Viral Genome Organization

All biological organisms have a genome

- The genome can be either DNA or RNA
- Encode functions necessary
    - to complete its life cycle
    - interact with the environments

Variation

- Common feature when comparing genomes

Genome sequencing projects

- Uncovering many unique features
- These were previously known.

General Features of Viruses

Over 4000 viruses have been described

- Classified into 71 taxa
- Some are smaller than a ribosome

Smallest genomes

- But exhibit great variation

Major classifications

- DNA vs RNA

Sub classifications

- Single-stranded vs. double-stranded

Segments

- Monopartitie vs. multipartite

ssRNA virus classifications

- RNA strand found in the viron
    - positive (+) strand
        - most multipartite
    - negative (-) strand
        - many are multipartite

Virus replication by

- DNA viruses
    - DNA polymerase
        - Small genomes
            - Host encoded
        - Large genomes
            - Viral encoded
- RNA viruses
    - RNA-dependent RNA polymerase
    - Reverse transcriptase (retroviruses)

Genome sizes

- DNA viruses larger than RNA viruses

ss viruses smaller than ds viruses

- Hypothesis
- ss nucleic acids more fragile than ds nucleic
- drove evolution toward smaller ss genomes

Examples:

ssDNA viral genomes

- As small as 1300 nt in length

ssRNA viral genomes

- Smallest sequenced genome = 2300 nt
- Largest = 31,000 nt

Mutations

- RNA is more susceptible to mutation during transcription than DNA during replication
- Another driving force to small RNA viral genomes

Largest genome

- DNA virus
- *Paramecium bursaria* Chlorella virus 1
    - dsDNA
    - 305,107 nt
    - 698 proteins

Table 1. General features of sequenced viral genomes. The data was collected from NCBI (http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html) and represents the October 13, 2014 release from NCBI.

Viral class (# of segments; range of # protein) **Examples**	**# Completed genomes**	**Size range (nt)**	**# proteins**
dsDNA	1803		
Bovine polyomavirus		4,697	6
Pandoravirus salinus		2,473,870	2541
ssDNA	666		
Circoviridae SFBeef		858	1
Cellulophaga phage phi48:2		11,703	29
dsRNA	198		
Saccharomyces cerevisiae killer virus M1		1,801	1
Fiji disease virus		29,339	12
ssRNA	1126		
Blueberry mosaic associated virus		1,934	1
Marburg marburgvirus		38,225	14

DNA Viruses

Classified as

- ‘small’ genome
- ‘large’ genomes

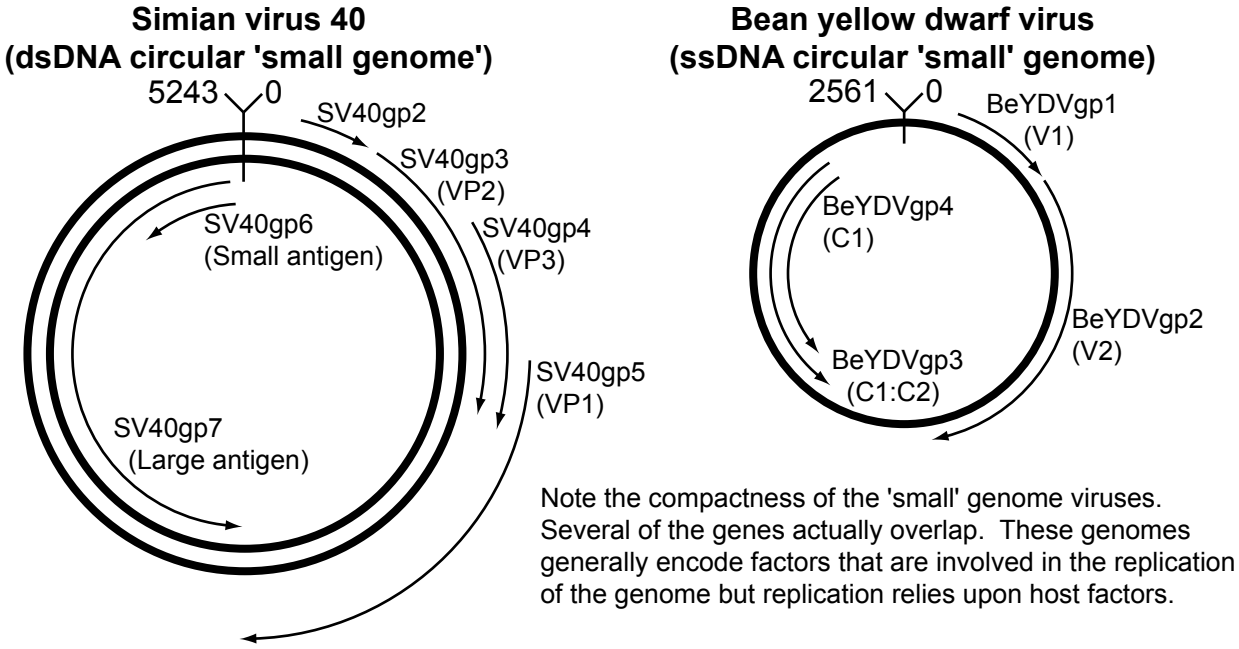
Typical observation of all viruses

- Nearly all DNA used for genes
- Different genomes differ in the number of proteins encoded

DNA Viral Genomes

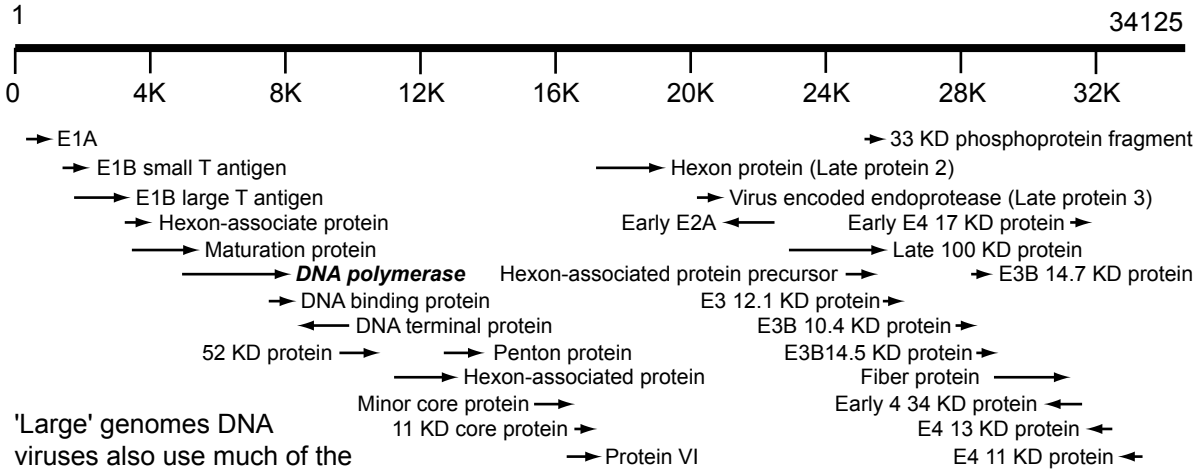
A. The Concept

DNA viruses are a major class of this biological entity. The viruses can be either ***double- or single-stranded***. In general, the single stranded genomes are smaller than those that are double-stranded. Among the double-stranded genomes, these can either have ***'small' or 'large' genomes***. One major difference between the two genomes is the mechanism of DNA replication. Small genomes use ***host polymerase activities***, whereas large genomes ***encode a DNA polymerase***.



Note the compactness of the 'small' genome viruses. Several of the genes actually overlap. These genomes generally encode factors that are involved in the replication of the genome but replication relies upon host factors.

Human adenovirus A (dsDNA linear 'large genome')



'Large' genomes DNA viruses also use much of the genome. A major difference is that these genomes do not have extensive overlaps between the genes. These genomes also encode a DNA polymerase (italicized above) that is used for genome replication.

Figure 1. Organization of DNA viral genomes.

Clustering is a common feature of 'small' genome viruses

Simian Virus 40

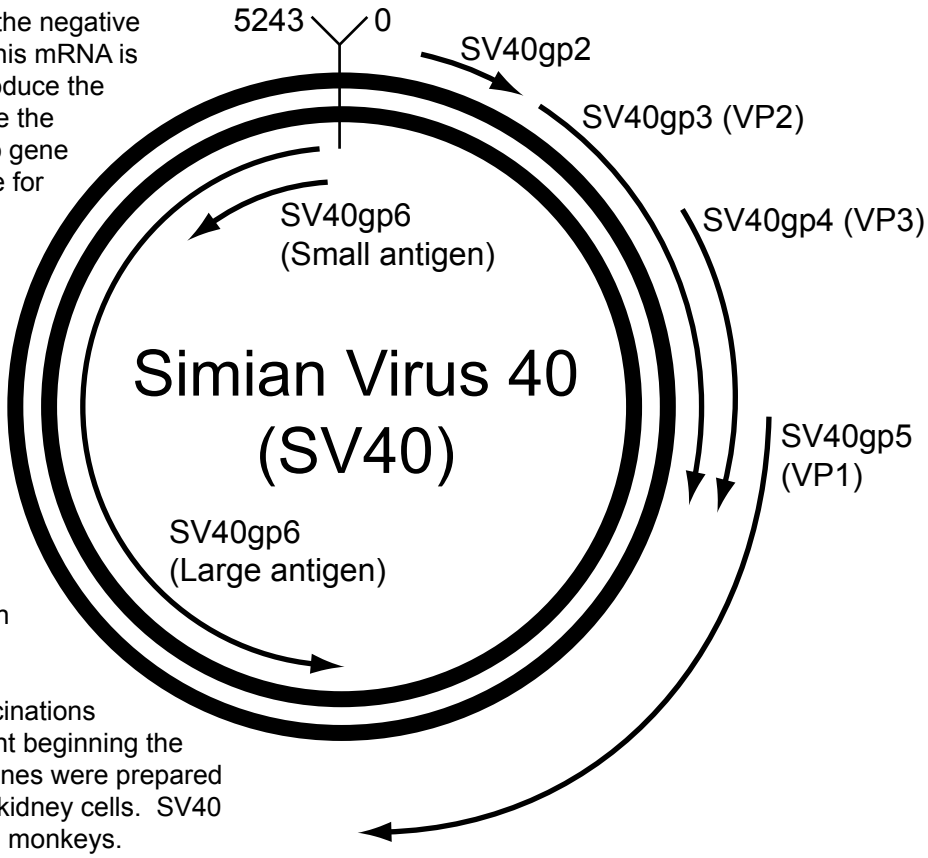
- Good example of a small genome
- Shows how a small genome can be extensively utilized
- Features:
    - 5243 nt dsDNA genome
        - Both strands contains genes
            - Five of the six genes overlap
        - "Life-cycle specific" regions
            - Early genes
                - Negative strand genes important for early development of new virions
                - Genes transcribed in a single mRNA
                - mRNA is alternatively spliced
                - Encode the small and large T-antigens
                - Proteins critical replication of the genome
            - Late genes
                - Encoded on the positive strand
                - Two genes (VP2 and VP3) overlap
                - A single mRNA for these genes
                - Alternative splicing produces unique mRNAs
                - Proteins critical for the structure of the virion

DNA Viruses: The SV40 Example

A. The Concept

DNA viruses genomes range from very small to very (1,360 nt) to very large (305,107 nt). These genomes encode a few (one) to many proteins (698). They also show a complex pattern of gene organization and gene expression. A good example is the ds DNA Simian Virus 40. This monkey pathogen has a small genome (5243 nt), encodes seven protein products, cotains overlapping genes, some of which are the result of alternate splicing.

1. Early gene expression. The first SV40 genes to be expressed are SV40pg6 and SV40gp7. Thes encolde the small and large T antigens, respectively. A single mRNA is transcribed counterclockwise using the negative strand as a template. This mRNA is alternately spliced to produce the mRNAs used to translate the two proteins. These two gene products are responsible for unwinding the DNA and making it ready for replication.



2. Late gene expression. The products of the SV40gp3, SV50gp4, and SV40gp5 genes produce the three proteins found in the viral capsid. These genes encode the VP2, VP3, and VP1 proteins, respecitvely. Alternate splcing of a single mRNA produces the mRNA used for VP2 and VP3 translation.SV40pg5 overlaps the coding region for SV40gp3 and SV40gp4. The transcription of these genes is in a clockwise orientation using the positive strand.

3. Note of interest. SV40 has received much attention because of its assocaition with polio vaccinations. Polio vaccinations were a major social event beginning the mid 1950s. These vaccines were prepared using Rhesus monkeys kidney cells. SV40 is a pathogen of Rhesus monkeys. Subsequents studies showed that vaccines developed from these cells were contaminated with SV40. This was of concern because SV40 can cause cancer. From 1955-1963, between 10 and 30 million individuals recevied the contaminated vaccines. These indiiduals may be infected with the virus.

Figure 2. Organization of SV40 viral genomes.

Large vs small DNA genomes

- Major difference
    - Small genomes
        - Use host factors for replicatioin
    - Large genomes
        - Encode a DNA polymerase

Large genomes

- More genes encode more proteins
- Genome organization is less complex
- Overlap of genes is less frequent

Human adenovirus A

- Genome size
    - dsDNA
- 34,125 nt
    - Number of genes
- 29 genes
    - Overlapping genes
        - Five
    - Complementary strand genes
        - Five

Table 2. Human adenovirus A genes and gene locations.

