sequence have been produced for build 34 of the NCBI's genome annotation [see [documentation](#)].

On Oct 7, 2003 this sequence version replaced gi:[29800185](#). The DNA sequence is part of the second release of the finished human reference genome. It was assembled from individual clone sequences by the Human Genome Sequencing Consortium in consultation with NCBI staff.

COMPLETENESS: not full length.

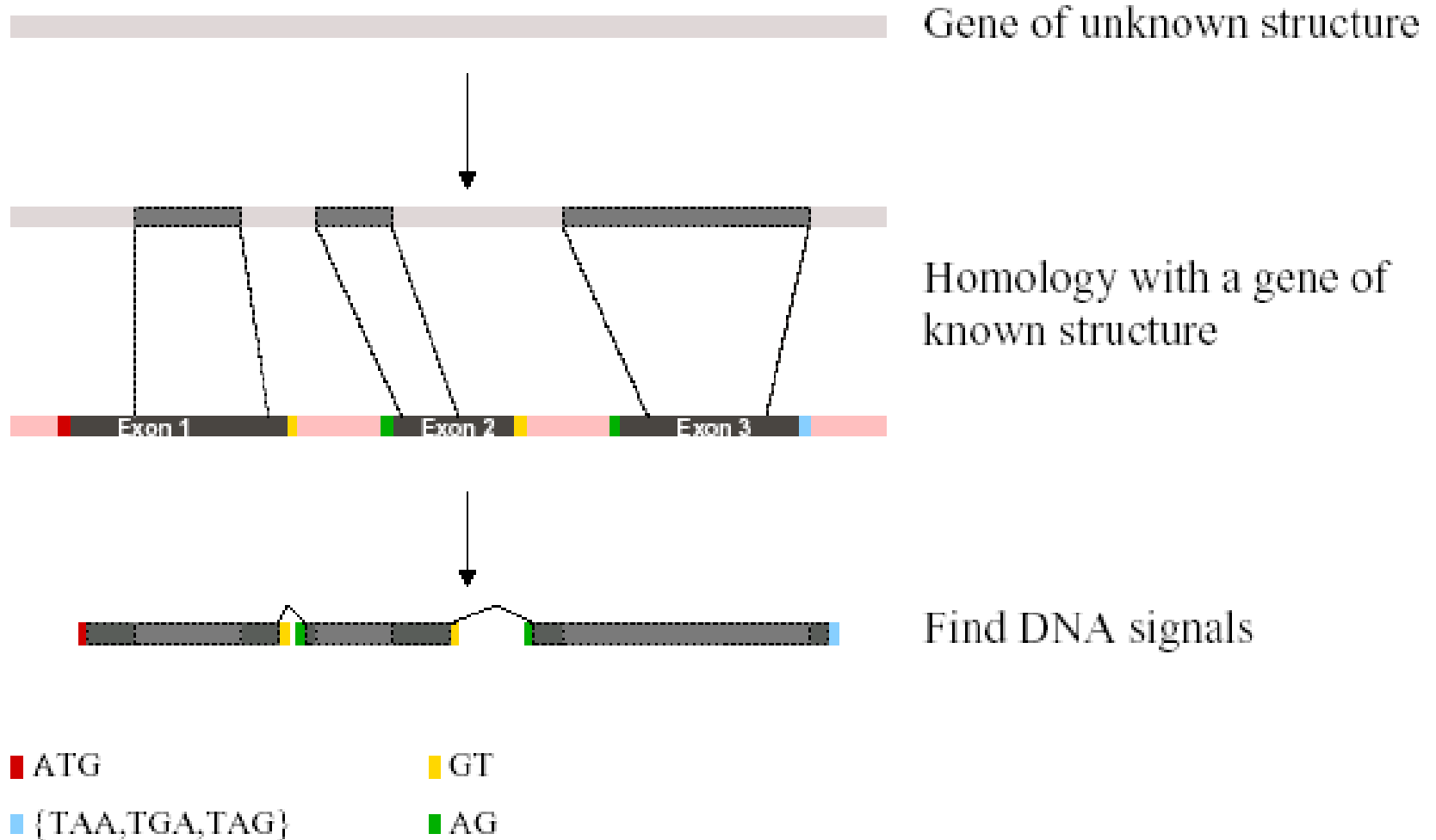
Homology method

Principles of the homology method:

- Coding regions evolve slower than non-coding regions, i.e. local sequence similarity can be used as a gene finder.
- Homologous sequences reflect a common evolutionary origin and possibly a common gene structure, i.e. gene structure can be solved by homology (mRNAs, ESTs, proteins, domains).
- Standard homology search methods can be used (BLAST, Smith-Waterman, ...).
- Include "gene syntax" information (start/stop codons, ...).

Homology methods are also useful to confirm predictions inferred by other methods

A simple view



Inference by homology

- For exon finding, we need to find matches to
 - mRNA/cDNA sequences
 - ESTs
 - Known exons

ESTs can be helpful in confirming a gene model

Genome sequence

Predicted
exons



ESTs should match exons

may need to fill in gaps by RT PCR and often need to obtain the whole cDNA sequence

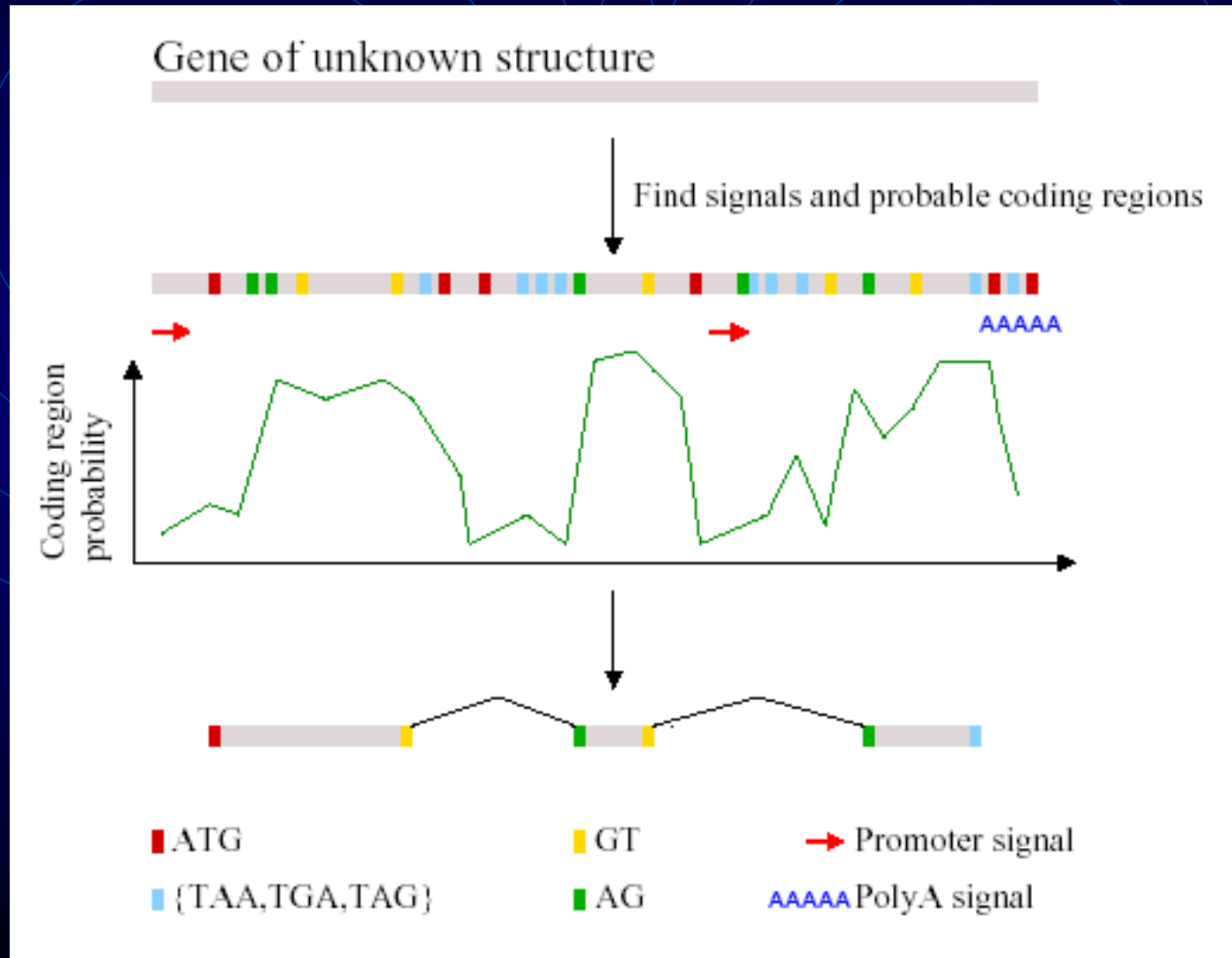
same mRNA may be spliced differently in different tissues giving a different protein or mRNA may be edited to change sequence

