ICB Fall 2009

G4120: Introduction to Computational Biology

Oliver Jovanovic, Ph.D.
Columbia University
Department of
Microbiology & Immunology

Lecture 9
Genomics
November 18, 2009

Genome

"Chromosomes from different sets (or **genoms**) of *Triticum vulgare* show affinity toward each other."
– *Cytologia* I. 14, 1930

"The inviability of deficient **genomes** in the haploid generation serves to some extent as an alternative distinction between mutation and deficiency."
– *Proc. 6th Int. Congr. Genetics* I. 275, 1932

"There are two species having **genoms** resembling *C. neglecta."*
– *Proc. 6th Int. Congr. Genetics* II. 5, 1932

"The appearance of such terms as gene-complex and **genome** (denoting a set of chromosomes as a working unity) testify to the movement towards holism in genetics."
– C. P. Blacker *Eugenics* x. 243, 1952

"Among organisms with chromosomes, each species has a characteristic set of genes, or **genome**. In diploids a **genome** is found in each normal gamete. It consists of a full set of the different kinds of chromosomes."
– A. M. Srb et al. *Gen. Genetics* (ed. 2) vii. 190, 1965

"The human **genome**...consists of perhaps as many as 10 million genes."
– *Scientific American*. Oct. 19, 1970

Genomics

Genetic Maps
Restriction Fragment Length Polymorphisms (RFLPs), Variable Number of Tandem Repeat Polymorphisms (VNTRs), Microsatellite Polymorphisms, Single Nucleotide Polymorphisms (SNPs), Linkage Analysis

Physical Maps
Chromosomal Maps, Radiation Hybrid Maps, Expressed Sequence Tags (ESTs), Simple Sequence Length Polymorphisms (SSLPs), Random Sequence Maps

Genome Sequences
Sequencing, Assembly, Gene Prediction, Annotation

Functional Genomics
Microarrays, Genomic Transcription Analysis, Proteomics, Disease, Pharmacogenomics

Comparative Genomics
Phylogenomics, Paleogenomics, Metabolic Reconstruction

Genomics in the Future
Direct Sequencing, DNA Computing, Nanotechnology

A Short History of Genomics

1977	øX174 genome (5,386 bp) sequenced. First complete viral genome.
1995	*Haemophilus influenzae* genome (1.8 Mbp) and *Mycoplasma genitalium* genome (0.58 Mbp) sequenced and assembled using whole genome shotgun sequencing by The Institute for Genomic Research (TIGR). First complete microbial genomes.
1996	*Saccharomyces cerevisiae* genome (12.1 Mbp) sequenced (first complete eukaryotic).
1997	*Escherichia coli* genome (4.7 Mbp) published.
1998	*Caenorhabditis elegans* genome (100 Mbp) is published (first complete multicellular).
1999	*Deinococcus radiodurans* genome (2.6 Mbp) sequenced.
2000	*Pseudomonas aeruginosa* genome (6.3 Mbp) published. *Arabidopsis thaliana* genome (100 Mbp) sequenced. *Drosophila melanogaster* genome (180 Mbp) sequenced by Celera, Inc.
2001	Human genome (3.4 Gbp) published separately by Celera, Inc. and the Human Genome Project. It appears to have approximately 25,000 genes (not 10 million).
2002	Mouse genome (3.4 Gbp) published.
2004	*Legionella pneumophila* genome (3.4 Mbp) published.
2005	*Pan troglodytes* genome (3.6 Gbp) published.
2006	*Apis mellifera* genome (234 Mbp) published.
2007	Rhesus genome (3.5 Gbp) published.