Genes	**nt location**
E1A (HAdVAgp01)	503-1099
E1B, small T-antigen (HAdVAgp02)	1542-2033
E1B, large T-antigen (HAdVAgp03)	1847-3395
Hexon-associated protein (HAdVAgp04)	3374-3808
Maturation protein (HAdVAgp05)	3844-5202
DNA polymerase (HAdVAgp06)	4953-8138
DNA binding protein (HAdVAgp07)	7602-8219
DNA terminal protein (HAdVAgp08)	8312-10131 (complement)
52 KD protein (HAdVAgp09)	10428-11549
Hexon-associated protein (HAdVAgp10)	11570-13318
Penton protein (HAdVAgp11)	13394-14887
Minor core protein (HAdVAgp12)	15500-16543
11 KD core protein (HAdVAgp13)	16568-16786
Protein VI (HAdVAgp14)	16843-17640
Hexon protein, Late protein 2 (HAdVAgp15)	17740-20499
Virus encoded endoprotease, Late protein 3 (HAdVAgp16)	20525-21145
Early E2A (HAdVAgp17)	21215-22669 (complement)
Late 100 KD protein (HAdVAgp18)	22695-25043
33 KD phophoprotein fragment (HAdVAgp19)	25202-25558
Hexon-associated protein precursor (HAdVAgp20)	25612-26313
E3 12.1 KD protein (HAdVAgp21)	26313-26630
E3B 10.4 KD protein (HAdVAgp22)	28207-28482
E3B 14.5 KD protein (HAdVAgp23)	28479-28811
E3B 14.7 KD protein (HAdVAgp24)	28804-29190
Fiber protein (HAdVAgp25)	29368-31131
Early E4 17 KD protein (HAdVAgp26)	31183-31407
Early E4 34 KD protein (HAdVAgp27)	31436-32311 (complement)
E4 13 KD protein (HAdVAgp28)	32244-32606 (complement)
E4 11 KD protein (HAdVAgp29)	32613-32963 (complement)

RNA Viruses

General Features

- Minimal Genome Size
- Encode a limited number of proteins
    - Often encodes a RNA-dependent RNA polymerase (RdRp)
        - Essential for the replication
        - Both positive and negative strand ssRNAs use such an enzyme
        - A function of dsRNA genomes also
        - A gene in both monopartite and multipartite genomes
- Number of proteins
    - Range from 1-13 proteins
- + vs – strand ssRNA
    - Polymersase is contained within ss (-) virion
    - Polymerase immediately translated from the RNA of the ss(+) RNA

Monopartite ssRNA viruses

- Genome can encode a single polyprotein
- Processed into a number of small molecules
- Each critical to complete the life cycle of the virus

Multipartite ssRNA Viruses

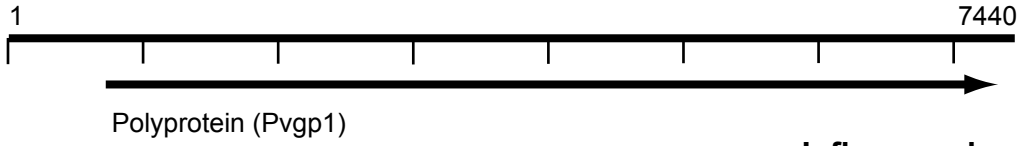
- Each segment generally contains a single gene

RNA Viral Genomes

A. The Concept

RNA viral genomes can be either ***single- or double-stranded***. In addition, these can be ***multipartite***, meaning they consist of several RNA molecules. The ssRNA molecules are also classified as ***positive- or negative-strand or retrovirues***. The + and - strand ssRNA genomes are replicated by a RNA-dependent RNA polymerase that is encoded by their genomes. Retroviruses are replicated as DNA following the conversion of the RNA into DNA by a ***reverse transcriptase***.

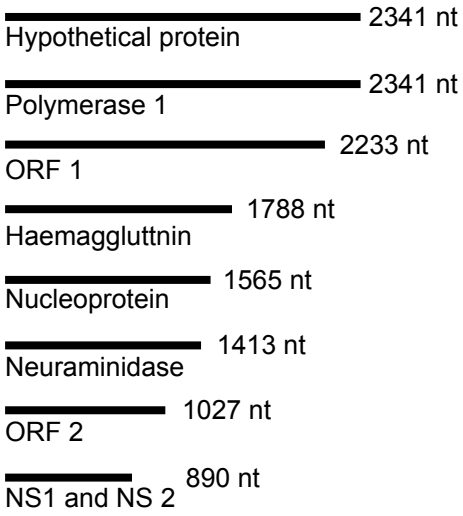
Poliovirus A (monopartite positive strand ssRNA)



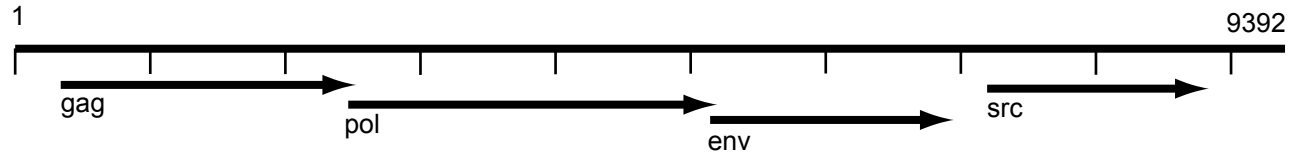
ssRNA viruses are compact. Monopartitie genomes generally have only a few genes. This masks the actually coding capacity. The Poliovirus A mRNA is used to translate a single **polyprotein**. This polyprotein is then cleaved into the 11 proteins necessary for the function of the virus. Multipartite ssRNA genomes partion the protein genes to several RNAs as seen here for Influenza Virus A. Both of these viruses encode a **RNA-dependent RNA polymerase**.

Retroviruses are also ssRNA viruses. The basic gene set for these viruses is *gag* (that encode the structural proteins, *pol* that (encodes the reverse transcriptase), and *env* (proteins that attach to the viron surface). In addtion, they can encode other genes. Oncogenes, such as Rous Sarcoma Virus *src* gene, can induce the cancerous state. Many of the retroviruses genes also encode polyproteins. The eight HIV I genes actually encode 22 different protiens. Retroviruses are replcated via a DNA intermediate. The reverse transcriptase creates a DNA copy of the genome that is used for replication purposes.

Influenza virus A (multipartite negative strand ssRNA)



Rous sarcoma virus (retrovirus)



Human immunodeficiency virus I (retrovirus)

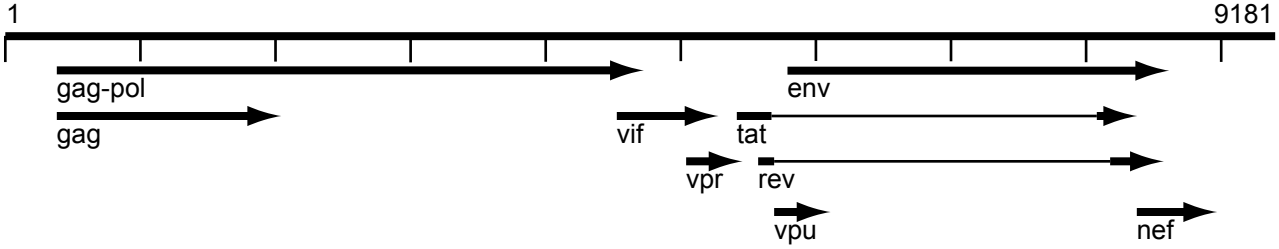


Figure 3. Organization of RNA viral genomes.

Retroviruses

- Minimal genome sizes
- Basic gene set
    - *Gag*
        - encode structural proteins
    - *Pol*
        - reverse transcriptase
    - *Env*
        - proteins embedded in the viral coat

Reverse transcriptase

- Converts RNA into a DNA copy
- Used to replicate the genome

Other Retroviral Genes

- Oncogenic retroviruses
    - Rous sarcoma virus
        - Fourth gene
            - Tyrosine kinase
            - Viral gene a mutated version of host gene
            - Gene causes uncontrolled cell growth
- Oncogenes generally are involved in cell growth and division

Human Immunodeficiency Virus I

- Causative agent of AIDS
- Contains the *gag/pol*/*env* suite of genes
- Example of a retrovirus that accumulated multiple genes
- A good model of retroviral evolution
- Genes expressed as polyproteins that are processed
- Additional genes
    - Affect other processes
        - Viral infectivity (*vif*)
        - Transcription activation (*tat*)
        - Replication (*vp*r, *vpu*, *nef*)
        - Regulation of virion protein expression (*rev*)

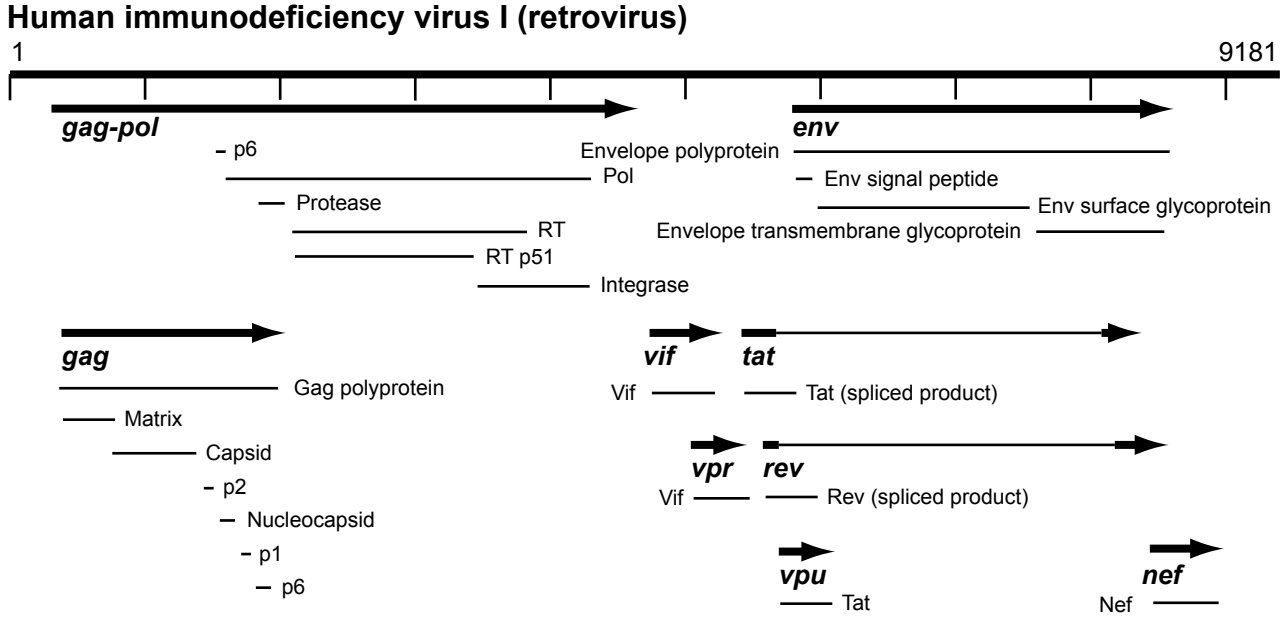
Table 3. Human immunodeficiency virus 1 control regions, genes, mature proteins and genetic locations.

Control region or gene **Mature protein(s)**	**Location (nt)**
polyA signal	73-78
5’ UTR	97-181
Primer binding	182-199
Gag-pol gene (HIV1gp1) Gag-pol transframe peptide (p6) Pol (unprocessed Pol polyprotein) Protease Reverse transcriptase Reverse transcriptase p51 subunit Integrase	336-4642 1632-1637, 1637-1798 1655-4639 1799-2095 2096-3775 2096-3415 3776-4639
Gag (HIV1gp2) Matrix (p17) Capsid (p24) p2 Nucleocapsid (p7) p1 p6	336-1838 336-731 732-1424 1425-1466 1467-1631 1632-1679 1680-1835
Vif (HIV1gp3) Vif (p23)	4587-5165 4587-5165
Vpr (HIVgp4) Vpr (p15)	5105-5396 5105-5396
Tat (HIV1gp5) Tat (p14)	5377-7970 5377-5591,7925-7970 (spliced)
Rev (HIV1gp6) Rev (p19)	5516-8199 5516-5592, 7925-8199 (spliced
Vpu (HIV1gp4) Vpu (p16)	5608-5856 5608-5856
Env (HIV1gp8) Envelope polyprotein Env signal peptide Envelope surface glycoprotein (gp120) Envelope transmembrane glycoprotein (gp41)	5771-8341 5771-8341 5771-5854 5855-7303 7304-8338
Nef (HIVgp9) Nef (p27)	8343-8963 8343-8963
3’ UTR	8631-9085

HIV I Genome and Proteins

A. The Concept

Viruses pack a lot of genetic information into a small amount of genomic space. A good example is the human immunodeficiency virus I. HIV I is the causal agent of AIDS. The small virus has eight gene which encode 22 proteins.



The HIV I virus is a good example of how a small genome can encode a large amount of genetic information. These eight HIV I genes encode 22 different genes. Three of the genes, gag-pol, gag, and env each encode a polyprotein that is processed into a collection proteins. It is relatively common for the *gag-pol*, *gag*, and *env* genes of retroviruses to encode multiple proteins. The most important protein for replication of the genome, the reverse transcriptase (RT), is part of the Pol polyprotein. The *gag-pol* polyprotein is cleaved by the protease encoded by the *gag-pol* gene. p2 cleaves the *gag* polyprotein.

One way to make the most from a limited size genome is to used the same sequence for multiple genes. Four of the genes, *ta*t, *rev*, *vpu*,and *nef* each share sequences with the *env* gene.

HIV 1 also has a feature in common with other retroviruses. It contains the required gag, pol and genes. It has functions,though, are defined by a series of other genes. These genes are necessary for viral infectivity (*vif*), transcription activation (*ta*t), replication (*vpr*, *vpu*, *nef*), and regulating virion protein expression (*rev*),

Figure 4. The genes and proteins of the human immunodeficiency virus I genome.

Bacterial Genomics

The Bacterial Kingdom

First organisms on earth

- 3 billion years ago
- 60% of earth's biomass
- Very diverse metabolic processes

Important to the biological world

- Donated mitochondria to eukaryotes
    - Alpha proteobacteria
- Created the atmosphere we live in today
    - Photosynthetic bacteria
        - Cyanobacteria
    - Evolved O_2 during the light reaction of photosynthesis
- Important ancestors of plants
    - Donated the photosynthetic system to eukaryotes

Found in all environments

- Soil
    - Interact with plants
        - Nitrogen fixation
            - *Rhizobium* species
- High-salt conditions
    - Halophiles
        - Great Salt Lake
- High temperatures
    - Thermal vents
        - *Themus aquaticus*
            - DNA polymerase used in PCR reactions

Mostly known as a pathogen

- Human diseases
    - Pneumonia
    - Blindness
    - Tuberulosis
    - Cholera
    - Plague (Black Death)
        - 25% of European population killed

Food industry

- Dairy products
    - Milk, cheese yogurt

Size varies

- Small
    - Mycoplasmas
        - 5X size of ribosome
- Large
    - *Thiomarginata namibiensis*
        - Size of a fruit-fly eye

Bacterial Clades (Based on Molecular Systematics)

Protobacteria

- Gram negative
- Photoautotrophs
- Chemoautotrophs
- Chemoheterotrophs

Alpha proteobacteria

- Associate with eukaryotic hosts
    - Rhizobium
- Rickettsias (Rocky Mountain Spotted Fever)
    - Endosymbionts

Beta Proteobacteria

- Some soil bacteria
    - Nitrogen recycling
        - Ammonium (NH_4^+) to nitrate (NO_2^-)

Gamma Proteobacteria

- Photoautotrophs and chemoheterotrophs
    - Non-oxygen evolving photobacteria
        - Sulfur bacteria
- Diseases
    - *Legionella* (Legionnaires' disease)
    - Intestinal (enteric) bacteria
        - *E. coli*
    - Cholera (*Vibrio cholerae*)
    - Food poisoning (*Salmonella*)

Delta proteobacteria

- Some are colony forming
- Develop fruiting bodies
    - *Myxobacteria*
    - *Chondromyces*
- Bacterial predators
    - *Bdellovibrio bacteriophorus*

Epsilon proteobacteria

- Close to the delta proteobacteria
- Diseases
    - Stomach ulcers
        - *Helicobacter pylori*

Chlamydias

- Gram negative
- Obligate animals heterotrophs
- Use host as ATP source
- Diseases
    - *Chlamydia trachomatis*
        - Most common sexual transmitted disease in US
        - Blindness

Spirochetes

- Can be:
    - Helical heterotrophs
    - Free-living
- Diseases
    - Syphillis (*Treponema pallidum*)
    - Lyme Disease (*Borrelia burgadorferi*)

Gram-positive Bacteria

- All gram-positive bacteria
  - Plus some gram-negative
- Nearly as diverse as Proteobacteria
- Most are free-living
- Diseases
  - Anthrax *(Bacillus anthracis*)
  - Botulism (*Clostribium botulinum*)
- Decompose organic matter in soil
- Antibiotic source
  - *Streptomyces*
- Some are colony formers
- Diseases
  - Tuberculosis
    - *Mycobacterium tuberculosis*
  - Leprosy
    - *Mycobacterium leprae*
- Mycoplasmas
  - Smallest bacteria
  - 5X size of ribosomes

Cyanobacteria

- Photoautotrophs
- Solitary and colonial
- Abundant in water

What genetic material do you find inside a bacterial cell?

