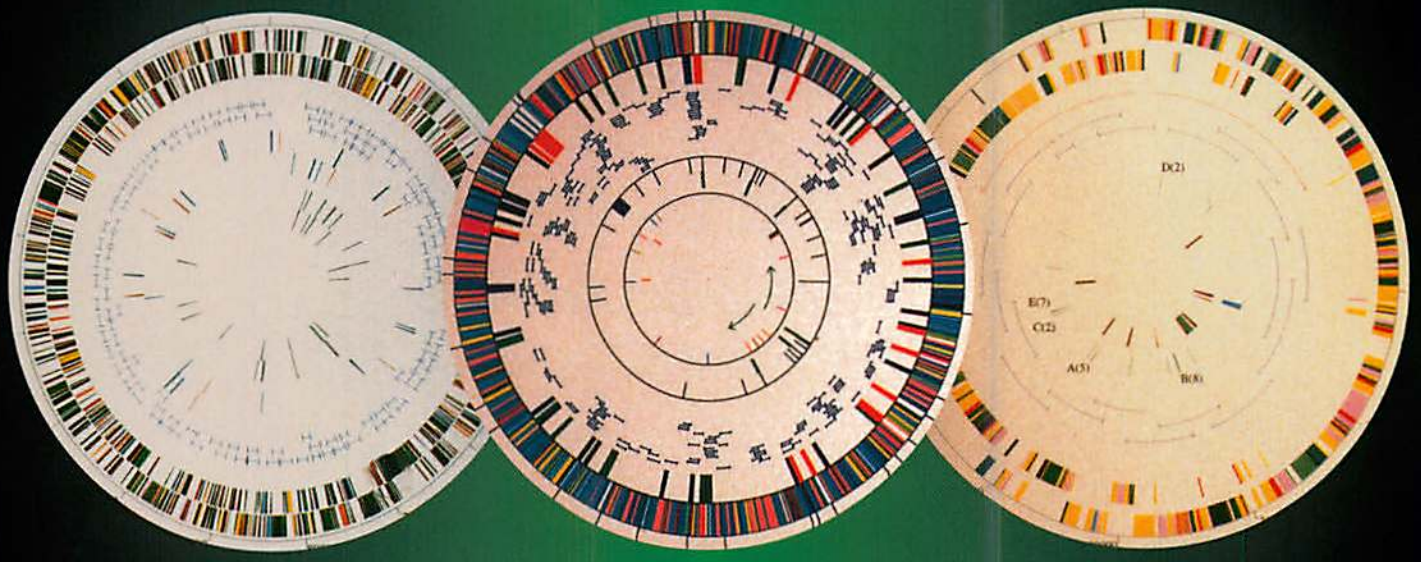
Volume 3, Number 4, 1998
ISSN 1090-6592

# Microbial & Comparative Genomics

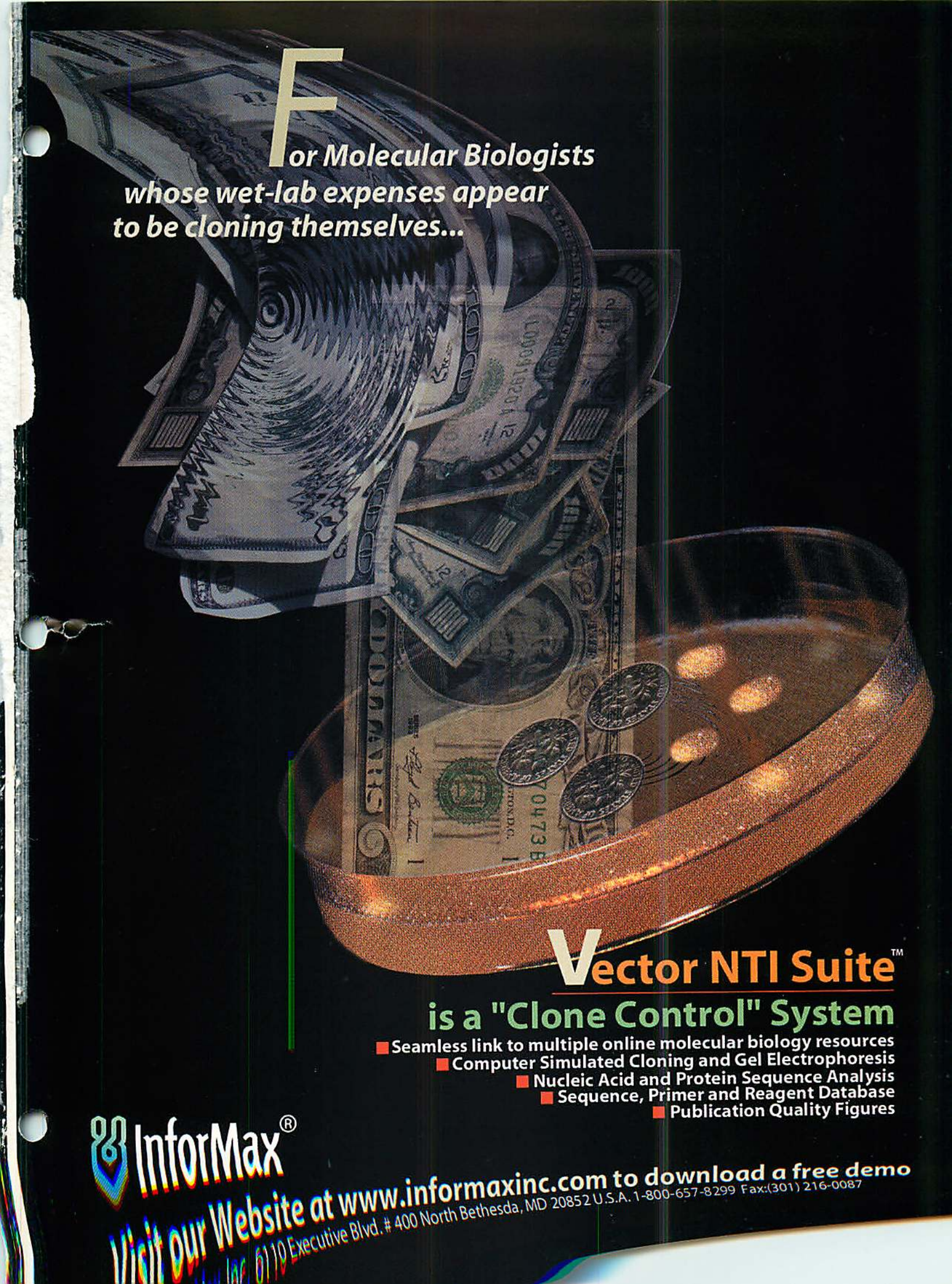


## Program & Abstracts



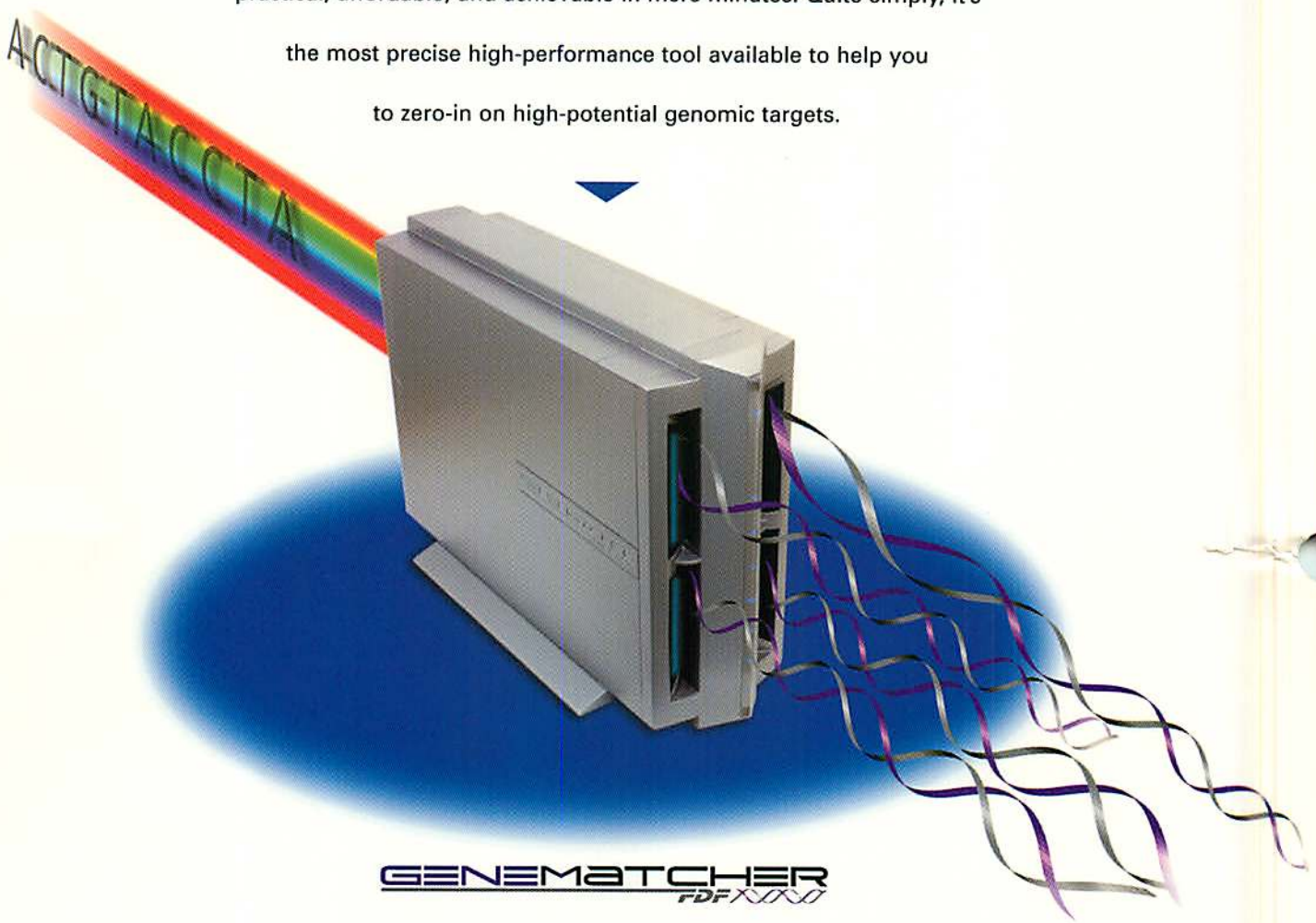
Microbial Genomes III: Sequencing, Functional Characterization and Comparative Genomics
January 29 - February 1, 1999

Mary Ann Liebert, Inc. publishers


For Molecular Biologists
whose wet-lab expenses appear
to be cloning themselves...
Vector NTI Suite™
is a "Clone Control" System
Seamless link to multiple online molecular biology resources
Computer Simulated Cloning and Gel Electrophoresis
Nucleic Acid and Protein Sequence Analysis
Sequence, Primer and Reagent Database
Publication Quality Figures
InforMax®
Visit our Website at www.informaxinc.com to download a free demo
6110 Executive Blvd. # 400 North Bethesda, MD 20852 U.S.A. 1-800-657-8299 Fax:(301) 216-0087

# THINK FAST.
# NOW, THINK FASTER.

In terms of sensitivity, selectivity and speed, no other analytical search engine even comes close to the remarkable Paracel FDF 4 GeneMatcher. Now, huge database searches are practical, affordable, and achievable in mere minutes. Quite simply, it's the most precise high-performance tool available to help you to zero-in on high-potential genomic targets.




GeneMatcher, Paracel's newest genetic data analysis system, answers your need for uncompromised homology searches.

Scalable up to more than 6,000 parallel processors, implemented in an Application Specific Integrated Circuit (ASIC), GeneMatcher provides unprecedented computing power for the most sensitive and selective search algorithms.

PARACEL™
80 S. LAKE AVE., SUITE 650, PASADENA, CA 91101-2616