Genomes Sequenced

Year	Genomes	Total
1994	0	0
1995	2	2
1996	2	4
1997	8	12
1998	7	19
1999	6	25
2000	14	39
2001	40	79
2002	39	118
2003	46	174
2004	128	233
2005	115	348
2006	138	486
2007	208	694
2008	205	899
2009 (to date)	189+	1,088+

Large sequenced genomes (bacteria, archaea, and eukaryotes) in the Kyoto Encyclopedia of Genes and Genomes (KEGG). Depending on the species, however, the function of ***30% to 50%*** *of the predicted genes in these genomes remains unknown.*

Intraspecies Genomic Variation

Organism	Size	# of genes	Coding density	%G+C
E. coli K-12 isolate W3110	4,636,552 bp	4,085	1,135 bp/gene	51%
E. coli K-12 isolate MG1655	4,639,221 bp	4,397	1,055 bp/gene	51%
E. coli O157:H7 substrain EDL93	5,529,376 bp	5,283	1,047 bp/gene	51%
E. coli O157:H7 substrain RIMD 0509952	5,498,450 bp	5,361	1,026 bp/gene	51%

- In *E. coli,* the size of the genome varies from 4.6 to 5.5 Mbp.
- The number of genes varies from 4,085 to 5,361 and the coding density varies from 79% to 88%.
- The G+C content does not vary, however. All four strains of *E. coli* are 51% G+C.

Interspecies Genomic Variation

Prokaryotes

- Genome size varies from 0.58 Mbp (*Mycoplasma genitalium*) to 9 Mbp (*Nostoc punctiforme*). The bacterial endosymbiont *Carsonelaa ruddii* is only 159.7 Kbp.
- Number of predicted genes varies from 500 to 8,000, and is closely related to the size of the genome.

