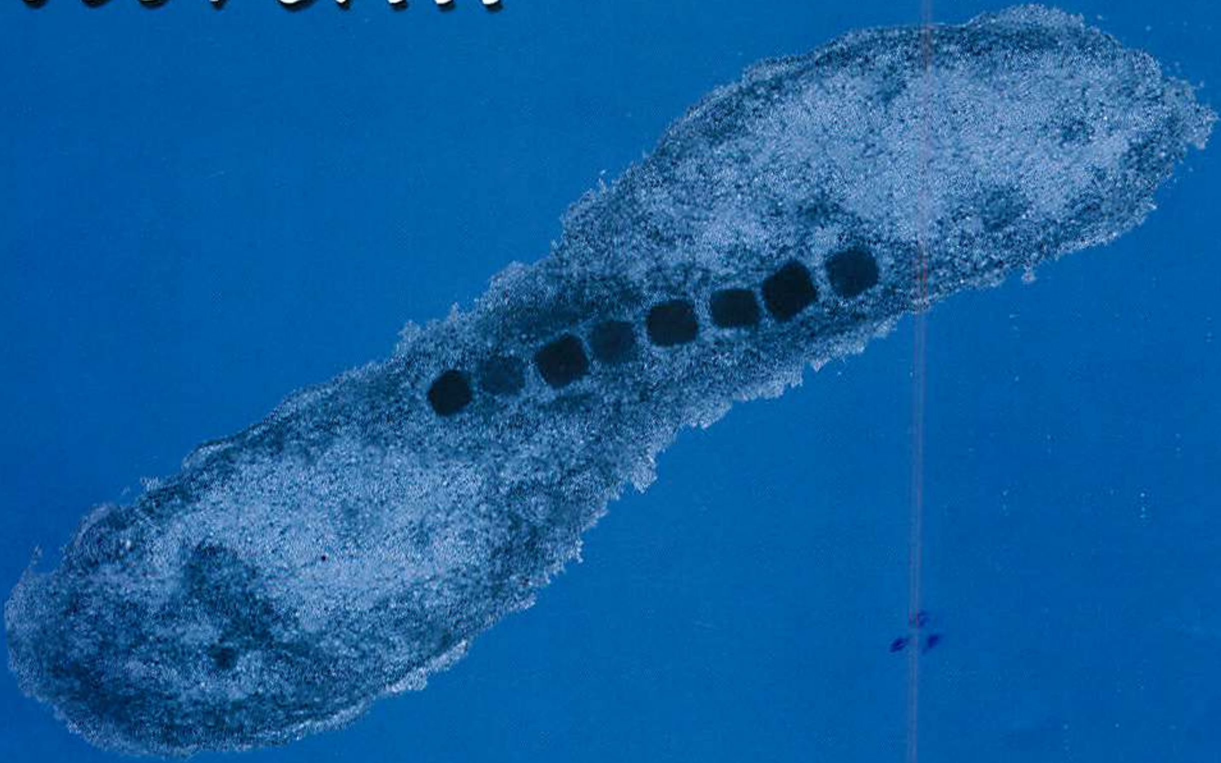
Discover...



Fourth Annual Conference on Microbial Genomes

February 12 - 15, 2000
Westfields Marriott, Chantilly, VA

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

CONFERENCES, EDUCATION AND TRAINING
THE INSTITUTE FOR GENOMIC RESEARCH

Program & Abstract Book Sponsored by:
DUPONT
The miracles of science™

Upcoming Conferences

February 19 – 20, 2000	**2000 Genome Seminar – "Genomic Revolution in the Fields: Facing the Needs of the New Millennium"** as part of 2000 AAAS Annual Meeting and Science Innovation Exposition, Marriott Wardman Park, Washington, DC
September 11 - 14, 2000	**12th Annual Genome Sequencing and Analysis Conference** Fontainebleau Hilton, Miami Beach, Florida Chairs: **Craig Venter,** Celera Genomics **Richard Gibbs,** Baylor College of Medicine **Mathias Uhlen,** Royal Institute of Technology
November 16 - 19, 2000	**Computational Genomics IV** Renaissance Harborplace Hotel in Baltimore, MD Chairs: **Steven Salzberg**, The Institute for Genomic Research **Tony Kerlavage,** Celera Genomics

Please check our Web site (http://www.tigr.org)
For agenda and conference updates

For additional information contact:
Conferences, Education and Training
The Institute for Genomic Research
9712 Medical Center Drive, Rockville, MD 20850-3319
301-610-5959 301-838-0229 (FAX) seqconf@tigr.org (E-MAIL)

Distinguished Leaders in Science

National Academy of Sciences Auditorium
2101 Constitution Avenue, NW, Washington, DC

Free and open to the public

Thursday, March 23, 2000 - 5:00 pm
Claude Canizares
Bruno Rossi Professor Experimental Physics & Director
Center for Space Research
Massachusetts Institute for Technology
"Exploring the Violent Universe with the Chandra X-ray Observatory"

Thursday, April 13, 2000 - 5:00 pm
Christopher Chyba
Carl Sagan Chair for the Study of Life
In the Universe SETI Institute
"Europa and the Rebirth of Exobiology"

Thursday, May 11, 2000 - 5:00 pm
Richard C. Canfield
Research Professor
Department of Physics
Montana State University
"Space Age"

For additional information please contact the TIGR web site:
http://www.tigr.org
or call the Conference, Education & Training Office at 301-610-5959

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

THE INSTITUTE FOR GENOMIC RESEARCH
9712 Medical Center Drive
Rockville, MD 20850

The Institute for Genomic Research

Fourth Annual Conference on Microbial Genomes

Program and Abstract Book

Westfields Marriott, Chantilly, VA

February 12 – 15, 2000

2000 Conference Cochairs:

Claire M. Fraser, Ph.D. - The Institute for Genomic Research
Siv Andersson, Ph.D. - University of Uppsala - Sweden
Jennie C. Hunter-Cevera, Ph.D. - University of Maryland Biotechnology Institute

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

CONFERENCES, EDUCATION AND TRAINING
THE INSTITUTE FOR GENOMIC RESEARCH

Contents

Acknowledgements

TIGR would like to thank the following sponsors for their support:

Compugen

U.S. Department of Energy

Dupont

USDOE grant no: DE-FG02-94ER61892.A008

*Preliminary Agenda**

*All information is subject to change. Please see addenda for final agenda.

Conference Cochairs:
Claire M. Fraser, Ph.D. - The Institute for Genomic Research
Siv Andersson, Ph.D. - University of Uppsala
Jennie C. Hunter-Cevera, Ph.D. - University of Maryland Biotechnology Institute

Saturday, February 12, 2000

Noon – 5:00 pm	Exhibit Set-up
3:00 pm	Registration Opens
3:00 – 5:45 pm	Poster Set-up
6:00 – 8:00 pm	Open Exhibits and Posters
6:30 - 8:00 pm	Welcoming Reception

Sunday, February 13, 2000

7:00 am	Breakfast
8:30 am - Noon	**Plenary Session I: Comparative Genomics**
8:30 am	Robert Fleischmann, The Institute for Genomic Research, Rockville, MD
9:00 am	Frederick Blattner, University of Wisconsin, Madison, WI
9:30 am	Burkhard Tuemmler, Klinische Forschergruppe, Hannover, GERMANY
10:00 am	Richard Moxon, Oxford University, UK
10:30 - 11:00 am	Break
11:00 am	Lee Ann McCue, NY State Department of Health, Albany, NY

11:30 am	Ian Paulsen, The Institute for Genomic Research, Rockville, MD
10:00 am - 2:00 pm	Posters and Exhibits Open
12:30 pm	Lunch
2:00 - 5:30 pm	**Plenary Session 2: Genome Projects**
2:00 pm	William Nierman, The Institute for Genomic Research, Rockville, MD and Janine Maddock, University of Michigan, Ann Arbor, MI
2:30 pm	Timothy Read, The Institute for Genomic Research, Rockville, MD
3:00 pm	Siv G.E. Andersson, University of Uppsala, SWEDEN
3:30 - 4:00	Break
4:00 pm	Julian Parkhill, The Sanger Centre, Cambridge, UK
4:30 pm	Susan Douglas, Institute for Marine Biosciences, Halifax, Nova Scotia
5:00 pm	Derek Lovley, University of Massachusetts, Amherst, MA
6:00 pm	Dinner
7:30 - 9:00 pm	Exhibits and Posters Session

Monday, February 14, 2000

7:30 am	Breakfast
8:30 am - Noon	**Plenary Session 3: Genome Biology**
8:30 am	Les Baillie, Biomedical, Porton Down, Salisbury, UK
9:00 am	Rino Rapuoli, Chiron Corporation, Siena, ITALY

9:30 - 10:00 am	Break
10:00 am	R. Frank Rosenzweig, University of Florida, Gainesville, FL
10:30 am	David Alland, Division of Infectious Disease, Bronx, NY
11:00 am	Michael Laub, Standford University, Palo Alto, CA
11:30 am	Pierre Legrain, Hybrigenics, Paris, FRANCE
9:30 am - 2:00 pm	Posters and Exhibits Open
Noon	Lunch
Noon - 2:00 pm	Exhibits and Posters Session
2:00 - 5:00 pm	**Plenary Session 4: Genome Analysis**
2:00 pm	George M. Garrity, Bergey's Manual Trust, East Lansing, MI
2:30 pm	Steven Salzberg, The Institute for Genomic Research, Rockville, MD
3:00 pm	Sean Eddy, Washington University School of Medicine, St. Louis, MO
3:30 - 4:00	Break
4:00 pm	Peter Karp, SRI Internationale, Menlo Park, CA
4:30 pm	William Pearson, University of Virginia, Charlottesville, NC
5:00 pm	Jonathan Eisen, The Institute for Genomic Research, Rockville, MD
6:00 pm	Dinner
	Free Evening

Tuesday, February 15, 2000

7:00 am	Breakfast
9:00 am - Noon	Exhibits Open (No Poster Sessions)
9:00 - 11:30 am	**Plenary Session 5: Patterns and Processes in Genome Evolution** Ford Doolittle, University of Halifax, Halifax, Nova Scotia Richard Lenski, Michigan State University, East Lansing, MI Christos Ouzonis, European Molecular Biology Laboratory (EMBL), Cambridge, UK Mary Carrington, National Cancer Institute-FCRDC, Frederick, MD
Noon - 3:00 pm	Breakdown Exhibit Area
12:30 pm	Lunch Meeting Adjourns

Plenary Speakers and Chairs

David Alland M.D.
Assistant Professor of Medicine
MonteFiore Medical Center
Department of Medicine/Infect. Disease
111 East 210th Street
Bronx, NY 10467
P: 718-920-2971
F: 718-920-2746
dalland404@aol.com

Siv G. Andersson Ph.D.
Professor
Uppsala University
Molecular Evolution
Evolution Biology Center
Norbyvagen 18C
Uppsala, 752 36
SWEDEN
P: 46-18-471-4370
F: 46-18-557-723
Sivc.Andersson@ebc.uu.se

Les Baillie Ph.D.
CBD, Porton Down
Biomedical Sciences Department
Building 7
Salisbury, Wiltshire, ST4 0JQ
UNITED KINGDOM
P: 44-1980-613881
F: 44-1980-613284
lesbaillie@hotmail.com

Fred R. Blattner Ph.D.
Professor of Genetics
University of Wisconsin, Madison
Laboratory of Genetics
445 Henry Mall
Madison, WI 53706
P: 608-262-2534
F: 608-263-7459
fred@genome.wisc.edu

Mary Carrington Ph.D.
Senior Scientist
National Cancer Institute - FCRDC
Laboratory of Genomic Diversity
PO Box B
Frederick, MD 21702
P: 301-846-1390
F: 301-846-1909
carringt@mail.nicfcrf.gov

Ford Doolittle Ph.D.
Professor
Dalhousie University
Biochemistry and Molecular Biology
5859 University Avenue
Halifax, NS B3H 4H7
CANADA
P: 902-494-3569
F: 902-494-1355
ford@is.dal.ca

Susan M. Douglas Ph.D.
Senior Research Officer
Institute for Marine Biosciences
1411 Oxford Street
Halifax, NS B3H 3Z1
CANADA
P: 902-426-9441
F: 902-426-9413
susan.douglas@nrc.ca

Sean R. Eddy Ph.D.
Assistant Professor
Washington University School of Medicine
Department of Genetics
4566 Scott Avenue
St. Louis, MO 63110
P: 314-362-7666
F: 314-362-7855
eddy@genetics.wustl.edu

Jonathan Eisen Ph.D.
Assistant Investigator
The Institute for Genomic Research
Department of Microbial Genomics
9712 Medical Center Drive
Rockville, MD 20850
P: 301-838-3507
F: 301-838-0208
jeisen@tigr.org

Robert D. Fleischmann Ph.D.
Investigator
The Institute for Genomic Research
Microbial Genomics
9712 Medical Center Drive
Rockville, MD 20850
P: 301-838-3508
F: 301-838-0208
rdfleisc@tigr.org

Claire M. Fraser Ph.D.
President/CEO
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
P: 301-838-3500
F: 301-838-0209
cmfraser@tigr.org

George M. Garrity Sc.D.
Editor-in-Chief, Treasurer
Michigan State University
Department of Microbiology
Bergey's Manual Trust
152 Giltner Hall
East Lansing, MI 48824
P: 517-432-2459
F: 517-432-2458
garrit@msu.edu

Jennie C. Hunter-Cevera Ph.D.
Head-Center for Environmental
Lawrence Berkeley National Lab.
Life Sciences
1 Cyclotron Road, Building 70A-3317
Berkeley, CA 94720
P: 510-486-7359
F: 510-486-7152
jchunter-cevera@lbl.gov

Peter D. Karp Ph.D.
Senior Scientist
SRI International
Bioinformatics Research Group
333 Ravenswood Avenue, EK223
Menlo Park, CA 94025
P: 650-859-4358
F: 650-859-3735
pkarp@ai.sri.com

Michael Laub
Stanford University
Developmental Biology
Stanford Medical Center, Beckman
279 Campus Drive
Palo Alto, CA 94304
P: 650-723-5685
F: 650-725-7739
laub@leland.stanford.edu

Pierre Legrain Ph.D.
Vice President
Hybrigenics
Science and Technologies
180 Avenue Daumesnil
Paris, 75012
FRANCE
P: 33-170-912302
F: 33-170-912949
plegrain@hybrigenics.fr

Richard E. Lenski Ph.D.
Hannah Professor
Michigan State University
Center for Microbial Ecology
Plant & Soil Science Building, Room
East Lansing, MI 48824
P: 517-355-3278
F: 517-353-3955
lenski@pilot.msu.edu

Derek R. Lovley Ph.D.
Distinguished Professor& Department
University of Massachusetts
Department of Microbiology
203 Morrill Science Ctr., IVN
Amherst, MA 01003
P: 413-545-9651
F: 413-545-1578
dlovley@microbio.umass.edu

Janine Maddock Ph.D.
Assistant Professor
University of Michigan
Department of Biology
830 North University
Ann Arbor, MI 48109-1048
P: 734-936-8068
F: 734-647-0884
maddock@umich.edu

E. Richard Moxon Ph.D.
Professor and Head of Department
University of Oxford
Paediatrics Department
Institute of Molecular Medicine
John Radcliffe Hospital
Oxford, OX3 9DU
UNITED KINGDOM
P: 44-1865-221074
F: 44-1865-221889
richard.moxon@paediatrics.ox.ac.uk

Christos Ouzonis Ph.D.
Group Leader
EMBL-EBI
Research Department
Wellcome Trust Campus
Cambridge, CB10 1SD
UNITED KINGDOM
P: 44-1112-494653
F: 44-1223-494471
ouzounis@ebi.ac.uk

Julian Parkhill Ph.D.
Computer Biologist
The Sanger Centre
Pathogen Sequencing Unit
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SA
UNITED KINGDOM
P: 44-122-383-4075
F: 44-122-349-4919
parkhill@sanger.ac.uk

Ian Paulsen Ph.D.
Assistant Investigator
The Institute for Genomic Research
Functional Genomics
9712 Medical Center Drive
Rockville, MD 20850
P: 301-838-3531
F: 301-838-0208
ipaulsen@tigr.org

William R. Pearson Ph.D.
Professor
University of Virginia
Dept. of Biochemistry & Molecular Gen.
Jordon Hall, UVA Health Systems
Box 800733
Charlottesville, VA 22908-0733
P: 804-924-2818
F: 804-924-5069
bmp2h@virginia.edu

Rino Rappuoli Ph.D.
Head of Research
IRIS, Chiron SpA
Via Fiorentina 1
Siena, 53100
ITALY
P: 39-0577-243414
F: 39-0577-243564
rino_rappuoli@biocine.it

Timothy Read Ph.D.
Assistant Investigator
The Institute for Genomic Research
Microbial Genomics
9712 Medical Center Drive
Rockville, MD 20850
P: 301-838-3554
F: 301-838-0208
tread@tigr.org

Frank Rosenzweig Ph.D.
Associate Professor
University of Florida
Molecular Genetics and Microbiology
PO box 100266
Gainesville, FL 32611-0695
P: 352-392-7077
F: 352-846-2042
frosenzw@mgm.ufl.edu

Steven Salzberg Ph.D.
Director of Bioinformatics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
P: 301-315-2537
F: 301-838-0208
salzberg@tigr.org

Burkhard Tummler Ph.D.
Medizinische Hochschule Hannover
Klinische Forschergruppe, OE6710
Carl Neuberg Str. 1
Hannover, D-30623
GERMANY
P: 49-511-532-2920
F: 49-511-532-6723
tuemmler.burkhard@mh-hannover.de

Speaker Abstracts

Plenary Session 1: Comparative Genomics

Sunday, 8:30 am – Noon

Sequencing of The *M. Tuberculosis* Genome; Comparison of a Recent Clinical Isolate with the Laboratory Strain

Robert D. Fleischmann[1], Liane Carpenter[1], Arthur Delcher[2], Jeremy Peterson[1], William Bishai[3], David Alland[4], Claire Fraser[1], [1]The Institute for Genomic Research, Rockville, MD, [2]Celera Genomics, Rockville, MD, [3]The Johns Hopkins University School of Medicine, Baltimore, MD, [4]Albert Einstein College of Medicine, New York, NY

Significant differences have been demonstrated between genomes of laboratory strains with long histories of passage and recent clinical isolates. The H37Rv laboratory strain of Mycobacterium tuberculosis was first isolated in 1905 and has been passaged for many decades. Of greater importance is that H37Rv is of unknown virulence in humans. The selection of a clinical isolate, CDC1551, a strain involved in a recent cluster of tuberculosis cases (that is, known to be transmissable and virulent in humans) ensures that the sequence of the genome of a fully virulent M. tuberculosis strain will be available.