VOICE 626.744.2000 • FAX: 626.744.2001 • WWW.PARACEL.COM

3rd Annual
Microbial Genomes Conference

# Addenda

## Session Chairs and Monday Speaker Change

### Saturday, January 30, 1999

**8:30 am - 12:05 pm** **Plenary Session I: Genome Biology**
Chair: **André Louis Goffeau**–Université Catholique de Louvain

**2:00 - 5:30 pm** **Plenary Session 2: Genome Projects I**
Chair: **Frank Kunst**–Institut Pasteur

### Sunday, January 31, 1999

**8:30 am - Noon** **Plenary Session 3: Comparative Genomics**
Chair: **Monica Riley**–Marine Biological Laboratory

**2:00 - 5:00 pm** **Plenary Session 4: Genome Biology II**
Chair: **Brendan W. Wren**–St. Bartholomew's Hospital

### Monday, February 1, 1999

**9:00 - 11:10 am** **Plenary Session 5: Genome Analysis I**
Chair: **Terry Gaasterland**–The Rockefeller University

**11:45 am** **Updates**
Chair: **Claire M. Fraser**–The Institute for Genomic Research

**Philip Harriman**, National Science Foundation replaces Mary Clutter

*The sponsors of the Distinguished Leaders in the Life Sciences lecture series (TIGR, US Department of Energy, and National Research Council) are pleased to announce that the February 17 speaker is:*

**Jane Goodall, Ph.D.**
The Jane Goodall Institute
"Of Chimpanzees, Humans, and Habitat"

*These lectures are free and open to the public and are held in the auditorium of the National Academy of Sciences. Dr. Goodall's lecture will begin at 7:00 pm.*

## Conference Event Location
**(Second Floor unless Specified)**

**Welcome Reception 6:00-8:00 pm** - Promenade
This reception is the conference-sponsored meal on Friday evening. Dining room service is available to conference attendees at their own expense.