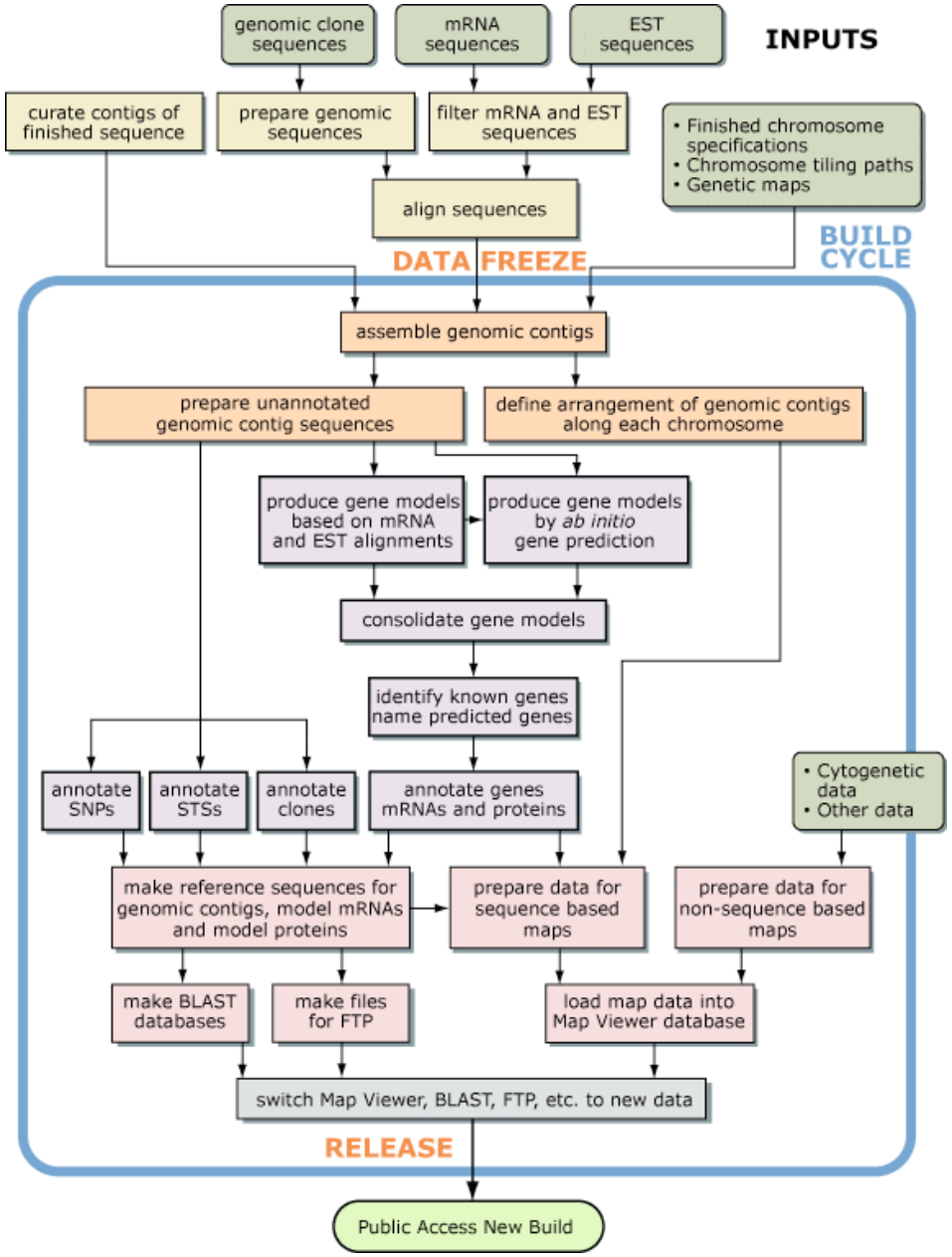
Eukaryotes

- Genome size varies from 12.6 Mbp (the green alga *Ostreococcus tauri*) to 3.2 Gbp (*Homo sapiens*) to 129 Gbp (the marbled lungfish *Protopterus aethiiopicus*) to 670 Gbp (*Amoeba dubia*) to 1.37 Tbp (the free-living amoeba *Chaos chaos*). The cryptomonad nucleomorph of *Guillardia theta* is only 0.55 Mbp while the plant-parasitic nematode *Pratylenchus coffeae* is 19 Mbp.
- The number of predicted genes in a genome varies from 2,000 to 100,000, but is not necessarily related to the size of the genome.
- In eukaryotes, there appears to be no clear correlation between the characteristic genome size of a species (C-value) and the apparent complexity of the species or number of predicted genes. The *Fugu rubripes* genome, which is one tenth the size of the human genome, appears to have the same number of genes.
- Eukaryotes exhibit isochores, which are long segments of uniform G+C content which can be classified into distinct families (L1, L2, H1, H2, H3, etc.). Some isochores are associated with coding regions, i.e. H3 isochores.
- Eukaryotic chromosomes vary in their coding density. They can be gene-rich or gene-poor. Only 5-10% of a vertebrate genome consists of coding regions.

Causes of Genomic Variation

Duplications	Genome duplication Chromosomal duplication Replication slippage Unequal crossing over Rolling circle DNA amplification
Insertions	Retroviral retroposons Long interspersed nuclear elements (LINEs) Short interspersed nuclear elements (SINEs) DNA transposons Plasmids
Deletions	Replication slippage Unequal crossing over
Rearrangements	Chromosomal rearrangements Transposition Recombination
Point Mutations	Synonymous Nonsynonymous

Human Genome Assembly and Annotation Process



Lecture 9
Genomics
November 18, 2009

Sequencing Strategies

Directed Cloning

A ordered series of overlapping fragments is prepared, typically by a series of deletions from one end of a larger fragment. The assembly is trivial, but construction is slow, and sequencing redundancy is low.

Primer Walking

A series of sequencing reactions is performed, each based on information derived from the prior round of sequencing, with new sequencing primers designed after each round. Although the assembly is trivial, this approach is slow, and regions that are difficult to prime or accurately sequence can cause problems.

Shotgun Sequencing

Randomly generated short fragments, typically created by shearing or restriction, are sequenced with high redundancy. The complete sequence is then derived by computational assembly. This approach is generally the fastest and most cost effective, and currently the most popular.

Reading Sequence Data

Tracking

Need to correctly identify the lanes in the sequencing gel. This is typically automatically handled by the sequencing unit.

Base Calling

Need to correctly interpret the electropherogram peaks into sequence data by properly calling each base. Quality values are assigned to each base called. Miscalls, spurious deletions and spurious insertions are possible problems.
Software: Sequencher, phred, Acembly

Read Size

Currently typical read size from a single sequencing run is 500 to 800 bp. Large templates, such as those generated by large vectors (e.g. YACs) are harder to read directly, so smaller vectors (e.g. M13) are generally used.