The completion of both the H37Rv and CDC1551 genomes provides the first opportunity for a complete comparison between two closely related organisms of the same bacterial species and the chance to correlate differences in phenotype with genome content and organization. Types of polymorphic comparisons include the number and distribution of the IS6110 element (four in CDC1551 vs 16 in H37Rv), insertions/deletions and gene duplications (~ every 90,000 bp), differences in copy number of tandem repeats (~ every 75,000 bp), and the frequency of single nucleotide polymorphisms (~ every 3,200 bp). The comparison afforded by this opportunity provides the potential to recognize the genetic basis for successful human colonization, infectivity, and fully fledged transmission of a pathogen.

We have developed a hybridization method which allows us to rapidly probe hundreds of clinical isolates for polymorphisms identified by the H37Rv vs CDC1551 comparison. A collection of linked and unlinked strains from Baltimore and New York City have been investigated in this manner. The comparison afforded by this opportunity provides the potential to recognize the evolutionary relationships between strains and the genetic basis for successful human colonization, infectivity, and fully fledged transmission of a pathogen.

Bacterial Pathogen Genomics

Frederick R. Blattner, Nicole Perna, Jason Gregor, George Mayhew, Gyorgy Posfai, Guy Plunket III, Ying Shao, Debra Rose, Robert Mau and Valerie Burland, University of Wisconsin, Madison, WI

We are sequencing the genomes of a group of Gram-negative bacterial pathogens that are related to *E. coli*, making use of the completed *E. coli* K-12 sequence to accelerate assembly and analysis of the new genomes. The pathogens we have selected include members of *E. coli* human diarrheagenic and extraintestinal strains (enteropathogenic, enterotoxigenic, enteroaggregative, uropathogenic *E. coli*, and K1, cause of neonatal sepsis and meningitis). The diarrheagenic strains cause diseases of major importance in developing countries, whereas the extraintestinals are frequently responsible for nosocomial or community-acquired infections in North America and Europe. In addition to these *E. coli* genomes we will determine thegenomic sequences of *Yersinia pestis* KIM, causative agent of plague, *Shigella flexneri* 2a, the principle agent of dysentary, and *Salmonella typhi* Ty2, causing Typhoid fever.

In a parallel project, the genome of enterohemorrhagic *E. coli* O157:H7 EDL933 has been sequenced. This bacterium was the cause of recent outbreaks of acute diarrhea and hemolytic uremic syndrome. The genome has been assembled on *E. coli* K-12 as a backbone, and has served as a model for the genomics approach. This comparative method revealed that O157:H7 is a complex mosaic of about 200 novel genetic elements forming insertions in a relatively conserved K-12-like backbone. The novel regions, totalling 1 Mb, vary in size from a few base pairs up to at least 60 kb, and are distributed throughout the genome. Since many of the virulence determinants found in these novel regions, and indeed, many "backbone" genes, are shared among the pathogens, we expect this multigenomic approach to allow rapid identification of a "pathosphere" of the virulence genes that make up the pathogenic potential of this group of bacteria. Using the partially completed sequence of O157:H7 we have begun PCR based experiments to evaluate variation among natural populations of *E. coli* from outbreaks and standard *E.coli* strain collections.

Comparative Biology Of Pseudomonas Species

B. Tummler[1], C. Kiewitz[1], K. Larbig[1], A.-S. Limpert[1], C. Weinel[1], L. Wiehlmann[1], A. Christmann[2], H.-J. Fritz[2], G. Gottschalk[2], U. Romling[3], H. Hilbert[4], A. Dusterhoft[4], K.E. Nelson[5], C.M. Fraser[5], [1]Klinische Forschergruppe, OE 6710, Medizinische Hochschule Hannover, Germany; [2]Goettingen Genomics Laboratory, Goettingen; [3]GBF, Braunschweig, Germany; [4]QIAGEN, Hilden, Germany; [5]The Institute for Genomic Research, Rockville, MD

All members of the genus Pseudomonas (rRNA group I) are nutritionally versatile, bioactive, and prolific colonisers of surfaces, but differ in their pathogenic potential for animals and plants. For example, the opportunistic pathogen Pseudomonas aeruginosa and the non-pathogenic Pseudomonas putida share numerous metabolic pathways and inanimate habitats, but are distinct in their repertoire of virulence factors, membrane proteins, exoproducts and transport systems. The number and chromosomal localization of rrn operons are species-specific features. The genome of the type species P. aeruginosa (5-7 Mb) comprises a mosaic of species-specific, clone-specific and strain-specific blocks of DNA whose relative impact for habitat-specific survival is currently being investigated by STM technology. Species-specific DNA is characterized by conserved gene contigs and low nucleotide substitution rates (0.3 %). Sequencing of 100 kb blocks of strain-specific DNA in hypervariable regions revealed gradients in GC contents and CAI and an overrepresentation of orphan genes and mobile genetic elements.

Use of Whole Genome Sequence of *Neisseria Meningitidis* Serogroup (Menb) Strain Mc58 to Facilitate Understanding of The Molecular Basis of its Pathogenicity

Richard Moxon, University of Oxford, on behalf of the MenB Research Consortium; The Institute for Genomic Research (TIGR), Rockville, USA, IRIS, Chiron S.p.A, Siena, Italy, and Molecular Infectious Disease Group, Institute of Molecular Medicine, University of Oxford, UK

The genome of MenB, strain MC58, a causative agent of meningitis and septicemia, contains 2,158 predicted coding regions in its 2,272,325 base-pairs. Of these putative genes, 1,158 (54%) were assigned a suggested biological role. These included 31 novel candidate virulence factors, many of which were apparently acquired through horizontal transfer events, as well as the more than 70 previously described genes associated with pathogenicity. In addition, 345 (16%) predicted genes had matches to proteins of unknown function in other species and 532 (25%) had no matches to sequences found in publicly available data bases.

A striking feature of MC 58 genome organisation and content is the large number (more than 60) of known and predicted phase variable (contingency) genes, the number exceeding any of the other bacterial pathogens studied to date. The surface variation and adaptive potential mediated by these contingency genes emphasises the challenge of identifying conserved, surface exposed antigens that could be effective vaccines.

Functional Classification Of cNMP-Binding Proteins and Nucleotide Cyclases

Lee Ann McCue, Kathleen A. McDonough, and Charles E. Lawrence, Wadsworth Center, NY State Dept. of Health, Albany, NY

We analyzed the cyclic nucleotide-binding protein and nucleotide cyclase superfamilies using Bayesian computational methods of protein family identification and classification. In addition to the known types of cNMP-binding proteins - cNMP-dependent kinases, cNMP-gated channels, cAMP-guanine nucleotide exchange factors, and bacterial cAMP-dependent transcription factors - new functional groups of cNMP-binding proteins were identified, including putative ABC-transporter subunits, translocases, and esterases. Classification of the nucleotide cyclases revealed subtle differences in sequence conservation of the active site that distinguish five classes of cyclases: the multicellular eukaryotic adenylyl cyclases, the eukaryotic receptor-type guanylyl cyclases, the eukaryotic soluble guanylyl cyclases, the unicellular eukaryotic and prokaryotic adenylyl cyclases, and putative prokaryotic guanylyl cyclases. Phylogenetic distribution of these proteins was also analyzed, with particular attention to the 22 complete archaeal and eubacterial genome sequences. Among these species, *Mycobacterium tuberculosis* H37Rv and *Synechocystis* PCC6803 each encoded many cNMP-binding proteins and cyclases, whereas the archaeal species appeared not to encode these proteins at all. In addition, our results suggest a possible horizontal transfer event of a cyclase gene between eukaryotes and prokaryotes.

Comparative Analysis of Microbial Transport Proteinslan

T. Paulsen, The Institute for Genomic Research, Rockville, MD

All organisms for which complete genome sequences are available were analysed for their content of cytoplasmic membrane transport proteins. The transport systems present in each organism were classified according to (1) putative membrane topology, (2) protein family, (3) bioenergetics, and (4) substrate specificities. The overall transport capabilities of each organism were thereby estimated. The number of transporters identified in each organism varied drammatically, but was approximately proportional to genome size. Over eighty distinct families of transport proteins were identified. Two superfamilies, the ATP-binding cassette (ABC) and major facilitator (MFS) superfamilies account for nearly 50% of all transporters in each organism, but the relative representation of these two transporter types varied over a

50-fold range, depending on the organism. Differences in their reliance on primary vs. secondary transport and the range of transporter substrate specificities in each organism were found to generally correlate with the respective ecological niches and metabolic capabilities of each organism. The complete complement of predicted transporters in each completely sequenced organism are available on my WWW site (http://www-biology.ucsd.edu/~ ipaulsen/transport/).

Plenary Session 2: Genome Projects
Sunday, 2:00 – 5:30 pm

The Caulobacter Crescentus Genome Sequencing Project

William C. Nierman[1], **Janine Maddock**[2], [1]The Institute for Genomic Research, Rockville, MD, [2]Department of Biology, University of Michigan, Ann Arbor, MI

Caulobacter crescentus is a member of the alpha subclass of the proteobacteria which also include Rickettsia, Rhizobium, Agrobacterium and Brucella species. It is the most prevalent non-pathogenic bacterium in nutrient-poor fresh water streams and is also found in marine environments. It is one of the organisms responsible for sewage treatment. Caulobacters are being modified for use as a bioremediation agent for removing heavy metals fromwastewater.

Caulobacter crescentus has been extensively studied because it exhibits a well-defined developmental pattern that is independent of environmental stress. The free-swimming morphologically distinct swarmer cell progresses to an anchored stalked cell, the only cell type capable of genome replication and cell division. Cell division of the stalked cell splits out a swarmer daughter cell.

C. crescentus has a genome size of 4 Mb, with G+C content of about 66.5%. Tremendous power for genome assembly was brought to this project through the use of a 2 and 10 kb insert size 2 plasmid library strategy. In sequencing this organism at TIGR, 65,588 random sequence reads from both ends of plasmid clones were used to assemble the genome into only three groups comprising essentially all of the genome sequence. A preliminary review of *C. crescentus* ORFs revealed by the sequence is provided.

Bacillus Anthracis Genome Sequencing Project

Timothy Read and Scott Peterson, The Institute for Genomic Research, Rockville, MD

The sequencing of a plasmid-cured *Bacillus anthracis* Ames strain is underway using a whole-genome random-shotgun strategy. The initial phase of the project, sequencing a 2-3 kb sheared insert library has been completed, yielding 77,093 sequences with an average readable length of 562 nt. Assuming a genome size of ~5.0 Mb, this provided an 8.7-fold average base coverage. The average G+C content of this DNA sequence was 36.3%. Efforts are being directed currently to linking assemblies generated from the small-insert library data using both PCR techniques and a 7-10 kb insert library as a 'genomic scaffold'

A preliminary gene list derived from the sequence data indicates that at least 60% of B. anthracis ORFs have homologs to known genes from other *Bacillus* species. This includes many spore coat and spore germination determinants believed to play in important role in virulence. Also notable was the presence in the genome of 64 copies of a conserved 16 bp palidrome known to regulate expression of extracellular virulence factors in *B. thuringiensis*. The information from the genome sequence will provide an invaluable resource to aid the design of vaccines, detection methods and novel therapies to counter the *B. anthracis* biowarfare threat.

B. anthracis sequence data is available through the TIGR microbial database site www.tirg.org/tdb/mdb/mdb.html).

Comparative Genomics of Intracellular Parasites and Symbionts: Rickettsia, Bartonella, Francisella and Buchnera

Siv G.E. Andersson[1], Cecilia Alsmark[1], Bjorn Canback[1], Asa Sjogren[1], Ivica Tamas[1], Jan Karlsson[2], Richard Titball[3], Jennifer Wernegren[4], Nancy Moran[4], Charles Kurland[1]. [1]Department of Molecular Evolution, Uppsala University, SWEDEN; [2]National Defense Research Establishment, Umea, SWEDEN; [3]Defence Evaluation and Research Agency, Porton Down, UK; [4]Department of Ecology and Evolutionary Biology, University of Arizona, Tuscon, AZ

We study the evolutionary forces that drive the delicate balance between genome shrinkage and expansion. Initially, we have focused on intracellular parasites associated with the rickettsioses disease complex, such as Rickettsia and Bartonella. The 1.1. Mb genome sequence of Rickettsia prowazekii, the causative agent of epidemic, louse-borne typhus was published in 1998. Here, we present a detailed comparative analysis of the metabolic profiles of Rickettsia and Bartonella based on the 2 Mb genome sequences of Bartonella henselae, the causative agent of cat scratch disease (CSD) and Bartonella quintana, the causative agent of trench fever. Parallels are drawn to the 2 Mb genome of Francisella tularensis, the causative agent of tularemia. More recently we have shifted our attention to Buchnera aphidicola, which are obligate endosymbionts of aphids and have genomes about 600 kb in size. The effects of reductive evolution on the metabolic profiles are markedly different for obligate intracellular bacteria that have specialized as either parasites or symbionts. To study the mechanisms of gene degradation we have examined the evolution of pseudogenes and noncoding sequences in closely related strains and species. Evidence is presented which shows that the junk DNA in the R. prowazekii genome represents degraded remnants of ancestral, inactivated genes. The basic principles of gene elimination that are being explored in our genomic work are likely to be of major importance for our understanding of the way in which microbial genomes evolve.

The *Guillardia theta* (Cryptophyceae) Nucleomorph Sequencing Project

Susan E. Douglas, Program in Evolutionary Biology, Canadian Institute for Advanced Research and National Research Council Institute for Marine Biosciences, Halifax, Nova Scotia

Nucleomorphs are vestigial eukaryotic nuclei that persist in plastids of two groups of algae, the cryptomonads and the chlorarachniophytes, after the incomplete reduction of algal secondary endosymbionts. These cells thus contain four genomes - nuclear, nucleomorph, plastid and mitochondrial – the latter three of which have undergone severe reduction in gene content through gene losses and transfers to the nucleus.