All **Plenary Sessions** - Grand Dominion Ballroom

**Exhibits** - Jeffersonian II, IV, VI

**Posters (Saturday)** - Jeffersonian II, IV, VI; and I, III, V

**Posters (Sunday and Monday)** - Jeffersonian II, IV, VI; and Promenade

**Sunday Lunch** - Washingtonian Ballroom, 1st Floor

**TIGR's Superbowl Party** - Washingtonian Ballroom, 1st Floor. This party is the conference-sponsored dinner on Sunday evening. Dining room service is available to conference attendees at their own expense.

*Meals are served in the dining rooms on the second floor unless otherwise specified.*

## Speaker Abstract

### Phyletic and Functional Patterns of ORF Distribution among Whole Genomes

**Terry Gaasterland**[1] and Mark A. Ragan[2]. [1]The Rockefeller University, NY; and [2]IMB-NRC, CANADA.

Using the methodology introduced in Gaasterland & Ragan, 1998, we assign genomic signatures to ORFs across 10 microbial genomes. We augment the tables of signatures with functional category annotations, gene product annotations, genomic location, and metabolic pathway information. The resulting cross-genome database provides a rich environment for computing answers to statistical queries about phyletic and functional patterns of ORF distribution. It also enables the generation of tables of ORFs that meet particular phyletic distributions, organized by functional category or by pathway. A genome is contained in a genomic signature for an ORF at some score level if it contains a detectable counterpart for that ORF at or above that level. Users define a level by imposing threshold constraints on similarity matches (e.g., on probability, percentage of query sequence involved in the match, percent identity). We consider an ORF to be characteristic of a set of genomes if it has a detectable counterpart in each of those genomes and in no other genome. For each ORF in 10 genomes, we computed and queried the signatures at three increasingly strict levels to identify gene products characteristic of bacteria, characteristic of archae, characteristic of prokaryotes, and characteristic of prokaryotes and yeast. Using this cross-genome database, we

assessed over-representation and under-representation of functional categories among ORFs in each group. Detailed results are given in two companion papers (Gaasterland & Ragan, 1999; Ragan & Gaasterland, 1999).

This talk uses the *Aquifex aeolicus* genome to illustrate specific results of our study. The cross-genome database leading to our results is available through the website maintained at http://genomes.rockefeller.edu.

T. Gaasterland and M. Ragan, (1998), Constructing Multigenome Views of Whole Microbial Genomes, Journal of Microbial and Comparative Genomics, 3(3):177-192.

T. Gaasterland and M. Ragan, (1999), Microbial Genescapes: Phyletic and Functional Patterns of ORF Distribution among Prokaryotes, Journal of Microbial and Comparative Genomics, 3(4):199-217.

M. Ragan and T. Gaasterland, (1999), Microbial Genescapes: A Prokaryotic View of the Yeast Genome, Journal of Microbial and Comparative Genomics, 3(4): 219-235.

# Poster Abstracts

## GenomeList

L. Jones[1], I. Moszer[2], M. Klaerr-Blanchard[2], C. Medigue[3], A. Viari[4], A. Danchin[2]. [1]Institut Pasteur, Service Informatique Scientifique, Paris; [2] Institut Pasteur, Unité de l'Expression Genetique, Paris; [3]Atelier de BioInformatique et Institut Pasteur, Unité de l'Expression Genetique, Paris; and [4]Atelier de BioInformatique, Université de Paris VII, Paris, FRANCE.

GenomeList is a Web based microbial genome browser. It allows for the search in a microbial genome by various criteria. These criteria include the search by gene name, by keyword phrases, by region (either a defined region (physical location or mapped location) or a region around a given gene), by functional category or classification, by blast or fasta, and by pattern or motif. An extended search capability has been added which permits the search based on multiple criteria including the size and length, the location (mapped or physical), molecular weight, and specified text field search. The results are displayed in form of a gene list with user selected display of gene data or a drawing of a region. The user can display or download the protein or DNA sequence for a list of genes or the DNA sequence of a region.

There are three genomes currently available:

| | |
|---|---|
| *Bacillus subtilis* | http://www.pasteur.fr/Bio/SubtiList |
| *Escherichia coli* | http://www.pasteur.fr/Bio/Colibri |
| *Mycobacterium tuberculosis* | http://www.pasteur.fr/Bio/TubercuList |

## Genomics of the Radioresistant Bacterium *Deinococcus radiodurans*

Kira S. Makarova[1,2], Michael J. Daly[1], L. Aravind[2], O. White[3], Eugene V. Koonin[2] and Kenneth W. Minton[1]. [1]Uniformed Services University of the Health Sciences, Bethesda, MD; [2]The National Center for Biotechnology Information, The National Institutes of Health, Bethesda, MD; and [3]The Institute of Genome Research, Rockville, MD.

The bacterium *Deinococcus radiodurans* is the most resistant to ionizing radiation organism discovered to date. Although it is known that this resistance stems from superlative DNA repair processes, the mechanisms of this repair remain only partially understood, despite intensive investigation. Towards this end the entire genome sequence of this organism has recently been obtained by The Institute for Genome Research (TIGR). We have annotated this sequence, with special attention to the genes that might mediate novel DNA repair functions. Features noted to date include several proteins with unique domain organization and proteins have been horizontally transferred and discovered in bacteria for the first time. Given to domain composition some of these proteins might be involved in DNA-repair and recombination processes. Furthermore, we found three distinct desiccation-resistance proteins similar those seen in plants. This is of particular interest, as there is known positive correlation between deinococcal desiccation-resistance and radioresistance. Finally, an expansion of several protein families, including phosphatases, proteases, acyl transferases and MutT pyrophosphohydrolases family, was detected. In addition, *Deinococcus radiodurans* genome is extraordinarily rich in repetitive sequences, suggesting a mechanism of recombinational repair that involves these repeats.

## Comparative Phylogenetic Analysis across Genomes

Gerard Manning, Laurie Thorbjornsen and Glen Wilder. Molecular Applications Group, Palo Alto, CA.

All genes have relatives, and an evolutionary history. We wish to place all genes within families and study their relatedness to predict functional attributes; we are treating evolution as a multi-billion year experiment in mutation and selection, whose results can be read in modern-day genomes.

We are clustering related genes from a set of 15 microbial genomes, using standard homology-detection tools, hidden Markov models and threading, to create a hierarchical set of gene families, from strict orthologs to super-families, containing most genes in these genomes.

From these we create phylogenetic trees relating family members. These trees contain both an evolutionary trace of each species represented, as well as gene-specific events, such as gene loss and duplication, horizontal transfer, and the fusion and splitting of genes. We are building tools to tease apart these various forces that guide the evolution of all genes.

Once gene-specific events are accounted for, we create a consensus evolutionary tree derived from hundreds of gene families. This acts as an estimated species tree, and allows comparisons to individual gene trees to determine relative rates of evolution.

Tree comparison gives a measure of the evolutionary plasticity of each tree, and highlights branches of unusual length, indicating strong functional pressure for change within a specific lineage.