Ab initio method

Principles of the *ab initio* methods

- Integration of **signal detection** and **coding statistics**
- **Signal detection** and **coding statistics** are deduced from a training set
- Probabilistic frameworks are used to infer a probable gene structure
- A solid scoring system can be used to evaluate the predictions

A simple review



Sequence model

Sequence models: probabilistic model

- iid: niid and eiid
- Markov chain
- weight matrix
- regular expression (consensus sequence/ pattern)

Genetic model:

$$\diamond Y = \mu + G + e \quad (\text{Johannsen, 1909})$$

$$\diamond Y = \mu + (A + D + I) + e \quad (\text{Fisher, 1918})$$

$$\diamond Y = \mu + (A + D + (AA + AD + DD)) + e \quad (\text{Cockerham, 1954})$$

Probabilistic framework

We can express all of these weak signals as probabilities

- $P(\text{exon} \mid \text{length})$
- $P(\text{exon} \mid \text{hexamer composition})$
- $P(\text{exon} \mid \text{CpG content})$
- $P(\text{exon} \mid \text{adjacent “splice” signals})$
- etc. etc.

HMM

Markov Model (MM)

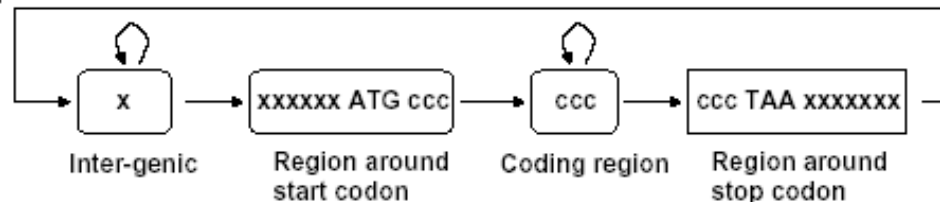
- Biological sequences can be modeled as the output of a stochastic process in which the probability for a given nucleotide to occur at position p depends on the k previous positions. This representation is called k -order Markov Model.

$$P(x_i | x_1, x_2, \dots, x_{i-1}) = P(x_i | x_{i-k}, x_{i-(k-1)}, \dots, x_{i-1})$$



Hidden Markov Model (HMM)

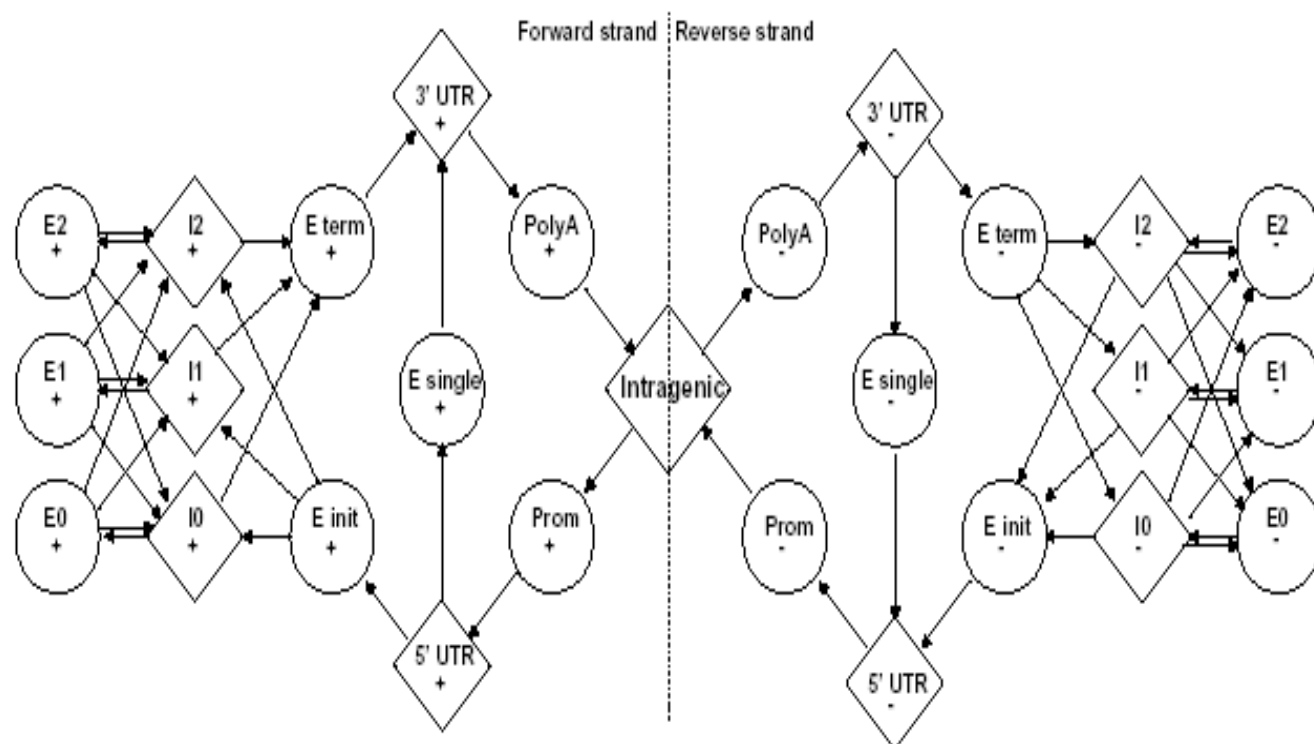
- In a HMM the biological sequences are modeled as the output of a stochastic process that progresses through a series of discrete states. Each state model correspond to a Markov Model.



(Krog, 1998)

Example of the GHMM in GENSCAN (Burge and Karlin, 1997)

- Each state can correspond to a position weight matrix, a neural network, ...
- Each state has an associated length distribution (time) → there are no transitions from the state to the state itself.
- The underlying (hidden) model of GENSCAN:



GENSCAN

- A Semi-Markov Model
 - Explicit model of how long to stay in a state (rather than just self-loops, which must be exponentially decaying)
- Tracks “phase” of exon or intron (0 coincides with codon boundary, or 1 or 2)
- Tracks strand (and direction)

GenScan parameters

- Initial probabilities for being in each state
- All transition probabilities
- A set of length distributions for all states
- A set of “sequence generating” models for each state.
- Fitting Semi-Markov processes is much more computationally complex
 - use explicit length distributions only when necessary
 - others are made exponentially decreasing like HMMs.