Since it is impossible to purify nucleomorphs of the cryptomonad *Guillardia theta* and nucleomorph DNA comprises only 0.1% of total cellular DNA, it is extremely difficult to recover nucleomorph DNA-containing recombinants from libraries constructed from unfractionated cellular DNA. Small amounts of DNA were purified from the three nucleomorph chromosomes resolved by pulsed field gel electrophoresis and from bisbenzimide-cesium chloride gradients, and nucleomorph-enriched libraries constructed in plasmid vectors using a variety of restriction enzymes.

The complete sequence of the three nucleomorph chromosomes (196, 181 and 174 kb) reveals the highest gene density for any cellular genome (one per 0.8 kb). Surprisingly, genes transcribed by different RNA polymerases overlap their neighbours and there are very few introns. Both ends of all three chromosomes are identical over 12 kb, and contain the rRNA repeats and aberrant telomeres. Although most genes are for gene expression or protein degradation, several genes have been detected that are targetted to the plastid, thus providing the raison d'etre for the retention of the nucleomorph genome.

Genome of Geobacter Sulfurreducens

B.A. Methe, L. Banerjei, W.C. Nierman, and **D.R. Lovley**, University of Massachusetts, Amherst, MA

The complete genome sequence of dissimilatory metal-reducing microorganism, Geobacter sulfurreducens, is currently being determined in order to better understand the metabolic potential of this environmentally significant genus. A combination of geochemical and microbiological studies have suggested that respiration with Fe(III) may have been the first form of microbial respiration on Earth and microbial metal reduction is important in the cycling of carbon and metals in pristine and contaminated modern environments. Both molecular and culturing studies have suggested that Geobacter species closely related to G. sulfurreducens are the most numerous metal-reducing microorganisms in a variety of subsurface environments. The biochemistry of metal reduction in G. sulfurreducens is being intensively investigated and a genetic system for this organism has recently been developed. The determination of the whole genome of G. sulfurreducens is being accomplished using a random whole genome shotgun approach to provide eight-fold coverage of the 1Mb genome followed by closure of remaining physical or sequence gaps. TIGR assembler and other computer programs developed by The Institute for Genome Research are being used to assemble the genome, to aid in gap closure, and to finish the annotation process. It is expected that the genome sequence will provide information crucial to the further understanding of metal reduction and other key metabolic processes in subsurface environments.

Plenary Session 3: Genome Biology
Monday, 8:30 – Noon

Bacillus Anthracis, A Bug with Attitude

Les Baillie, Biomedical Sciences, DERA Porton Down, Salisbury, UK

The US Department of Defence describes anthrax spores as the top choice in biological weapons for warfare (www.anthrax.osd.mil/AVIP.htm). They are easy to produce, resistant to most vicissitudes and cause disease via the aerosol route (mortality > 80%). The biology of the organism, its ability to cause disease and the nature of its virulence factors will be described. Vaccination is the most cost effective form of mass protection. The current US and UK licensed human anthrax vaccines have been in use for many decades and have been shown to be effective in non-human primates. These vaccines where developed using 1950's technology and as a consequence are expense to produce and use a process which is not amenable to large scale production. Recent advances in biotechnology have enabled researchers to developed improved vaccine expression systems. One such system, based on Bacillus subtilis will be described. What is the future of anthrax vaccines? Access to the genome sequence the organism (www.tigr.org) will enable researchers to better understand the biology of the organism and through this understanding identify new vaccine targets.

Novel Proteins For Vaccine Development From N. Meningitidis Genome

Rino Rappuoli, IRIS, Chiron SpA, Siena, Italy, on behalf of The Institute for Genomic Research, Rockville, MD, and Molecular Infectious Disease Group, Institute of Molecular Medicine, University of Oxford, UK

Meningococcal meningitis and sepsis are devastating diseases that can kill children and young adults within hours despite the availability of effective antibiotics. The diseases are caused by Neisseria meningitidis, a Gram-negative, capsulated bacterium, that has been classified into five major pathogenic serogroups (A, B, C, Y, and W135) on the basis of the chemical composition of distinctive capsular polysaccharides. Vaccines against serogroups A, C, Y and W135 were developed in the 1960s by using the purified capsular polysaccharide as antigen. Second generation, conjugated vaccines are now being introduced. This approach could not be used for serogroup B because the capsular polysaccharide is a polysialic acid, identical in structure to a self antigen. As a consequence, today there are no effective vaccines available for the prevention of serogroup B N. meningitidis (MenB) disease, which is responsible for approximately 50% of all cases. To identify novel vaccine candidates, we determined the genome sequence of the virulent strain MC58. DNA fragments were analyzed to identify open reading frames (ORFs) that potentially encoded novel surface-exposed or exported proteins. These genes were expressed, the recombinant proteins were purified and used to immunize mice. Immune sera were then tested in ELISA and Fluorescence Activated Cell Sorter (FACS) analyses to detect proteins that were present on the surface of the bacterium. In addition, the immune sera were tested for bactericidal activity, as this assay correlates with protection in humans. Approximately 600 proteins were identified by computer screening. Of these 350 were expressed in Escherichia coli and used to immunize mice. The screening allowed to discover 85 new surface-exposed proteins, 25 of which induced bactericidal antibodies. Most of the newly-identified proteins are conserved in sequence across a set of strains selected to represent the sequence diversity of the group B meningococcus population. Genomic studies of bacterial pathogens have greatly increased our knowledge. However, they have not yet led to new advances in therapeutic or preventive measures. Here we report a case where a genomic approach has provided candidates that will be the basis for the clinical development of a vaccine against an important pathogen.

Molecular Beacons in Multiplex Formats: Susceptibility Testing and Indentification in M. Tuberculosis

David Alland, Division of Infectious Diseases, Department of Medicine, Montefiore Medical Center, Bronx, NY

Molecular beacons are fluorogenic reporter molecules that can be used in closed-tube PCR assays to detect point mutations, insertions and deletions in M. tuberculosis genes that are associated with drug resistance. The simplicity and robust nature of these assays have enabled us to use molecular beacons for rapid susceptibility testing, large population-based genetics studies, and to test new hypotheses on the genetics of resistance to isoniazid. Recent advances in molecular beacon design now enable us to use five different molecular beacon probes simultaneously in the same assay well.

The multiplexed molecular beacon assays offer higher throughput and increased reliability. A new form of "sloppy" molecular beacon that can be "programed" to hybridize with a wider range of target sequences is also under development. When used in multiplex assays, partially hybridizing sloppy molecular beacons generate flourescent "signature spectra"of short DNA sequences that uniquely identify a particular DNA, even if the actual DNA sequence is not known. This method may be applied to the development of simple, rapid, and sensitive polymerase chain reaction (PCR) assays that can identify multiple pathogens in a single reaction well.

Dissecting the Temporal Regulation of Caulobacter Cell Cycle Progression with DNA Microarrays

Michael Laub, Stanford University, Palo Alto, CA

Progression through the cell cycle requires the precise coordination and timing of multiple events, each critical to survival and proliferation. In the bacterium Caulobacter crescentus, this includes DNA replication, DNA methylation, chromosome segregation, cell division, and the biogenesis of polar organelles such as a stalk, flagellum, and pili. Each of these events, which must occur at a specific stage of the cell cycle, requires the expression and function of a discrete set of genes. Using the nearly complete Caulobacter genomic sequence, we developed DNA microarrays of nearly 3000 predicted ORFs as a tool to comprehensively identify and catalog these sets of genes and their timing of expression. Large populations of G1-staged Caulobacter cells were isolated and allowed to progress synchronously through the cell cycle with RNA isolated at time points throughout for expression analysis on microarrays. Analysis of the expression profiles led to the identification of 462 genes whose transcripts varied in a cell cycle-dependent fashion. This included 31 genes encoding two component signal transduction proteins, a class of regulatory molecules known to play a pivotal role in controlling cell cycle progression in Caulobacter. The results of this first, genome-wide analysis of a bacterial cell cycle will be presented along with recent microarray experiments designed to place cell cycle-dependent regulatory molecules into the context of the genetic network that controls the Caulobacter cell cycle.

Functional Proteomics On Microbial Genomes

J.-C. Rain and **P. Legrain,** Hybrigenics S.A., Paris, FRANCE

Large scale DNA-sequencing leads to the primary structure prediction of the proteome. New tools are needed to link these linear data to well-defined biological functions. Since protein-protein interactions are key determinants of the cell's life-cycle and many proteins function as parts of protein complexes, knowledge of protein-protein interactions is a strong tool for predicting protein function.

Hybrigenics, a functional proteomics company, has developed a highly selective and standardized procedure that allows networks of protein-protein interactions (Protein Interaction Maps, or PIMs®) to be constructed from genomic data.

Hybrigenics' technology of exhaustive two-hybrid screens of highly complex libraries solves many of the problems that arise when the classical matrix approach (testing pair-wise a collection of proteins, even at a large scale). False negatives occur when few combinations of bait and prey are tested for a given pair of proteins. Indeed, many chimeric, heterologous proteins are not correctly expressed, folded or located in a yeast host cell. Our strategy allows the screening of millions of interactions for every single bait, and many baits can be screened in parallel within several weeks. False positives occur when a given protein exhibits a surface that either auto-activates transcription or has a tendency to stick to other proteins. Such properties are easily detected and can be precisely measured and compared when screening the same library with reproducible protocols.

Decreasing the number of false negatives to a minimum while at the same time identifying false positives constitutes a key advantage of the exhaustive genome-based screening methodology when compared to other approaches available. This patented method, when applied to the yeast genome, has already been shown to be a powerful tool for functional analysis (1). Our technology also leads to the identification of interacting domains and can be used to isolate compounds capable of modulating specific protein-protein interactions.

Using Hybrigenics proprietary technology, we have performed an genome-wide analysis on the ulcer-provoking bacterium *Helicobacter pylori*. To date, over 1200 protein-protein interactions linking more than 800 proteins constitute the PIM® of *Helicobacter pylori*. This PIM® reveals new factors involved in known pathways, suggests the existence of novel complexes and ascribes novel biological functions to proteins. We are currently applying similar approaches to other bacterial pathogens, *Staphylococcus aureus* and *Streptococcus pneumoniae*, as well as the model bacterial organism *Escherischia coli*. This opens the way to the new field of comparative functional proteomics.

1. Fromont-Racine, M., Rain, J.C. and Legrain, P. (1997) 'Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. Nature Genetics 16, 277-82.

Plenary Session 4: Genome Analysis

Monday, 2:00 – 5:00 pm

Seeing The Forest Despite All The Trees: Microbial Classification In The Genomic Era

George M. Garrity, Bergey's Manual Trust and Department of Microbiology, Michigan State University, East Lansing, MI, Timothy Lilburn, Ribosomal Database Project and Center for Microbial Ecology, Michigan State University, East Lansing, MI

The field of systematic bacteriology has undergone tremendous change in the past two decades, due in large part to the widespread application of DNA sequence data (16S rDNA) to the derivation of a natural classification. The new classification serves as the basis of the forthcoming second edition of Bergey's Manual of Systematic Bacteriology and represents the evolution in our collective thinking about the "true" relationships among all prokaryotes. While this new classification represents a major improvement over earlier schemes, it should be viewed as an interim solution rather than a final one. In compiling the comprehensive outline of the valid prokaryotic taxa, one of the more challenging problems was visualizing the "biological landscape". While phylogenetic trees can provide a satisfactory view of the relationships among small numbers of strains, such graphs are severely constrained when applied to large data sets with many thousands of sequences. Alternative approaches of exploratory data analysis are available that permit construction of scalable 2-D and 3-D maps of the "biological terrain" with a fixed orientation and accurately reflect the variance among all sequences. These maps can be overlayed with a variety of other types of taxonomic and genetic data and allow easy assessment of the effects of different sequence alignments, the resolving power of different evolutionary models and hypothesis testing.

Algorithms for Whole Genome Analysis

Steven Salzberg, The Institute for Genomic Research, Rockville, MD

This talk will highlight two new systems: (1) MUMmer, a system for aligning whole genome sequences, and (2) improvements in the Glimmer bacterial gene finding system. MUMmer uses an efficient data structure called a suffix tree, which allows it very rapidly to align sequences containing millions of nucleotides. Its use will be demonstrated on several pairs of strains of bacteria, ranging from very highly homologous organisms to much more distantly related organisms. For very similar organisms, such as Mycobacterium tuberculosis strains H37Rv and CDC1551, the system allows one to quickly catalog all SNPs and all significant insertions. For more distantly related organisms the system helps provide a mapping between the genes and also identifies significant rearrangements. For example, it was recently used to identify a major chromosomal duplication between two *different* chromosomes of Arabidopsis thaliana. Glimmer 2.0, the newest version of the Glimmer system for microbial gene identification, now finds approximately 97--98% of all genes in a genome when compared with published annotation. Glimmer uses interpolated Markov models (IMMs) as a framework for capturing dependencies between nearby nucleotides in a DNA sequence. When we consider only those genes that have significant homology to genes in other organisms, Glimmer's accuracy rises to better than 99%. Some recent improvements will be described as well as an evaluation on ten completed genomes.

Web references: http://www.tigr.org/~salzberg, http://www.tigr.org/softlab/glimmer/glimmer.html

Computational Screens For Noncoding RNA Genes

Sean Eddy, Washington University School of Medicine, St. Louis, MO

Some genes produce functional RNAs instead of encoding proteins. Current genefinding approaches focus almost exclusively on protein-coding genes, so the diversity of the "modern RNA world" is an open question. We are developing probabilistic modeling approaches -- specifically, hidden Markov models and stochastic context-free grammars -- to identify noncoding RNA genes in genome sequence data. We have been using these methods to study the diversity of small nucleolar RNAs (snoRNAs), which are responsible for guiding specific nucleotide modifications of eukaryotic ribosomal RNAs. In collaboration with Pat Dennis's group at the University of British Columbia, we have recently found that Archaeal genomes also have numerous snoRNA genes. Many of our predictions have been confirmed both experimentally and by comparative analysis among three available Pyrococcus genome sequences. Because snoRNAs are still unknown in Bacteria, this result provides another shared character between the Archaeal and Eukaryotic lineages.

Knowledge-Based Modeling of the E. coli Metabolic Network

Peter D. Karp, Bioinformatics Research Group, SRI International, Menlo Park, CA

A knowledge-based model of an organism consists of a set of static facts about the molecular components of the organism, plus a set of rules of inference for deriving new relationships that are implicit in those static facts. If we consider a genome database to be a knowledge-based model, some key questions about that genome database are: What range of static facts can the database encode? And what new relationships can it infer? The ontology (schema) of the Pathway Tools software can encode a wide variety of information about the genes, gene products, metabolic pathways, transporters, and genetic-regulatory circuitry of an organism. For example, the EcoCyc DB describes the full genome and metabolic-pathway complement of E. coli, as well as many of its transporters and operons. The ontology of a genome database is of central importance because a poorly designed ontology will distort the information that the database encodes. The Pathway Tools software can infer a wide variety of relationships regarding the genome and the biochemical network of the organism that are implicit in the static facts within a Pathway Tools database. Each rule of inference is in a sense a predefined database query that infers a biological relationship of interest. We present a connectivity analysis of the E. coli metabolic network that consists of statistics on many of these inferred relationships.

FASTA: The Next Generation Sequence Alignment Tool

William R. Pearson, Aaron J. Mackey, and Ming-qian Huang, Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA

The FASTA package of programs has evolved considerably since the FASTP program was written in the fall of 1983. In addition to providing rapid methods for searching protein and DNA databases (FASTA), programs are available for comparing translated DNA sequences to protein databases (FASTX/FASTY), for comparing protein sequences to translated DNA databases (TFASTX/TFASTY), and for searching with short peptide sequences (FASTS/TFASTS) and mixed peptides (FASTF/TFASTF). The FASTX/FASTY/TFASTX/ TFASTY programs allow alignments with frameshifts, and thus are particularly effective for analyzing low accuracy single-pass sequence data. FASTF and FASTS can be used to identify proteins from minimal (8 - 20 residues) sequence data. The programs in the FASTA package are considerably slower than the corresponding BLAST programs, largely because FASTA seeks to calculate an approximate similarity score for every sequence in a database, while BLAST focuses its computation on those database sequences that are likely to be homologous. Because it calculates scores for tens of thousands of unrelated sequences in every database search, FASTA can estimate very accurately the statistical properties of virtually any set of local similarity scores from unrelated sequences, whether the scores are calculated for protein:protein, DNA:DNA, or protein:translated DNA alignments. More accurate statistical estimates and more sensitive DNA and translated-DNA searches are major strengths of the FASTA program. The FASTA package continues to evolve to provide more effective searches in the presence of low-complexity sequences, and to search with position-specific scoring matrices, and to provide more intelligent strategies for searching for closely related sequences, and for distinguishing paralogs from orthologs.