## Genomic Sequencing of *Chlamydia trachomatis* Mouse Pneumonitis Serovar Reveals Potential Virulence Determinants

T. Read[1], R.C. Brunham[2], G. McClarty[2], T. Utterback[1], S. Gill[1], B. Craven[1], C. Fraser[1] and the [1]TIGR Sequencing Core Facility. [1]The Institute for Genomic Research, Rockville, MD; and [2]The University of Manitoba, Winnipeg, Manitoba, CANADA.

The genome sequence of the *Chlamydia trachomatis* Mouse Pneumonitis (MoPn) serovar is more than 95% completed. A 1.06 Mb sequence, including two gaps, encodes about 900 ORFs. We compared MoPn data with the recently published genome sequence of *C. trachomatis* serovar D (1.04 Mb), which causes human trachoma, in order to identify genetic variation that might relate to the strains' altered pathogenicity. Most differences between MoPn and serovar D were localized to a region between the conserved *ycfV* (ABC transporter ATPase) and *dsbB* (disulphide bond reductase) genes. In this region, the MoPn strain contained four large ORFs homologous to *Escherichia coli* cytotoxin B as well as two genes encoding purine nucleoside synthesis enzymes, absent in *C.trachomatis* serovar D. Conversely, the serovar D strain had a partial tryptophan biosynthesis operon not present in the MoPn genome. The *ycf dsbB* region also contains several divergent genes encoding produc with a conserved HKD motif, possibly of the phospholipase D (PLD) superfamily. There are four such paralogs at this position in the serovar D, seven in the MoPn strain. Outside the *ycfV-dsbB* region

the MoPn genome is remarkably similar to serovar D in terms of ORF order and identity. Differences can be attributed to rare insertion of genes apparently unrelated to pathogenicity in MoPn (DNA helicase, uracil phosporibosyltransferase) and gene duplications. Type III secretion chaperone, translocase and effector genes, likely virulence determinants, are conserved between the two *C. trachomatis* strains. Although the function of most of the genes in the *ycfV-dsbB* region is unknown, their presence in this variable location of otherwise highly conserved *C. trachomatis* genomes suggests involvement in strain-specific pathogenicity.

### Rapid Detection of Antibiotic Resistance-associated Mutations in 10 Gene Targets in *Mycobacterium tuberculosis* Using the OpenGene™ System

Diane Nuesca[1], Sara Jacob[1], Frances Jamieson[2], George Broukhanski[2] and **Rob Shipman**[1]. [1]Visible Genetics Inc.; and the [2]Ontario Ministry of Health, Toronto, Ontario, CANADA.

Rapid and accurate detection of antibiotic resistance in *Mycobacterium tuberculosis* would greatly improve both patient treatment and outcome. Amplification and cycle sequencing primer sets have been developed for the detection and analysis of antibiotic resistance-associated mutations in defined regions of the rpoB (rifampin), katG (isoniazid), oxyR-ahpC PR (isoniazid), fabG (isoniazid), rpsL/s12 (streptomycin), 16S/rrs (streptomycin), embB (ethambutol), pncA (pyrazinamide), gyrA (ciprofloxacin) and 23S (azithromycin) genes. Using these primer sets and the OpenGene™ automated DNA sequencing system, a protocol has been developed which permits both the rapid identification of *M.tuberculosis* and the detection of antibiotic resistance-associated mutations in a series of gene targets. In a panel of five phenoptypic streptomycin-resistant *M.tuberculosis* isolates, antibiotic resistance-associated mutations were detected in the rpsL/s12 gene in four isolates. Parallel antibiotic resistance genotypes to rifampin, isoniazid, ethambutol, pyrazinamde and ciprofloxacin were also identified in the five steptomycin-resistant isolates. The protocol has been designed for use with both AFB smear-positive and smear-negative samples which should enable firstly, the identification of *M.tuberculosis* and secondly, the detection of antibiotic resistance-associated mutations in these samples in advance of standard culture-based methods.

## GSAC XI

The Eleventh International Genome Sequencing and Analysis Conference will be held Saturday, September 18 through Tuesday, September 21, 1999 at the Fontainbleau Hotel in Miami Beach, Florida.

If you plan to attend, we suggest that you make your room reservation now if you wish to stay at the Fontainbleau. We have a limited number of rooms in our room block and many were reserved early and quickly last year. The phone number for room reservations is: 305-538-2000 and the FAX number is: 305-674-4607.

Please check our web site; http://www.tigr.org for conference updates in a few weeks.

## Poster Abstract Correction

ADD Bruce Vincent as a co-author to the poster of Bradley Slaven, Poster Number 51, *Distributed Homology Searches using PVM and WU-Blast on a Heterogenous Computer Network*

## Poster Assignments

### FORMAL POSTER SESSION
### Saturday, January 30, 1999
### 7:30 pm

### THE FOLLOWING ABSTRACTS ARE POSTED IN THE EXHIBIT AREA:

| | Poster Number |
|---|---|
| **Alterovitz**, Gil | 01 |
| **Amano**, Naoki | 02 |
| **Azuma**, Yoshinao | 03 |
| **Bentley**, Stephen | 04 |
| **Brettin**, Thomas S. | 05 |
| **Cahill**, Dolores | 06 |
| **Clayton**, Rebecca A. | 07 |
| **Eroshkin**, Alexey M. | 08 |
| **Gill**, Steven R. | 09 |
| **Goodner**, Brad W. | 10 |
| **Govindarajan**, Sridhar | 11 |
| **Guiliano**, David B. | 12 |
| **Harris**, David | 13 |
| **Hartsch**, Thomas | 14 |
| **Heilig**, Roland | 15 |
| **Hickey**, Erin K. | 16 |

### THE FOLLOWING ABSTRACTS ARE POSTED IN THE BALLROOM ADJOINING THE EXHIBIT AREA AND WILL MOVE TO THE PROMENADE STARTING SUNDAY MORNING:

| | Poster Number |
|---|---|
| **Hide**, Winston | 17 |
| **Hihara**, Satoshi | 18 |
| **Hinkle**, Gregory | 19 |
| **Hogue**, Christopher | 20 |
| **Holtzapple**, Erik | 21 |
| **Jacobi**, Carsten | 22 |
| **Jones**, L. | 23 |
| **Kawarabayasi**, Yutaka | 24 |
| **Ketchum**, Karen A. | 25 |
| **Khouri**, Hoda M. | 26 |
| **Larimer**, Frank | 27 |
| **Lawson**, Daniel | 28 |
| **Li**, Guangshan | 29 |
| **Liu**, Jia Yeu | 30 |
| **Machida**, Masayuki | 31 |
| **Makarova**,Kira S. | 32 |
| **Gerard Manning** | 33 |
| **Michalickova**, Katerina | 34 |
| **Mungall**, Karen | 35 |
| **Murphy**, Lee | 36 |
| **Ohfuku**, Yuko | 37 |
| **Pallen**, Mark J. | 38 |
| **Parkhill**, Julian | 39 |
| **Peterson**, Scott | 40 |
| **Pham**, Xuan-Quynh T. | 41 |
| **Qi**, Rong | 42 |
| **Qin**, Haiying | 43 |

