Genomes
DSE Bioinformatics , sem V

General Variables of Genomes

- Prokaryote versus Eukaryote versus Organelle
- Genome size:
  - Number of chromosomes
  - Number of base pairs
  - Number of genes
- GC/AT relative content
- Repeat content
- Genome duplications and polyploidy
- Gene content

Eukaryotic genes in the genomes

- introns
- Retro- transposons (Ty1 elemnets of yeast)
- Conseuqneces of RNA activity of genomes
  - Exon shuffling
  - Formation of retrogenes
  - Exonization of mobile elements

- **Genetic acquisition by transfer**
- Intracellular, Interorganellar transfer (NUMT, NUPT)
- Horizontal gene transfer
- Introgression of large chromosomal segments

Elements of eukaryotic genomes (nuclear)

Chromosomes: linear, centromeres, telomeres, origins of replication, replicons

Protein-coding genes and **spliceosomal introns**

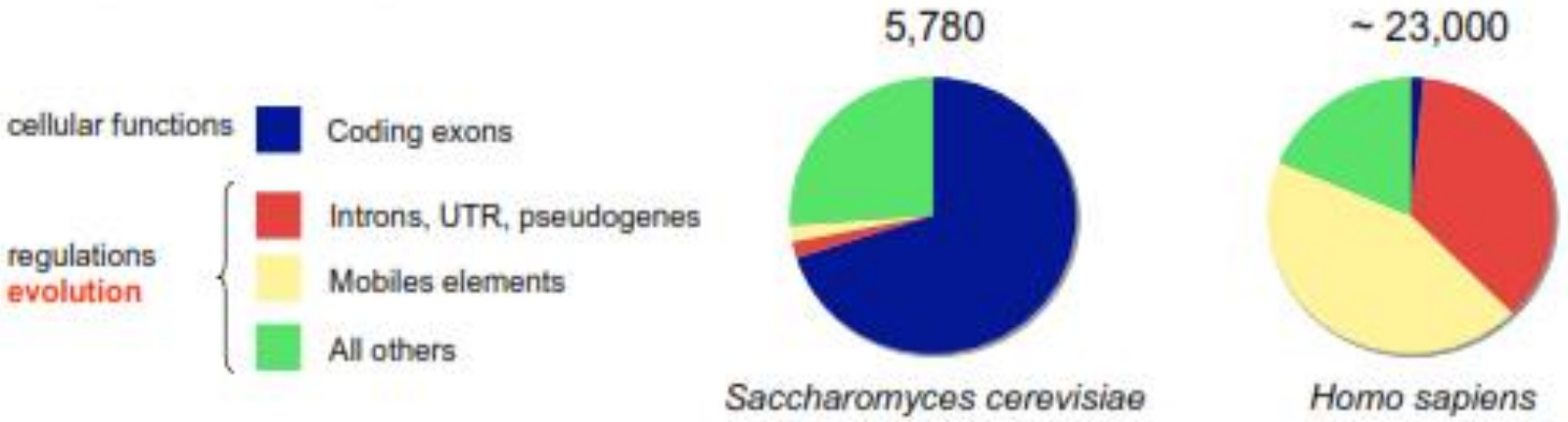
Genes for non coding RNAs: rRNAs, tRNAs, snoRNAs, snRNAs, microRNAs ….

Mobile genetic elements: and their remnants

Pseudogenes: and processed pseudogenes

Satellite DNAs: micro-, minisatellites, repeated sequences

Fragments of organellar DNAs: NUMTs and NUPTs



Duplicated genes and regions, acquired DNA sequences, newly-created genes : > intense dynamics of genome modification and evolution

Genome projects

- Complete genomes:
  - Eubacteria (*Escherichia coli, Mycoplasma genitalium)*
  - Archaea (*Methanococcus jannaschii)*
  - *Viral genomes*
  - Eukarya:
    - *Saccharomyces cerevisiae*
    - *Arabidopsis thaliana*
    - *C elegans*
    - *D melanogaster*
    - *Homo sapiens*

- **Prokaryotic genome**

(chromosome + plasmids)

Eukaryotic genomes

(Nuclear Chromosomes + Mitochondrial genome + Chloroplast genome)

Eubacterial genomes: *E. coli*

- 4288 protein coding genes:
  - Average ORF 317 amino acids
  - Very compact: average distance between genes 118bp
- Numerous paralogous gene families: 38 – 45% of genes arisen through duplication
- Homologues:
  - *H. influenzae* (1130 of 1703)
  - *Synechocystis* (675 of 3168)
  - *M. jannaschii* (231 of 1738)
  - *S. cerevisiae* (254 of 5885)

Publication:

The Complete Genome Sequence of Escherichia coli K-12

Frederick R. Blattner,* Guy Plunkett III,* Craig A. Bloch, Nicole T. Perna, Valerie Burland,
Monica Riley, Julio Collado-Vides, Jeremy D. Glasner, Christopher K. Rode, George F. Mayhew,
Jason Gregor, Nelson Wayne Davis, Heather A. Kirkpatrick, Michael A. Goeden, Debra J. Rose,
Bob Mau, Ying Shao

Science, Vol 277, 1453-1462

Update on *E. coli* genome ; publication

The goal of this group project has been to coordinate and bring up-to-date information on all genes of *Escherichia coli* K-12.