Plenary Session 5: Genome Analysis

Tuesday, 9:00 – 11:30 am

Assessing Horizontal Gene Transfer In Microbial Species.

Camilla Nesbo, Yan Boucher, Yuji Inagaki, **W. Ford Doolittle**, Canadian Institute for Advanced Research, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, NOVA SCOTIA

Clearly, genome data reveal lateral gene transfer (LGT) to be more frequent and important than we had imagined, at least among prokaryotes. We will will review some recent analyses from this laboratory bearing on transfer between bacteria and archaea, and then address some general questions: (i) is the attempt to reconstruct a universal tree fatally compromised by LGT?, (ii) is the attempt to reconstruct the genome of the last common ancestral cell also compromised?, (iii) are there new, proper and useful ways to redefine what phylogeneticists do?

Dynamics of Genomic and Phenotypic Evolution: A 20,000-Generation Experiment with E. coli

Richard E. Lenski, Center for Microbial Ecology, Michigan State University, East Lansing, MI

Twelve replicate populations of E. coli have been propagated in a simple laboratory environment for 20,000 generations. During this time, the populations have undergone substantial phenotypic evolution, including: improved competitive fitness relative to their common ancestor; changes in cell size and shape; and, in some cases, loss of methyl-directed mismatch repair. After reviewing these phenotypic changes, I will discuss recent collaborative work to link the phenotypic changes to genomic evolution. Three of the approaches we are using include: (1) RFLP analyses with IS elements as probes (with M. Blot & D. Schneider, Grenoble); (2) mini-transposon insertions to create new linkages between markers and beneficial mutations; and (3) sequencing representative regions of the genome (with M. Riley, Yale). Results from the first approach indicate that rates of phenotypic and genomic change are discordant, as has often been suggested but not previously demonstrated. The second approach shows that beneficial mutations, even 1-bp mutations, can be found, but the work is labor intensive. The third approach is only beginning, but it offers the opportunity to quantify rates and patterns of genomic evolution in a system wherein the time of divergence is known exactly.

Metabolic Networks in the Last Common Ancestor

Christos Ouzounis, European Molecular Biology Laboratory, Cambridge, UK

The phylogenetic distribution of completely sequenced genomes can provide us with insights into the processes of genome evolution. We have compiled a list of the distribution of the proteins for the first completely sequenced archaeal species, /Methanococcus jannaschii/. The number of proteins present in at least one other species from both Bacteria and Eucarya is 324. This set is relatively small, non-redundant and well-characterized: there exist 301 functions of which 246 are unique. This 'universal set' contains mostly genes coding for energy metabolism or information processing, and can be viewed as an estimate for the genome of the Last Universal Common Ancestor. Technical aspects of the data handling, technological developments in computational genomics and a simple format for genome sequence annotation (GATOS) will also be presented.

Genetic Susceptibility to Infectious Diseases in Humans

Mary Carrington, National Cancer Institute, Frederick, MD

Host genetic influences on microbial pathogenesis in humans is highly complex with few examples of nucleotide variation resulting in clear and predictable outcomes. A growing number of host genetic associations have been identified for HIV-1 disease, including those which relate to cellular receptors for the virus and those involved in immune response to the virus. Genetic effects on HIV-1 due to variation in the viral coreceptor gene CCR5 and the human major histocomplatibility class I genes are the clearest examples of such associations with HIV-1. We are now at the point with this disease that the possibility of synergistic interaction between genetic variants in disease pathogenesis must be considered. One example that will be discussed involves genes encoding natural killer (NK) cell receptors molecules which recognize specific HLA class I ligands, thereby controlling NK cell inhibition and activation. We are testing the possibility that the diverse KIR haplotypes may interact epistatically with the unlinked HLA class I alleles affecting pathogenesis of HIV-1 and HCV disease.

Index Poster Abstracts

Jason Hinds	P-11	Construction Of Whole Genome DNA Microarrays For Campylobacter Jejuni and Mycobacterium Tuberculosis
Andre Johann	P-12	Sequencing The Whole Genome Of *Methanosarcina Mazei* Strain Gö1
Yutaka Kawarabayasi	P-13	Genome Sequencing and Genome Comparison of Three Thermophilic Archaea
Thomas W. Kephart	P-14	Codon Redundancy and Comparative Genomics
Hans-Peter Klenk	P-15	Inference Of Microbial Evolution By Whole-Genome Sequence Comparison
Daniel Kosack	P-16	Extending The Capabilities And Capacity Of Genomic Assembly At TIGR
A. Malykh	P-17	Sequencing Directly off Sub-Microgram Quantities of Bacterial Genomic DNA
Scott N. Peterson	P-18	Analysis Of Competence Genes *From S. Pneumoniae* Using DNA Microarrays
Leigh A. Riley	P-19	Microbial Genome Resources At NCBI
Sylvie M. Rodriguez	P-20	Large-Scale Genome Diversity Sequencing Project
Carsten I. Rosenow	P-21	Escherichia Coli High Density Oligonucleotide Array Studies. Comparison of Two RNA Sample Preparation Techniques: Direct 5' End Labeling of RNA and c-DNA Synthesis Employing Random Priming

Index to Poster Abstract

Poster Abstracts

P-01
The Pseudomonas Putida Kt2440 Genome Sequencing Project

Hoda Khouri[1], Karen Nelson[1], Erik Holtzapple[1], Jeff Buchoff[1], Michael Rizzo[1], Azita Moazzez[1], Kelly Moffat[1], Kevin Tran[1], Hean Koo[1], P. Chris Lee[1], Daniel Kosack[1], Bradley Slaven[1], Helmut Hilbert[2], Claire M. Fraser[1], [1]The Institute for Genomic Research, Rockville, MD, [2]Qiagen, Hilden, GERMANY, Burkhard Tuemmler, Medizinische Hochschule Hannover, GERMANY

Pseudomonas putida is a soil bacterium with significant potential for bioremediation. To determine the genome sequence of strain KT2440, a joint sequencing project was initiated between The Institute for Genomic Research and a German Consortium (see http://www.tigr.org/tdb/mdb/mdb.html) in January 1999. The 6.1 Mbp genome is being sequenced by the random shotgun method, with multiple sized large insert libraries (10 kb and 40 kb) acting as scaffolding. At the end of random sequencing, there were a total of 392 sequencing gaps. The high GC content of the genome is highlighted in long stretches of G's and C's encountered in both sequencing and physical gaps, and through which we have had problems sequencing by traditional methods. Sequencing of short reads that point into gaps, dye primer chemistry, transposon mutagenesis on selected spanning clones, as well as ET chemistry were used to resolve most difficult areas. Multiplex PCR and micro-library construction, have also assisted in resolving and ordering the RNA operons. Grouper and autoprimer (TIGR softwares) were modified to deal with the size of the genome. Ultimately, the P. putida genome sequence will reveal the organisms potential in various biotechnological areas including the production of natural compounds, and remediation of polluted habitats.

P-02
An Initial Genome Comparison Of Pseudomonas Putida Strains Via PCR Based Subtractive Hybridization

Esperanza T. Nunez, Joel A. Malek, Karen E. Nelson, Artis Hicks, The Institute for Genomic Research, Rockville, MD

Bacteria belonging to the genus Pseudomonas are especially diverse in their range of metabolic capabilities. Specifically, P. putida has great potential as a model organism for microbial degradation of various toxic compounds. Presently, more than 200 strains of P. putida have been identified, each with it's own unique set of capabilities. Generally, differences between bacterial strains are identified through Southern Blotting or by sequencing of the related genome. With the availability of the nearly complete genome sequence of strainKT2440 (see http://www.tigr.org/tdb/mdb/mdb.html), and 1000 sequences of strain PRS1, we attempted to identify differences between these two strains. The inexpensive option, 'Subtractive Hybridization', was investigated for it effectiveness in identifying differences between environmentally relevant organisms. The CLONTECH PCR-Select Bacterial Genome Subtraction Kit(CLONTECH, # K1809-1) was used to subtract PRS1 against KT2440 andKT2440 against PRS1. Initially, 120 unique sequences were identified by the subtractions. Assignment of sequences was based on BLAST searching the subtracted sequences to sequences appearing in public databases. Important findings include sequences having significant hits at the amino acid level with genes coding for a Benzoate transport protein, regulators and core proteins among others.

P-03
Whole Genome Sequencing and Initial Analysis of Enterococcus faecalis Strain V583

Linda C. Banerjei[1], Karen A. Ketchum[2], Bao Tran[1], Jyoti Shetty[1], Jessica Vamathevan[1], Keita Geer[1], Hoda Khouri[1], Haiying Qin[1], Thomas S. Hansen[1], Jonathan F. Upton[1], Diana Radune[1], Cheryl L. Bowman[1], Lisa A. McDonald[1], Teresa R. Utterback[1], Brian A. Dougherty[3], Claire M. Fraser[1], [1]The Institute for Genomic Research, Rockville, MD, [2]Celera Genomics, Rockville, MD, [3]Department of Applied Genomics, Bristol-Myers Squibb, CT

Enterococci are the leading cause of nosocomial bacteremia, surgical wound infection, and urinary tract infection. They are notable among the lactic acid bacteria for their hardiness and resistance to concentrations of agents that would inhibit or kill other bacteria. The rising interest in enterococci in recent years is due to their increasing resistance to a broad spectrum of antibiotics. Enterococcus faecalis is responsible for 80-90% of human enterococcal infections. E. faecalis strain V583, the first vancomycin resistant clinical blood isolate in the United States, was selected for genome sequencing. We will highlight our progress on this project and initial analysis of sequence data.

P-04
The Cryptosporidium Parvum Genome Sequencing Project

Gregory A. Buck[1], Giovanni Widmer[2], Saul Tzipori[2], Mitchell S. Abrahamsen[3], Minnesota Ping Xu[1], Yingping Wang[1], Xu Wang[1], Donna Akiyoshi[2], Michael A. Buckholt[2], Xiaochuan Feng[2], Stephen M. Rich[2], Kim M. Deary[2], Cindi A. Bowman[2], [1]Microbiology and Immunology, Virginia Commonwealth University, Richmond, VA, [2]Division of Infectious Diseases, Tufts University School of Veterinary Medicine, North Grafton, MA, [3]Veterinary Pathobiology, University of Minnesota, St. Paul, MN

C. parvum, an apicomplexan parasite associated with AIDS and other immunocompromised states, is a worldwide cause of diarrhea in humans and animals. Effective therapies are unavailable largely due to a poor understanding of the biology of these protozoan pathogens. A two-project multi-lab consortium has been formed to sequence the ~10.4 mbp C. parvum genome. Project 1 of the consortium will focus on the genotype 1 NEMC1 isolate; Project 2 will focus on the genotype 2 Iowa isolate. Although both cause disease, genotype 1 is most prevalent in humans. Genotype 1 is less well studied because it has been refractory to laboratory propagation. We have recently succeeded in reproducibly propagating the NEMC1 genotype 1 isolate. Thus, Project 1 will generate ~5X shotgun sequence coverage, a large insert BAC library, and a least tiling path BAC contig of the of the NEMC1 genome. The results of a preliminary genome survey project of ~2% of the genome of NEMC1 is described herein. Project 2 of the consortium will generate a >5X shotgun sequencing coverage of the Iowa genotype 2 isolate. A comparison of these genotypes will provide clues to their host range, pathogenicity and other phenotypic differences.

P-05
Gateway Cloning System: A High-Throughput Gene Transfer Technology For Functional Analysis and Protein Expression

James Hartley, Gary Temple, **David Cheo**, Michael Brasch, Life Technologies, Rockville, MD

As a result of numerous ongoing genome sequencing projects, large numbers of candidate open reading frames are being identified, many of which have no known function. The analysis of these genes typically involves transfer of various DNA segments into a variety of vector backgrounds for protein expression or functional analysis. We describe a new method called the Gateway Cloning System that uses *in vitro* site-specific recombination to transfer DNA segments between vector backbones. We present results that demonstrate the use of this approach for efficient, directional cloning of PCR products. Such cloned PCR products, or other DNA segments flanked by recombination sites, can then be "automatically" transferred into new vector backgrounds, simply by adding the desired "Destination" vector and recombinase. By incorporating appropriate selections, the desired subclones are recovered at high efficiency , typically >90%, following introduction into *E. coli.* The Gateway Cloning method is fast, convenient, and can be automated, allowing numerous DNA segments to be transferred in parallel into many different vector backgrounds, in a single experiment. Resulting subclones maintain reading frame register, permitting the generation of amino and carboxy translation fusions. Approaches for optimization of protein expression, rapid functional analysis, and the integration of numerous expression technology platforms will be discussed.

P-06
The Bioknowledge™ Library – A Collection Of Curated Model Organism Proteome Databases With Applications For Comparative And Functional Genomics

Matthew E. Crawford, Ann M. Fancher, Maria C. Costanzo, Jodi Lew-Smith, Brian P. Davis, Kevin J. Roberg-Perez, Peter E. Hodges, Jennifer D. Hogan, Michael Cusick, Michael Tillberg, Carol A. Lingner, James I. Garrels, Proteome, Inc., Beverly, MA

The BioKnowledge Library is a collection of Proteome Databases, or volumes, that encapsulates the existing knowledge for all the genes and proteins of specific model organisms. The Yeast Proteome Database (YPD™), used by thousands of academic and corporate researchers, has become the standard for in-depth genome annotation. Knowledge is extracted from the research literature by Proteome's expert curators and integrated with yeast genomic and functional genomic datasets. YPD is used by functional genomics researchers who can quickly evaluate lists of genes from expression experiments, by drug discovery researchers who must select and validate targets, and by bioinformaticists who use the yeast genome to help assign functions to unknown genes from other organisms. With the ongoing development of PombePD, Proteome is expanding the Fungal BioKnowledge Collection. PombePD will include thorough coverage of the scientific literature for proteins of the fission yeast Schizosaccharomyces pombe. Also CalPD, a database for the pathogen Candida albicans, will be expanded into a collection of databases including information about all known proteins of the major fungal pathogens of humans. This collection will contain comprehensive curation of the relevant literature for model fungal organisms and pathogenic fungi. These databases will be fully integrated with the other volumes of BioKnowledge library, which currently contains YPD, CalPD and WormPD. Software tools allow queries of protein function, sequence similarity, molecular interactions, and pathway information across species lines. Using the databases within the BioKnowledge Library, proteins implicated in infectious diseases of human can be traced to homologs in model organisms, where experiments to test for drug effects can easily be designed and implemented. YPD and WormPD are accessible at http://www.proteome.com.

P-07
Pneumocystis Genome Project Update

Melanie T. Cushion[1], A. George Smulian[1], James R. Stringer[1], Chuck A. Staben[2], Jonathan Arnold[3], John Wunderlich[3], Michael Weise[3], [1]University of Cincinnati College of Medicine, Cincinnati, OH, [2]University of Kentucky, Lexington, KY, [3]University of Georgia, Athens, GA

Pneumocystis carinii (Pc) are a collective group of fungal organisms that inhabit the lungs of mammals and cause a lethal pneumonia when the host becomes immunocompromised. In humans with AIDS, Pc pneumonia remains the most common opportunistic infection and resistance to standard therapies has been recently reported. These fungi are intractable to long term culture and little is understood about their basic biology. In 1999, a genome project was initiated to map and sequence the genomes of Pc from rats and humans, with an EST project for the rat form. The EST database was established this year with 4800 clones using a l ZAPII cDNA library of rat Pc. ~1758 unique ESTs with an average read of 584bp were generated representing 12.5% of the 7.7 Mbp genome. The surface glycoprotein family of genes unique to Pc (MSG/MSR) were in ~2% abundance as were the protease/kexin-like genes (PRT1/KEX1). The 1557 sequences that identified genes new to Pc had the greatest homology to S. pombe (75%) and yeast (19%) while 782 sequences had no known homologies by BLASTX. Mapping has begun using a combination of cosmid-end sequencing and hybridization strategy with the cDNA library and pWEB library of 2200 clones.