**Qiu**, J. 44
**Raleigh**, Elisabeth A. 45
**Read**, T. 46
**Reynolds**, Thomas R. 47
**Shipman**, Rob 48
**Simons**, Guus 49
**Slaven**, Bradley E. 50
**Stoker**, Neil G. 51
**Movahedzadeh**, Farahnaz 52
**Takami**, Hideto 53
**Tatusova**, T. 54
**Vamathevan**, Jessica J. 55
**Ussery**, David 56
**Viari**, A. 57
**Whiteside**, Simon T. 58
**Williams**, Kerstin 59
**Wood**, Todd C. 60
**Wu**, Thomas D. 61
**Zengel**, Janice M. 62

## Additional Registrants

**Daniel Drell**
Biologist
Office of Health and Environmental Research
US Department of Energy
ER-72/GTN
19901 Germantown Road
Germantown, MD 20874-1290
301-903-4742
FAX: 301-903-8521
daniel.drell@oer.doe.gov

**Robin Cline**
Research Associate III
Functional Genomics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319
301-838-0200
FAX: 301-838-0208
rtcline@tigr.org

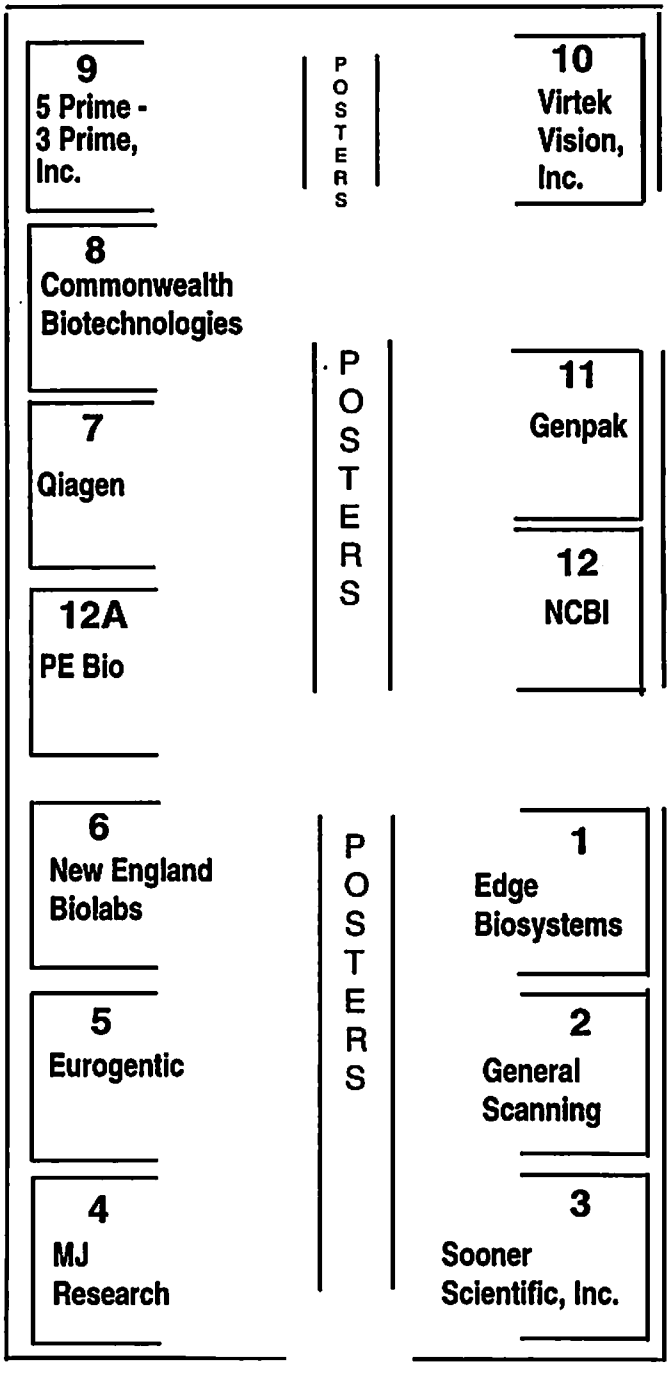
**Anthony Jung**
GenBank Indexer
National Center for Biotechnology Information
8600 Rockville Pike
Bethesda, MD 20894
301-496-2475
FAX: 301-480-9241
info@ncbi.nih.gov

**Leigh Riley**
GenBank Indexer
National Center for Biotechnology Information
8600 Rockville Pike
Bldg. 38A/Room 8N803
Bethesda, MD 20894
301-496-2475
FAX: 301-480-9241
info@ncbi.nlm.nih.gov

**Gulriz Kurban**
Student
Computer Science
Rockefeller University
1230 York Avenue
Gaasterland Lab
New York, NY 10021
212-327-7785
FAX: 212-327-7765
gulriz@cs.uchicago.edu

**Alexander Sczyrba**
Research Assistant
Computation Genomics
Rockefeller University
1230 York Avenue
New York, NY 10021
212-327-7785
FAX: 212-327-7765
asczyrba@willa.rockefeller.edu

# Exhibit Area Layout with Posters



# Exhibit Descriptions

**Commonwealth Biotechnologies, Inc. (CBI)** is a full service Contract Research Organization specializing in total project design and management for the biotech industry. CBI has the most comprehensive array of technologies in the industry today, all supported by a highly skilled, innovative group of research scientists. CBI is especially pleased to be introducing its newest product at this conference, a lane tracking technology for DNA sequencing gels which will be marketed as AccuTrac™.

**Contact person:** Thomas Reynolds
Sr. Vice President, Co-Founder
601 Biotech Drive
Richmond, VA 23235
800--735-9224
804-648-2641 FAX

**Edge Biosystems** has Edge Seq-To-Clone which will send you the 5 longest clones found for the lowest fee in the market. We also do Custom cDNA Libraries or offer UNAMPLIFIED Libraries at the price of Amplified Libraries. Also, we make 96-well Gel Filtration Blocks as well as cartridges to remove Dye Terminator and a new Quick Step PCR Purification Kit.

**Contact person:** Bill Miller
Sales Manager
19208 Orbit Drive
Gaithersburg, MD 20879
301-990-2685
301-990-0881

**Eurogentec** is a company in the field of custom tools for the life sciences, with a product portfolio containing custom DNA arrays, custom oligonucleotide synthesis, custom peptide synthesis, and custom antibody production. Services include a DNA sequencing service, transgenic services and contract research. Eurogentec is present mainly in Europe, with subsidiaries in France, Belgium, Germany, the Netherlands, Switzerland, Austria, the UK and Scandinavia. The company has recently opened an office in the US.

**Contact person:** Erik M. Stokhof
Public Relations Manager
Parc Scientifique du Sart Tilman
Seraing 4102 BELGIUM
32 4 366 61 35
32 4 365 51 03 FAX
info@eurogentec.be

**General Scanning, Inc.** **ScanArray 3000®** With 30 years of experience in the development of laser systems, General Scanning presents the ScanArray® 3000, a confocal fluorescent imaging system that captures the full potential of microarray biochips. General Scanning has specifically designed the ScanArray 3000 to meet the challenges of multi-color microarray detection and analysis.