Observations on GenScan

- Overall probability of a sequence (parse) depends not only on the probability of a particular exon, but also how well it fits with adjacent states.
- Exon *score* is probability of exon normalized by the probability of alternative parses (e.g. an intron).
- Model was forced to be “strand symmetric”
 - Corresponding transition probabilities set equal
 - Initiation of new gene on particular strand set to 50%

GENSCAN output

●

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.00	Prom	+	1653	1692	40							-1.16
1.01	Init	+	5215	5266	52	0	1	83	75	151	0.925	12.64
1.02	Intr	+	5395	5562	168	2	0	89	75	163	0.895	15.02
1.03	Intr	+	11738	11899	162	0	0	74	113	101	0.990	11.15
1.04	Intr	+	12188	12424	237	0	0	71	86	197	0.662	15.39
1.05	Intr	+	14288	14623	336	0	0	82	98	263	0.986	22.19
1.06	Intr	+	17003	17203	201	0	0	116	86	102	0.976	12.06
1.07	Intr	+	17741	17859	119	0	2	78	109	51	0.984	6.38
1.08	Intr	+	18197	18264	68	1	2	103	72	81	0.541	5.70

>02:36:44|GENSCAN_predicted_peptide_1|448_aa

MCRAISLRRLLLLLLQLSQLLAVTQGKTLVLGKEGESAE L PCESSQKKITVFTWKFS DQR
KILGQHKGVLIRGGSPSQFDRFDSKKGAWEKGSFPLIINKLKMEDSQTYICELENRKEE

...

Gn.Ex : gene number, exon number (for reference)

Type : Init = Initial exon (ATG to 5' splice site)

Intr = Internal exon (3' splice site to 5' splice site)

Term = Terminal exon (3' splice site to stop codon)

Sngl = Single-exon gene (ATG to stop)

Prom = Promoter (TATA box / initiation site)

PlyA = poly-A signal (consensus: AATAAA)

S : DNA strand (+ = input strand; - = opposite strand)

Begin : beginning of exon or signal (numbered on input strand)

End : end point of exon or signal (numbered on input strand)

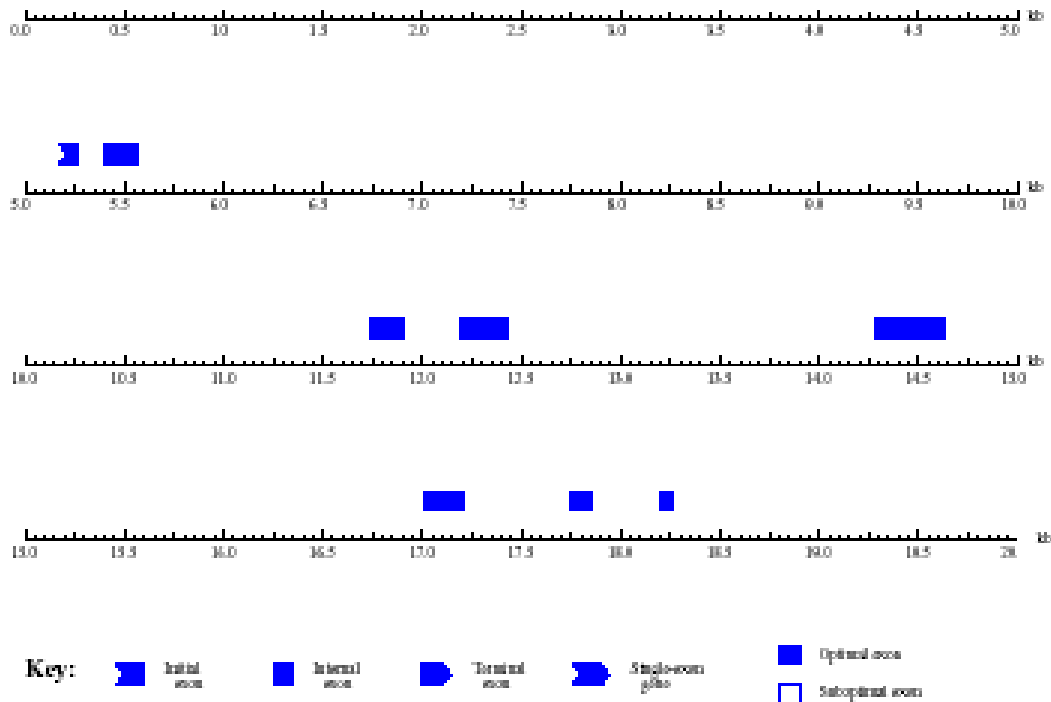
Len : length of exon or signal (bp)

Fr : reading frame (a forward strand codon ending at x has frame $x \bmod 3$)

Ph : net phase of exon (exon length modulo 3)
 I/Ac : initiation signal or 3' splice site score (tenth bit units)
 Do/T : 5' splice site or termination signal score (tenth bit units)
 CodRg : coding region score (tenth bit units)
 P : probability of exon (sum over all parses containing exon)
 Tscr : exon score (depends on length, I/Ac, Do/T and CodRg scores)

Graphical output

GENSCAN predicted genes in sequence 02:36:44



Fgenesh

(www.softberry.com)

- Fgenes (Find genes) is the multiple gene prediction program based on dynamic programming;
- Fgenesh: Hidden Markov Model (HMM)-based gene prediction program (Salamov and Solovyev 2000, Genome Res)
- Fgenesh+: is a version of Fgenesh, which uses additional information from the available protein homolog. When exons predicted by Fgenesh show high similarity to a protein from the database, it is often advantageous to use this information to improve the prediction accuracy.

GENESH-2.1 Prediction of potential genes in genomic DNA
 Time: Thu Jan 28 19:25:51 1999.
 Seq name: HUMHBB 73308 bp DNA PRI 20-JAN-
 length of sequence 73308bp G+C content: 39 Isochore: 1
 number of predicted genes 7 in +chain 7 in -chain 0
 number of predicted exons 18 in +chain 18 in -chain 0

Gn	S	Type	Start	End	Score	ORF	Len
1	+	TSS	19447		-7.15		
1	+	CDSf	19541 -	19632	16.12	19541 - 19630	90
1	+	CDSi	19755 -	19977	14.12	19756 - 19977	222
1	+	CDSl	20833 -	20961	2.99	20833 - 20961	129
1	+	PolA	21055		1.05		
2	+	TSS	34437		-7.15		
2	+	CDSf	34531 -	34622	15.25	34531 - 34620	90
2	+	CDSi	34745 -	34967	20.74	34746 - 34967	222
2	+	CDSl	35854 -	35982	5.59	35854 - 35982	129
2	+	PolA	36043		1.05		
3	+	TSS	39373		-7.15		
3	+	CDSf	39467 -	39558	15.25	39467 - 39556	90
3	+	CDSi	39681 -	39903	20.74	39682 - 39903	222
3	+	CDSl	40770 -	40898	5.74	40770 - 40898	129
3	+	PolA	40959		1.05		
4	+	TSS	44415		-8.75		
4	+	CDSf	45995 -	46151	16.01	45995 - 46150	156
4	+	CDSl	46997 -	47100	2.71	46999 - 47100	102
4	+	PolA	47243		1.05		
5	+	TSS	54703		-4.45		
5	+	CDSf	54790 -	54881	13.41	54790 - 54879	90
5	+	CDSi	55010 -	55232	14.20	55011 - 55232	222
5	+	CDSl	56131 -	56259	3.87	56131 - 56259	129
5	+	PolA	56365		1.05		
6	+	TSS	62100		-6.65		
6	+	CDSf	62187 -	62278	13.59	62187 - 62276	90
6	+	CDSi	62409 -	62631	19.50	62410 - 62631	222
6	+	CDSl	63482 -	63610	10.23	63482 - 63610	129
6	+	PolA	63718		1.05		
7	+	TSS	68088		-9.45		
7	+	CDSf	68183 -	68428	14.87	68183 - 68428	246
7	+	PolA	68509		1.05		