P-08
Web-Based Visualization Tools For Bacterial Genome Alignments

Liliana Florea, Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA, Cathy Riemer, Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA, Scott Schwartz, Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA, Michael McClelland, Sidney Kimmel Cancer Center, San Diego, CAWebb Miller, Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA

With the increased flow of bacterial sequence data, both whole genome and contigs, visual aids for comparison and analysis studies are becoming imperative. We describe three web-based tools for visualizing alignments of bacterial genomes. The *Menteric* web server computes and displays nucleotide-level multiple alignments of sequences from several related genomes in a 1Kbregion surrounding a user-specified address. The alignment is rendered as a PostScript or PDF document in which annotations of ORFs, promoters and protein binding sites are color-coded. The *enteric* server produces a graphical, hypertext view of the alignments of the *Escherichia coli* genome with related bacteria covering 20Kb around a user-specified position. Clicking on *E. coli* gene names links to the WIT database. Pointing at alignments exposes discontinuities in gene order, the name of the matching sequence contigs, and the length of sequences lacking an *E. coli ortholog. Maj* combines features from the other modes, including multiplepercent identity plots with color-coding, the corresponding nucleotide-level multiple alignment, and hyperlinked annotations, all in an interactive graphical display with zoom-in capabilities, via a Java applet. These facilities are available for public use at http://globin.cse.psu.edu/ or via the Salmonella genome sequencing project at http://genome.wustl.edu/gsc/bacterial/newlistdisplay.pl.

P-09
GMP-Tool-Box: A Software Package For Bacterial Genome Project From Shotgun Follow-Up To Annotation

Lionel Frangeul, Philippe Glaser, Farid Chetouani, Christophe Rusniok, Carmen Buchrieser, Frank Kunst, GMP, Institute Pasteur, Paris, FRANCE

GMP-Tool-box (GMPTB) is a software package developed for the computational part of a genome project. The modules of GMPTB are designed for the Listeria Monoytogenes genome project but are also applicable to other genomes. The modules are classed into 3 groups corresponding to 3 steps of a genome project : shotgun follow-up, closure phase and annotation. Follow-up: During the shotgun phase, GMPTB extracts from the result file (Phrap format) all characteristics of the assembly (nber of contigs, nber of sequences.) and displays them in a table. This table can be used by other programs to create graphs of the progress of the shotgun phase. Moreover, GMPTB compares each assembly result with the former assembly and creates an HTML page to explain the relationship between old and new contigs (fusion, creation...). Closure phase: GMPTB contains tools to predict links between contigs. GMPTB searches for all the inserts with its ends in two contigs and marks this insert as linking-insert or as a misassembly according to orientations of each end and the distance between the 2 ends. These results can be obtained simultaneously with the different categories of libraries used in the shotgun phase (small, medium, BAC...). GMPTB can also predict links by searching similarities between ends of the contigs and other sequences. Annotation: GMPTB allows to start annotation during the finishing phase. Actually, GMPTB creates an Individual Protein File (IPF) for each ORF of an assembly. This IPF is a text file with a specific format which contain 3 categories of fields : - The "minimum fields" contain identification number, version number, location and sequences. The nucleotide sequence exported correspond to the sequence of the ORF with 500 additional bases before the first stop condon and 200 additional bases after the second stop. - The "automatic fields" contain results added by different programs to the IPF. These results can concern the ORF itself (homology, domains.) but also for instance the research of RBS, promotors or terminators (before and after the ORF). - The "manual fields" contain the results and comments

sequences could answer long unsolved phylogenetic and taxonomic questions. Several novel bioinformatic procedures have been developed for the analysis of these large and complex data sets. For this presentation we compared the complete sets of protein-coding genes from all publicly available genome sequences. Our analysis procedure enabled us to screen the resulting database for answers to specific phylogenetic questions, e.g. the existence of sets of genes specific for certain taxonomic groups (prokaryotes, archaea, eocytes,...), and for sets of genes linked to certain phenotypic features, e.g. genes linked to extreme thermophilia. The most astonishing result of the comparisons was the diversity in the evolution of several fractions of the microbial genomes. As much as these results may disappoint, having not solved some critical phylogenetic and taxonomic questions, they will also inspire us to formulate new hypotheses about the course and mechanism of evolution.

P-16
Extending The Capabilities And Capacity Of Genomic Assembly At TIGR

Daniel Kosack, Bradley Slaven, The Institute for Genomic Research, Rockville, MD

In order to meet the growing needs of the genomic community, methods of extending the capability of TIGR's home-grown assembly tool, TIGR Assembler, have been explored. A foremost priority is increasing the sequence capacity of the software, allowing for assembly of larger genomes. Regarding time optimization, attention is directed toward utilizing distributed processing and multi-threading. The redesign prototype assembler can assemble 60% or larger genomes on existing hardware. Other features have been added to generate the scoring graphs multithreaded. Furthermore, using new transitive closure software on the scoring graph produced by the assembler, related groups of sequences can be clustered and assembled in a distributed computing environment. Resulting contigs are used to initialize whole genome assembly.

P-17
Sequencing Directly off Sub-Microgram Quantities of Bacterial Genomic DNA

A. Malykh, O. Malykh, N. Polouchine, A. Slesarev and S. Kozyavkin, Fidelity Systems, Inc., Gaithersburg, MD

Robust sequencing off genomic templates presents a new challenge in technology development. The problems associated with the use of standard oligonucleotides as primers in genomic cycle sequencing protocols include insufficient specificity of primer annealing, non-specific amplification, low sensitivity and premature truncation at secondary structures in template DNA.

To overcome these problems we have developed a new method to generate combinatorial libraries of chemically modified oligonucleotides (fimers). The method is based on the use of our proprietary monomers containing MOX or SUC reactive moieties. We assessed the effects of modifications on DNA melting, electrophoretic mobility and DNA-protein interaction for individual oligonucleotides and their small libraries. We have developed rapid procedure for modification, deblocking and purification of fimers in 96-well plate format. Different design strategies for fimers have been tested with ThermoFidelase-2A, -2B and -2C, deaza-dGTP and dGTP in various thermal cycling protocols. We found that fimer design eliminates many restrictions on choosing primer sequence. Our results demonstrate feasibility of suppressing non-specific PCR amplification and primer-dimer formation after 100-400 cycles, synergy of chemical and enzymatic tools to sequence through strong stop and long simple repeats and sequence directly off sub-microgram quantities of bacterial genomic templates. Enhanced reaction chemistry has allowed us to overcome major obstacles in bacterial genomic sequencing associated with high florescent background and low signal. We obtained high quality reads from as low as 100-300 ng genomic template. Major applications of direct genomic sequencing including the discovery of novel genes and characterization of bacterial populations will be discussed.

Supported in part by DOE and NIH (DE-FG02-98ER82557 and 2R44GM55485-02).

P-18
Analysis Of Competence Genes From *S. Pneumoniae* Using DNA Microarrays

Robin T. Cline, Donald A. Morrison, Carissa Horst and **Scott N. Peterson**, The Institute for Genomic Research, Rockville, MD

We have constructed a microarray containing genes known to be induced during competence in *Streptococcus pneumoniae.* The precision of competence gene regulation and the signaling through two-component regulators are hallmarks of gene regulation events expected to be observed in the analysis of *S. pneumoniae* pathogenicity. The molecular events that accompany competence are complex and highly efficient and include uptake of DNA, degradation of a single strand of the DNA, and finally recombination of the DNA into the recipient chromosome. Exposure of *S. pneumoniae* to the competence stimulating peptide (CSP), activates a two-component signaling pathway, resulting in the coordinated control of several, perhaps as many as 40 competence-inducible genes. Purified CSP can be used to induce competence in the laboratory at non-saturating cell densities. Our microarray data is consistent with a large body of information accumulated over the years that have documented the level of induction of proteins and/or RNA by direct or indirect methods such as β-galactosidase assays using lac Z reporter constructs. The great majority of genes induced during competence are undetectable 5 minutes after induction with CSP but reach a peak just 10-15 minutes after CSP exposure. Here we report an analysis of wild-type and "competence-defective" mutant expression profiles.

P-19
Microbial Genome Resources At NCBI

Leigh A. Riley, Tatiana Tatusov, Jonathan Kans, Ilene Mizrachi, Jim Ostell, National Center for Biotechnology Information, NIH, Bethesda, MD

A centralized resource for obtaining and analyzing microbial sequences is a useful tool for the microbial genomics community. NCBI has two resources available: The Microbial Genomes Blast Databases and Entrez Genomes. The Microbial Genomes Blast Databases--http://www.ncbi.nlm.nih.gov/BLAST/unfinishedgenome.html--is a centralized location for BLASTing against complete or unfinished microbial genomes. Contigs from unfinished genomes and complete sequences can be searched using tblastn, blastn and tblastx tools. In addition to the hits themselves, 1000 bases of sequence on either side of the hits found in unfinished genomes can be viewed. Unfinished databases are updated as frequently as biweekly, with sequencestatistics available on the About the Databases page as well as organism and institution specific credits pages. The sequences from finished microbial genomes are available from the Entrez Genomeswebsite--http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html. Entrez Genomes includes a collection of completely sequenced and annotated microbial genomes. Submissions to Entrez Genomes can be made using the Sequin submission tool. Sequin contains an import table feature which allows submitters to annotate all of their features using a simple tab-delimited table.

P-20
Large-Scale Genome Diversity Sequencing Project

Sylvie M. Rodriguez[1], D.Ashley Robinson[1], Mary E. Golden[1], Ellson Y. Chen[2], Chun-Nan Chen[2], Scott Peterson[3], Elliot J. Lefkowitz[1], John I. Glass[1], Susan K. Hollingshead[1], [1]University of Alabama at Birmingham, Birmingham, AL, [2]Celera Genomics, Foster City, CA, [3]The Institute for Genomic Research, Rockville, MD

A genome diversity sequencing project is underway at the University of Alabama in collaboration with the Celera Genomics group in Foster City, CA and The Institute for Genomic Research in Rockville, MD. The goal is to provide data relating to intra-species diversity within *Streptococcus pneumoniae.* One component of the project is to build a sequence database comprising 10 kbp regions sampled from a set of highly diverse strains of *S. pneumoniae.* Fourteen strains were selected based on their genetic distance as judged from multilocus sequence typing (MLST) or IS1167-Box typing. Twenty key chromosomal regions of 10 kbp in size were chosen for sequencing (currently at 13). The obtained sequences are edited and assembled for public release. All the data in this project is released to a database maintained at the web site http://genome.microbio.uab.edu/strep/. As preliminary results, we observed that for a subset of 20 "housekeeping" genes the nucleotide sequence divergence levels were only a few percent. Analyses of the *aro* genes in the chorismate biosynthetic pathway showed that this region had a mosaic organization. Finally, the diversity observed in the 60 first amino-acids of the *comD* locus is directly correlated with the presence of comC1 and comC2.1 .

P-21
Escherichia coli High Density Oligonucleotide Array Studies. Comparison of Two RNA Sample Preparation Techniques: Direct 5' End Labeling of RNA and c-DNA Synthesis Employing Random Priming

Carsten I. Rosenow, Infectious Disease, Affymetrix, Santa Clara, CA, Rini Mukherjee Saxena, Infectious Disease, Affymetrix, Santa Clara, CA, Thomas R. Gingeras, Infectious Disease, Affymetrix, Santa Clara, CA

The development of high-density oligonucleotide arrays has offered the possibility to determine changes in RNA levels simultaneously for all transcripts in a microbial cell. The availability of the whole *Escherichia coli* (*E.coli*) genome sequence has allowed for the development of a high density oligonucleotide array encoding probes which interrogate all currently annotated open reading frames and all intergenic regions of the *E.coli* genome. Quantitative or qualitative evaluation of the *E.coli* RNA requires either direct labeling of the RNA or the use of a cDNA synthesis method using a pool of random oligonucleotide primers to generate cDNA which is subsequently 3' end labeled using deoxy-terminal transferase. Both sample preparation methods have been used to label *E.coli* total RNA as well as enriched RNA (depleted for r-RNA). Analysis of the enriched RNA sample indicated a 4 times reduction in r-RNA compared to the total RNA. Duplicate samples prepared by the same direct labeling or cDNA synthesis method show 86% and 96% concordance, respectively. However, results derived from comparing direct labeling to cDNA show significant discordance 45%. In addition both methods show differences in sensitivity and in specificity for genes expressed at high expression levels versus low expression levels as well as long versus short transcripts. Results from the comparison analysis have been validated using classic Northern blot techniques.

P-22
Single Point Mutation Found Responsible for Loss of Methoxymycolate Production in BCG Vaccine Strains Obtained After 1927

Benjamin G. Schroeder[1], Richard A. Slayden[2], Jacqueline N. Brinkman[3], Clifton E. Barry[2], and Marcel A. Behr[3], [1]The Institute for Genomic Research, Rockville, [2]Tuberculosis Research Section, Laboratory of Host Defenses, NIAID, Rockville, MD, [3]McGill University Health Centre, Montreal, CANADA

Bacille Calmette-Guérin (BCG) is currently the only available vaccine against Mycobacterium tuberculosis, the causative agent of tuberculosis. BCG vaccines are substrains of Mycobacterium bovis derived by attenuation in vitro. After the original attenuation (1908-1921), BCG strains were maintained by serial propagation in different BCG laboratories (1921-1961). As a result, various BCG substrains developed which are now known to differ in a number of genetic and phenotypic properties. However, to date, none of these differences has permitted a direct phenotype-genotype link. Since BCG strains differ in their ability to synthesize methoxymycolic acids and recent work has shown that the mma3 gene is responsible for O-methylation of hydroxymycolate precursors to form methoxymycolic acids, we analyzed methoxymycolate production and mma3 gene sequences for a genetically defined collection of BCG strains. We found that BCG strains obtained from the Pasteur Institute in 1927 and earlier produced methoxymycolates in vitro, but those obtained from the Pasteur Institute in 1931 and later all fail to synthesize methoxymycolates and furthermore, that their mma3 sequence differs from that of Mycobacterium tuberculosis H37Rv by a point mutation at base pair 293. Site-specific introduction of this guanine to adenine mutation into wild type mma3 (resulting in the replacement of glycine 98 with aspartic acid) eliminated the ability of this enzyme to produced O-methylated mycolic acids when cloned in tandem with mma4 into Mycobacterium smegmatis. These findings indicate that between 1927 and 1931, a point mutation in mma3 occurred with this mutant population becoming the dominant clone of BCG at the Pasteur Institute.

P-23
Omnione: Using a PVM Cluster to Conduct Homolog and Ortholog Searches on All 25 Fully Sequenced Microbial Genomes

Bradley E. Slaven, Jeremy D. Peterson, Hanif G. Khalak, Daniel S. Kosack, Erin K. Hickey, Owen White, The Institute for Genomic Research, Rockville, MD

The need for automated and consistent curation standards is becoming increasingly evident as the number of fully sequenced genomes continue to grow. Databases which contain consistent intra-organism homolog and inter-organism ortholog comparisons are essential and increasingly difficult to unify as the world-wide genomic sequencing efforts accelerate. In an effort to provide consistent homolog and ortholog databases, we used a Parallel Virtual Machine (PVM) cluster of Linux workstations to conduct All verses All homology searches on all 25 fully sequenced microbial genomes. Open reading frames for each of the genome are searched using a parallel version of the Blast- Extend-Repraze script, which conducts preliminary whole genome homology comparisons using WU-Blast and the Praze implementation of the Smith-Waterman Algorithm. Auto-BYOB was used to conduct preliminary automated annotation determinations. Lastly, parallel Hidden Markov Model (HMMER) search scripts are used to ensure reliable homolog and ortholog results from the preliminary Blast-Extend-Repraze All versus All open reading frames searches. The results of these searches are publicly available on our web site. We will report on the implementation and preliminary results of these annotation methods.