**Contact person:** Michael Megginson
Sales Representative
500 Arsenal Street
Watertown, MA 02472
617-924-1010
617-924-7250 FAX

**Genpak** is a specialist supplier of DNA sequencing and mapping reagents. half TERM™, half BD™ and half TSLL™ reduces the cost of your sequencing reactions on **all** the major gel and capillary sequencing platforms, yet accuracy is maintained and in many instances the quality can be improved. We also supply reagents for Genotyping and Differential Display and PCR Cloning.

**Contact person:** Kingsley Cox
25 East Loop Rd.
Stony Brook, NY 11790-3350
516-444-6625
516-444-6626 FAX
k.cox@genpakdna.com

**MJ Research** will display MJR's full line of Peltier Thermal Cyclers for DNA amplification and cycle sequencing, including the new high-capacity Tetrad™ cycler (four microplates), the versatile DNA Engine™ with interchangeable blocks, the portable MiniCycler™, and the venerable PTC-100™. Instruments are available to fit 96-well, 0.5ml & 0.2ml tubes, and microscope slides for in-situ reactions.

**Contact person:** Steve Clougherty
Marketing Assistant
149 Grove Street
Watertown, MA 02472
617-972-8143
617-923-8080 FAX
steve@mjr.com

**National Center for Biotechnology Information (NCBI)** provides integrated access to DNA and protein sequence data, associated mapping data, protein structures, and MEDLINE. Demonstrations of the GenBank database, the Entrez retrieval system, PubMed for MEDLINE searching, BLAST and VAST similarity searches for sequence and structures, and the BankIt and Sequin sequence submission software will be provided.

**Contact person:** Barbara Rapp
Operations Research Analyst
National Center for Biotechnology Information (NCBI)
National Library of Medicine, National Institutes of Health
8600 Rockville Pike, Bldg. 38A, Rm. 8N803
Bethesda, MD 20894
301-496-2475
301-480-9241 FAX
info@ncbi.nlm.nih.gov

**New England Biolabs** is the leading supplier of restriction enzymes in the USA, offering a full range of reagents for Molecular Biology and cell signaling analysis. Products include nine 8-base cutters; enzymes for DNA sequencing; phosphospecific antibodies; and innovative kits e.g.GPS-1 (Genome Priming System) for generating sequencir templates in vitro, IMPACT T7 (Intein based protein expression).

**Contact person:** Dr. Barton Slatko
Director
DNA Sequencing Group
New England Biolabs, Inc.
32 Tozer Road
Beverly, MA 01915
978-927-5054
978-921-1350 FAX
dnaseq@neb.com

**PE Applied Biosystems** offers the most comprehensive range of products for life science research. As the leader in PCR technology and automated genetic analysis systems, our extensive product line of instruments and reagents enable scientists to understand the structure and function of protein and DNA molecules.

**Contact person:** Stacy Montgomery
Research Scientist
850 Lincoln Centre Drive
Foster City, CA 94404
650-638-5064
650-638-6333 FAX

**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.

**Contact person:** Nancy Stover
Product Manager
28159 Avenue Stanford
Valencia, CA 91355
800-426-8157 [ext. 543]
805-295-7652 FAX

**Sooner Scientific, Inc.** will have on display the New 96EZLoad System designed specifically to load Applied Biosystems 96 lane upgrade. PrismPette™ and PrismStrip™ Tips for loading 24, 36, 48 and 64 lane gels on Applied Biosystems 377 and 373 DNA Sequencers. SeaQuate® pre-mixed gel solutions, custom sequencing accessories including Combs, Spacers and Borosilicate Glass Plates.

**Contact person:** Jerry Coffey
President
PO Box 1800
Garvin, OK 74736
580-286-9408
580-286-7047 FAX

**Virtek Vision Inc.** is leading the world in precision laser applications. Our new ChipReader™ is a laser confocal DNA microarray analytical tool for high sensitivity detection of multiple fluorescent dyes. Computer-controlled, the ChipReader™ provides high through-put, has a small footprint and accommodates a variety of substrates including glass microscope slides.

**Contact person:** Beverly Giammara
300 Wildwood Avenue
Woburn, MA 01801
781-933-3456
781-933-3461 FAX

**The Institute for Genomic Research**
Genomic Science Series

**Conference on**

# Microbial Genomes III:
## Sequencing, Functional Characterization and Comparative Genomics

**January 29 - February 1, 1999**

**Westfields Marriott**
**Chantilly, Virginia**

**1999 Conference Co-Chairs**

Claire M. Fraser, Ph.D.
The Institute for Genomic Research
Rockville, Maryland

André Louis Goffeau, Ph.D.
Université Catholique de Louvain
Louvain la Neuve, Belgium

# Acknowledgments

The organizers would like to acknowledge the significant support given to the Microbial Genomes Conference by the following sponsors:

**US Department of Energy**

**Mary Ann Liebert, Inc.**

**Schering Plough Research Institute**

**Compugen**

**New England Biolabs, Inc.**

# Agenda

## Friday, January 29, 1999

3:00 pm Registration Opens

8:00 pm Welcoming Reception

## Saturday, January 30, 1999

7:00 am Registration Re-opens

7:00 am Breakfast

**8:30 am - 12:05 pm** **Plenary Session I: Genome Biology**

The Long Road from Sequence to Function: Experiments with Yeast
**Bernard Dujon**, Institut Pasteur

The Yeast Proteome Database: Connecting Model Organism Knowledge to Comparative and Functional Genomic Data
**James I. Garrels**, Proteome, Inc.

Proteome Analysis of *Saccharomyces Cerevisiae*
**Stephen J. Fey**, Center for Proteome Analysis in Life Sciences

10:15 am Break

Expression Clustering and Motif Quantitation in *E.coli* and Yeast
**George M. Church**, Harvard Medical School

Systematic Analysis of Gene Function in *Bacillus subtilis*
**Dusko Ehrlich**, Institut National de la Recherche Agronomique

12:05 pm Lunch

**2:00 - 5:30 pm** **Plenary Session 2: Genome Projects I**

Genome Sequence of the Gastrointestinal Pathogen *Campylobacter jejuni*
**Brendan W. Wren**, St. Bartholomew's Hospital

Completion of the Genome Sequence of *Deinococcus radiodurans*
**Owen White**, The Institute for Genomic Research

Genomics and Engineering of a Radioresistant Bacterium
**Kenneth Minton**, Uniformed Services University of the Health Sciences

3:15 pm Break

New Insights from the *Thermotoga maritima* Genome
**Karen E. Nelson**, The Institute for Genomic Research

UNCONTrolled Rearrangements among Borrelia Linear Plasmids?
**Sherwood R. Casjens**, University of Utah

Chromosome 2 Sequence of the Human Malaria Parasite *Plasmodium falciparum*
**Malcolm J. Gardner**, The Institute for Genomic Research