Chromosome

- One chromosome (normally)
    - Multiple chromosome genomes discovered
        - 1989
            - *Rhodobacter sphaeroides*
                - Two chromosomes
- Circular (normally)
    - Linear genomes discovered
        - 1990s
            - *Borelias* and *Streptomyces*

Agrobacterium has a circular and linear chromosome

Chromosome exists as a bacterial nucleoid

- Core
  - Protein
    - HU, IHT, H1
    - Not absolutely required
    - Mutants of these genes viable
  - RNA
    - Function not known

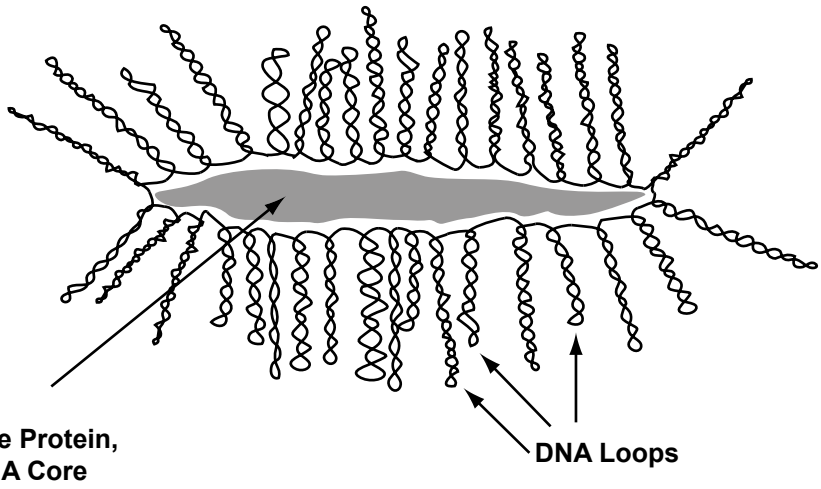
DNA loops

- *E. coli* example
  - Contour length larger than size of cell
    - DNA must be condensed
  - DNA loops major form of condensation
    - 40 kb in length
    - Negative supercoiling
    - Each loop independent

Bacterial Nucleoid

A. The Concept

The bacterial chromosome is generally a single, ***circular molecule***. Unlike the eukaryotic chromosome, it does not have a higher order. Thin-section electron micrographs show a structure in which the DNA is formed into loops. The actual physical features are not known. The DNA appears to be associated with a dense core of unknown nature. Collectively, the structure is called the ***bacterial nucleoid***.



The dense, internal region of the nucleoid consists of both protein and RNA. A number of proteins have been isolated (HU, IHF, H1). Each of these have a role in some DNA functions, but no mutant of any of these proteins is fatal. Therefore, it is unclear what are the critical protein components of the nucleoid. The function of the RNA component of the dense core is unknown

The contour length of a bacterial chromosome, for an organism such as *E. coli*, is about 1100 μm. This is much larger (1-2x) than the length for the entire cell. Therefore, the DNA must be condensed in some manner. ***DNA loops*** are the major condensation feature. The entire chromosome consists of 50-100 loops. The loops are generally 40 kb in length. Each loop cosists of negatively supercoiled DNA, and each loop is structurally independent of the other loops. This means if the supercoiled structure of one loop is released using enzymatic treatment, the neighboring loops will still be supercoiled.

An Old Adage Dispelled. Historically, it was generally reported that bacteria have a single, circular chromosome. This is not a universal truth. Although this is generally true, exceptions do exist. The first exceptions to the "single chromosome" rule was reported in 1989; *Rhodobacter sphaeroides* was discovered to have two large circular chromosomes. The "linear chromosome" rule was conclusively disproven in 1990 with the discovery that the genomes of genera *Borelias* and *Streptomyces* were linear. A dramatic exception is *Agrobacterium tumefaciens*: it contains two circular and two linear chromosomes.

Bacterial Plasmids

- Autonomous molecules in bacterial genomes
- Structure
    - Circular
    - Linear
- Size variation exists
    - Megaplasmid
        - Hundreds of kilobases in size
        - Contain hundreds of genes
        - Example
            - *Agrobacterium* (plant pathogen/plant transformation vector)
                - pAT
                    - 543 kb, 547 genes
                - pTI
                    - 214 kb, 198 genes
    - "Miniplasmid"
        - A few to tens of kb
        - A few to tens of genes
- Function of plasmids
    - Virulence
        - Attack other organisms
        - *Agrobacterium*
    - Drug resistance
    - Toxin production
    - Conjugation (transfer of DNA) with other bacteria
    - Metabolic degradation of molecules
        - Environmenal importance

Bacterial Genome Examples

Escherichia coli

- Circular chromosome
    - strain specific sizes
        - 4.5-5.5 megabases
        - 4300-5300 genes

Agrobacterium tumefaciens

- Chromosomes
    - Circular chromosome
        - 2.8 megabases
        - 2721 genes
    - Linear chromosome
        - 2.0 megabases
        - 1833 genes
- Megaplasmids
    - plasmid pAT
        - 543 kilobases
        - 547 genes
    - plasmid pTi
        - 214 kilobases
        - 198 genes

Borrelia burgdorferi

- Linear chromosome
  - 911 kilobases
  - 850 genes
- 21 plasmids
  - 12 circular plasmids
    - 9-31 kilobases
    - 11-45genes
  - 9 linear mini plasmids
    - 5-54 kilobases
    - 6-76 genes

Microbial Genome Sequencing

The Beginnings

- *Hemophilius influenzae*
  - 1995
  - First microbe genome sequenced
  - Shotgun sequence approach used
    - Showed utility of shotgun sequencing

Genomes sequenced

- September 2003
  - 136 species publicly released
    - 120 bacteria
    - 16 archaea
  - TIGR (http://www.tigr.org/tdb/mdb/mdbcomplete.html)
    - The Institute for Genome Research
    - 32 genomes sequenced
    - Focuses on microbes
    - Promoted importance of microbe sequence information
      - Useful in bioterrorism defenses
  - Private industry
    - Don’t know how many sequenced
    - Industrial uses are being investigated

Table 1. Features of selected bacterial genomes. Data compiled from http://www.cbs.dtu.dk/services/GenomeAtlas/Bacteria/index.html

Species	Interest	Size (nt)	G+C (%)	Coding density	bp/gene	# Genes (% unique)*
Mycoplasma genitalium		580,074	32	90	1208	480 (20)
Bruchnera aphidicola		609,132	26	81	1222	504 (1)
Rickettsia conorii	Spotted fever	1,268,755	33	80	923	1374 (36)
Haemophilus influenzae	Flu	1,830,138	39	85	1070	1709 (14)
Lactococcus lactis	Cheese starter	2,365,589	36	84	1043	2266 (25)
Clostridium tetani	Tetnus	3,799,251	29	85	1179	2373
Clostridium perfringens	Gangrene	3,031,430	29	83	1139	2660
Mycobacterium leprae	Leprosy	3,268,203	58	76	1201	2720 (27)
Clostridium acetobutylicum	Solvent producer	3,940,880	31	85	1073	3672 (26)
Vibrio cholerae (Total)	Cholera	4,033,464	48	86	1053	3828 (20)
Chromosome 1		2,961,149	48	87	1082	2736
Chromosome 2		1.072,315	47	84	982	1092
Mycobacterium tuberculosis	Tuberculosis	4,411,529	66	90	1126	3918 (20)
Escherichia coli	Model					
K-12 MG1655	organism	4,639,221	51	87	1081	4289
K-12 W3110		4,641,433	51	87	1057	4390
CFT073		5,231,428	51	87	972	5379
O157 RIMD0509952		5,498,450	51	87	1025	5361
O157 EDL93		5,528,970	51	87	1033	5349 (25)
Yersinia pestis	Plague	4.653,728	48	83	1161	4008 (17)
Xanthomonas campestris	Citrus canker	5,076,188	66	84	na	4181
Vibrio parahaemolyticus	Gastroenteritis	5,165,770	46	86		4832
Chromosome 1		3,288,558	46	86	1067	3080
Chromosome 2		1,877,212	46	86	1071	1752
Bacillus anthracis	Anthrax	5,227,293	36	84	800	5,508
Agobacterium tumefaciens	Transformation	5,673,563	60	89	1070	5299
Chromosome 1 (circular)	vector	2.841,581	60	88	1040	2721 (16)
Chromosome 2 (linear)		2,074,782	60	90	1131	1833
plasmid pAT		542,869	58	85	992	547
plasmid pTi		214,331	57	86	1082	198
Pseudomonas syringae	Plant pathogen	6,397,126	59	85	1169	5471
Anabaena nostic	Photosynthesis	7,211,893	42	82		6129
Streptomyces avermitilis	Antibiotics	9,025,608	71	86	1191	7575
Bradyrhizobium japonicum	N2 fixation	9,105,828	65	86	1094	8317

*Unique sequences are those not listed as part of a COG (Cluster of Orthologous Genes)

General Features of Bacterial Genomes

- 16X differences in genome sizes
  - *Mycoplasma genitalium* vs. *Bradyrhizobium japonicum*
- Wide range of G+C content
- Low degree of intergenic DNA (10-24%)
- Gene sizes are similar
- Increased genome size means more genes
  - For genomes in Table 1
    - r = 0.98 between genome size and number of genes

Unique Genes

COG

- Cluster of Orthologus Genes
  - If gene similar to two other genes, it is considered part of a cluster
  - Determines if a gene is unique
- Most genes part of a known class
  - Unique genes may define unknown functions specific to a species
- Novel ORFs appear in each newly sequenced genome
  - Source???
    - Rapidly evolving genes
    - Genes of lineage specific function

Relationship Among Genes and Proteins

Essential terms

- Homolog
    - two genes that are related by descent
    - Important to note
        - genes are homologous or they are not homologus
        - there is not percentage homology
        - Proper way of expressing the relationship
            - "Genes A and B are homologous and share X% amino acid (or nucleotide) identity."
        - For amino acids
            - Proteins can be identical or similar
                - Identity
                    - % identical amino acids
                - Similar
                    - % similar amino acids
                    - similar amino acids share similar biochemical properties

Ortholog

- ***Two genes from different species*** that are identical by descent

Paralog

- ***Two genes from the same species*** that have arisen by gene duplication

Table 2. A classificatioin scheme for functional genomics developed by TIGR (http://www.biochem.ucl.ac.uk/~rison/FuncSchemes/Tables/tigr.html)

1 AMINO ACID BIOSYNTHESIS
- 1.1 Other
- 1.2 Serine family
- 1.3 Pyruvate family
- 1.4 Histidine family
- 1.5 Glutamate family
- 1.6 Aspartate family
- 1.7 Aromatic amino acid family

2 AUTOTROPHIC METABOLISM
- 2,1 Chemoautotrophy
- 2.2 Photoautotrophy

3 BIOSYNTHESIS OF COFACTORS, PROSTHETIC GROUPS, AND CARRIERS
- 3.1 Pantothenate
- 3.2 Pyridine nucleotides
- 3.3 Pyridoxine
- 3.4 Riboflavin
- 3.5 Thiamine
- 3.6 Other
- 3.7 Molybdopterin
- 3.8 Biotin
- 3.9 Folic acid
- 3.10 Glutathione
- 3.11 Heme and porphyrin
- 3.12 Lipoate
- 3.13 Menaquinone and ubiquinone

4 CELL ENVELOPE
- 4.1 Other
- 4.2 Surface structures
- 4.3 Lipoproteins
- 4.4 Degradation of polysaccharides
- 4.5 Biosynthesis of surface polysaccharides and lipopolysaccharides
- 4.6 Biosynthesis of murein sacculus and peptidoglycan

5 CELLULAR PROCESSES
- 5.1 Other
- 5.2 Transformation
- 5.3 Toxin production and resistance
- 5.4 Protein and peptide secretion
- 5.5 Detoxification
- 5.6 Chaperones
- 5.7 Cell division

6 CENTRAL INTERMEDIARY METABOLISM
- 6.1 Other
- 6.2 Sulfur metabolism
- 6.3 Polyamine biosynthesis
- 6.4 Phosphorus compounds
- 6.5 Nitrogen metabolism
- 6.6 Nitrogen fixation
- 6.7 Amino sugars

7 DNA METABOLISM

- 7.1 Restriction/modification
- 7.2 Degradation of DNA
- 7.3 DNA replication, recombination, and repair
- 7.4 Chromosome-associated proteins

8 ENERGY METABOLISM

- 8.1 Methanogenesis
- 8.2 Pentose phosphate pathway
- 8.3 Polysaccharides
- 8.4 Pyruvate dehydrogenase
- 8.5 Sugars
- 8.6 TCA cycle
- 8.7 Other
- 8.8 Glycolysis/gluconeogenesis
- 8.9 ATP-proton motive force interconversion
- 8.10 Aerobic
- 8.11 Amino acids and amines
- 8.12 Anaerobic
- 8.13 Electron transport
- 8.14 Entner-Doudoroff
- 8.15 Fermentation

9 FATTY ACID AND PHOSPHOLIPID METABOLISM

- 9.1 Biosynthesis
- 9.2 Degradation
- 9.3 Other

10 HYPOTHETICAL

- 10.1 General

11 PURINES, PYRIMIDINES, NUCLEOSIDES, AND NUCLEOTIDES

- 11.1 Other
- 11.2 Sugar-nucleotide biosynthesis and conversions
- 11.3 Salvage of nucleosides and nucleotides
- 11.4 Pyrimidine ribonucleotide biosynthesis
- 11.5 Purine ribonucleotide biosynthesis
- 11.6 Nucleotide and nucleoside interconversions 11.7 2'-Deoxyribonucleotide metabolism

12 REGULATORY FUNCTIONS

- 12.1 General

13 TRANSCRIPTION

- 13.1 Other
- 13.2 Transcription factors
- 13.3 RNA processing
- 13.4 Degradation of RNA
- 13.5 DNA-dependent RNA polymerase

14 TRANSLATION

- 14.1 Other
- 14.2 tRNA modification
- 14.3 Translation factors
- 14.4 Ribosomal proteins: synthesis and modification
- 14.5 Amino acyl tRNA synthetases
- 14.6 Degradation of proteins, peptides, and glycopeptides
- 14.7 Nucleoproteins
- 14.8 Protein modification

15 TRANSPORT AND BINDING PROTEINS

- 15.1 Other
- 15.2 Unknown substrate
- 15.3 Porins
- 15.4 Nucleosides, purines and pyrimidines
- 15.5 Amino acids, peptides and amines
- 15.6 Anions
- 15.7 Carbohydrates, organic alcohols, and acids
- 15.8 Cations

16 OTHER CATEGORIES

- 16.1 Other
- 16.2 Transposon-related functions
- 16.3 Phage-related functions and prophages
- 16.4 Adaptations and atypical conditions

Table 3. Cluster of Orthologous Genes (COG) functional groups.