Assembling Sequence Data

Assembly

Computationally reconstructing the complete sequence from the sequenced fragments. Although a simple process for bacterial genomes, this can be difficult for genomes with extensive repeats (e.g. *Homo sapiens*) or genomes with extreme G+C content that are difficult to sequence accurately (e.g. *Dictyostelium discoideum*).
Software: Sequencher, phrap, Acembly

Primer Design

Primer design is critical to walking strategies, and may be used with other strategies to close gaps, test quality, improve quality, or verify a particular region.
Software: Primer, Oligo, MacVector

Genome Analysis

Fundamentals

Quality Control: test accuracy, compare sequence data to overlapping clones
Annotation: name, source, library, method and references
Mapping: generate predicted restriction maps and compare to existing data, compare sequence data to genetic maps

Compositional Analysis

G+C content: G+C content graphs, isochore identification
Complexity analysis: identify regions of high or low information content
Linguistic analysis: CpG island identification, expected vs. observed word frequencies, hexamer frequency analysis
DNA structure prediction: bendability mapping, triple helix predictions, nucleosome mapping, chromatin structure modeling

Repeat Analysis

Masking: can mask repeats with RepeatMasker or other software to ease assembly or analysis
Extracting: extract repeats to ease assembly or reconstruct ancestral pre-insertion sequences

Genome Analysis and Annotation

Gene Finding

Coding Regions: identify coding and non-coding regions
Intron-Exon Stucture: identify splice sites and assemble a properly spliced product
Control Elements: identify promoters, enhancers, repressor sites and other control elements