Predicted protein(s):

```

>ID 1 3 exon(s) 19541 - 20961 147 aa, chain +
MVHFTAEEKAAVTSLSWKMNVEEAGGEALGRLLVVYPWTQRFDFSFGNLSSPSAILGNPK
VKAHGKKVLTSLFGDAIKNMDNLKPAFAKLSELHCDKLHVPENFKLLGNVMVILATHFG
KEFTPEVQAAWQKLVSAVAIALAHKYH
>ID 2 3 exon(s) 34531 - 35982 147 aa, chain +
MGHFTTEEDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRFDFSFGNLSSASAIMGNPK
VKAHGKKVLTSLGDAIKHLDDLKGTFAQLSELHCDKLHVPENFKLLGNVIVTVLAIHFG
KEFTPEVQASWQKMTGVASALSSRYH
>ID 3 3 exon(s) 39467 - 40898 147 aa, chain +
MGHFTTEEDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRFDFSFGNLSSASAIMGNPK
VKAHGKKVLTSLGDAIKHLDDLKGTFAQLSELHCDKLHVPENFKLLGNVIVTVLAIHFG
KEFTPEVQASWQKMTAVASALSSRYH
  
```

Fgenesh output:

G - predicted gene number, starting from start of sequence;

Str - DNA strand (+ for direct or - for complementary);

Feature - type of coding sequence: CDSf - First

(Starting with Start codon), CDSi - internal (internal exon), CDSl - last coding segment, ending with stop codon);

TSS - Position of transcription start (TATA-box position and score);

Start and End - Position of the Feature;

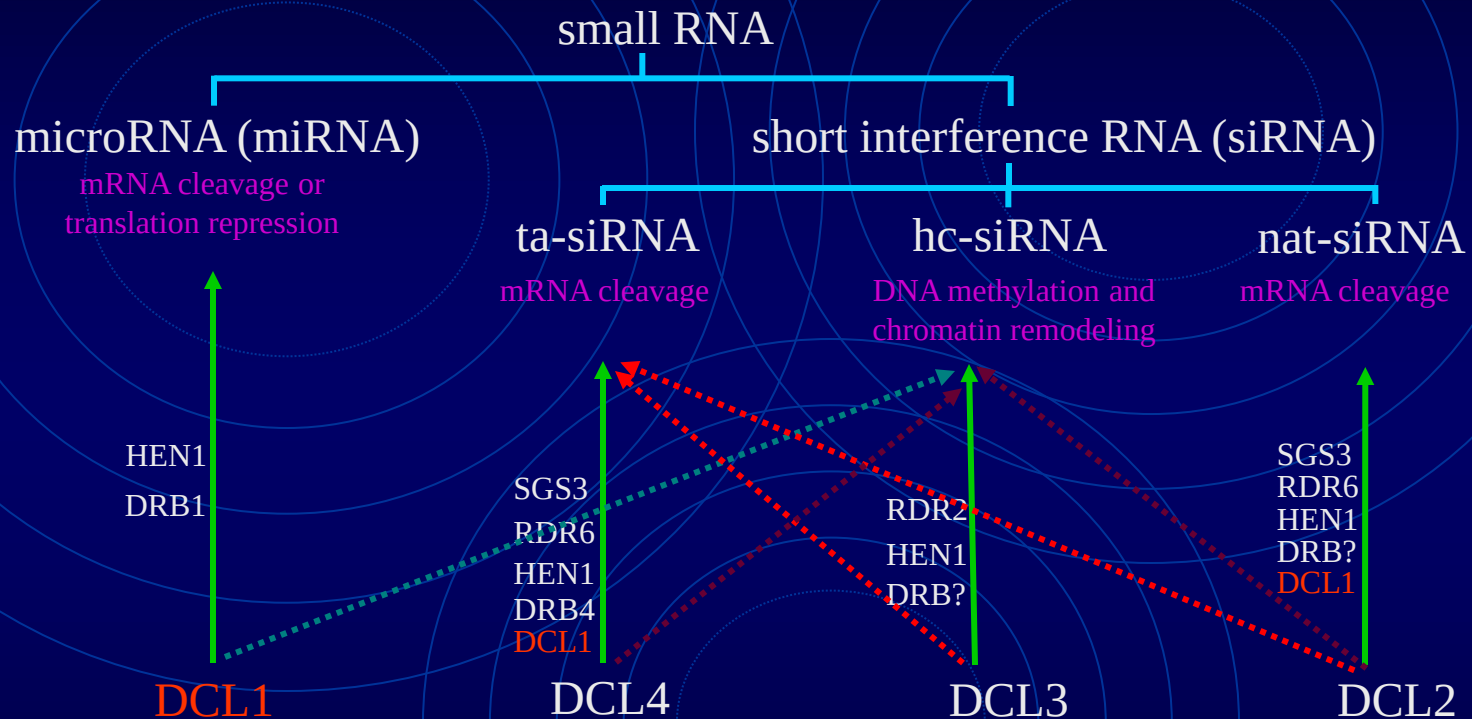
Weight - Log likelihood*10 score for the feature;

ORF - start/end positions where the first complete codon starts and the last codon ends.