Nucleic Acids Res. 2006; 34(1): 1–9.
Published online 2006 Jan 5. doi: 10.1093/nar/gkj405
PMCID: PMC1325200

Escherichia coli K-12: a cooperatively developed annotation snapshot—2005

Monica Riley,[*] Takashi Abe,[1] Martha B. Arnaud,[2] Mary K.B. Berlyn,[3] Frederick R. Blattner,[4] Roy R. Chaudhuri,[5] Jeremy D. Glasner,[4] Takashi Horiuchi,[6] Ingrid M. Keseler,[7] Takehide Kosuge,[1] Hirotada Mori,[8,9] Nicole T. Perna,[4] Guy Plunkett, III,[4] Kenneth E. Rudd,[10] Margrethe H. Serres, Gavin H. Thomas,[11]

E.coli

- Important in biosphere
- Facultative anaerobe
- Colonizes the lower gut of animals
- Widely disseminated in the new host after release in the natural environment
- Pathogenic strains cause enteric, urinary, pulmonary and nervous system infections

Escherichia coli strain K-12

- *Escherichia coli* strain K-12 is arguably the single organism about which the most is known. So it was chosen for whole-genome sequencing.
- The ancestral strain had been isolated from the stool of a convalescent diphtheria patient in 1922 and given the designation ‘K-12' when deposited in a strain collection at Stanford in 1925.
- Originally isolated in 1922, it was catapulted to prominence by the discovery of strain K-12's ability to carry out genetic recombination by conjugation & soon after, by generalized transduction.
- The strain K-12 has been widely distributed to laboratories across the world.
- It became the primary model organism for basic biology, molecular genetics and physiology of bacteria, and was the

Why MG1655 ??

- Maintained as a laboratory strain with minimum genetic manipulations
- Cured of temperate bacteriophage lambda by UV light
- Cured of F plasmid by acridine orange
- All K-12 derivatives including MG1655 have IS5 insertion in rfb-50 resulting in absence of O-antigen synthesis in lipopolysaccharide

genome revealed:

- 4284 protein coding genes (before publication of this genome sequence 1853 proteins were already known)

Which was 4464 in 2006 (2nd release), The increase is due in large part to identifying small proteins and small RNAs

- 122 structural RNA genes
- Noncoding repeat sequences
- Regulatory elements
- Transcription /translation guides
- Transposases
- Prophage remnants
- Insertion sequence elements
- Patches of unusual composition, likely to be foreign elements introduced by horizontal gene transfer
- Comparison with six other completed genomes representing three major kingdoms

Protein repertoire

- Largest class of proteins is the enzymes accounting to 30% of total genes
- Many enzymatic functions are shared by more than one protein. Some of these sets of functionally similar enzymes are very closely related and appear to have arisen by duplication, either in *E. coli* or in an ancestor or donor species
- Other sets of functionally similar enzymes have very dissimilar sequences and differ in specificity, regulation or intracellular location.
- Several features of *E. coli's* array of enzymes give it a versatile metabolic competence, which allow it to grow and compete under varying conditions

Table 4. Distribution of *E. coli* proteins among 22 functional groups (simplified schema).

Functional class	Number	Percent of total
Regulatory function	45	1.05
Putative regulatory proteins	133	3.10
Cell structure	182	4.24
Putative membrane proteins	13	0.30
Putative structural proteins	42	0.98
Phage, transposons, plasmids	87	2.03
Transport and binding proteins	281	6.55
Putative transport proteins	146	3.40
Energy metabolism	243	5.67
DNA replication, recombination, modification, and repair	115	2.68
Transcription, RNA synthesis, metabolism, and modification	55	1.28
Translation, posttranslational protein modification	182	4.24
Cell processes (including adaptation, protection)	188	4.38
Biosynthesis of cofactors, prosthetic groups, and carriers	103	2.40
Putative chaperones	9	0.21
Nucleotide biosynthesis and metabolism	58	1.35
Amino acid biosynthesis and metabolism	131	3.06
Fatty acid and phospholipid metabolism	48	1.12
Carbon compound catabolism	130	3.03
Central intermediary metabolism	188	4.38
Putative enzymes	251	5.85
Other known genes (gene product or phenotype known)	26	0.61
Hypothetical, unclassified, unknown	1632	38.06
Total	4288	100.00*

*Total of these rounded values is 99.97%.