P-24
Analysis of *Pneumocystis Carinii* Ests: Problems In Assembly Of Gene Families

A. George Smulian[1], Jonathan Arnold[2], Michael Weise[2], John Wunderlich[2], Melanie T. Cushion[1], [1]University of Cincinnati, Cincinnati, OH, [2]University of Georgia, Athens GA

An EST database was created for rat Pc using a lambda ZAPII cDNA. Assembly of the ESTs using PHRAP generated 708 contigs containing >2 sequences and 1050 unique Pc ESTs. Contigs of three classes were examined to verify the robustness of clustering: single copy genes (B-tubulin, EF-3, mkp2); a multigene family (PRT/kexin protease genes) and the complex repetitive multigene family (MSG/MSR). Assembly of multiple ESTs (4-8 copies) encoded by single copy genes appeared appropriate. Three PRT contigs were generated but appeared redundant as each contig contained most of the 43 PRT ESTs sequences. Although appearing as a single contig, 20 distinct PRT genes were transcribed at different frequencies. Sixteen MSG/MSR contigs were generated containing 315 members, however, resolution of redundant sequences reduced this to 145 unique ESTs. 37 MSG ESTs arose from cDNAs primed from internal A-rich sequences, indicative of problems with AT-rich genomes, while 95 unique MSG/MSR ESTs primed from the polyA tai l of MSG mRNA. Analysis of these 95 MSG ESTs by phylogenetic software bore no resemblance to the original MSG contig assignment by PHRAP. Assembly of ESTs arising from multiple closely related members of gene families is problematic using PHRAP and alternatives should be developed.

P-25
Use Of ESTs to Study Expression of Gene Families In The Pathogenic Fungus Pneumocystis Carinii

Scott P. Keely, Melanie T. Cushion, A. George Smulian, **James R. Stringer**, University of Cincinnati, Cincinnati, OH

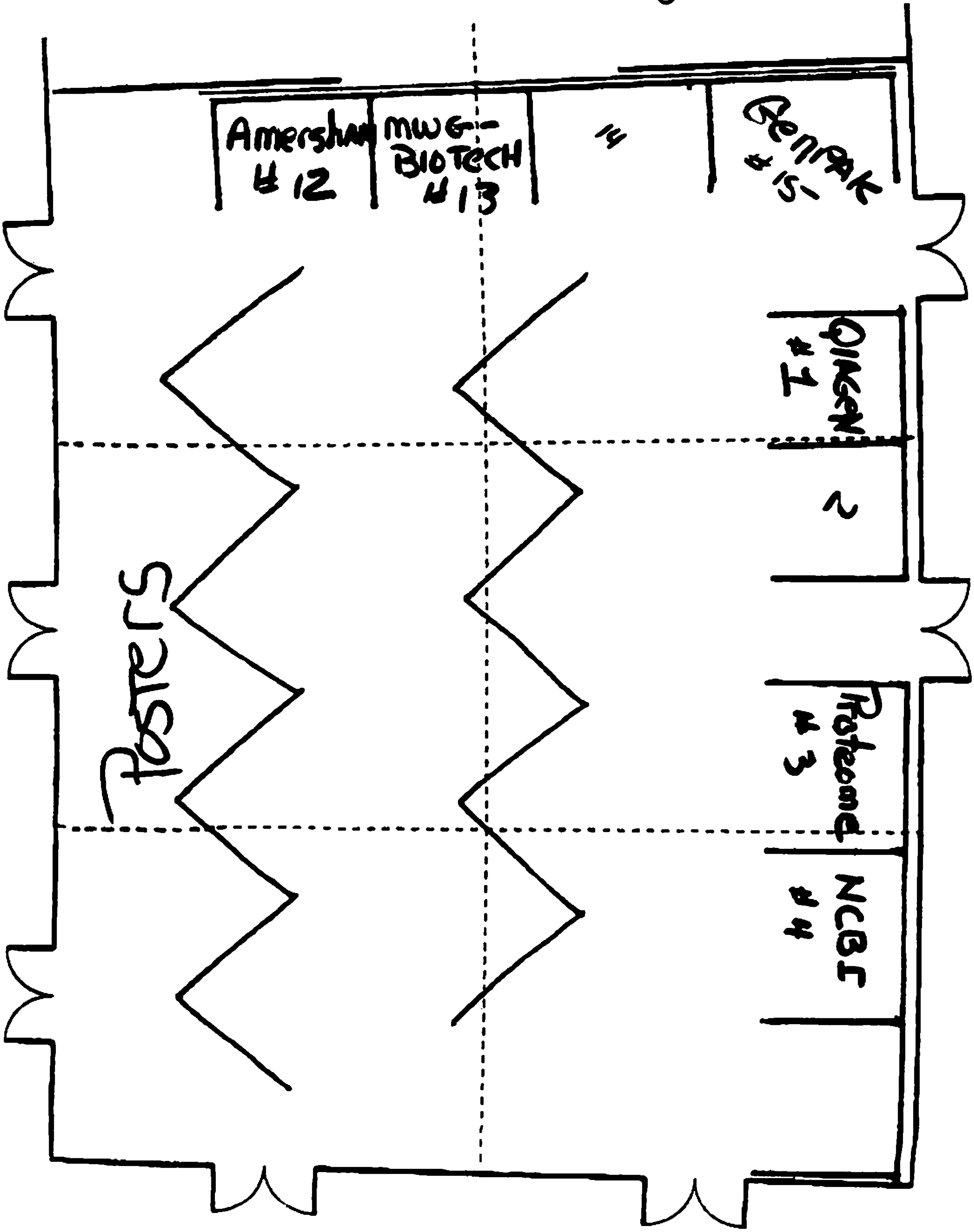
Pneumocystis carinii contains three gene families, MSG, MSR and PRT. Only one MSG gene is expressed in a given organism, and different organisms express different MSG genes. Expression of MSR and PRT families is not understood. It's possible that expression of PRT and MSR is coordinated with that of MSG because members of these three families are clustered in the genome. The EST database presented an opportunity to better understand expression of each family and to assess the possibility of coordinate regulation. We analyzed 36 MSG, 43 MSR, and 43 PRT ESTs. The MSG, MSR and PRT ESTs were comprised of 18, 16, and 18 different sequences, respectively. In all three sets of ESTs, different sequences were not equally abundant. For example, 9 of the 18 different MSG sequences were each in only one EST. A tenth sequence was most abundant, being in 6 ESTs. The other 8 sequences were of intermediate abundance. MSR sequences were similarly distributed into high, medium and low abundance classes. PRTs were distributed less evenly, with three high abundance sequences and the other 15 sequences all of low abundance. High abundance MSR and PRT genes are hypothesized to be linked to high abundance MSG genes.

P-26
Mapping the Bacteroides Thetaiotaomicron Genome

Jian Xu[1], Catherine C. Baublite[2], Cindy M. Foushee[2], Lora V. Hooper[1], Erica Donaldson[2], Lydia M Thompson[2], Magnus Bjursell[1] and Astra Zeneca R and D Molndal[3], Jason Himrod[1], Sam LaBrie[2], David A. Smoller[2], Jeffrey I. Gordon[1], [1]Washington Univ. School of Medicine, St. Louis, MO, [2]Genome Systems, St. Louis, MO, [3]SE-431 83 Molndal, Sweden

Bacteroides thetaiotaomicron is an abundant member of the normal mouse and human intestinal microflora. We have created a gnotobiotic mouse model that uses this anaerobe to study the molecular basis of a commensal relationship, Hooper et al. PNAS 96:9833, 1999. We are monitoring host responses to colonization with DNA microarrays. To characterize microbial responses, an academic-industrial collaboration was initiated by our groups at Washington University and Astra Zeneca to sequence the entire ~4.8 MB *Bacteroides thetaiotaomicron* genome and identify all of its ORFs. To support efficient production of a finished genome sequence, we decided to generate a BAC-based physical map of the genome that will be anchored to whole-genome shotgun data via BAC end sequences. 1059 BAC clones, with an average insert size of 90 kb, were digested with Hind III. Fingerprint analysis of the resulting fragments was conducted using FPC software, Soderlund et al., Comput. Appl Biosci 13:523, 1997. 742 clones have been built into 27 contigs, with an estimated coverage of 4.8 MB. The largest contig contains 108 clones and covers 624 KB. Based on the relative locations of the contigs, a minimal tiling path of the genome has been built comprised of 79 BACs.

Exhibit Hall Layout



Exhibitors

BOOTH #:(s): 15
Genpak LTD.
Kingsley Cox
25 East Loop Road
Stony Brook, NY 11790
P: 631-444-6625
F: 631-444-6616

BOOTH #:(s): 12
Amersham Pharmacia Biotech Inc.
Liz Crespo
800 Centennial Avenue
P.O. Box 1327
Piscataway, NJ 08855
P: 732-457-8111
F: 732-235-2201
liz.crespo@am.apbiotech.com

BOOTH #:(s): 13
MWG-Biotech, Inc.
Randy Gooch
4170 Mendenhall Oaks Parkway
Suite 160
High Point, NC 27265
P: 336-812-9995
F: 336-812-9983
sflood@mwgbiotech.com

BOOTH #:(s): 4
Nat'l Ctr. for Biotechnology Information
Vyvy Pham
Bldg. 38A, Room 8N805
8600 Rockville Pike
Bethesda, MD 20894
P: 301-496-2475
F: 301-480-9241
info@ncbi.nlm.nih.gov

BOOTH #:(s): 2
Proteome, Inc.
Ann Fancher
100 Cummings Center, Suite 435M
Beverly, MA 01915
P: 978-922-1643
F: 978-922-3971
amf@proteome.com

BOOTH #:(s): 1
QIAGEN, Inc.
Yolanda Reed
28159 Avenue Stanford
Valencia, CA 91355
P: 661-294-7900
F: 661-295-7652
seminars-us@qiagen.com

Exhibit Descriptions

Amersham Pharmacia Biotech provides the tools necessary for the discovery, development and delivery of tomorrow's biotechnology-based drug therapies. The company's expertise in DNA Sequencing, Microarray Analysis, Protein Expression, Chromatography, Drug Screening and Large-Scale Biomolecular Separations offer an invaluable service to customers in all phases of biotechnology research.

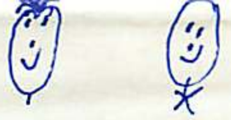
Genpak specializes in the development and supply of high quality reagents and equipment for use in all aspects of DNA analysis. Recent developments include: Automated DNA sequencing reagents, sequencing grade glass plates, fluorescent markers for genotyping and a fully automated high specification gene arraying robot.

MWG-Biotech is the world's fastest growing provider of high quality genomic services and instrumentation for molecular biology applications. MWG's sequencing facility guarantees our customers >99% accuracy while its proprietary salt-free PSF and HPSF processes provides the highest quality oligos available. MWG's molecular biology instruments fulfill today's needs and tomorrow's demands.

NCBI provides integrated access to DNA and protein sequence data, associated mapping data, protein structures, and MEDLINE. Demonstrations of the GenBank database, Entrez retrieval system, PubMed for MEDLINE searching, BLAST and VAST similarity searches for sequences and structures, and the Bankit and Sequin sequence submission software will be provided.

Proteome has developed the BioKnowedge Library^{TM}, a collection of comprehensively curated proteome databases for model organisms. Each volume includes essentially all experimental knowledge of protein properties and functions, organized as user-friendly Protein Reports and relational tables. Accompanying software tools allow powerful applications of the Library for analysis of functional genomics and proteomic experiments.

QIAGEN is dedicated to the development and distribution of high-quality products and services for molecular biology, clinical, and gene therapy research. QIAGEN offers a wide range of innovative products for PCR, transfection and the purification of DNA, RNA, and proteins. QIAGEN specializes in products optimized for both manual and automated processing.



www

D[illegible]her
Salzberg
Peterson
Leanne Camp
DAllenA

RFleischmann
- ① TB important ② H37RV ③ CDC1551
- IS6110 v. diff in copy #

F.Blattner
"Alternalogs" - codon usage differences - seem to be newly transferred DNA

TB
CJ
MY

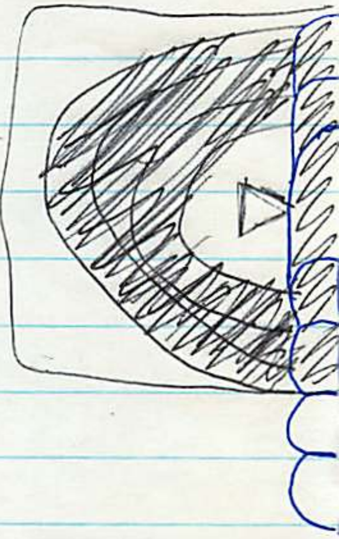
How good a correlation between T/ and unusual GC.

- 3 closely related to E. coli
EC / O157, Shigella, uropathogenic coli

- O157 1.3 mb novel relative to coli

0.5 Mb in K12 not O157
1.2 in O157 not K12
4.2 in both

- composition varies more in new regions of genome



- polymorphism - most clinical isolates v. similar to O157

- micromirror array

Pseudomonas

type I - Pseudomonas

P. aeruginosa   Ps. stutzeri.

P. fluores.

P. syringae

P. syringae

- high variability of pathovars syringae

P. stutzeri

- nitrate → N_2

- eight genomovars

P. putida

P. aeruginosa

- ubiq. in aquatic habitat

- pathogen of animals, plants, humans

- conserved gene - v. low variability

all clones in linkage equilibrium

- strong strand bias -

- all genes have high CAI

RMoxon

Even w/ whole genome sequences work on pathogens very hard.

N. meningitidis

showed plot of best hits

- still v. lethal even if diagnosis
- not known in any other host
- in respiratory tract - usually commensal

Annotated w/ homolog + non-homolog based

lots of repetitive DNA

Ignorance index 0.41   16% Cons. Hypo or unknown
25% no homology

MenB vs Other

.91% of genes likely orthologs
- most of the differences are in hypotheticals

obviously dinucleotide signature they came from somewhere else.

- 3 islands of horizontal transfer ... prob. not pathogenicity island - not in menA

Phenotypic variability

Liann McCue - wadsworth Ctr

Probe program → Classifyer

cyclic NMP Binding Proteins

-207 proteins in family

-2 proteins of known structure

classification is NOT phylogenetic

in one class - many Mtb proteins

Nucleotide cyclases

RV1625c ... similar to mammalian

Ian Paulsen

- # of transporters varies greatly between species

- most of this due to genome size

- Basca + E.coli somewhat higher. (6/100Kb vs 3/100Kb mean)

- 76 [illegible] families

B. Nierman

Caulobacter

- grows well in nutrient poor conditions

S. Andersson

Genome Reductions

- low influx
- high efflux
- high mutation /

- obligate vs facultative intrac- parasites
- .

R. prowazeki:

Bartonella henselae - cat scratch fever
- many repeats... close to complete

Francisella tularensis

Buchnera aphidicola

I just reviewed a paper by a japanese group on this genome.

Pseudogenes - can reconstruct likely ancestral genes

Mutational degradation
= continuous outflow of genes

Can it be compensated by transfer?
- v. little variation in GC, etc

- v. few examples of gene transfers

J. Parkhill

M. leprae is done - genome appears degraded

Things we found

- surprising the M. tb is highly stable

Campylobacter

- w/in shotgun - enormous variation in SSM repeats
- where are these tracts
- only variation in homopolymeric tracts
  NOT di- or tri- -repeats
- Campy + Helpy v. distant
  - rapid Δ from common ancestor

Neisseria

- lots of repeats

Caulobacter microarray

- predicted 3000 orfs
- cell cycle time course
- 462 cell cycle reg. transcripts
- Flagellar proteins

Induced

6 ribo prots
meth. metab
4 phosphate tport (G1)
flag. biosyn.
regul. prots
DNA metab.
pili biosyn.
cell division
unknown/conserved
metabolic enzymes/tporters

CtrA

looking for good looking CTRA boxes
a very good looking helix turn helix motif
that was hugely reproducible

G. Garrity

Defining higher taxonomic structure of prokaryotes

Two large scale trees

RDP - 7000 aligned sequences
ARB W. Zillig -

Want to know higher structure

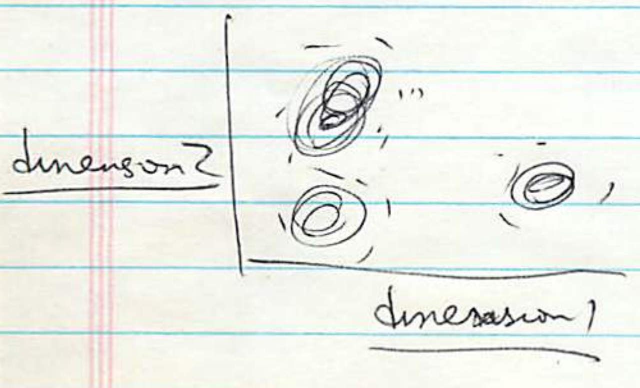
	Total	Bacteria	Archaea
Domains	2	1	1
Phyla	25	23	2
Classes	39	31	8

but does it work

Principal component analysis
- find taxa that are correlated w/ each other

has it ever been simulated

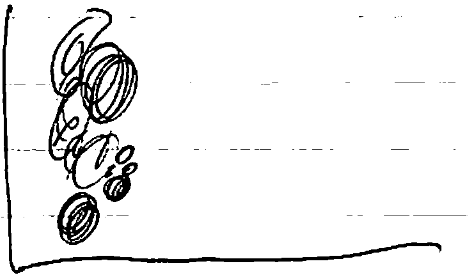
Methods
① 2444 rRNAs
② 184 families
③ evolution models
- many distances -
- pulled out matrices -- fed into ST



- dimension 1, 2 account for >85% of variance

-reduce # of taxa. -model works well

But this won't deal w/ branch



long branches

Two clades of spirochetes
leptospiras vs Treponemes

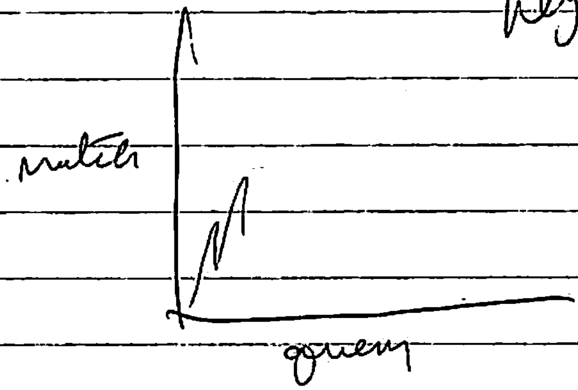
Automated classification for gene families

Exploratory Data Analysis

8.