Code	Function	# Pathways/ functional systems
Information storage and processing		
J	Translation, ribosomal structure and biogenesis	4
K	Transcription	3
L	DNA replication, recombination and repair	2
Cellular processes		
D	Cell division and chromosome partitioning	-
O	Posttranslational modification, protein turnover, chaperones	-
M	Cell envelope biogenesis, outer membrane	1
N	Cell motility and secretion	2
P	Inorganic ion transport and metabolism	1
T	Signal transduction mechanisms	-
Metabolism		
C	Energy production and conversion	7
G	Carbohydrate transport and metabolism	4
E	Amino acid transport and metabolism	10
F	Nucleotide transport and metabolism	5
H	Coenzyme metabolism	11
I	Lipid metabolism	2
Q	Secondary metabolites biosynthesis, transport and catabolism	-
Poorly characterized		
R	General function prediction only	-
S	Secondary metabolites biosynthesis, transport and catabolism	-

Table 4. Translation factors and enzymes involved in translation

Gene	Category	COG	Function
TIF6	[J]	COG1976	Eukaryotic translation initiation factor 6 (EIF6)
TufB	[JE]	COG0050	GTPases - translation elongation factors
eRF1	[J]	COG1503	Peptide chain release factor eRF1
Def	[J]	COG0242	N-formylmethionyl-tRNA deformylase
Map	[J]	COG0024	Methionine aminopeptidase
Fmt	[J]	COG0223	Methionyl-tRNA formyltransferase
Pth	[J]	COG0193	Peptidyl-tRNA hydrolase
PrfA	[J]	COG0216	Protein chain release factor A
PrfB	[J]	COG1186	Protein chain release factor B
Frr	[J]	COG0233	Ribosome recycling factor
FusA	[J]	COG0480	Translation elongation and release factors (GTPases)
EFB1	[J]	COG2092	Translation elongation factor EF-1beta
Efp	[J]	COG0231	Translation elongation factor P/translation initiation factor eIF-5A
Tsf	[J]	COG0264	Translation elongation factor Ts
SUI1	[J]	COG0023	Translation initiation factor (SUI1)
nfB	[J]	COG0532	Translation initiation factor 2 (GTPase)
InfA	[J]	COG0361	Translation initiation factor IF-1
InfC	[J]	COG0290	Translation initiation factor IF3
GCD7	[J]	COG1601	Translation initiation factor eIF-2, beta subunit/eIF-5 N-terminal domain
GCN3	[J]	COG0182	Translation initiation factor eIF-2B alpha subunit
GCD2	[J]	COG1184	Translation initiation factor eIF-2B delta subunit
SUI2	[J]	COG1093	Translation initiation factor eIF2alpha

COGs

- Based on bacterial and yeast genomes
- September 10, 2003 information
- Total number
  - 3307 genes
    - Two species COGs
      - 115 genes
    - Three species COGs
      - 493 genes
    - 26 species
      - 84 genes
  - A **few genes** are widely common
  - **Most genes** are shared with only a few other species

see: http://www.ncbi.nlm.nih.gov/COG/
http://www.ncbi.nlm.nih.gov/cgi-bin/COG/palox?num=all

Examples of COGs

	# prot	Species	Symbol	Categ		Gene
-	22	--m-k---vd-lb-efgh---j----	RsmC	[J]	COG2813	16S RNA G1207 methylase RsmC
-	64	-------qvdrlbcefghsnujxit-	RsuA	[J]	COG1187	16S rRNA uridine-516 pseudouridylate synthase and related pseudouridylate synthases
-	21	a-mpkz-qvd--b-ef-----j----	LigT	[J]	COG1514	2'-5' RNA ligase
3	59	aompkz-qvdr-bcefghsnujxit-	MiaB	[J]	COG0621	2-methylthioadenine synthetase

Table 5. Distribution of genes by functional classes for selected bacterial genomes.

	Mycoplasma genitalium	*Rickettsia conorii*	*Haemophilus influenzae*	*Escherichia coli K-12*	*Bradyrhizobium japonicum*
Total Proteins	484	1374	1709	4279	8317
Proteins in COG	385	876	1591	3587	6197
Translation	101	126	149	171	205
RNA procesisng and modification	0	0	1	2	0
Transcription	40	25	73	280	499
Replication, recombination, repair	14	71	111	220	369
Chromatin structure and dynamics	0	0	0	0	2
Cell cycle control, mitosis, meiosis	5	18	24	34	41
Nuclear structure	0	0	0	0	0
Defense mechanisms	8	23	19	48	101
Signal transduction mechanisms	3	22	37	134	328
Cell wall/membrane biogenesis	12	82	122	235	314
Cell motility	0	3	6	107	138
Cytoskeleton	0	0	0	0	0
Extracellular structures	0	0	0	0	0
Intracellular trafficking and secretion	6	32	27	37	74
Posttranslational modification, protein turnover, chaperones	20	56	87	128	233
Energy production and conversion	20	77	95	275	445
Carbohydrate transport and metabolism	26	35	104	368	440
Amino acid transport and metabolism	15	33	154	350	723
Nucleotide transport and metabolism	21	21	57	87	95
Coenzyme transport and metabolism	14	31	72	123	188
Lipid transport and metabolism	9	36	40	83	402
Inorganic ion transport and metabolism	17	22	91	191	298
Secondary metabolites biosynthesis, transport and catabolism	0	13	18	68	179
General function prediction only	40	91	146	338	627
Function unknown	14	59	158	308	496
not in COGs	99	498	118	692	2120

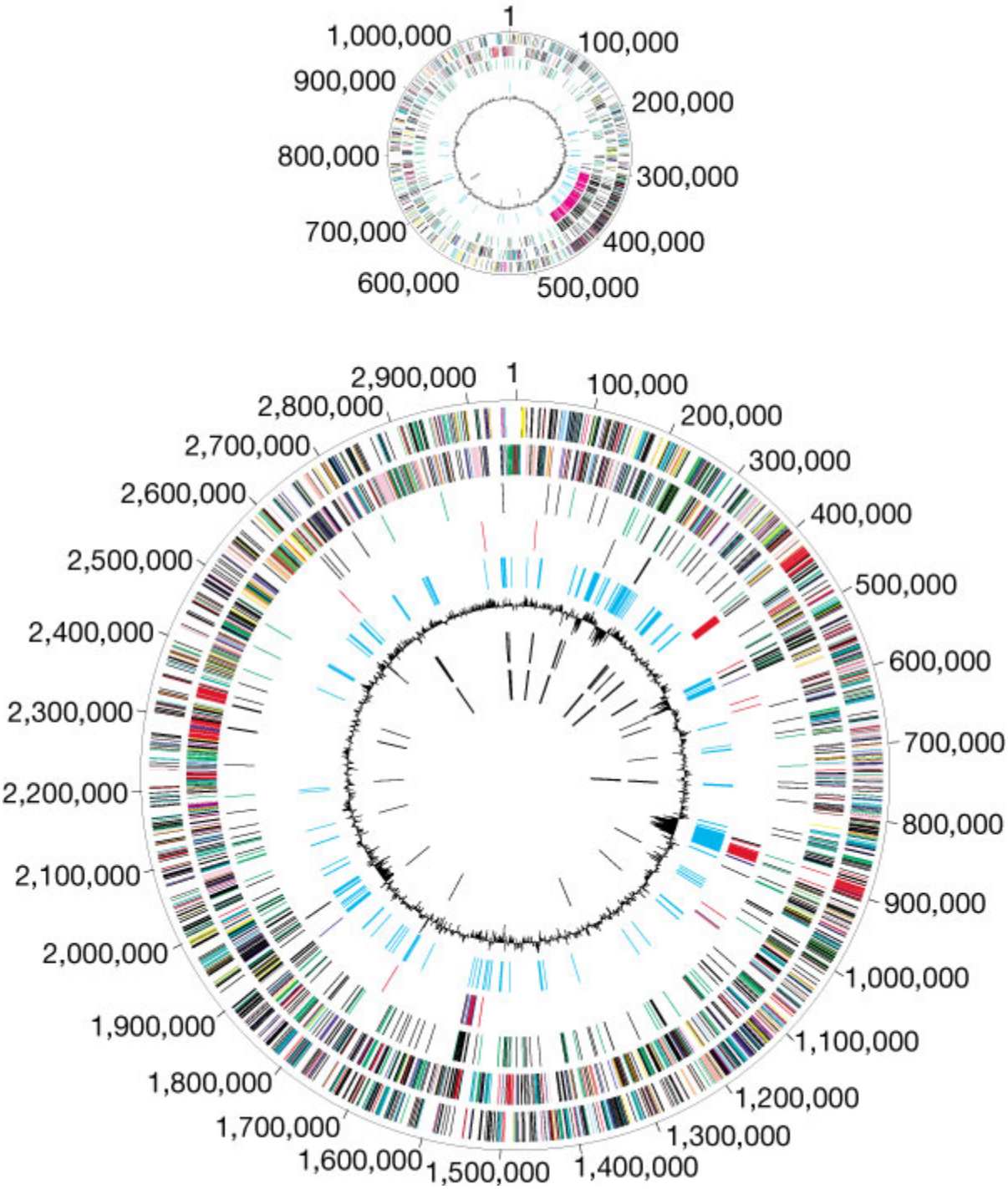


Figure 2 Circular representation of the *V. cholerae* genome. The two chromosomes, large and small, are depicted. From the outside inward: the first and second circles show predicted protein-coding regions on the plus and minus strand, by role, according to the colour code in Fig. 1 (unknown and hypothetical proteins are in black). The third circle shows recently duplicated genes on the same chromosome (black) and on different chromosomes (green). The fourth circle shows transposon-related (black), phage-related (blue), VCRs (pink) and pathogenesis genes (red). The fifth circle shows regions with significant χ^2 values for trinucleotide composition in a 2,000-bp window. The sixth circle shows percentage G+C in relation to mean G+C for the chromosome.The seventh and eighth circles are tRNAs and rRNAs, respectively.

Graphical Representation of Genomes

A. The Concept

Genomic sequencing generates a tremendous amount of sequence data. It is always a challenge to represent that in a manner that is digestable to the scientific public. Chromosome information is typically represented in a linear form. This also is true for genomes, such as for many bacterial speices that have circular chromosomes.

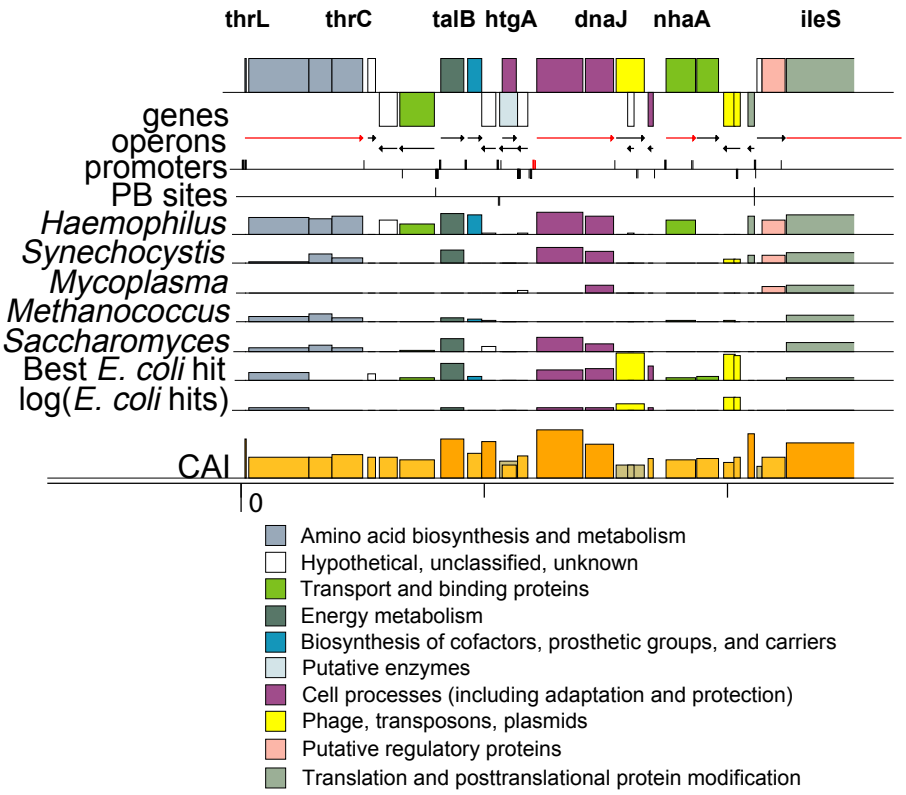


Fig. 3. (PDF). Map of the complete E. coli sequence, its features and similarities to proteins from five other complete genome sequences, proceeding from left to right in 42 tiers. The top line shows each gene or hypothetical gene, color-coded to represent its known or predicted function as assigned on the basis of biochemical and genetic data. Genes are vertically offset to indicate their direction of transcription. Space permitting, names of previously described E. coli genes are indicated above the line. The second line contains arrows indicating documented (red) and predicted (black) operons. Documented operons encoding stable RNAs are blue. Line 3, below the operons, contains tick marks showing the position of documented (red), predicted (black), and stable RNA (blue) promoter sequences. Line 4 consists of tick marks showing the position of documented (red) and predicted (black) protein binding sites. Lines 5 to 9 are histograms showing the results of alignments between E. coli proteins and the products encoded by five other complete genomes. The height of each bar is a simple index of similarity: the product of the percent of each protein in the pairwise alignment and the percent amino acid identity across the aligned region. Line 10 indicates similarity among proteins in E. coli in the same fashion. Line 11 histograms show the logarithm of the number of proteins in the E. coli genome that match a particular protein. Line 12 in each tier is a histogram that indicates the CAI of each ORF. Genes with intermediate CAI values are shown in orange, genes with high CAI values (>90th percentile) are a darker shade of orange, genes with low CAI values (<10th percentile) are light brown, and clusters of four or more genes with low CAI values (<0.25) are yellow. The final line in each tier is a scale showing position (in base pairs).

Figure 2. A graphical representation of the first seven genes of the *Eschericia coli* genome. (The legend is from the original manuscript: Science (1997) 277:1453.