4.4 Non-coding gene finding

- microRNA(miRNA) finding
- ta-siRNA finding

Endogenous Small RNAs and Their Biogenesis in Plants

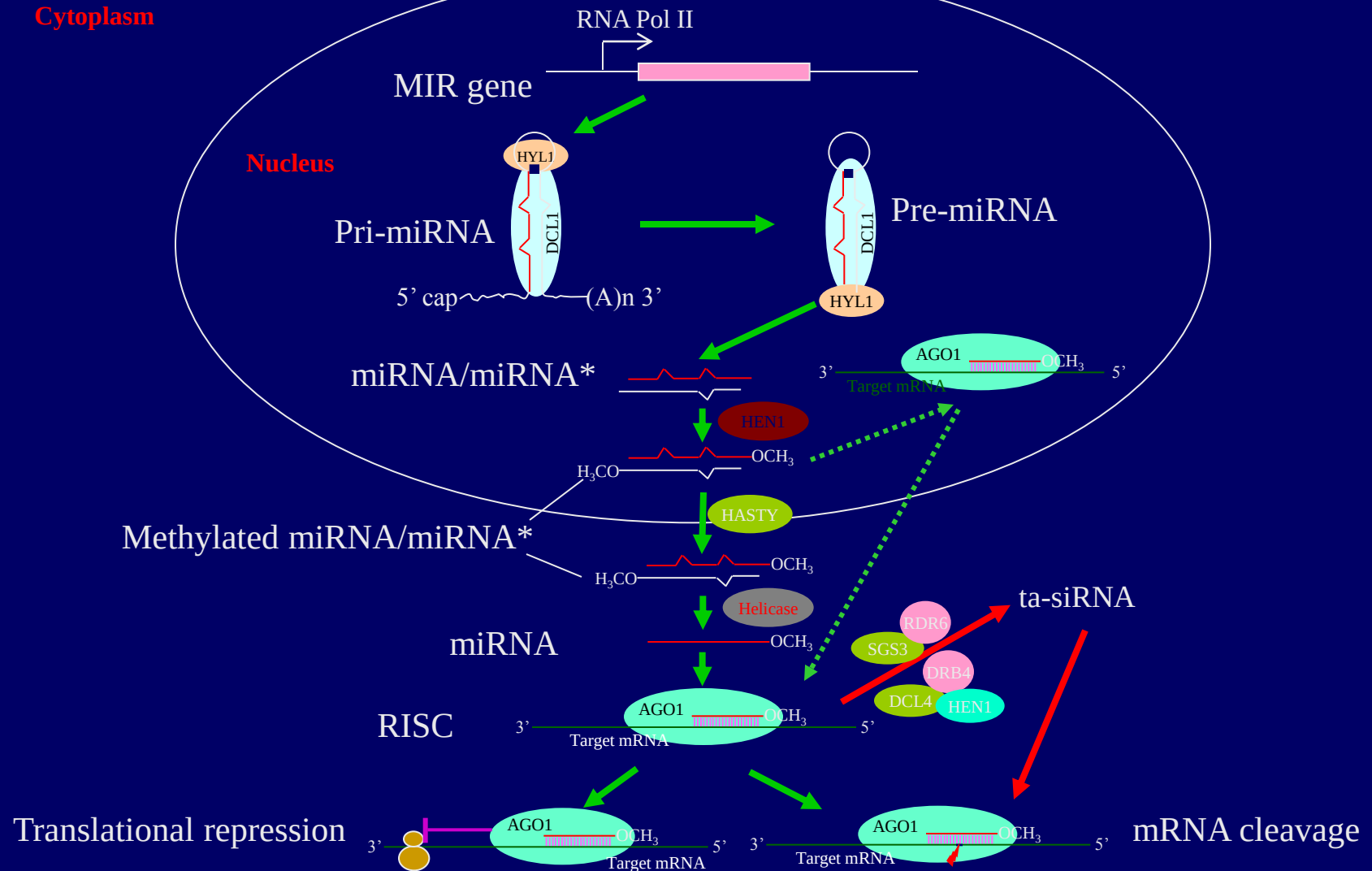


ta-siRNA: trans-siRNA; *na-siRNA*: natural *cis*-antisense siRNA; *hc-siRNA*: heterochromatin siRNA; *DRB*: dsRNA-BINDING PROTEIN; *HEN1*: HUA ENHANCER 1; *RDR*: RNA-DEPENDENT RNA POLYMERASE; *SGS3*: SUPPRESSOR OF GENE SILENCING 3

Comparison of miRNAs and siRNAs

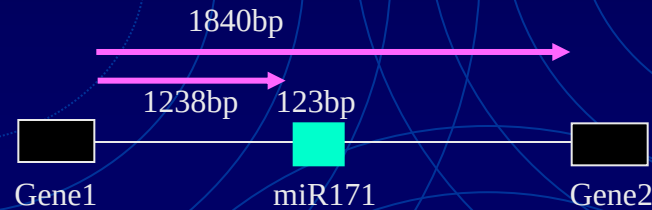
	miRNAs	siRNAs
Dicer	DCL1	DCL1-4
Size	21	21-24
5' end	Mostly U	Mostly A for 24-mers
Origin	Mostly intergenic region	Transposable-element, repeat sequence, heterochromatic sequence
Precursor	Hairpin pri-miRNA	Long dsRNA
Copy number	One pre-miRNA, one miRNA (one arm of the foldback structure)	One precursor, many siRNA (both strands of the precursor)
Conservation	Yes	No
Target gene	<i>trans</i>	<i>cis</i> and <i>trans</i>
Function	developmental regulation and stress response	Chromatin modification and stress response
Mechanism	mRNA cleavage or translational repression	DNA methylation and mRNA cleavage

miRNA Biogenesis and Pathways

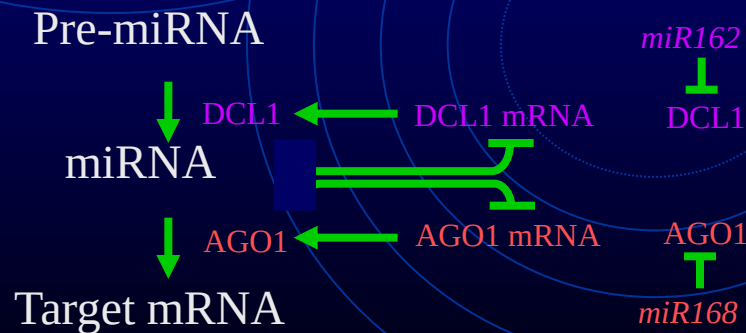


Regulation of miRNA Biogenesis

- Transcriptional regulation: miRNA genes have independent transcription units and are transcribed from their own promoters



- Post-transcriptional regulation
- Feedback control of DCL1 and AGO1 mRNAs



Diverse Functions of miRNAs

- ta-siRNA biogenesis (miR173 – *TAS1*, *TAS2*; miR390 – *TAS3*)
- miRNA biogenesis (miR162 – *DCL1*; miR168 – *AGO1*)
- Disease resistance (miR393 – targets mRNAs for the F-box auxin receptors *TIR1*, *AFB2* and *AFB3*, suppresses auxin signaling and increases bacterial resistance)
- Stress response (miR395 – induced over 100-fold in sulfate-starved plants; miR399 – induced by low-phosphate)
- Plant development (via regulation of transcription factors, hormone response and signal transduction)

TIR1: Transport Inhibitor Response 1; *AFB2* and *AFB3*: Auxin Signaling F-Box Proteins 2 and 3

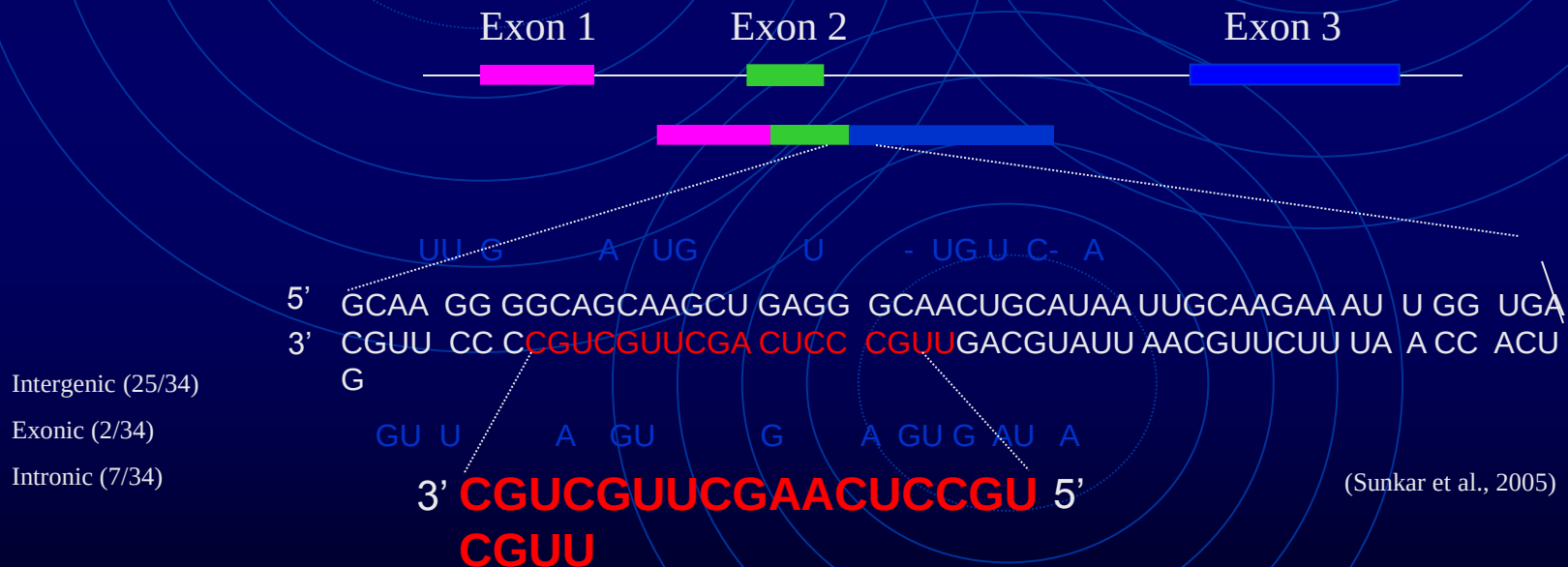
Developmental Regulation of miRNAs in Plants