6:00 pm Dinner

**7:30 pm Exhibits and Poster Session**

## Sunday, January 31, 1999

7:30 am Breakfast

**8:30 am - Noon Plenary Session 3: Comparative Genomics**

From Genome to T Cell: Searching for Agretopes
**Anne DeGroot**, International Health Institute

Monitoring Changes in Gene Expression of *Mycobacterium tuberculosis* using DNA Microarrays
**Michael A. Wilson**, Stanford University Medical School

Identification of Novel Virulence Factors in *Vibrio cholerae*
**John J. Mekalanos**, Harvard Medical School

10:00 am Break

A Phylogenomic Analysis of DNA Repair Processes and the Benefits of a Phylogenomic Perspective
**Jonathan Eisen**, The Institute for Genomic Research

Categorizing Functions of Gene Products
**Monica Riley**, Marine Biological Laboratory

Noon Lunch

**2:00 - 5:00 pm** **Plenary Session 4: Genome Biology II**

Genomic Sequence Comparison of Two Strains of *Helicobacter pylori*
**Richard Alm**, Astra Research Center Boston

Sequencing of the *M. tuberculosis* Genome: Comparison of a Recent Clinical Isolate with the Laboratory Strain
**Robert D. Fleischmann**, The Institute for Genomic Research

Genomic and Functional Analysis of the Hyperthermophilic Archaeon *Pyrococcus furiosus* and Related Species
**Frank T. Robb**, University of Maryland Center for Marine Biology

3:15 pm Break

Constructing Multi-genome Views of Whole Microbial Genomes
**Terry Gaasterland**, The Rockefeller University

New Methods for the Prediction of Protein Structure and Function from Sequence
**Jeffrey Skolnick**, The Scripps Research Institute

Comparative Genomics as a Tool for Understanding Protein Thermostability
**David Eisenberg**, UCLA

6:00 pm Superbowl Celebration

## Monday, February 1, 1999

7:00 am Breakfast

**9:00 - 11:10 am** **Plenary Session 5: Genome Analysis I**

How Good are Deep Phylogenetic Trees?
**Hervé Philippe**, Université Paris-Sud

Life's Third Domain: An Established Fact or an Endangered Paradigm?
**Radhey Gupta**, McMaster University

9:50 am Break

Identification of Bacterial Promoter Sequences
**David Ussery**, Danish Technical University

Novel Approaches for the Identification of Microbial Antigens
**Alexander von Gabain**, Institute of Microbiology and Genetics, Vienna Biocenter

**11:45 am** **Updates**

**Mary E. Clutter**, National Science Foundation

**Marvin E. Frazier**, US Department of Energy

**Michael Gottlieb**, National Institute of Allergy and Infectious Diseases, National Institutes of Health (invited)

Noon Meeting Adjourns

Noon - 2:00 pm Lunch

# Plenary Speakers and Chairs

**Richard Alm, Ph.D.**
Research Scientist and Genomics Project Leader
Molecular Biology and Microbiology
Astra Research Center Boston
128 Sidney Street
Cambridge, MA 02139-4239
617-234-2540
617-576-3030 FAX
richard.alm@arcb.us.astra.com

**Sherwood R. Casjens, Ph.D.**
Professor
Department of Oncological Sciences
University of Utah Medical Center
Salt Lake City, UT 84132-0001
801-581-5980
801-581-3607 FAX
sherwood.casjens@
qmserver.genetics.utah.edu

**George M. Church, Ph.D.**
Professor
Department of Genetics
Harvard Medical School
Warren Alpert Building, Room 513
200 Longwood Avenue
Boston, MA 02115
617-432-7562
617-432-7266 FAX
Church@arep.med.harvard.edu

**Mary E. Clutter, Ph.D.**
Assistant Director
Department of Biological Sciences
National Science Foundation
4201 Wilson Boulevard
Arlington, VA 22230
703-306-1400
703-306-0343 FAX

**Anne S. DeGroot, Ph.D.**
TB/HIV Research Laboratory
International Health Institute
Brown University
Box G/B 473
Providence, RI 02912-4799
401-863-1374
410-863-1243 FAX
Anne_DeGroot@Brown.edu

**Bernard Dujon, Ph.D.**
Biotechnologies
Institute Pasteur
25 Rue du Dr Roux
Paris Cedex 15
75724 FRANCE
33-1-45-68-84-82
33-1-40-61-34-56 FAX
bdujon@pasteur.fr

**Dusko Ehrlich, Ph.D.**
Laboratoire de Genetique Microbienne
Institut National de la Recherche Agronomique
Jouy en Josas Cedex
78352 FRANCE
33-1-34 65 25 11
33-1-34 65 25 21 FAX
ehrlich@biotec.jouy.inra.fr

**Jonathan Eisen, Ph.D.**
Assistant Investigator
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-652-5844
301-838-0208 FAX
jeisen@tigr.org

**David Eisenberg, Ph.D.**
Director
Department of Energy Lab of Structural Biology and Molecular Medicine
UCLA
Box 951570
Los Angeles, CA 90095-1570
310-825-3754
310-206-3914 FAX
david@mbi.ucla.edu

**Stephen J. Fey, Ph.D.**
Center for Proteome Analysis
Odense University
International Science Park Odense
Forskerparken 10B
Odense M, DK-5230
DENMARK
011-45-63-15-7239
011-45-63-15-72-40 FAX
sjf@cpa.spo.dk

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

**Claire M. Fraser, Ph.D.**
President
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-0200
301-838-0208 FAX
cmfraser@tigr.org

**Marvin E. Frazier, Ph.D.**
Director of Health Effects and Life Sciences Division
Office of Biological and Environmental Research
US Department of Energy
19901 Germantown Road
Germantown, MD 20874-1290
301-903-5368
301-903-8521 FAX
marvin.frazier@science.doe.gov

**Terry Gaasterland**
Laboratory for Computational Genomics
The Rockefeller University
1230 York Avenue, Box 250
New York, NY 10021-6399
212-327-7755
212-327-8544 FAX
gaasterland@rockefeller.edu

**Malcolm J. Gardner, Ph.D.**
Assistant Investigator
Eukaryotic Genomics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319
301-838-3519
301-838-0208 FAX
gardner@tigr.org

**James I. Garrels, Ph.D.**
President and CEO
Proteome, Inc.
100 Cummings Center, Suite 435M
Beverly, MA 01915
978-922-1643
978-922-3971 FAX
jg@proteome.com

**André Louis Goffeau, Ph.D.**
Professor
FYSA
Université Catholique de Louvain
Place Croix du Sud 2-20
Louvain la Neuve
B-1348 BELGIUM
32-10-47-36-14
32-10-47-38-72 FAX
goffeau@fysa.ucl.ac.be

**Radhey S. Gupta, Ph.D.**
Department of Biochemistry
McMaster University
Hamilton, Ontario L8N 3Z5
CANADA
905-525-9140 ext. 22639
905-522-9033 FAX
gupta@fhs.mcmaster.ca