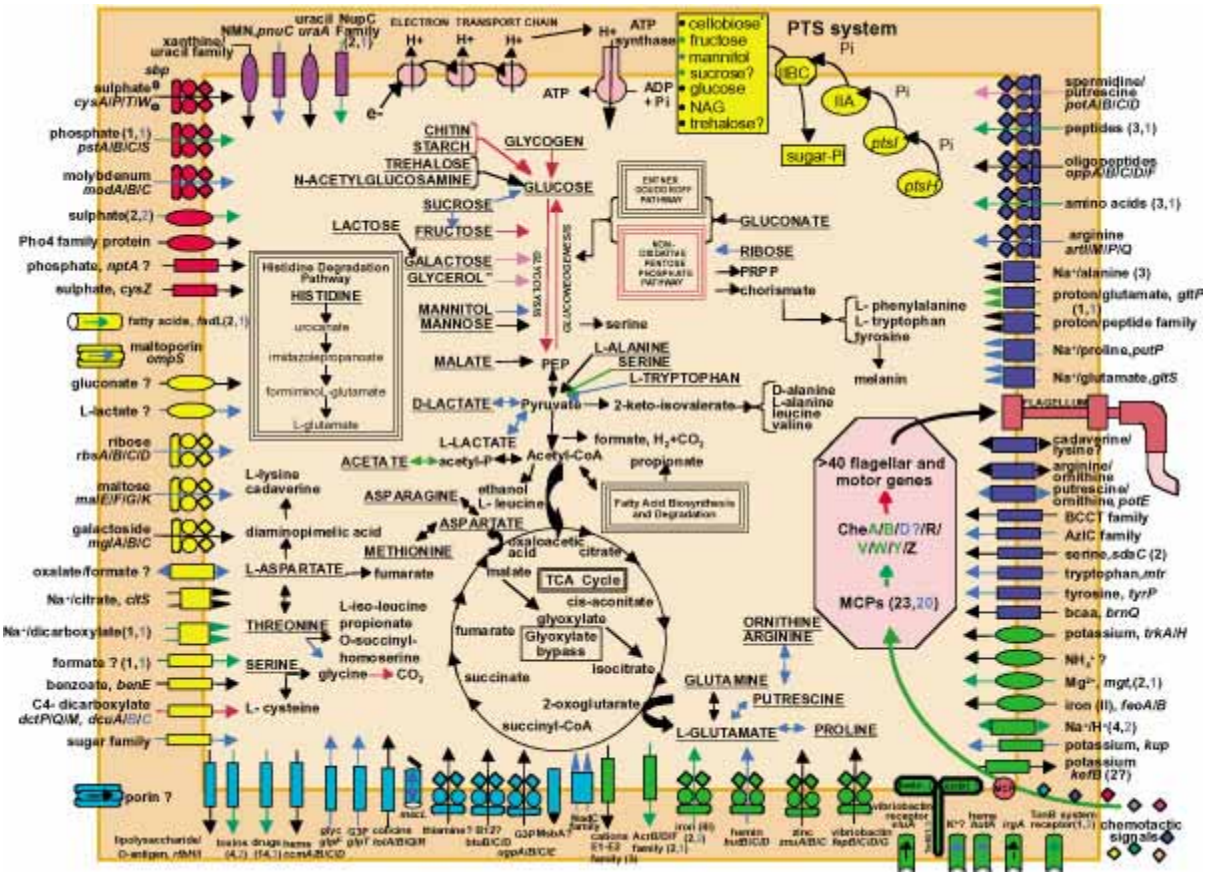


Figure 3 Overview of metabolism and transport in *V. cholerae*. Pathways for energy production and the metabolism of organic compounds, acids and aldehydes are shown. Transporters are grouped by substrate specificity: cations (green), anions (red), carbohydrates (yellow), nucleosides, purines and pyrimidines (purple), amino acids/peptides/amines (dark blue) and other (light blue). Question marks associated with transporters indicate a putative gene, uncertainty in substrate specificity, or direction of transport. Permeases are represented as ovals; ABC transporters are shown as composite figures of ovals, diamonds and circles; porins are represented as three ovals; the large-conductance mechanosensitive channel is shown as a gated cylinder; other cylinders represent outer membrane transporters or receptors; and all other transporters are drawn as rectangles. Export or import of solutes is designated by the direction of the arrow through the transporter. If a precise substrate could not be determined for a transporter, no gene name was assigned and a more general common name reflecting the type of substrate being transported was used. Gene location on the two chromosomes, for both transporters and metabolic steps, is indicated by arrow colour: all genes located on the large chromosome (black); all genes located on the small chromosome (blue); all genes needed for the complete pathway on one chromosome, but a duplicate copy of one or more genes on the other chromosome (purple); required genes on both chromosomes (red); complete pathway on both chromosomes (green). (Complete pathways, except for glycerol, are found on the large chromosome.) Gene numbers on the two chromosomes are in parentheses and follow the colour scheme for gene location. Substrates underlined and capitalized can be used as energy sources. PRPP, phosphoribosyl-pyrophosphate; PEP, phosphoenolpyruvate; PTS, phosphoenolpyruvate-dependant phosphotransferase system; ATP, adenosine triphosphate; ADP, adenosine diphosphate; MCP, methyl-accepting chaemotaxis protein; NAG, *N*-acetylglucosamine; G3P, glycerol-3-phosphate; glyc, glycerol; NMN, nicotinamide mononucleotide. Asterisk, because *V. cholerae* does not use cellobiose, we expect this PTS system to be involved in chitobiose transport.

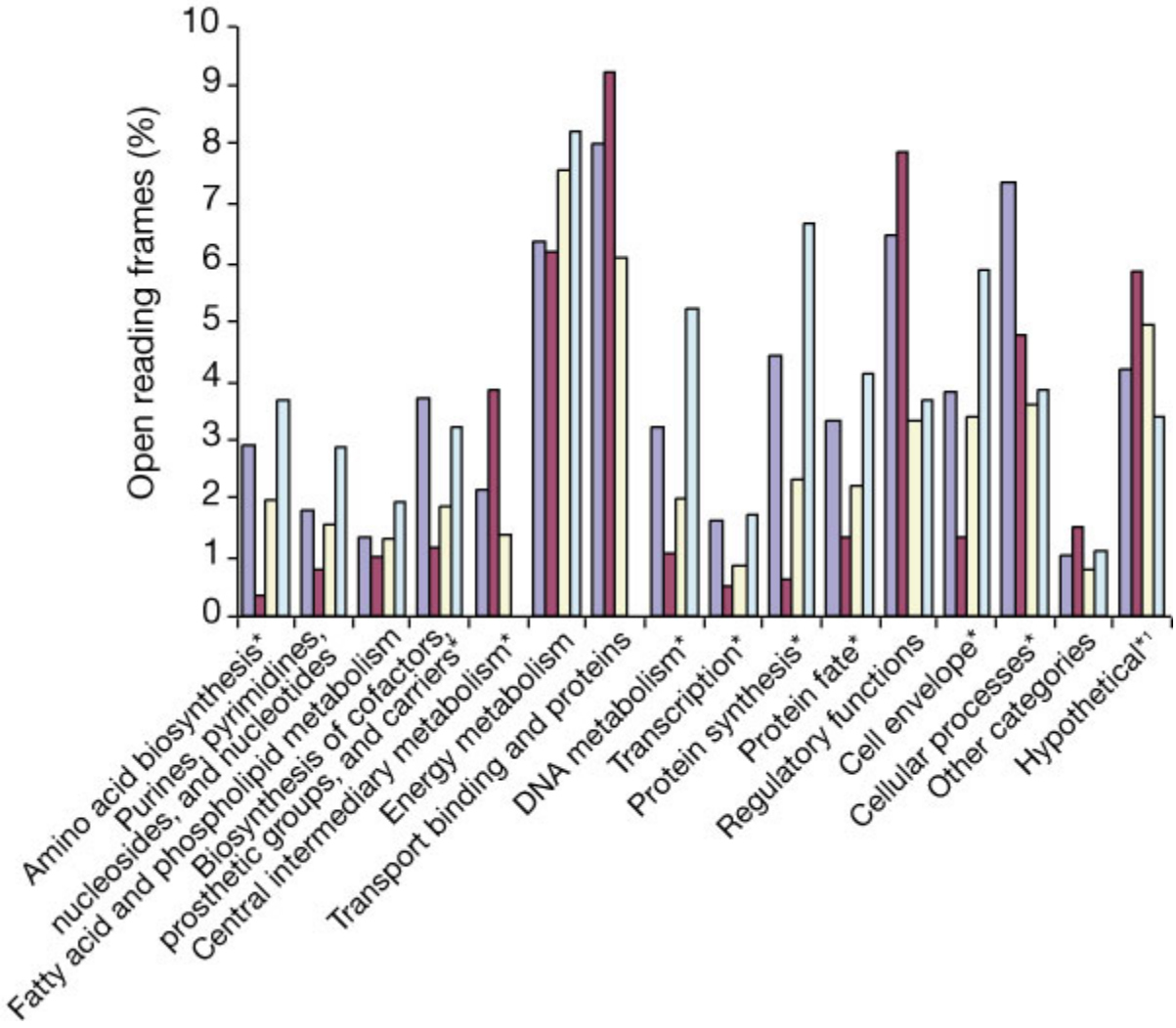


Figure 4 Percentage of total *Vibrio cholerae* open reading frames (ORFs) in biological roles compared with other γ-Proteobacteria. These were *V. cholerae*, chromosome 1 (blue); *V. cholerae*, chromosome 2 (red); *Escherichia coli* (yellow); *Haemophilus influenzae* (pale blue). Significant partitioning ($P < 0.01$) of biological roles between *V. cholerae* chromosomes is indicated with an asterisk, as determined with a χ^2 analysis. 1, Hypothetical contains both conserved hypothetical proteins and hypothetical proteins, and is at 1/10 scale compared with other roles.

Leading vs Lagging Strand

- Genes located on both strands
- Most genes on the leading strand
- But highly expressed genes preferentially on leading strand
    - Reason???
      DNA and RNA polymerase functions would collide on lagging strand

G+C Content

- Range
    - 22% *Wigglesworthis glossinidia* (tsetse fly endosymbiont)
    - 67% *Pseudomonas aeruginosa*
- Genome-wide
    - G+C not related to the thermal environment
    - But
        - For species living in elevated temperatures, structural RNAs have higher G+C content in ds regions
        - Aerobic genomes have higher G+C content than anaerobic bacteria
- Within a species
    - G+C content does not vary among genes
    - Genes with unusual G+C content are indicative of lateral transfer of genes

Operons

Definition

- Cluster of gene under the control of a single promoter that are expressed as a single mRNA

Components

- Promoter
- Operator
    - Repressor protein binds to this site
- Gene(s)

Example:

- Lac Operon
- Features
    - Promoter
    - Operator
    - LacZ (beta-galactosidase)
    - LacY (beta-galactoside permease)
    - LacA (beta-galactoside transacetylase)
    - Activation
        - Increased levels of lactose
        - Repressor released

E. coli Operons

Prediction method

- Distance between genes
    - Is there enough distance for a promoter???
        - No: then genes are part of the same operon

Total

- 392 known
- 2192 predicted
- one gene
    - 73% (surprisingly high)
- two genes
    - 16.6%
- three genes
    - 4.6%
- four or more genes
    - 6.0%

E. coli promoters

- 2584 operons
    - 2402 predicted promoters
- one promoter per operon
    - 68%
- two promoters per operon
    - 20%
- three or more promoters
    - 12%

Horizontal (or Lateral) Gene Transfer in Bacterial Genomes

How is DNA transferred between bacterial genomes?

- Mechanisms are known
  - Transformation
    - Free DNA is known to exist in the biological world
      - Bacteria are known to take DNA up from the environment (=transformation)
    - Influenced by high population density
    - Influenced by salt concentration
    - Practical use
      - Introducing foreign DNA into bacteria for cloning
  - Conjugation
    - Well studied biological function of bacteria
      - A pilus is formed between bacteria
      - DNA is transferred between cells via the pilus
      - Can occur between distantly related species
        - *E. coli* and cyanobacteria
        - *E. coli* and yeast
  - Transduction
    - Transferred mediated by viruses
    - Bacterial genes encapsulated in viral genome by mistake
    - Bacterial genes transferred along with the viral gene
    - Detected as “foreign bacteria genes” surrounded by phage sequences

Eurkaryotic Organelles

Role of Organelles

- Energy generating compartments

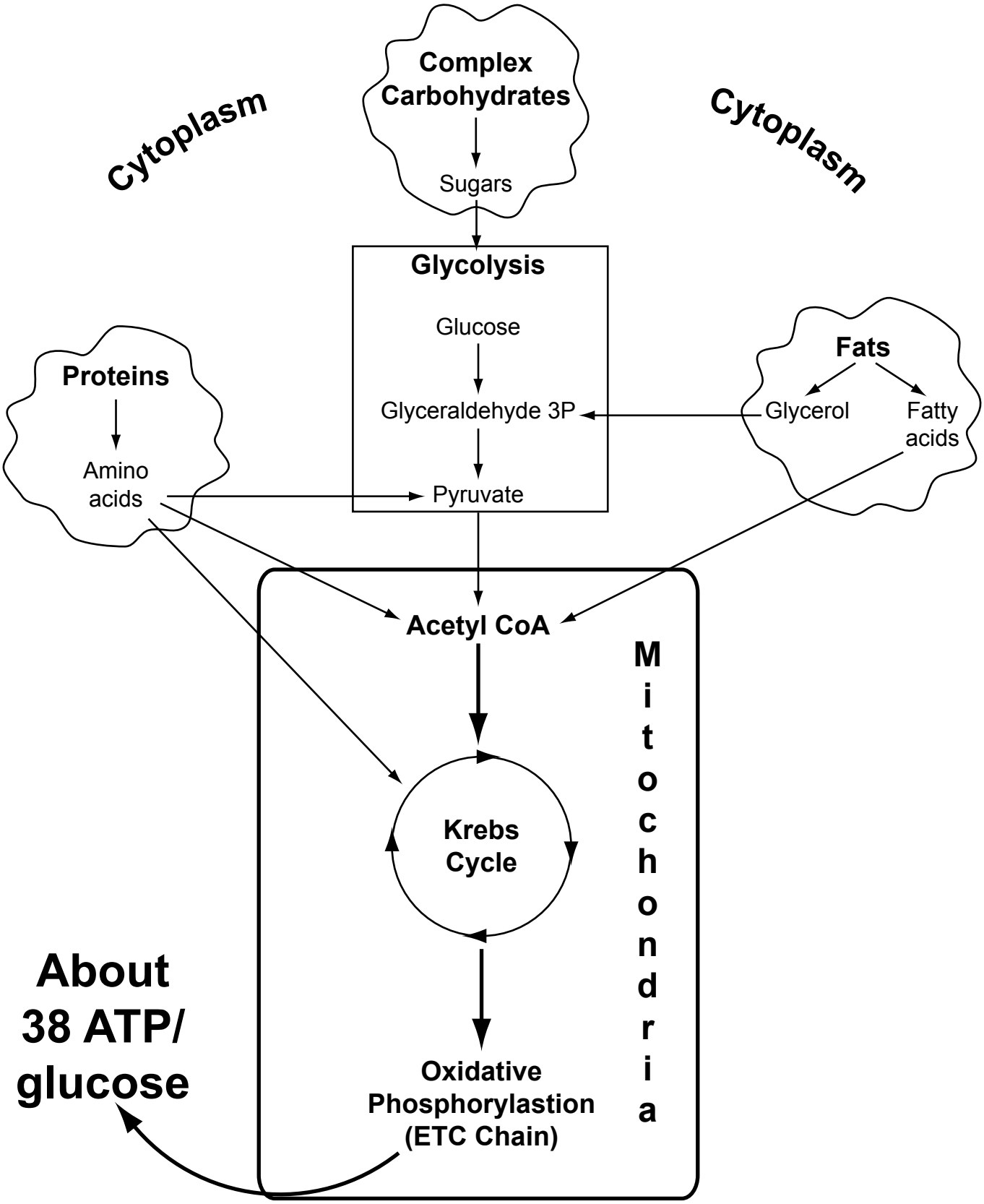
Mitochondria

- Convert "biological" nutrients into ATP
- Use nutrient molecules as substrates
    - Carbohydrates, fats, proteins
- Nutrients broken down into molecules that can be oxidized
    - Breakdown Products
        - Simple sugars, fatty acids, amino acids
- Simple molecules feed into the Krebs (or citric acid) cycle via intermediates
- Energy from nutrients stored in intermediates
- Intermediates feed oxidative phosphorylation chain
    - ATP is produced
    - ATP used to support life