miRNA	Target genes	Function of the target genes
miR156	<i>SPL2, SPL10</i>	Floral meristem identity and flowering time
miR159	<i>MYB33, MYB65</i>	Plant growth, anther development and flowering time
miR160	<i>ARF17</i>	Regulation cotyledon, leaf, root, and flower development via auxin homeostasis
miR164	<i>CUC1, CUC2, NAC1</i>	Control of boundary size in meristems, formation and separation of embryonic, vegetative and floral organs, control of petal number and lateral root development
miR165 and 166	<i>PHB, PHV, REV, ATHB8, ATHB15</i>	Embryo patterning, postembryonic meristem initiation, axial meristem initiation, vascular development, leaf polarity and meristem size regulation
miR171	<i>SCL6, SCL6-II, SCL6-III</i>	Radial patterning in roots
miR172	<i>AP2, TOE1, TOE2, TOE3</i>	Specification of flower organ identity and flowering time
miR319/jaw	<i>TCP2, TCP3, TCP4, TCP10, TCP24</i>	Cell division, leaf development and embryo patterning

AP2: APETALA2; ARF: AUXIN RESPONSE FACTOR; ATHB: ARABIDOPSIS THALIANA HOMEODOMAIN GENE; CUC: CUP-SHAPED COTYLEDONS; NAC: NAM/ATAF/CUC; PHB: PHABULOSA; PHV: PHAVOLUTA; REV: REVOLUTA; SCL: SCARECROW-LIKE; SPL: AQUAMOSA PROMOTER BINDING LIKE; TCP: TOXIN CO-REGULATED PILUS; TOE: TARGET OF EAT

Discovery of miRNAs in Plants

- Forward genetics: inefficiency due to the small size of miRNAs and functional redundancy among different miRNAs of the same family
- Computational prediction: conservation; secondary structure, thermodynamic stability of the hairpin and free energy of folding



- Cloning: preferred, sequencing size-fractionated cDNA libraries

Computational prediction of miRNAs and their targets

- Conserved miRNA:
 - Conservation/small RNA reads;
 - Secondary structure
 - Thermodynamic stability of the hairpin and free energy of folding
- Novel miRNAs
 - Small RNA reads
 - Secondary structure
 - Thermodynamic stability of the hairpin and free energy of folding
- Target prediction

ta-siRNA finding

- Conserved ta-siRNA
 - TAS3: ta-siARF
 - miR390 binding sites
- Novel ta-siRNA
 - Small RNA reads
 - Phase of small RNA reads
 - miRNA targeting sites

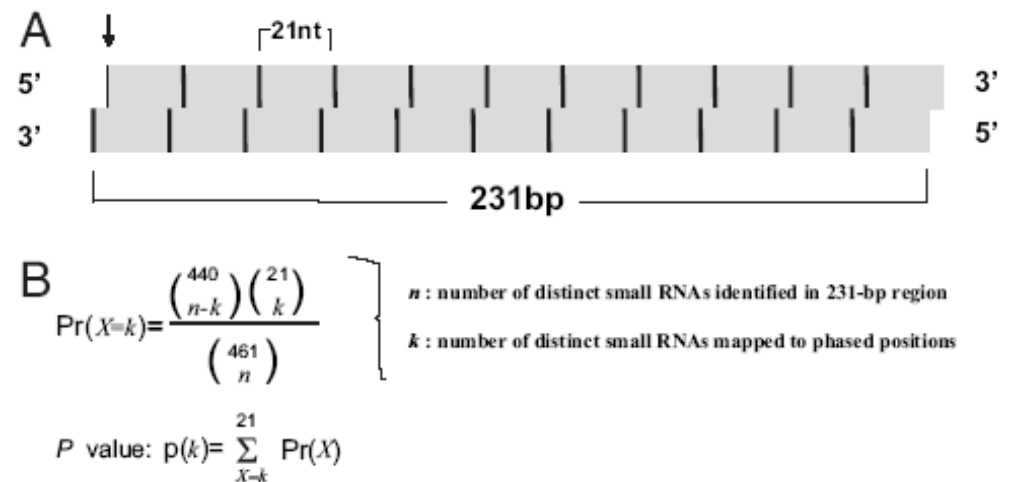
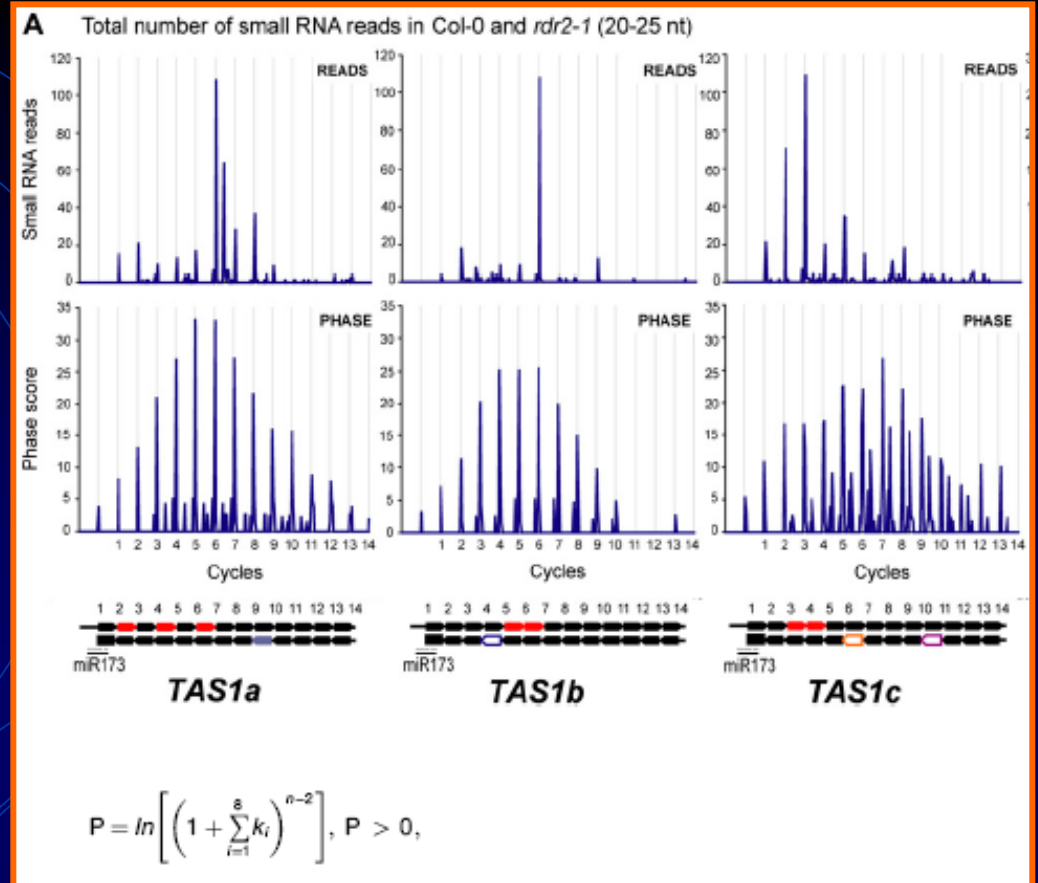
Methods