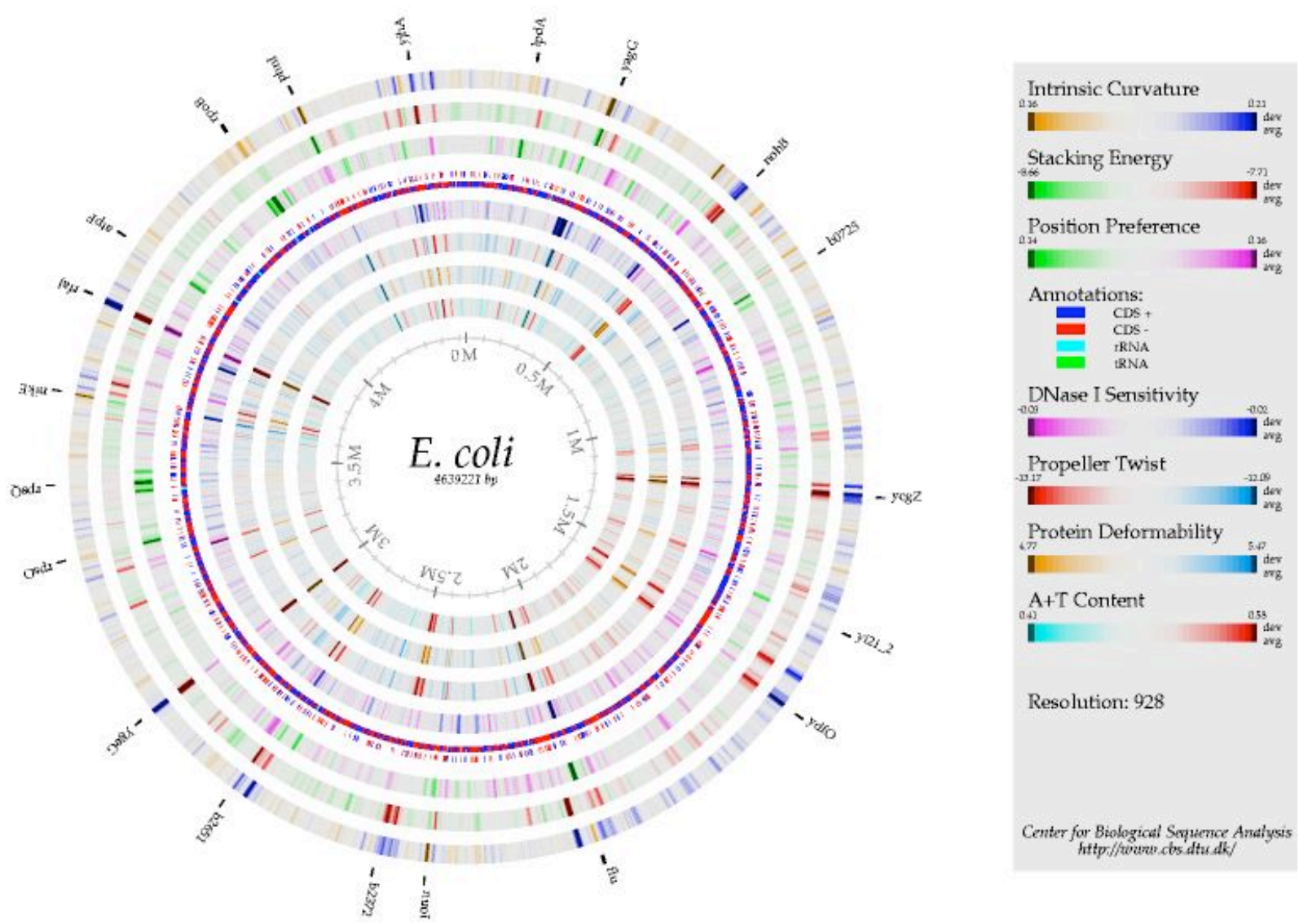
Comparative Genomics

Internal: look for related regions within the same genome
External: compare to other genomes, look for missed genes, construct phylogenies, reconstruct metabolic pathways

Publication

Final Annotation: check to make sure errors are not being propagated, including false positive or false negatives from database searches, and check genome, protein and organismal context carefully
Sequence Submission: BankIt (simple web submission to the NCBI), or Sequin (stand alone tool for electronic submission of large or complex sequences to the NCBI).