S Salzberg

height = % ID



take out chr2+4 of AT

Seddy

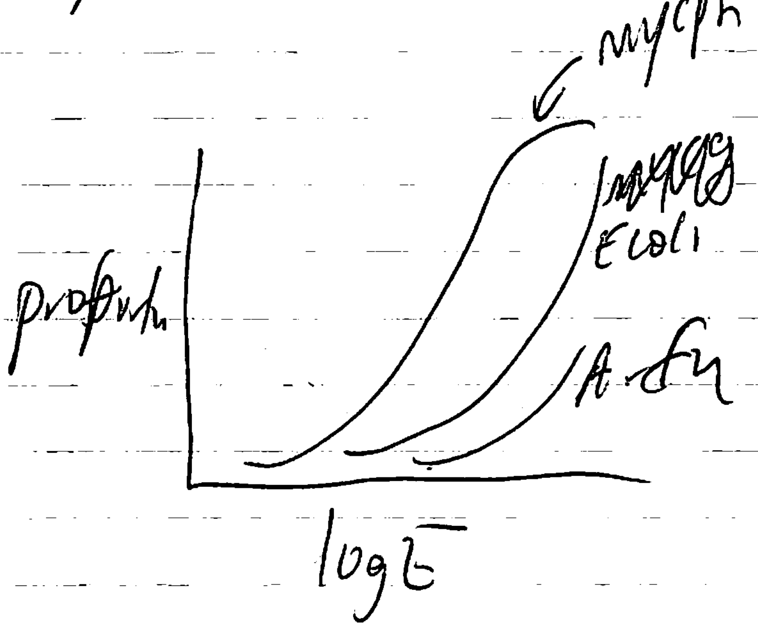
UHG - U22 host gene

- 8 snoRNAs in introns

PKarp Ecocyc

-Ecocyc   ecocyc.doubletwist.com

-Metacyc



MYCge as query

HM

LGT

How far can you run with it?

Brief history

Inferring LGT - the logic

Inferring - some examples

Redefining null hypothesis -

What are rules that govern transfer?

Is there a core of untransferable genes?

Radical model

Sfalkow 1975 - speculates role in biology

S. Sonea + M Panisset - universal gene pool

Horizontal transfer may be responsible for speciation

Inference

- want all together

- GC content, codon usage

- presence/absence - use parsimony inference

discordant trees

HMGCoA - trees, distribution, + signals

-

Taxtable

- bacterial genes in one archaea but not others

Redefining Null Hypothesis

- why estimates might be ~~high~~ low
  - close transfers hard to detect
  - amelioration
  - recurrent transfers not enough
  - distant transfers likely rarer… so we may be making low estimates

- rules governing transfer
  - importance of gene
  - informational vs. operational } but cell doesn't know
  - Marzels-sociology w[illegible]e- complexity/physical interactions
  - biochemical interactions
  - linkage to genes that are hard to transfer or easy to transfer … genes in pathway need to be transferred together
  - see Roth + Lawrence selfish operons essentiality … rare to be transferred b/c never disappear
  - susceptibility of product to Ab/toxins - SELECTION
  - neutral substitution
    - integration + random loss
  - genetic compatability
  - physiological compatability

- trees of genes in everything suggests there is no core.

-large scale phylogenetic patterns are created and maintained by LGT

All genes can be exchanged although at diff rates... close relatives exchange more frequently

Deep organismal phylogeny is impossible

-compared to ship of Thesaus.. replacing planks on ship of Thesues... when did it stop being same ship.

-reconstructing LCA genomes

-never needed to be a last common ancestral cell

RNA world - enclosed genomes - code - genes -

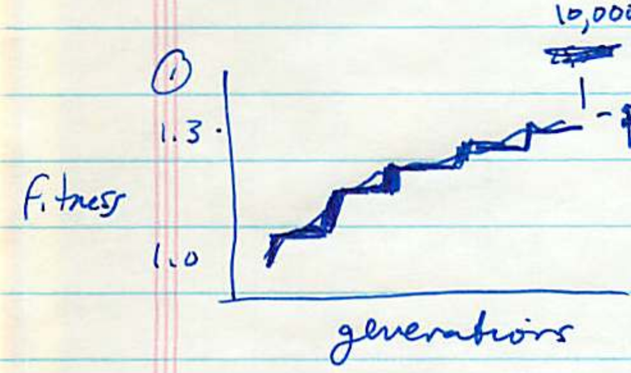
Richard Lenski - 20,000 E coli:

Evolutionary microscope

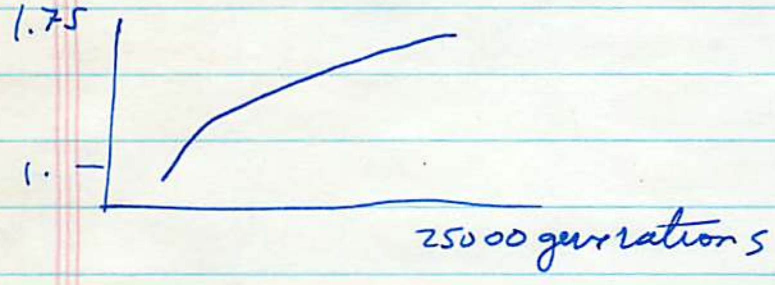
① Phenotypic Evolution

② Genomic Evolution

③ Future directions



- punctuated evolutionary dynamics prob. due to sweep of mutations



rate of improvement decreases over time

- in several populations you get mutator phenotypes
  - all caused by MMR mutations
  - mutator alleles do not confer direct advantage - they can hitchhike to fixation b/c more likely to generate beneficial mutation

Genome evolution

- RFLP analysis w/ IS's
- Genome evolution continues even as fitness changes decrease
- Rates of IS divergence varies even among strains w/ same fitness Δ's
- burst of IS50 -

Is fitness adaptive.

Some IS mutations go to fixation
- 11 characterized

All 12 lost ability to metabolize ribose - this is a hotspot for mutations

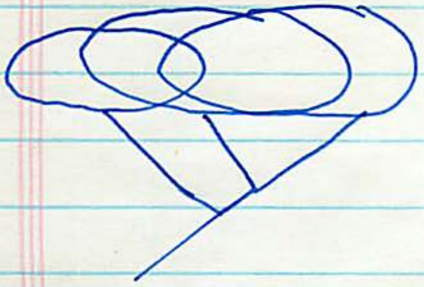
measuring rates of molecular evolution

- v. few pt. mutations found so far
- 26.5 subst. per genome per 1000 generations
0.9 IS tposition
0.4 IS recombination

How find adaptive mutations

Proteome Sets

-absent families count as characters



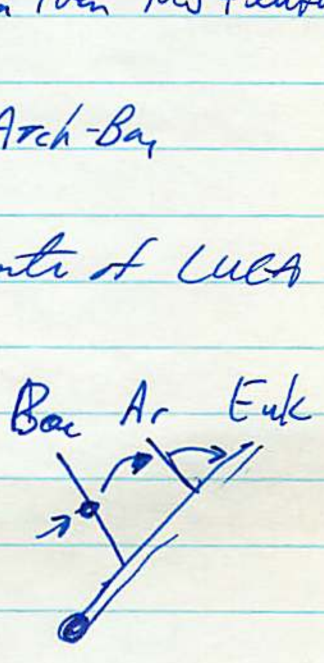
problem - this usually only deals w/ area of overlap among trees

~~Which life form contains characters most like LUCA~~

-obtain minimal gene set from LUCA

①history of studies

- TFIIB similar betw. euk + arch.
- phylogenetic distribution of positive patterns
- Met. jannasci
    - more genes shared betw. MJ + bacteria than MJ + eukarya
- functional classes
    - energy genes - either universal of Arch-Bac
    - information -
- universal families may provide estimate of LUCA



The *Streptomyces coelicolor* genome sequencing project at the Sanger Centre

S. Bentley, A-M Cerdeño, K. James, J. Parkhill, N. Thomson, K. Rutherford, S. Brown, A. Goble, D. Harris, J. Hidalgo, T. Hornsby, S. Howarth, L. Murphy, K. Oliver, S. Rutter, D. Saunders, K. Seeger,
R. Squares, S. Squares, K. Taylor, A. Cronin, M. Quail, J. Woodward and B. Barrell
The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK.

1) Abstract

Streptomyces coelicolor (A3)2 is widely accepted as the model organism for the genetic study of the high (G+C), Gram positive, filamentous actinomycetes. This group of microbes is of interest not only because they produce two-thirds of the known antibiotics of microbial origin but also for study of their complex life cycle and their ecological impact on soil microbiology. The project to sequence its 8 Mb linear genome makes use of an ordered set of cosmids which covers the entire chromosome. Here we describe the strategy applied to the sequencing and annotation of these cosmids and the dissemination of the information gathered. Some of the most interesting features and progress of the project are also presented.

2) The organism

Originally isolated from soil, *Streptomyces coelicolor* can be cultivated on media which allow the demonstration of some spectacular natural characteristics.

A typical colony takes about 4 days to mature, by which time it has produced a fluffy surface of aerial hyphae bearing spores which are often pigmented. The blue halo seen around these colonies is caused by secretion of the antibiotic actinorhodin.

Genetically manipulation of regulatory functions can cause increased levels of antibiotic synthesis leading to the formation of hydrophobic droplets on the colony surface.

Here actinorhodin appears as a red pigment (due to different media pH). The red areas are due to spontaneous deletions from the chromosome leaving colonies unable to produce the aerial hyphae which would normally mask the red colour.

Electron-microscopic examination reveals the chains of the spores which form at the ends of the aerial hyphae.

3) The genome

Streptomyces coelicolor has an extraordinary genome. Its 8 Mb chromosome is typical of the genus but unusually large compared to most other bacteria whose genomes tend be around 2 - 4 Mb. At about one gene per 1.1 kb, gene density is more typical giving a predicted total of more than 7000 genes - some 20% more than the eukaryotic yeast genome. Why the genome should be so large is a question which can hopefully be answered by this project.
Also unusual is the high G+C content (~71%) and the linear (rather than circular) structure of the chromosome with large terminal inverted repeats and a centrally positioned origin of replication.

4) Sequencing

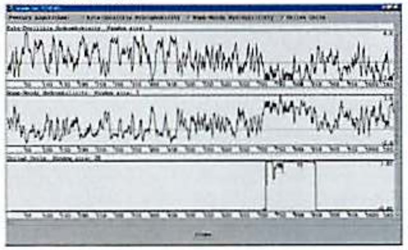
Cosmids proceed through a random (shotgun) phase followed by a directed (finishing) phase. To be accepted as "finished", sequence must have been produced from more than one subclone and be either double stranded or produced using two sequencing chemistries. The high G+C content of the DNA required some adaptation of protocols but the quality of sequence produced is such that an accuracy of at least 99.99% can be guaranteed. During sequence assembly any contig over 1000 bp is automatically added to the total DNA database which can be accessed via our FTP site or searched using a dedicated Blast server, both of which can be found on the project web pages at http://www.sanger.ac.uk/Projects/S_coelicolor

5) Annotation

Finished cosmid insert sequences are carefully analysed and annotated then immediately submitted to the EMBL databases at the European Bioinformatics Institute, an EMBL outstation and neighbour to the Sanger Centre. Various database searches and analysis programs are applied to the sequence. The output from many of these analyses can then be viewed directly using our sequence viewing/annotation software, Artemis (available for downloading from http://www.sanger.ac.uk/Software/Artemis/).

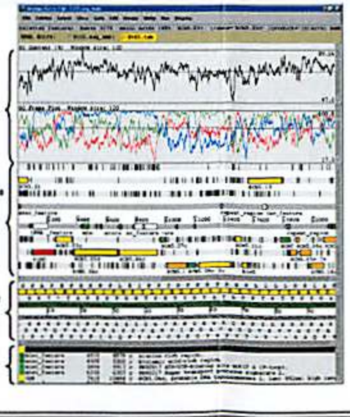
Artemis allows visualization of sequence features and the results of analyses within the context of the sequence and its six-frame translation. Written in Java, it reads EMBL format files, and can work on sequences of any size from a few kb to entire genomes of 5 Mb or more. Two views of the sequence are shown, both of which can be zoomed in to the base level, or zoomed out to display the entire sequence allowing features to be viewed in fine detail on in context with the flanking DNA and its features. There is also a list of features at the bottom of the window. In addition to this basic display, Artemis can plot the results of calculations on the sequence, or on any of the CDS features. The sequence plots are tied to the sequence display and scale with it as the zoom level is changed. For all of the plots the window size can be altered dynamically to suit the zoom level. In addition to the sequence viewing capabilities, Artemis can display the results of numerous analysis on top of the sequence; CDS predictions, BLASTN, in-frame BLASTX, tRNA and motif searches etc. can all be viewed and incorporated into the annotation. These are all run externally and the results parsed into EMBL-format, meaning that the results of any analysis can be easily incorporated without modification of the Artemis program itself. Artemis will also run analyses on sets of CDS features, such as FASTA and BLASTP searches, and allow the results to be viewed directly from the object selected.

Plots showing the hydrophobicity, hydrophilicity and coiled-coil potential across the protein sequence of the selected CDS.



Edit window for written annotation on the selected feature.

The main Artemis window.



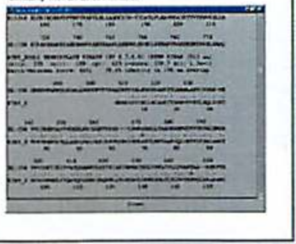
Plots of the results of calculations on the sequence showing %G+C content (upper window) and a Hidden-Markov Model prediction of coding sequences.

Overview window showing relative positions of genes (colour-coded), stop codons and various miscellaneous features such as Pfam and Prosite matches, repeat regions etc. Results of database searches and gene prediction programs can also be overlaid.

Base view is useful for looking at fine detail such as the position of start and stop codons, ribosome binding sites etc.

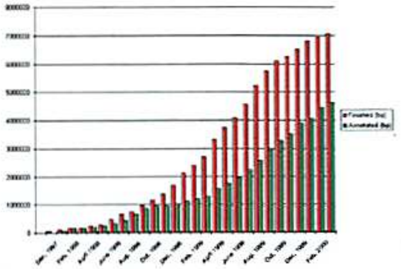
Feature table giving an ordered list of features marked on the sequence.

FastA output for the selected feature showing a region of similarity with an *E. coli* kinase.