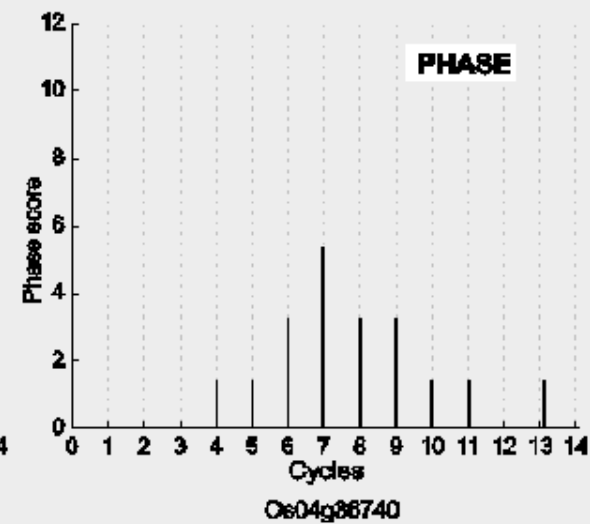
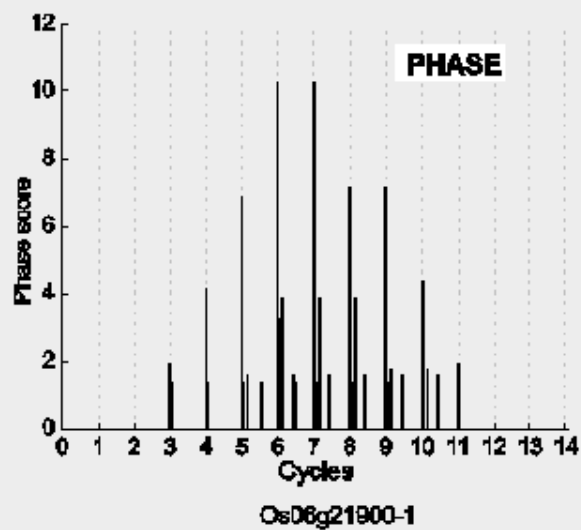
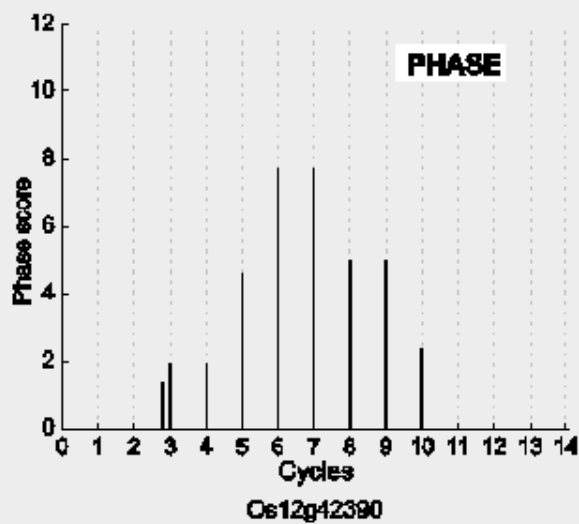
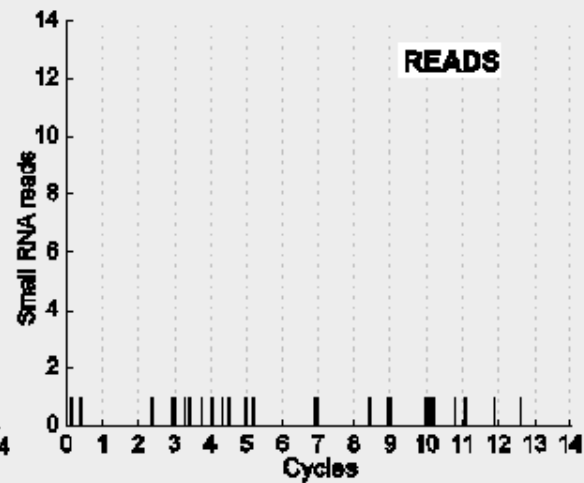
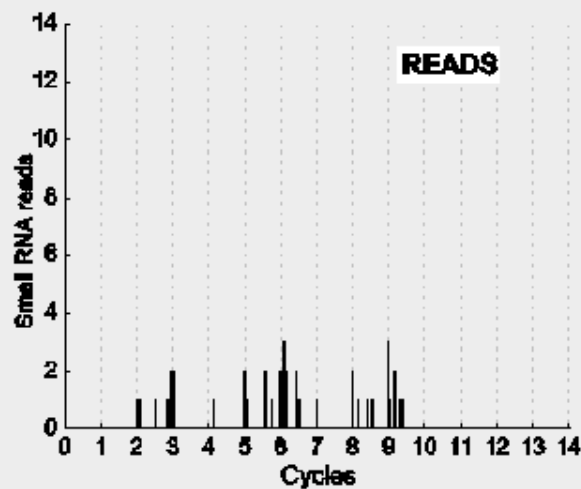
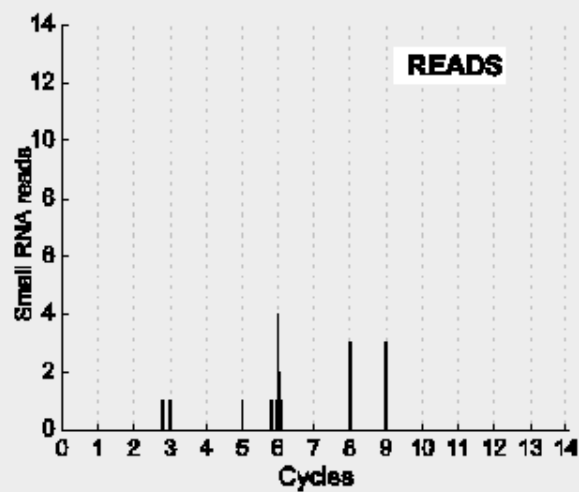
- Howell et al. 2007. Plant Cell