E. coli K12

Orange squares -forward genes

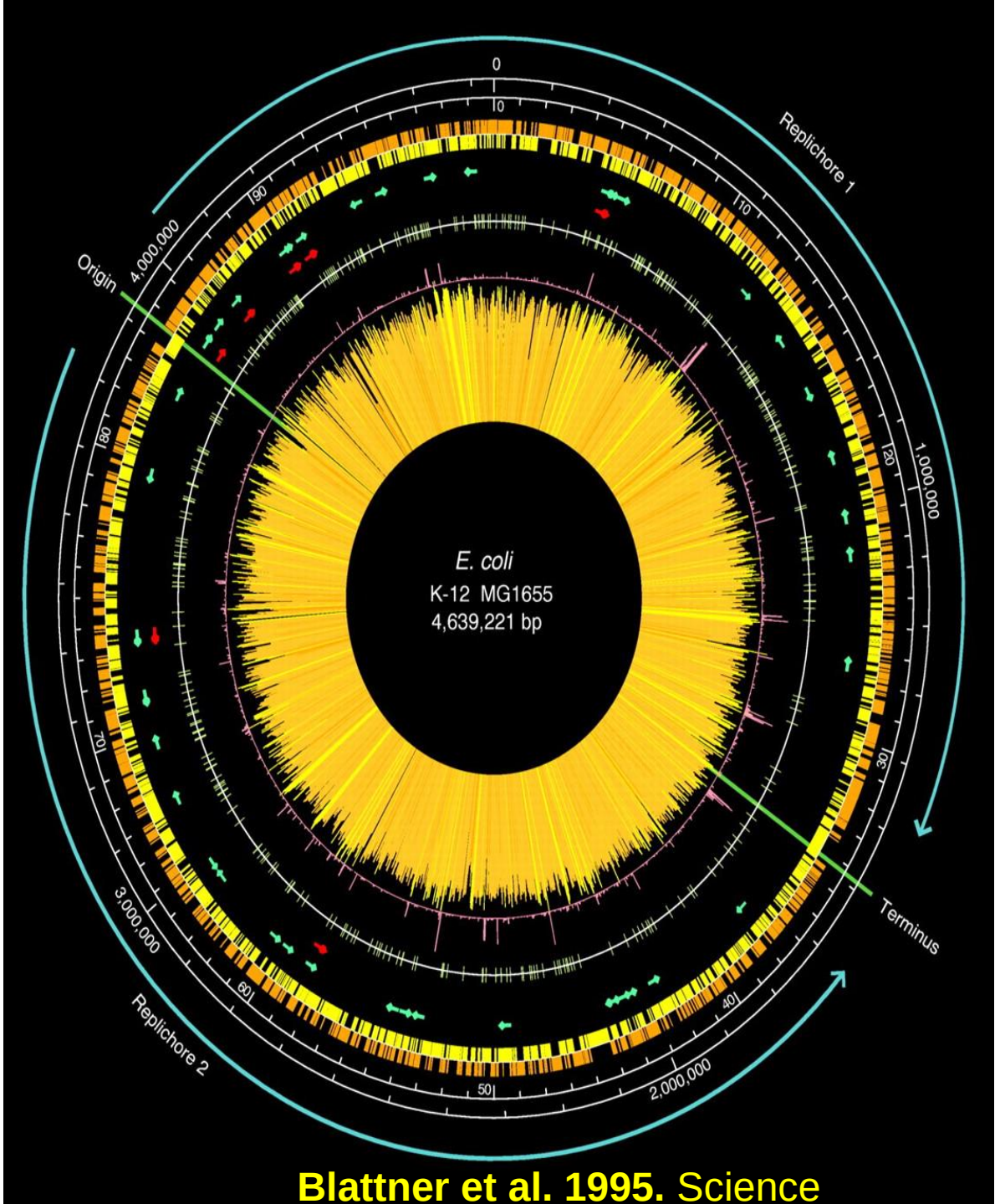
Yellow squares -complement genes

Red arrows -rRNA

Green arrows -tRNA

White ring -REP sequences

Blue ring - bacteriophage protein similarity



Blattner et al. 1995. Science

Conclusions:

- Took 6 years to complete sequencing
- Determine precise function of all genes by!!!!
  - global transcription analysis
  - Phenotypic analysis of mutants
  - Analysis of biochemical and catalytic properties of the expressed proteins
  - comparative genomics with other microbial genomes identify evolutionary relationships
  - Comparative genomics with genomes of pathogens to identify genes that confer unique detrimental or beneficial properties

These new studies available at E coli genome -Wisconsin web page

E. coli genome with sequence features