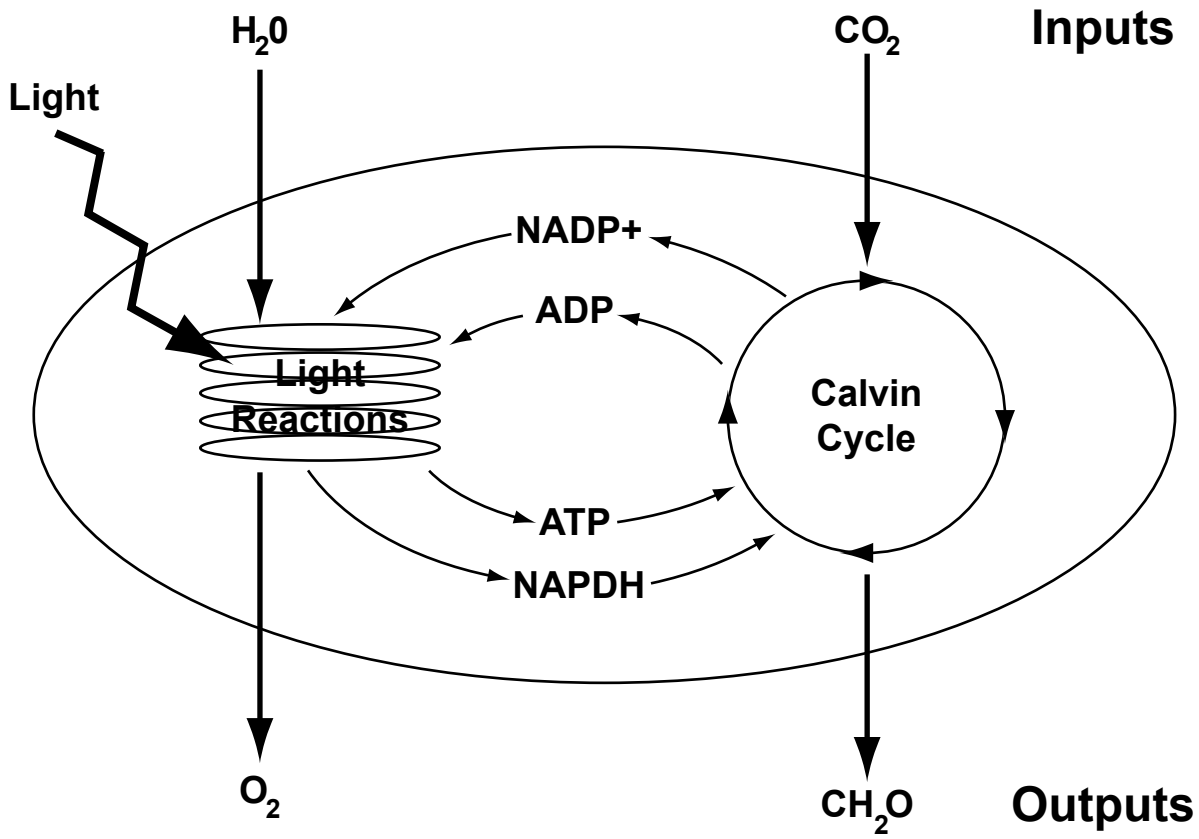
Chloroplast

- Converts electromagnetic energy into ATP and NADPH
    - Uses
        - H_20 as an electron source
        - Light energy to excite electrons
    - NADPH produced as by-product of electron flow
    - ATP synthesis coupled to H+ flow down a gradient
    - ATP and NADPH used for sugar synthesis
- Converts CO_2 into glucose
    - Uses ATP and NADPH as energy source

Flow of Nutrients Into Oxidative Phosphorylation



Inputs, Outputs and Energy Production in Chloroplasts



Where did these organelle genes come from?

Endosymbiotic Theory

- Bacterial cell was taken up by primative eukaryotic cell
  - Mitochondria
    - Relative of current alpha-proteobacteria
  - Chloroplast
    - Relative of current cyanobacteria
- Partioning of function
  - Some genes (and functions) retained by the organelle
  - Most genes (and functions) transferred to the nucleus

Organelle Genome Structure

Mitochondrial Genome Structural Features

- Almost exclusively circular molecules
- Exceptions
  - Linear molecules
    - *Plasmodium falciparum*
      - 6.0 kb
    - *Jacoba libera*
      - 100kb
  - Mini-circles
    - Mesozoan animal (*Dicyema misakiense*)
      - < 2000 bp in size
      - Each circle contains one or a few genes

Organization of Plant mt Genomes

- Master and subgenomic circles
- Repeat sequences on master circle
    - Mediate intrachromosomal recombination
    - Creates subgenomic circles
    - Examples
        - Brassica
            - Master circle
                - 218 kb
            - Subgenomic circles
                - 183 kb + 35 kb
        - Corn
            - Master circle
                - 570 kb
            - Subgenomic circles
                - 488 kb + 82 kb
                - OR
            - 503 kb + 67 kb
            - **Why????**
                - Multiple repeats
                - Generate different sub-genomic circles

mt DNA size range of sequenced genomes

- Protists
  - Very diverse biological group
    - 6-100 kb
- Animals
  - Mouse
    - 16.2 kb
  - Human
    - 16.6 kb
- Fungi
  - Yeast
    - 31 kb
  - Canidia
    - 40 kb
  - Podospora
    - 100kb
- Plants
  - Alagae
    - 56-67 kb
  - Liverwort (bryophyte)
    - 187 kb
  - Arabidopsis
    - 366 kb
  - Rice
    - 490 kb
  - Range by fragment analysis
    - 200- 2400 kb

What sequences and genes are retained in the organelles?

Genes and Sequences in Arabidopsis mtDNA (367 kb)

- Unknown function (49.0%)
- Protein coding genes (16.1%)
- Orf >100 aa (10.2%)
- Intron sequenes (8.8%)
- Nuclear origin (4.0%)
- Other repeats (3.9%)
- Recombination repeats (2.9%)
- Nuclear/mito homologs (2.2%)
- rRNAs (1.3%)
- Plastid origin (1.2%)
- tRNAs (0.4%)

Mitochondria DNA genes (general picture)

- Respiratory complex genes
  - *nad* genes (NADH oxidoreductase)
  - *sdh* genes (Succinate:ubiquinol oxidoreductase)
  - *cob* (Ubiquinol:cytochrome c oxidoreductase)
  - *cox* (Cytochrome c oxidase)
  - *atp* (ATP synthase)
- Translation
  - Ribsome
  - Large subunit
  - Small subunit
  - 5S rRNA (rarely)
  - tRNA (full or nearly full complement)
  - some ribosomal proteins (plants and protests only)
- Variation among genomes for gene content exists

Table . Genes in mitochondrial genomes.

	Species					
Genes	***Arabidopsis***	***Marchantia***	***Protheca***	***Chondrus***	***Podospora***	***H. sapiens***
Complex I						
nad1	X	X	X	X	X	X
nad2	X	X	X	X	X	X
nad3	X	X	X	X	X	X
nad4	X	X	X	X	X	X
nad4L	X	X	X	X	X	X
nad5	X	X	X	X	X	X
nad6	X	X	X	X	X	X
nad7	X(ψ)	X	X			
nad9	X		X			
Complex II						
sdh2				X		
sdh3		X		X		
sdh4	X	X		X		
Complex III						
cob	X	X	X	X	X	X
Complex IV						
cox2	X	X	X	X	X	X
cox2	X	X	X	X	X	X
cox3	X	X	X	X	X	X
Complex V						
atp1	X	X	X			
atp6	X	X	X	X	X	X
atp8	X	X	X	X	X	X
atp9	X	X	X	X		
Cyto c Biogenesis						
ccb206	X					
ccb256	X					
ccb453	X					
ccb 382	X	X				
ccb203	X	X				
Protein transport						
mttB	X	X	X	X		
Other ORFs						
orf25	X	X	X	X		

Table (cont). Genes in mitochondrial genomes.

	Species					
Genes	***Arabidopsis***	***Marchantia***	***Protheca***	***Chondrus***	***Podospora***	***H. sapiens***
Ribosomal RNAs						
5S	X	X	X			
srrn	X	X	X	X	X	X
lrrn	X	X	X	X	X	X
Ribosomal proteins						
rps1		X				
rps2		X	X			
rsp3	X	X	X	X		
rps4	X	X	X			
rps7	X	X	X			
rps8		X				
rps10		X	X			
rps11		X	X	X		
rps12	X	X	X	X		
rps13		X	X			
rps14	X(ψ)	X	X			
rps19		X	X			
rpl2	X	X				
rpl5	X	X	X			
rpl6		X	X			
rpl16	X	X		X		

Species: Arabidopsis (*Arabidopsis thaliana*, land plant); Marchantia (*Marchantia polymorpha*, liverwort); Prototheca (*Prototheca wickerhami*i, green algae); Condrus (*Chondrus crispus*, red algae); Podospora (*Podospora anserina*, fungus); H. sapiens (*Homo sapiens,* man)

From: Trends in Plant Science (1999) 4: 495.

ψ = psuedogene

Chloroplast Genome Structural Features

- Circular molecules
  - 130 – 170 kb
    - Why the difference??
    - Repeat region is lost in pea lineage
- Cytogenomic analysis (The Plant Cell (2001) 13:245)
  - Studied Arabidopsis, tobacco, and pea
    - Both complete linear and circular genomes observed
      - Contained one to four copies of the genome
    - Monomers the most prevalent form
    - Rearranged circles discovered
      - They contained incomplete genomes
    - ***Plasticity a feature of chloroplast genomes***
- Condensed genomes exists
  - Squawroot
    - 45 kb
  - Beechdrops
    - 71 kb
- Large version exist
  - Geranium
    - 217 kb
- Gene Order
  - Highly conserved

Chloroplast genes

- Photosynthesis complex genes
  - Photosystem II (some)
  - Photosystem I (many)
  - Cytochrome f
  - Cytochrome b6-f
  - NADH dehydrogenase (most)
  - ATP synthase
- Translation
  - Ribosome
  - 23S rDNA (large subunit)
  - 16S rDNA (small subunit)
  - 5S rDNA
  - 4.5S rDNA
  - Ribosomal proteins (many)
  - tRNAs
- Dark reactions
  - Large subunit of RUBISCO
- Arabidopsis
  - 87 genes
    - Proteases
    - Transcription
      - RNA polymerase
      - Transcriptional regulator (ycf2)

Other species

- Wheat
  - 99 genes
- Tobacco
  - 127 genes (most of algae or land plants)

Evolution of Organelle Genomes

Some information extracted from (Science 283:1476)

Genomics allows us to determine the evolution pattern

- Sequenced organelle genomes are becoming abundant
  - Mitochondria (9/22/03)
    - 629 genomes
  - Chloroplast (9/22/03)
    - 30 genomes

What approach is taken to determine evolutionary patterns?

- Comparative genomics
  - Information gained with one organism can be applied to other speices
    - Look for patterns and similarities
    - Deduce evolutionary events from the patterns

Comparative mt genomics

- Compare
  - Closest sequenced ancestor
  - Least derived sequenced mt genome
- Mitochondria
  - Closest ancestor
    - *Rickettsia prowazeki*i
    - 834 genes
  - Least derived
    - *Reclinomonas americana*
      - 97 genes
      - 18 unique to this species
      - RNA polymerase subunits
      - function not found in any other mt genome
  - Most derived
    - Animals

What do we know about the evolution of mt genomes?

- Two types of mt genomes
  - Ancestral
  - Derived

Features of ancestral genomes

- Extra genes compared to current animal mt gene set
  - *sdh* or ribosomal proteins, for example
- gene-rich genomes with little extra non-coding or intron sequences
- complete (or nearly complete) set of tRNA genes
- rRNA genes similar to those in eubacteria
- gene clusters similar to those in eubacteria
- standard genetic code

Features of derived genomes

- Greatly reduced gene set
- Reduced and/or non-standard tRNA gene set
- Diverged rRNA genes
- Greater divergence from eubacterial gene sequence than within ancestral genomes
- Introduction of intron and non-coding sequences

Derived and ancestral are fuzzy terms

- Some genomes
  - Have lost (or undergone mutation) of some genes typical of derived genomes
  - Other ancestral genome features are maintained
    - Example:
      - Plants
        - Content similar to the eubacterial-like protist genomes (ancestral feature)
        - But, recombinationally active repeats and introns were introduced (derived feature)
        - RNA editing features added (derived feature)

Single or Multiple Origin of Mitochondria

What evidence should be used?