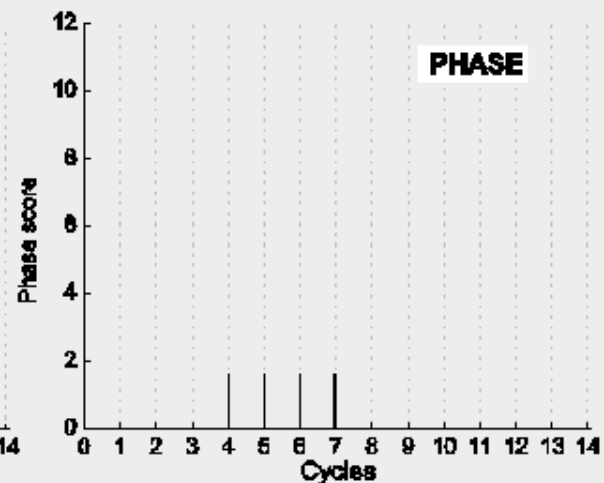
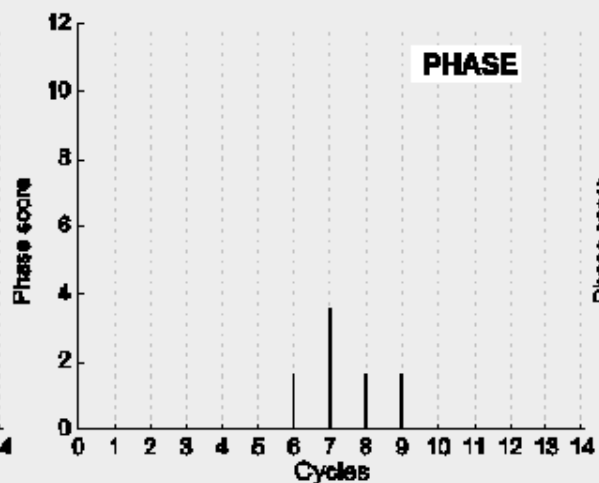
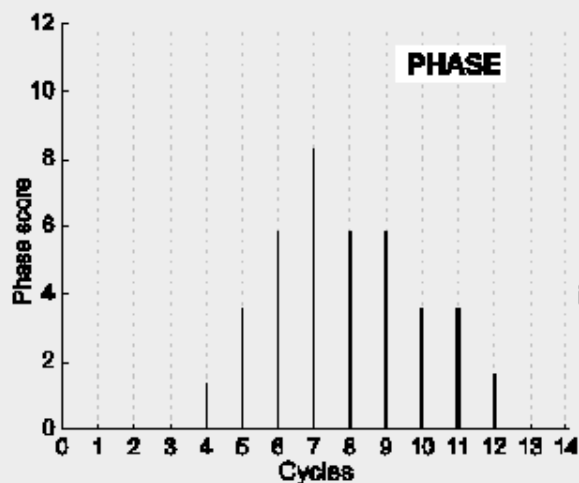
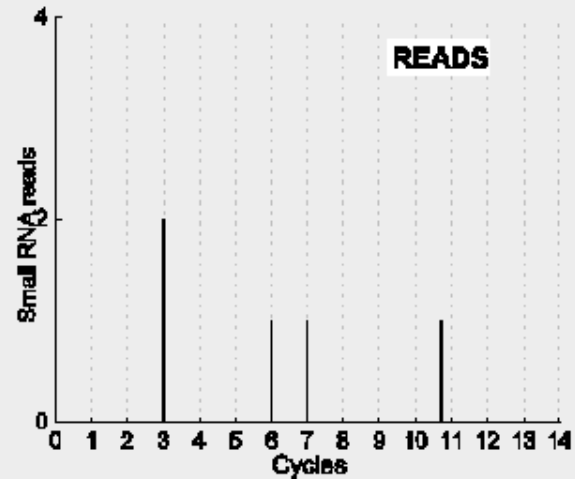
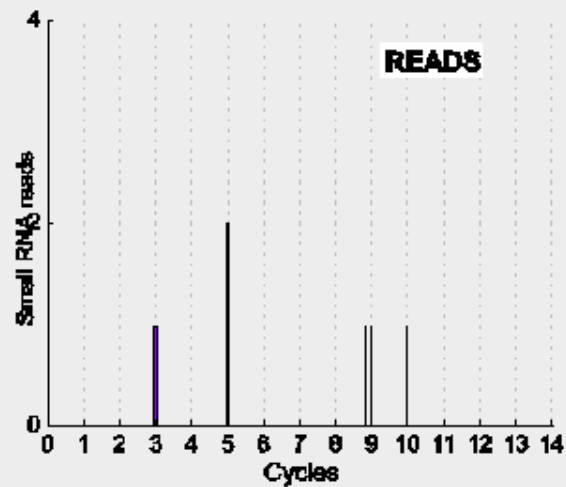
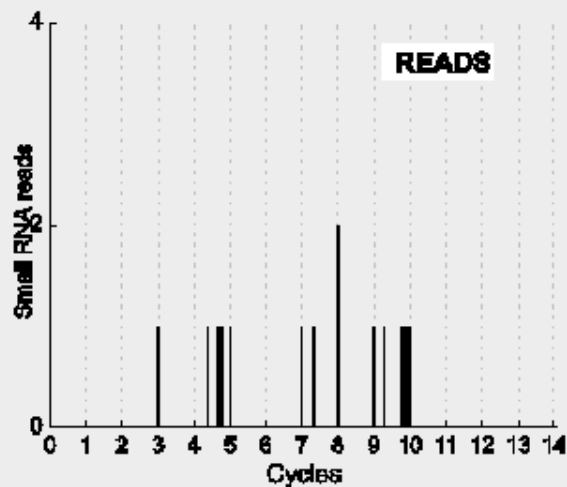
Phase peak value >1.4

- Chen et al. 2007. PNAS

P value <0.01







Os06g21900-2

Os02g39750 / Os02g89764

Os12g42380

Os06g21900-2

Os02g39750 / Os02g89764

Os12g42380

>TAS3_IGR1_Ch2:24015850,24016849 .

AGCACCACCCUUCUUCACCACACUUCUCAGCUCUACCACUCGCUACCCCCCUUUUCUUC
UUCUUCUCCUCUAGCCACCACCAUGCGUGACGACUCAAGCGGCAGCAUCGUCCACCAUUGCU
GCUGCUUCCUCCAUGGCCGCGCAAGCUUCCGGCCAUCAGCUGGUUCUUCUCGCUUGCCU
CCUUGGGGCGCCAUUGGAUCAGGAAGAGGAGGCUGAAGACUUGCAGCAAGCGGUGCGCCC
UUAUGUGAUGGCCCUAUC

CCCACCGAGCUUUUCUUCAC

cgecccuccucucguuc

||| ||||| | |||||

3' CCGCGAUAGGGAGGACUCGAA 5' **miR390**

uauuUCGGGAUGCUCUUGCAUUA^{GaggaauucuuucquuucgauGAAUGAUAGUUAAG}
CGGUGUCuucuuugaccuugcaagacuuuUCAUCGAAAAACACUCUUCUUcacuuuucug
uaauaucuuGCCAAAUUUCUCGUCCUCUCgcgucuaucacuccucccaucUUCUUUGU
CCCCGUCUAUC

|||||||:|||||||

3' CCGCGAUAGGGAGGACUCGAA 5' **miR390**

UUGUUAUUUAUAGAUGGCUAUGUAGUGAGUUGGUCUAUGUUUCUGAUUAUUCUUUUGAGUU
AUGAAUCAAUUAGUUGUAUGGUCGUCACUCGUCAGUAGUUCUUGCGUGGUGAUUGAUCU
CUCUCUAUAUAUAUAUGCUCUUGUUUUAUUGAUCCUAACUAACAGGGAAAGAAGAUUG
CCAUGCUAAGUAAAGCAAAGGGCAACGCUUGUGGUGUUUAAUACCUUGUGCUUGUGAUUGU
CCUCUCUCUCUCUCUCUCUCUCUCUCUCUCUCUCUCUCUCUCUCUCUCUUUA
ACCCUUGUUNCCUGUGUANUCACAGUGUGUACAUCAGUACAUGUUGGUGUUUCUGAUUCUC
UCUUUAUUUUUUUGUUGAUGCCGUCACGCGCUUGGACAGUUGGACGUGCAAAAUCUCUUGC
UAUAUAUAUAUAGCUGUGUGUUGUUUCGUCCGUGAGACUGUGAG

- initial and terminal phase
- tasiARF (Axtell et al. 2006)
- highly conserved phase