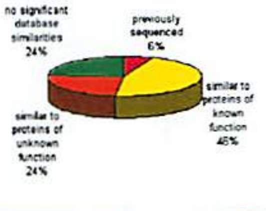
6) Progress

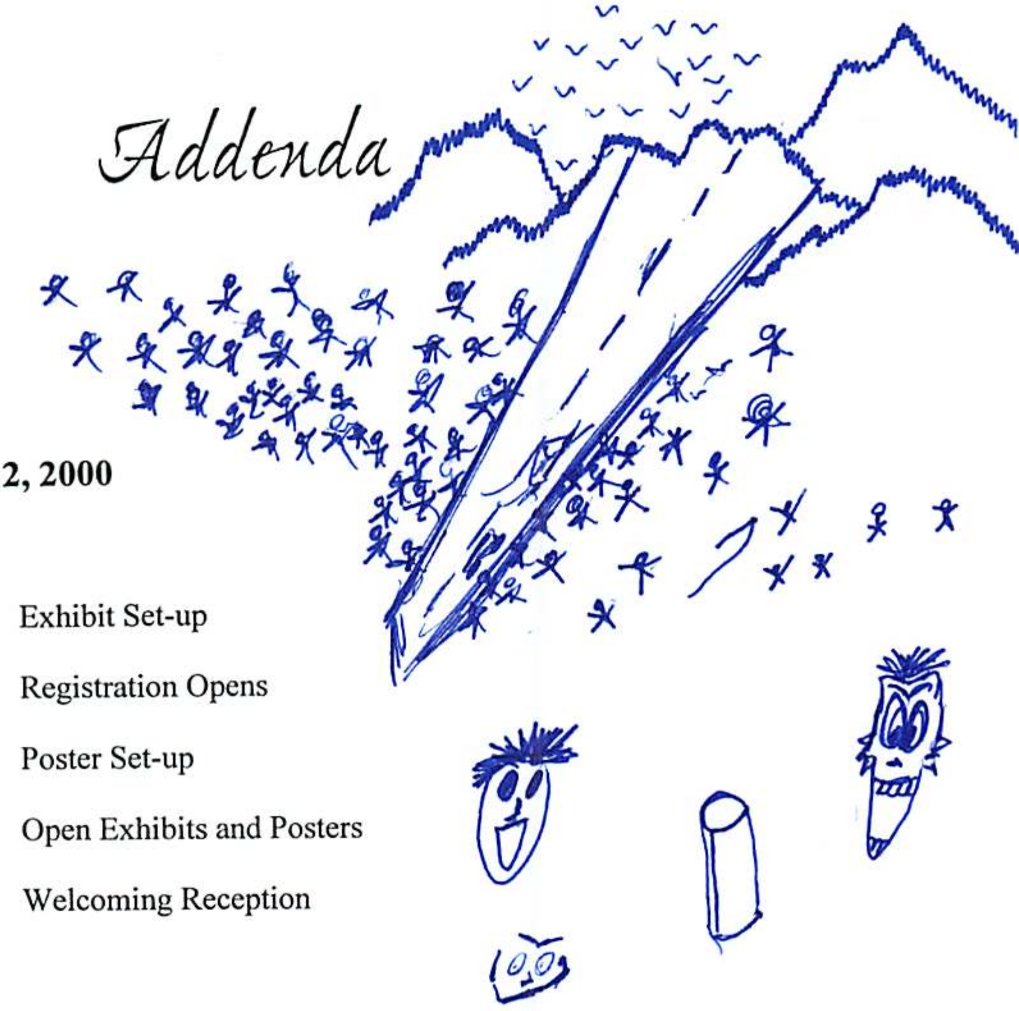
To date we have produced over 7 Mb of finished unique sequence, over 4.5 Mb of which is annotated and in the EMBL databases. The chart below shows the month-by-month progress of the project. The sequencing and annotation will be completed this year and will represent the largest bacterial genome to be sequenced so far.



7) Summary

With the chromosome being linear it is interesting to note that 57% of the genes are transcribed away from the centrally positioned origin of replication. The chromosome also seems to be arranged such that the more essential genes are located close the the origin with the chromosome ends appearing to have a decreased gene density.
Over 89% of the chromosome is coding DNA with an average CDS length of 975 bp (325 aa). The G+C content of coding DNA is slightly higher (72%) than that of non-coding DNA (69%). The chart below indicates the proportion of genes with regard to their database similarities. With over 7000 genes predicted, it is no surprise that a confident functional prediction is not possible for many of the annotated coding sequences. But why so many genes?
Perhaps this allows for greater diversity in enzymatic function enabling the organism to fully exploit the complex soil environment. A close interaction with the environment is also indicated by a high proportion of response regulators and membrane transport proteins.
One of the major rewards of genome sequencing is the information regarding "new" genes. Only 6% of the *S. coelicolor* gene have been previously sequenced so this project will produce sequence for over 6500 "new" genes.





Final Agenda

Saturday, February 12, 2000

Noon - 5:00 pm	Exhibit Set-up
3:00 pm	Registration Opens
3:00 - 5:45 pm	Poster Set-up
6:00 - 8:00 pm	Open Exhibits and Posters
6:30 - 8:00 pm	Welcoming Reception

Sunday, February 13, 2000

7:00 am	Breakfast
8:30 am – Noon	**Plenary Session I: Comparative Genomics** **CHAIR - Claire M. Fraser** **The Institute for Genomic Research, Rockville, MD**
8:30 am	**Robert Fleischmann,** The Institute for Genomic Research, Rockville, MD *Sequencing of The M. Tuberculosis Genome, Comparison of a Recent Clinical Isolate with the Laboratory Strain*
9:00 am	**Frederick Blattner,** University of Wisconsin, Madison, WI *Bacterial Pathogen Genomics*

9:30 am	**Burkhard Tummler**, Klinische Forschergruppe Hannover, GERMANY *Comparative Biology of Pseudomonas Species*
10:00 am	Break
10:30 am	**Richard Moxon**, Oxford University, UK *Use Of Whole Genome Sequence Of Neisseria Mexingetidid Serogroup (Menb) Strain Mc58 to Facilitate Understanding Of The Molecular Basis Of Its Pathogenicity*
11:00 am	**Lee Ann McCue**, NY State Department of Health, Albany, NY *Functional Classification of cNMP Binding Proteins and Nucleotide Cyclases*
11:30 am	**Ian Paulson**, The Institute for Genomic Research, Rockville, MD *Comparative Analysis of Microbial Transport Proteins*
10:00 am - 2:00 pm	Posters and Exhibits Open
12:30 pm	Lunch
2:00 - 5:30 pm	**Plenary Session 2: Genome Projects** **CHAIR - Frank Robb, University of Maryland, Baltimore, MD**
2:00 pm	**William Nierman**, The Institute for Genomic Research, Rockville, MD, **Janine Maddock**, University of Michigan, University of Michigan, Ann Arbor, MI *The Caulobacter Crescentus Genome Sequencing Project*
2:30 pm	**Timothy Read**, The Institute for Genomic Research, Rockville, MD *Bacillus Anthracis Genome Sequencing Project*
3:00 pm	**Siv G.E. Andersson**, University of Uppsala, SWEDEN *Comparative Genomics of Intracellular Parasites and Symbionts: Rickettsia, Bartonella, Francisella and Buchnera*
3:30 pm	Break
4:00 pm	**Julian Parkhill**, The Sanger Centre, Cambridge, UK
4:30 pm	**Susan Douglas**, Institute for Marine Bioseceinces, Halifax, Nova Scotia *The Guillardia theta (Cryptophyceae) Nucleomorph Sequencing Project*

5:00 pm	**Derek Lovley,** University of Massachusetts, Amherst, MA *Genome of Geobacter Sulfurreducens*
6:00 pm	Dinner
7:30 - 9:00 pm	Exhibits and Posters Session

Monday, February 14, 2000

7:30 am	Breakfast
8:30 am - Noon	**Plenary Session 3: Genome Biology** **CHAIR - Jennie C. Hunter-Cevera, University of Maryland Biotechnology Institute, College Park, MD**
8:30 am	**Les Baillie**, CBD, Porton Down, Salisbury, UK *Bacillus Anthracis, A bug with Attitude*
9:00 am	**Rino Rapuoli**, Chiron Corporation, Siena, ITALY *Novel Proteins For Vaccine Development From N. Meningitidis Genome*
9:30 am	**R. Frank Rosenzweig**, University of Florida, Gainesville, FL
10:00 am	Break
10:30 am	**David Alland**, Division of Infectious Disease, Bronx, NY *Molecular Beacons in Multiplex Formats: Susceptibility Testing and Identification in M. Tuberculosis*
11:00 am	**Michael Laub**, Stanford University, Palo Alto, CA *Dissecting the Temporal Regulation of Caulobacter Cell Cycle Progression with DNA Microarrays*
11:30 am	**Pierre Legrain**, HYBRIGENICS, Paris, FRANCE *Functional Proteomics On Microbial Genomes*
9:30 - 2:00 pm	Posters and Exhibits Open
Noon	Lunch

Noon - 2:00	Exhibits and Poster Session
2:00 - 5:00 pm	**Plenary Session 4: Genome Analysis** **CHAIR - Monica Riley, Marine Biological Laboratory, Woods Hole, MA**
2:00 pm	**George M. Garrity**, Bergey's Manual Trust, East Lansing, MI *Seeing The Forest Despite All The Trees Microbial Classification In The Genomic Era*
2:30 pm	**Steven Salzberg**, The Institute for Genomic Research, Rockville, MD *Algorithms for Whole Genome Analysis*
3:00 pm	**Sean Eddy**, Washington University School of Medicine, St. Louis, MO *Computational Screens for Noncoding RNA Genes*
3:30 pm	Break
4:00 pm	**Peter Karp**, SRI Internationale, Menlo Park, CA *Knowledge Based Modeling of the E. coli Metabolic Network*
4:30 pm	**William Pearson**, University of Virginia, Charlottesville, NC *FASTA: The Nest Generation Sequence Alignment Tool*
5:00 pm	**Jonathan Eisen**, The Institute for Genomic Research, Rockville, MD
6:00 pm	Dinner
	Free Evening

Tuesday, February 15, 2000

7:00 am	Breakfast
9:00 am - Noon	Exhibits Open (No Poster Sessions)
9:00 - 11:30 am	**Plenary Session 5: Patterns and Processes in Genome Evolution** **CHAIR - Siv Andersson - University of Uppsala - Sweden**

Addenda

9:00 am	**Ford Doolittle,** University of Halifax, Halifax, Nova Scotia *Assessing Horizontal Gene Transfer in Microbial Species*
9:30 am	**Richard Lenski,** Michigan State University, East Lansing, MI *Dynamics of Genomic and Phenotypic Evolution: A 20,000 Generation Experiment with E. coli*
10:00 am	Break
10:30 am	**Christos Ouzonis,** European Molecular Biology Laboratory (EMBL), Cambridge, UK *Metabolic Networks In the Last Common Ancestor*
11:00 am	**Mary Carrington,** National Cancer Institute- FCRDC, Frederick, MD *Genetic Susceptibility to Infectious Diseaes in Humans*
Noon - 3:00 pm	Breakdown Exhibit Area
12:30 pm	Lunch Meeting Adjourns

Poster Abstracts

P-14
Whole Genome Sequencing of *Shewanella Oneidensis*

J. Weidman, A. Wolf M. Placide, J. Vamatheven, H. Qin, H. Khouri, J.F. Heidelberg, The Institute for Genomic Research, Rockville, MD

Shewanella oneidensis grows both aerobically and anaerobically. In the anaerobic phase, it acts as a metal reducer. The use of metal reducing bacteria in bioremediation has several advantages over more standard respiring bacteria, including: 1) their substrates (iron oxides) are solids and thereby can be delivered to a contaminated site without diffusing away, 2) Iron oxides are specific substrates, so competition from other bacteria for the electron acceptor will be minimum, 3)in stratified aqueous environments, reduced iron should diffuse upward, be reoxidized by molecular oxygen in the oxic zone, and return to the anoxic zone by gravity, thus acting as a "pump" for oxidizing equivalents.

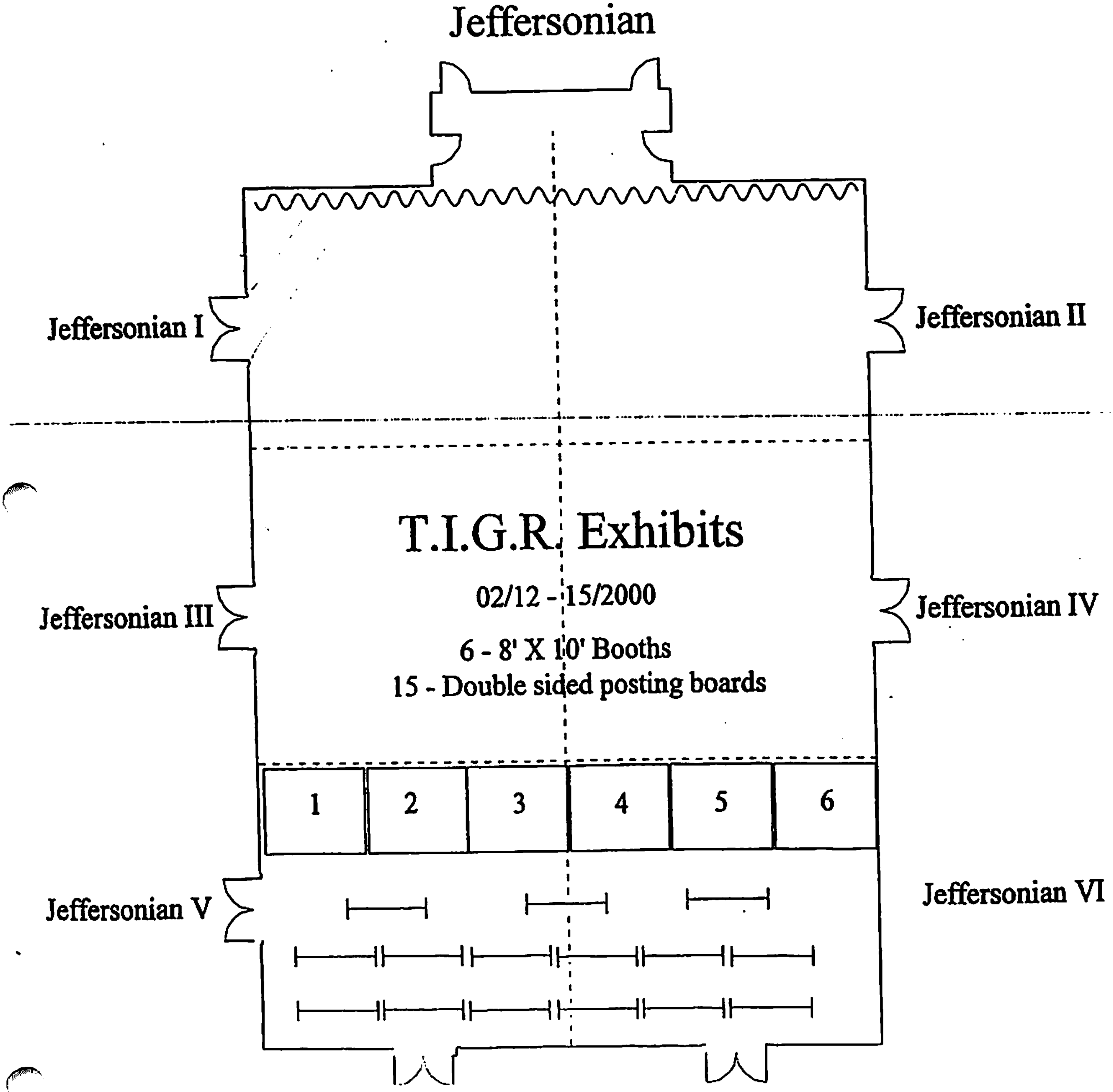
The 5 Mb *S. oneidensis* MR-1 genome is being sequenced by the random whole genome strategy used to complete the sequence of multiple bacterial genomes at TIGR. The random sequencing phase of the project resulted in 70,000 sequences. These were assembled into 125 linked contigs. Currently, gaps are being closed using a combination of PCR and small and large insert clone walking.

P-27
Intraspecific and Interspecific Fractal Analyses of Microbial Genome Sequences

Hanif Khalak, The Institute for Genomic Research, Rockville, MD

A fractal technique has been applied to analyze microbial DNA sequences in computing local composition and large-scale correlations within increasing windows of genomic sequence. This involved application of oligomer frequency operators along a genomic sequence yielding 1-D signals which are subsequently analyzed as time series using fractal methods.

Results from application to several completely sequenced genomes compare calculations of fractal dimensions and autocorrelation measures along the genomic sequence in relation to observed gene density, regulatory signals, and genomic structural features.



T.I.G.R. Exhibits

02/12 - 15/2000

6 - 8' X 10' Booths
15 - Double sided posting boards