- Gene order
  - But gene shuffling has occurred
- Single gene phylogenies
  - But this has only limited information
- Genome data
  - Multiple genes can be concatenated (considered as a large master gene)
  - Greatly increased amount of data for comparisons
  - Genes must be chosen carefully; must be vertically derived

Concatenated mt Protein Phylogeny

Uses four genes

- cytochrome b apoprotein
- cytochrome oxidase subunit 1
- cytochrome oxidase subunit 2
- cytochrome oxidase subunit 3
- Result
  - All current mitochondria genomes are related to a single proteobacteria ancestor
  - Said another way
    - All mitochondria genomes are monophyletic
      - ***Monophyletic*** = derived from a single ancestor

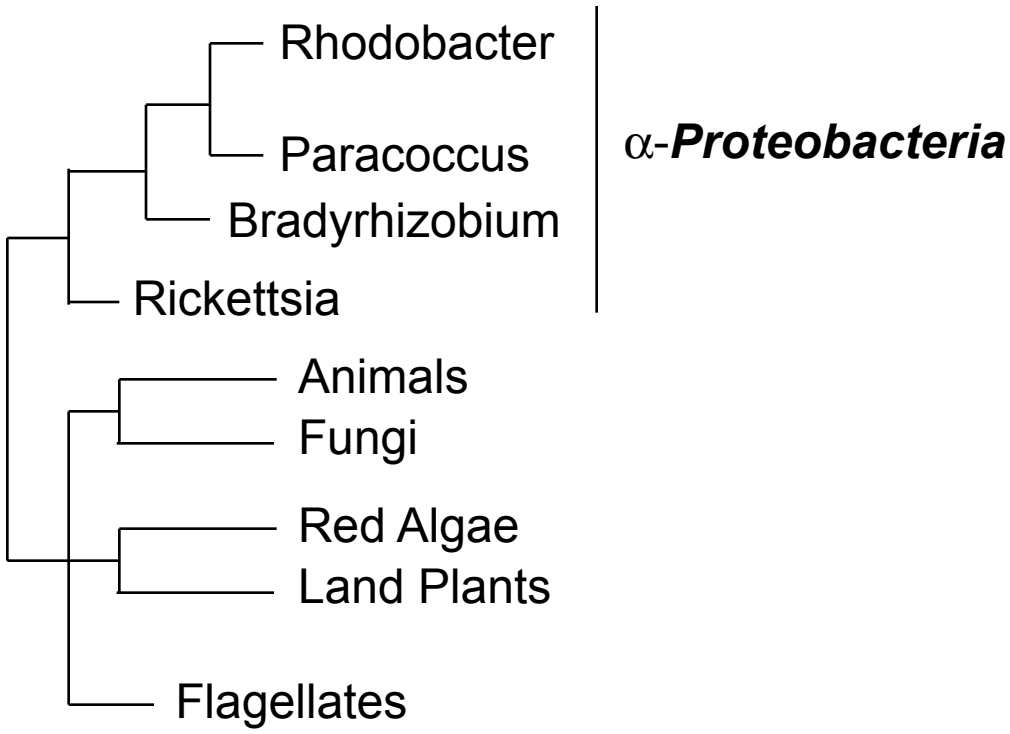
Supporting Evidence

- *Reclinomonas americana*
  - Genes are most like α–proteobacteria genes
  - All other mitochondria gene sets are a subset of the *R. americana* genes
    - Strong supporting evidence of a single, monophyletic origin

Concatenated Gene Phylogeny of Mitochondrial Genomes

A. The Concept

Concatenated phylogenies sum the sequence information over many genes to develop an understanding of the relationship between different organisms. This phylogeny of mtDNA genomes was designed to determine the relationship between several higher order biological classifications and the α-proteobacteria, the presumed ancestral species of the modern mitochondria genomes. This phylogeny was constructed using cytochrome b apoprotein protein sequence along with the cytochrome oxidase subunit 1, 2, and 3 protein sequences. This shows that



summarized from: Science (1999) 283:1476

Events After Endosymbiosis

Loss of genes from mitochondrial ancestor occurred

- How can we follow the events
    - Compare
        - Most mitochondrial-like α-proteobacteria
            - *Rickettsia prowaxekii* (834 protein-coding genes)
                - Parasitic bacteria
                - Reduced genome
                - Metabolic genes mostly lost
                - Metabolites provided by host
        - Least derived mitochondria genome
            - *Reclinomonas americana* (62 protein-coding genes)
                - What genes are retained?
                    - Partial set of respiratory genes maintained
                    - A few informational genes (translation) are maintained
                - All other functions are lost

How does this relate to the evolution of eukaryotic cells?

Evidence that needs to be included

- Nature of eukaryotic gene set
    - Contains genes homolgous with
        - Archaea genes
            - Information processes (replication, transcription, translation)
        - Eubacteria genes
            - Metabolic processes
- Amitochondriate eukaryotes exists
    - Some eukaryotes live anaerobically
        - Contain hydrogenosomes
            - Hydrogen producing "mitochondria" lacking a genome

Current Hypothesis

Two step process to develop mitochondria-containing eukaryotic cells

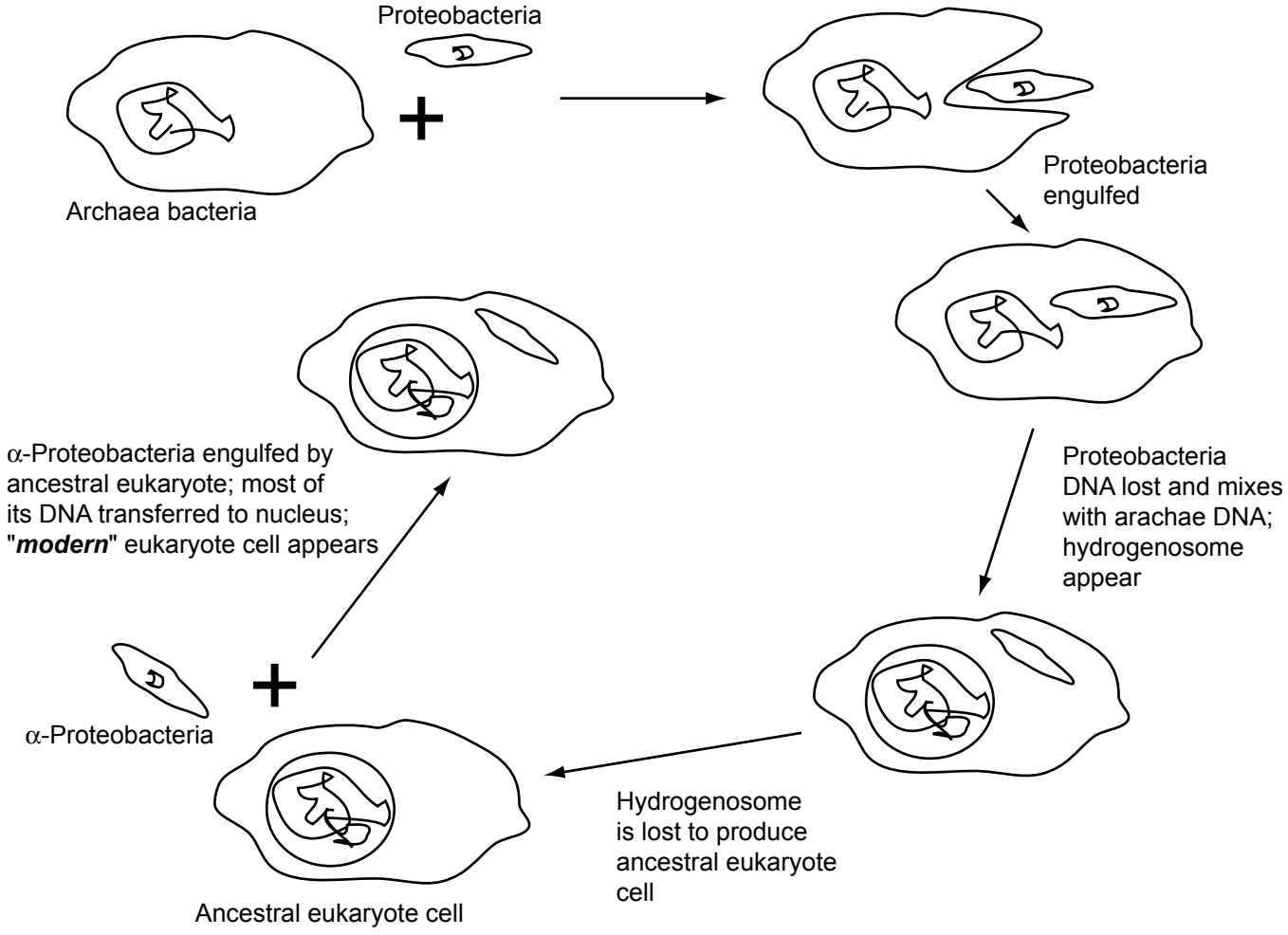
- **Step 1**
  - ***Hydrogen hypothesis***
    - Origin of original "eukaryotic-like" cell
    - Archaea bacteria and eubacteria merge
      - Archaea was a hydrogen requiring host cell
      - Eubacteria was a hydrogen producing symbiont
    - Hydrogenosomes appear
    - Eubacteria is fully engulfed
      - DNA is transferred to the archaea genome
      - This state is currently seen in amitochondriate protests
- **Step 2**
  - Hydrogenosome is lost
  - An α–proteobacteria is taken up
    - Provides aerobic mitochondrial functions
  - Endosymbiont loses many genes to the nucleus

Hydrogen Hypothesis of Evolution of Early Eukaryotic Cells

A. The Concept

The origins of eukaryotic cells is important to understanding the diversity of life on earth. A current proposal, called the ***Hydrogen Hypothesis***, suggests that an ancestral archaea-type bacteria was hydrogen requiring. To fulfill this requirement, it engulfed an ancestral protobacteria that produced hydrogen (thus the Hydrogen Hypothesis). Over time, all of the DNA from protobacteria was transferred to the archaea host, and along with the host DNA, a nucleus was formed. The proto-bacteria was converted into a hydrogenosome, a feature found in some current anaerobic protists.

Others have extended this hypothesis to describe events associated with the appearance of the "modern" eukaryotic-type cell. In their scenario, the hydro-genosome is lost, and an α-proteobacteria is taken up. Overtime, much of its DNA is transferred to the nucleus, and the bacteria is converted into the modern mitochondrial



The Modern Chloroplast-containing Eukaryote Story

Sources:
Trends in Plant Science (2000) 5:174
Current Opinion in Genetics and Development (1998) 8:655

Multiple endosymbiotic events
First event

- An early mitochondria-containing eukaryote engulfs a cyanobacteria
    - Eukaryote was probably a biciliate protozoan
    - Cyanobacteria contained
        - Phycobilins
        - Chlorophylls a and b
- Three lineages evolved from the original "photosynthetic" eukaryote
    - Rhodophhyte
        - Red algae lineage
        - Chlorophyll b lost
    - Chlorophyte/metaphyte
        - Green algae/land plant lineage
        - Phycobilins lost
        - Thylakoids appear
    - Glaucocystophyes
        - This lineage did not evolve
        - Chlorophyll b lost

Second events

- Green algae and red algae independently taken up by phagotrophs
    - These evolved independently
    - Modern plastid containing species appear

Where Did All the Endosymbiont DNA Go??

- Mitochondria
  - *Rickettsia prowaxekii*
    - 834 protein-coding genes
  - *Reclinomonas americna*
    - 62 protein-coding genes
- Chloroplast
  - *Synechocystis*
    - Cyanobacteria
      - About 3000 genes
  - Modern chloroplasts
    - Up to 200 genes
- Answer
  - Transferred to the nucleus

How Many Endosymbiont Genes Were Transferred??

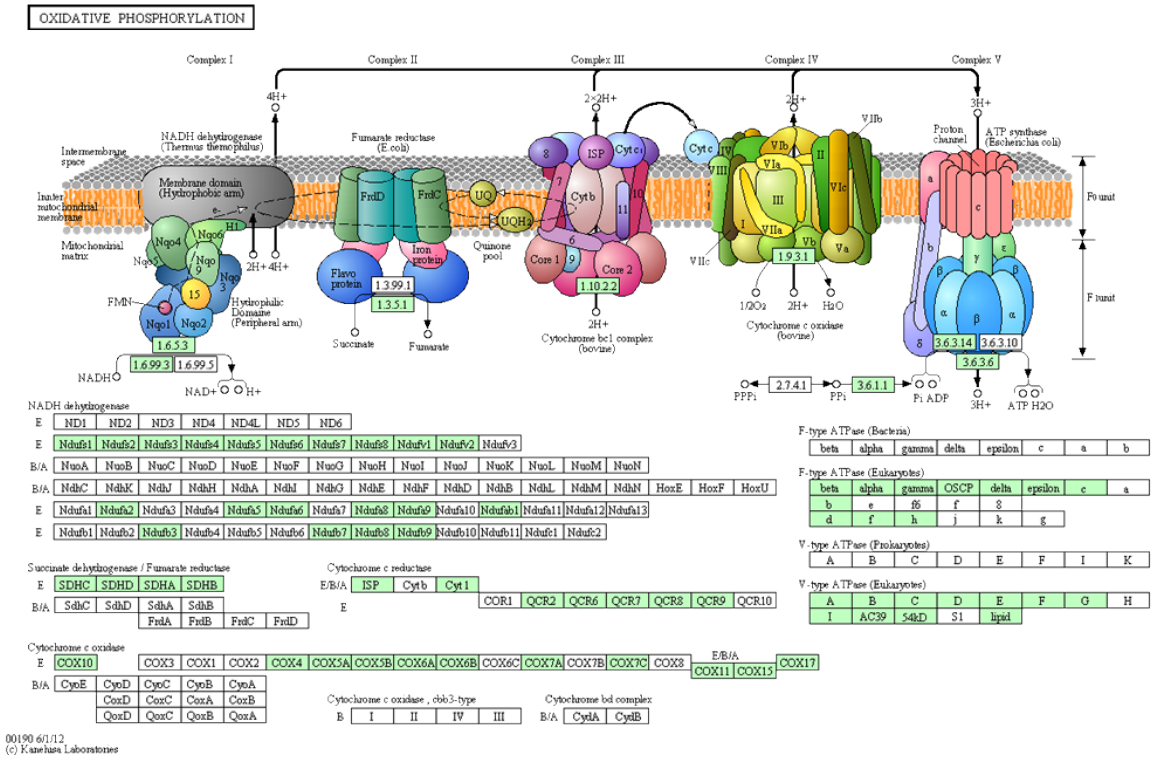
- Arabidopsis
  - PNAS (2002) 99:12246
  - What is the origin of its 25,000 genes?
    - 9400 genes are homologous to yeast or bacterial species
    - Most of these genes are like yeast
      - These are probably genes in host of the original endosymbiotic event that formed chloroplast-containing eukaryote
      - Why??
        - That cell is the last common ancestor of yeast and plant lineages
    - 1700 genes most like cyanobacteria
      - 866 only homologus to cyanobacteria
        - 18% of homologous gene set
      - Result of transfer events to the nucleus

Table . Functions of cyanobacterial genes found in *Arabidopsis* genome.