E. coli Genome Annotation



Intrinsic Curvature
0.16 0.23 dev avg

Stacking Energy
-8.66 -7.71 dev avg

Position Preference
0.34 0.36 dev avg

Annotations:
CDS +
CDS -
rRNA
tRNA

DNase I Sensitivity
-0.03 -0.02 dev avg

Propeller Twist
-13.17 -12.09 dev avg

Protein Deformability
4.77 5.47 dev avg

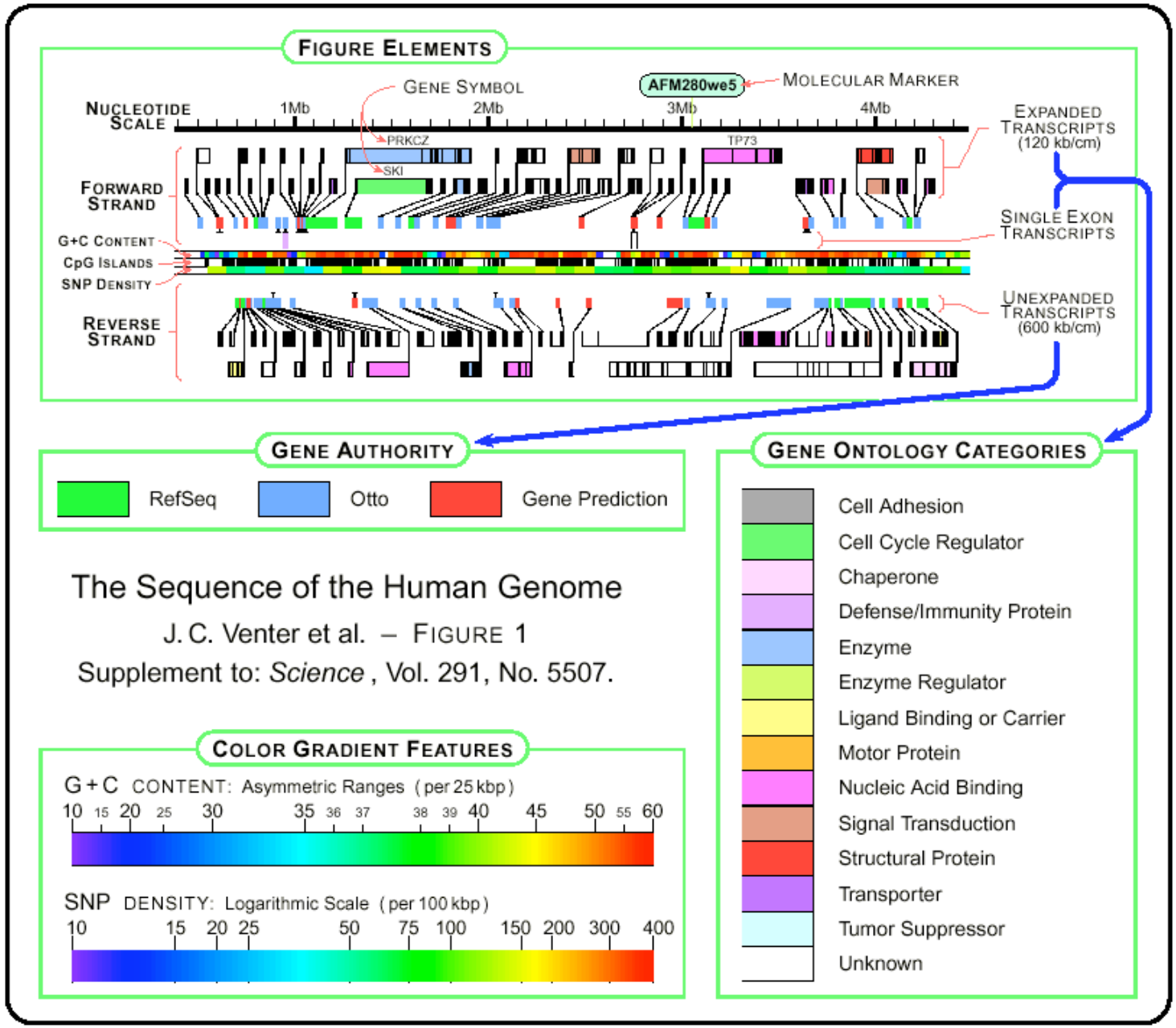
A+T Content
0.43 0.53 dev avg

Resolution: 928

Center for Biological Sequence Analysis
http://www.cbs.dtu.dk/

Lecture 9
Genomics
November 18, 2009

Human Genome Annotation



Lecture 9
Genomics
November 18, 2009

Genome Resources

PEDANT

Allows quick access and automatic and exhaustive analysis of genomic DNA and protein sequences, ranging from individual sequences to sets of sequences to complete genomes. Includes general, protein function, taxonomy and protein structure information.
http://pedant.gsf.de/

Kyoto Encyclopedia of Genes and Genomes (KEGG)

Includes general genome information, as well as detailed metabolic pathway charts and orthologous genes for many genomes.
http://www.genome.jp/kegg/

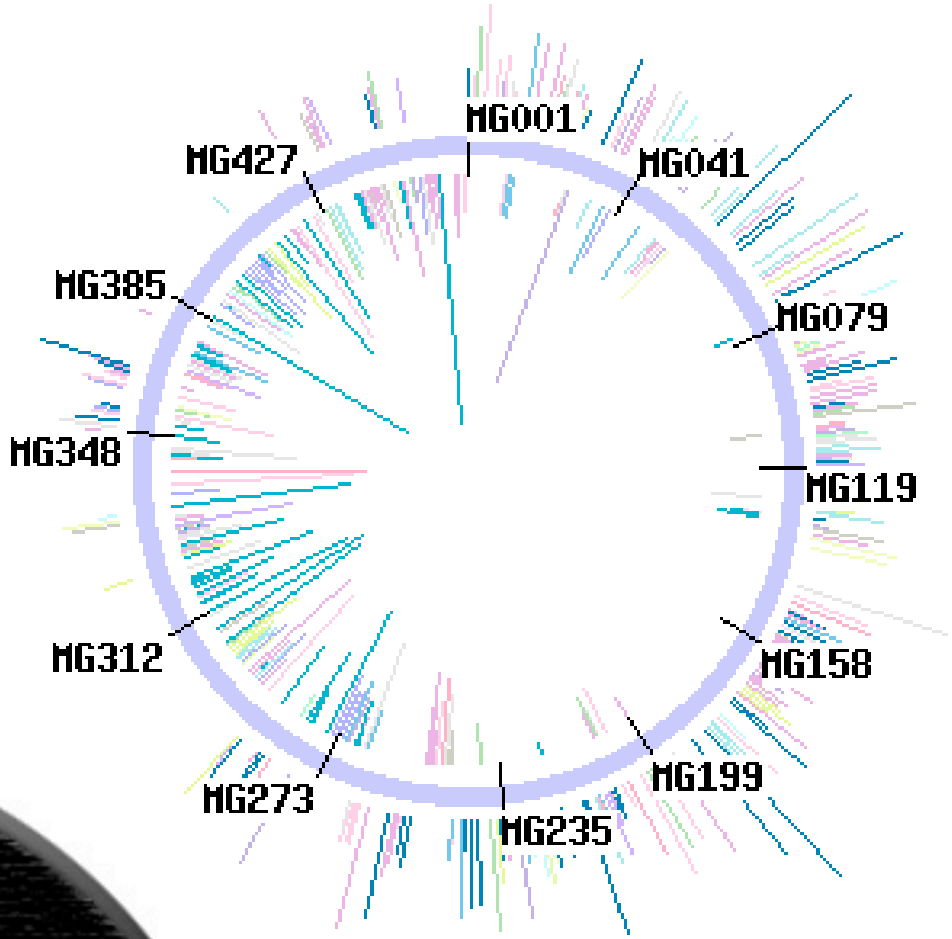
NCBI Clusters of Orthologous Groups (COG)

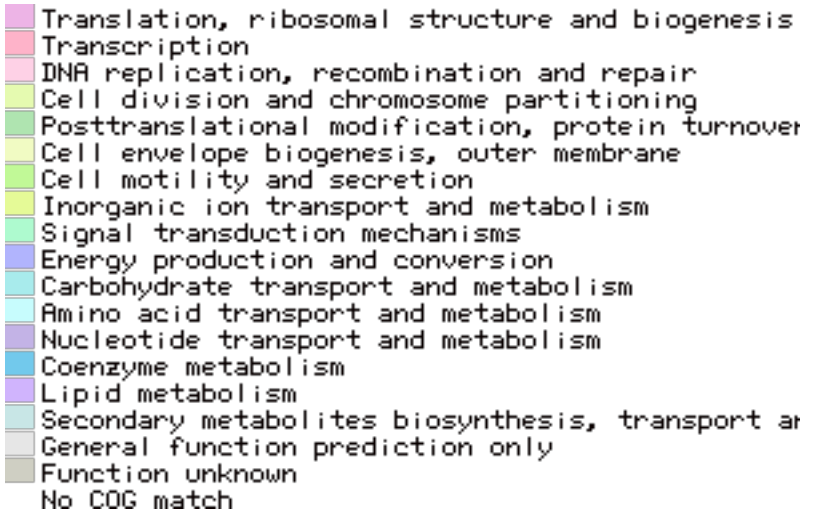
Clusters of orthologous groups of proteins delineated by comparing protein sequences encoded in 66 complete genomes. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain.
http://www.ncbi.nlm.nih.gov/COG/

Microbial Genome Database for Comparative Analysis (MBGD)

Allows for comparative analysis of microbial genomes, including searching for likely homologs among all sequenced microbial genomes, with homology assigned strictly by sequence similarity.
http://mbgd.genome.ad.jp/

NCBI COG





Mycoplasma genitalium

Genome References

A Quick Guide to Sequenced Microbial Genomes

A descriptive guide to over 180 fully sequenced microbial genomes published by The Genome News Network.

http://www.genomenewsnetwork.org/resources/sequenced_genomes/genome_guide_p1.shtml

NCBI Entrez Genomes and Genomic Biology Resources

Completely sequenced genomes and in progress genome sequences at NCBI. All three main domains of life – bacteria, archaea, and eukaryota – are represented, including over 300 microbes, as well as over 3,000 viruses and organelles.

http://www.ncbi.nlm.nih.gov/sites/entrez?db=Genome
http://www.ncbi.nlm.nih.gov/Genomes/

The J. Craig Venter Institute (JCVI/TIGR)

The JCVI (formerly TIGR) makes a comprehensive microbial genome resource available, as well as making a number of its genomic software tools available to academic researchers for free.

http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi
http://www.jcvi.org/cms/research/software/

Database of Genome Sizes

A list of known genome sizes maintained by the Center for Biological Sequence Analysis.

http://www.cbs.dtu.dk/databases/DOGS/index.php