Functional category	*Number of genes*
Biosynthesis and metabolism	562
Signal transduction	189
Cellular response	137
Energy generation	93
Cell organization	71
Protein synthesis	68
Protein destination	63
Transcription	54
Biogenesis	38
Transport facilitators	35
Cell growth and division	31
Intracellular transport	12
Homeostasis	5
Classification not-clear	39
Unclassified	303

Electron Transport in the Mitochondria

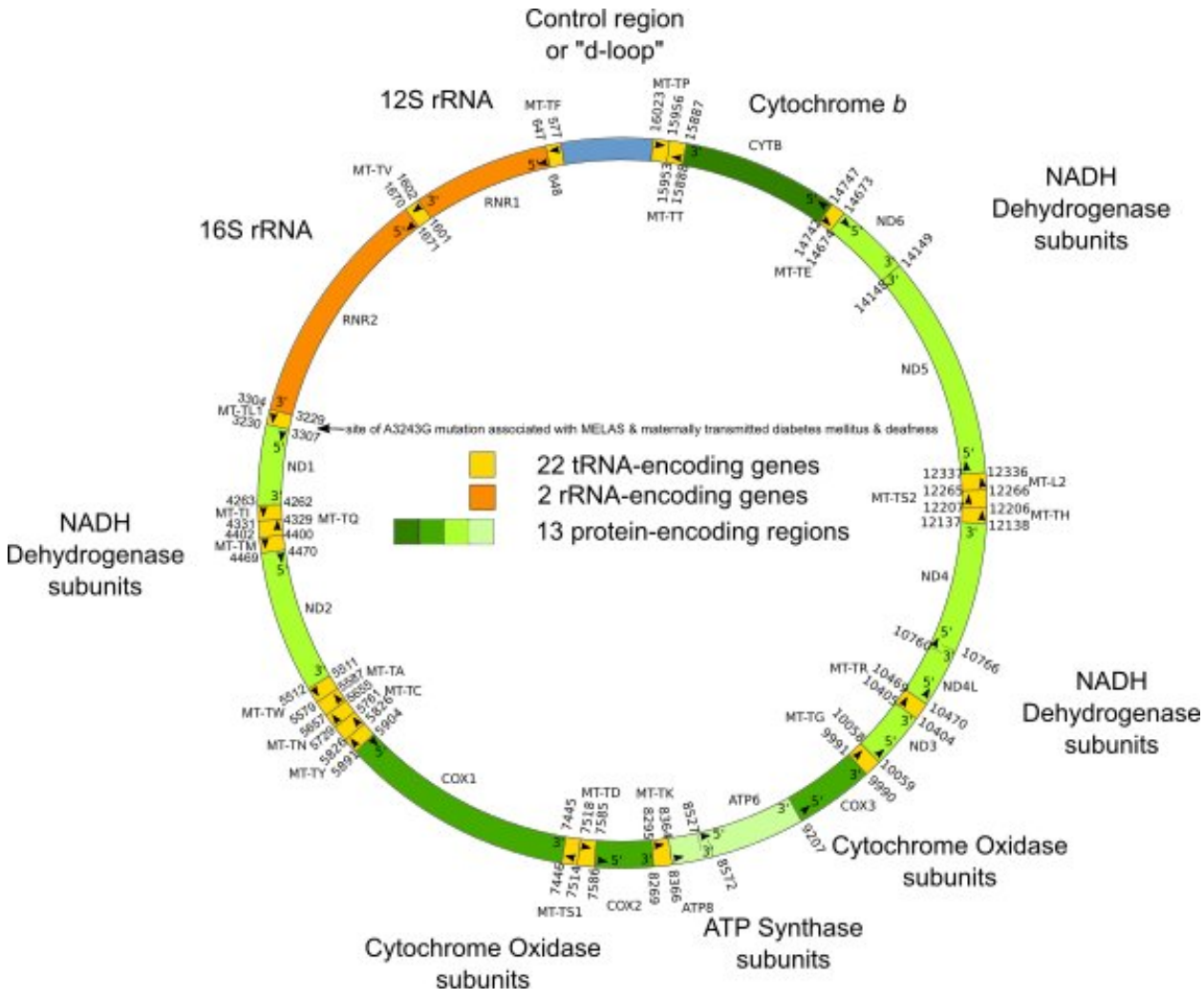
http://mitofun.biol.uoa.gr/images/oxidative_phosphorylation.png



Mitochondrial function

- Multiple protein complexes
  - Most proteins encoded by nuclear genes
  - Some encoded by mitochondrial genes

Human Mitochondrial Genome



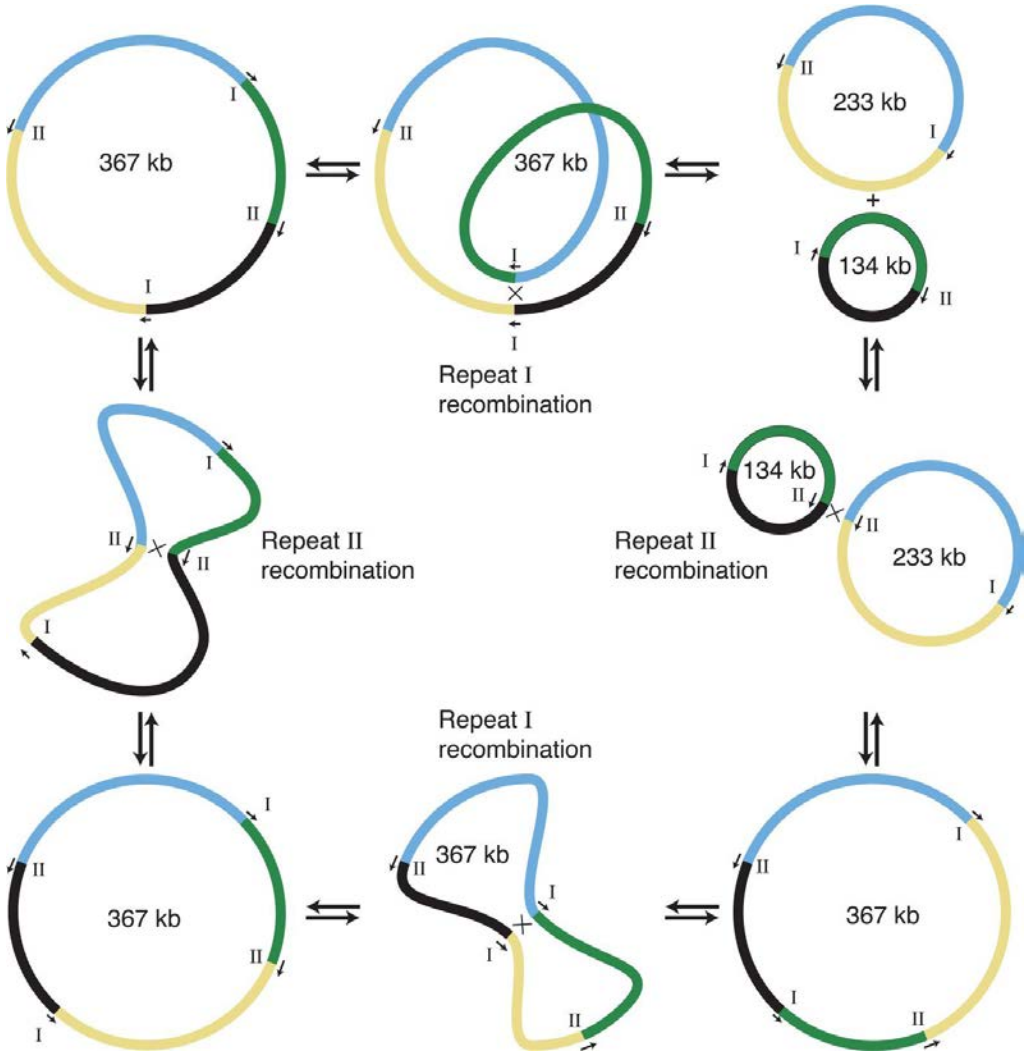
Human mitochondrial genome

- Small in size
  - 18kb
- Limited function
  - 13 protein encoding genes
  - Genes related to electron transport activity

Plant Mitochondrial Genomes

Structurally Plastic

http://6e.plantphys.net/topic12.06.html

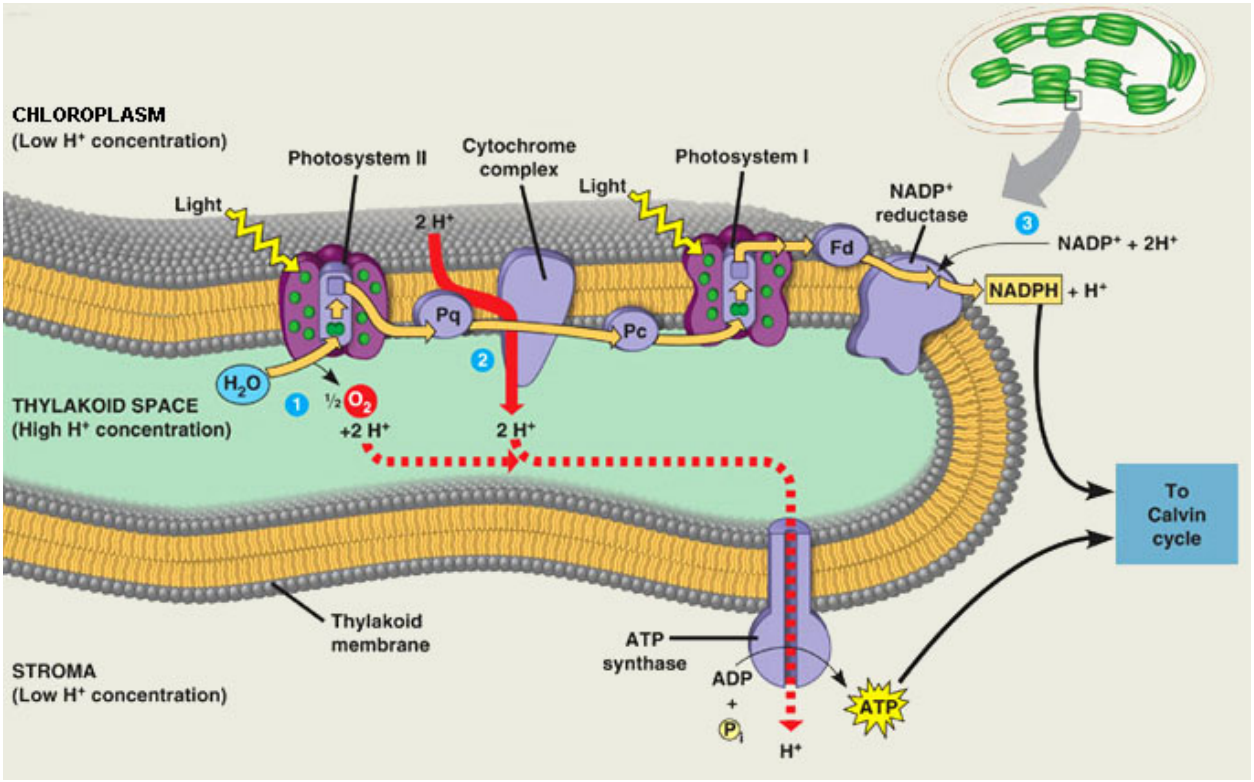


Arabidopsis genome

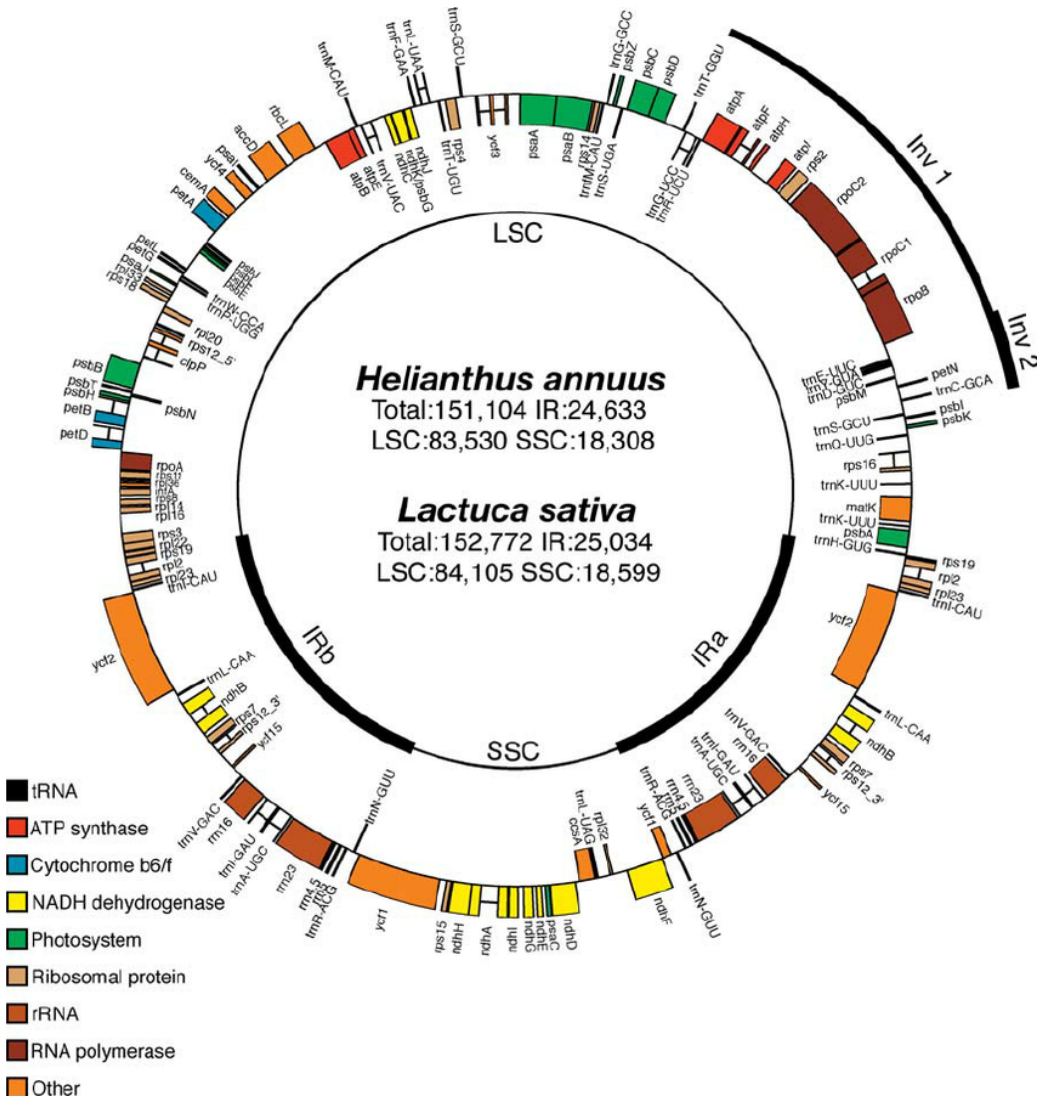
- Contains direct repeats
- Pairing of repeats can lead to structural changes through recombination
    - 367kb circular genome → 233kb + 134kb smaller circles
- Other plant mitochondrial genomes have more complex repeat/recombination patterns

Chloroplast Electron Transport

http://vle.du.ac.in/mod/book/print.php?id=10131&chapterid=16833



Chloroplast Genomes



Chloroplast genome genes

- Partial set of genes involved in photosynthesis
  - Light and dark reactions
  - Other genes encoded by nuclear genomes