**John J. Mekalanos, Ph.D.**
Professor and Chair
Department of Microbiology and Molecular Genetics
Harvard Medical School
200 Longwood Avenue
Building D1 - Room 421
Boston, MA 02115
617-432-1935
617-738-7664 FAX
jmekalan@warren.med.harvard.edu

**Kenneth Minton**
Professor
Department of Pathology
Uniformed Services University of the Health Sciences
4301 Jones Bridge Road
Bethesda, MD 20814-4799
301-295-3476
301-295-1640 FAX
kminton@usuhs.mil

**Karen E. Nelson, Ph.D.**
Collaborative Investigator
Prokaryotic Genomics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-3565
301-838-0208 FAX
kenelson@tigr.org

**Hervé Philippe, Ph.D.**
Laboratoire de Biologie cellulaire
Université Paris-Sud
Bat 444
Orsay Cedex 91405
FRANCE
33-1-69-156481
33-1-69-156803 FAX
hp@bio4.bc4.u-psud.fr

**Monica Riley, Ph.D.**
Senior Scientist
Molecular Evolution
Marine Biological Laboratory
7 MBL ST
Woods Hole, MA 02543
508-289-7612
508-540-6902 FAX
mriley@mbl.edu

**Frank T. Robb, Ph.D.**
Professor
University of Maryland Center
for Marine Biology
701 East Pratt Street, Suite 236
Baltimore, MD 21202
410-234-8870
410-234-8896 FAX
ROBB@mbimail.umd.edu

**Jeffrey Skolnick, Ph.D.**
Professor
Department of Molecular Biology
The Scripps Research Institute
10550 North Torrey Pines Road (TPC 5)
La Jolla, CA 92037
619-784-8821
619-784-8895 FAX
skolnick@scripps.edu

**David W. Ussery**
Associate Professor
Center for Biological Sequence Analysis
Institute of Biotechnology
Danish Technical University
Building 208
Lyngby DK-2800
DENMARK
45 4525 2488
45 4593 1585 FAX
Dave@CBS.dtu.dk

**Alexander von Gabain, Ph.D.**
CEO
Microbiology and Genetics
Vienna Biocenter
Dr. Bohrgasse 9
Vienna 1030
AUSTRIA
43-1-4277-54601
43-1-4277-9546 FAX
alex@gem.univie.ac.at

**Owen White, Ph.D.**
Deputy Director
Informatics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-3534
301-838-0208 FAX
owhite@tigr.org

**Michael Wilson**
Graduate Student
Microbiology/Immunobiology
Stanford University
B239 Beckman Center, S.U.M.C.
300 Pasteur Lane
Stanford, CA 94305
650-723-7026
650-723-1399 FAX
wilson@cmgm.stanford.edu

**Brendan W. Wren, Ph.D.**
Director of Research
Department of Medical Microbiology
St. Bartholomew's Hospital
West Smithfield
London EC1A 7BE
UNITED KINGDOM
44-1716-018411
44-1716-018409 FAX
b.w.wren@mds.qmw.ac.uk

# Speaker Abstracts

## Plenary Session 1: Genome Biology

**Dujon, Bernard**
not received in time for publication

### The Yeast Proteome Database: Connecting Model Organism Knowledge to Comparative and Functional Genomic Data

**James I. Garrels**, Peter H. Hodges, Andrew H. Z. McKee, Maria Costanzo, Michael Cusick, Kevin Roberg, and William E. Payne. Proteome, Inc., Beverly, MA .

The Yeast Proteome Database (YPD(tm)) is a curated database of protein properties and functions for the model organism, *Saccharomyces cerevisiae* (see http://www.proteome.com/YPDhome.html). YPD and similar databases under construction for other species are built by trained curators who review the entire research literature for the organism. This powerful knowledge resource, organized into a relational database and presented in a unified one-page per protein format, is used daily by most members of the yeast community and by many corporate users.

Each protein of yeast is represented in YPD as a Yeast Protein Report, of which the key elements are 1) a Title Line (best one-line description), 2) a formatted properties section, 3) a section containing free-text annotations with references, and 4) a full reference list. Accessed from each Protein Report are pop-up windows providing additional information on protein modifications, protein-protein interactions, gene regulation (including profiling data), and related genes of other species.

YPD is especially designed for use in comparative and functional genomics studies. The comprehensive yeast protein knowledge in YPD can be used for annotation of new genomic sequences from fungal, plant, and animal genomes. Furthermore, new expression profiling experiments in yeast or related organisms can best be analyzed using the comprehensive store of regulatory, pathway, and functional information already in YPD. Presentation and analysis tools help to understand common functions for co-regulated proteins.

Reference: P. H. Hodges, Andrew H. Z. McKee, William E. Payne, and J.I. Garrels. The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data (Nucleic Acids Res. Jan. 1999, in press).

**Fey, Stephen J.**
not received in time for publication

### Genomics and Engineering of a Radioresistant Bacterium

**George M. Church**, Pam Ralston, Martha Bulyk, Abby McGuire, Rob Mitra, Saeed Tavazoie, and Jason Hughes. Department of Genetics, Harvard Medical School, Boston, MA.

We have developed technologies for annotating genome sequences, including intergenic regions and regulon/operon comparisons. Performing enzymatic reactions on oligonucleotide chips or microarrays (of kbp-sized DNA) allowing us to replicate DNA chips using microcontact printing. RNA quantitations from chip, microarray and SAGE can be merged, clustered, and the motifs mechanistically responsible for the clusters of coregulated RNAs can be determined. Methods for measuring and modeling in vivo concentrations of protein, RNA, metabolite, protein interactions and mutant growth rates in response to diverse environments provide the foundations for a genome sequence function database.

Nature Biotech. 16:566-571; Nature Biotech. 16: 939-945; J. Molec. Biol. 284: 241-254. (http://arep.med.harvard.edu)

### Systematic Analysis of Gene Function in *Bacillus subtilis*

**S.D. Ehrlich**[1] and N. Ogasawara[2]. [1]Génétique Microbienne INRA, Jouy en Josas, FRANCE; and [2]NAIST, Ikoma, Nara, JAPAN.

Sequencing of the *B. subtilis* genome has revealed over 4000 putative protein coding sequences. On the basis of previous studies, which concerned only about a fourth these, and sequence comparisons with proteins known in other organisms, a function for about 2300 genes was proposed. Therefore, almost a half of the *B. subtilis* genes have no clear function. Furthermore, the genes that encode proteins similar to those found in other organisms, but that haven't been studied in *B. subtilis*, may have a different function than that predicted from comparisons. Systematic analysis of genes that have no established function is being carried out in two research consortia, one in Europe and one in Japan involving some 30 laboratories.

The main approach to assess gene function is the construction of mutants in the target genes and the analysis of the mutant phenotypes. An insertional vector was used to construct the mutants. Its integration within a target gene (i) inactivates the gene; (ii) places a reporter (-galactosidase gene under the control of the natural expression signals of the target gene; (iii) places the possible downstream genes under the control of an IPTG-dependent promoter, thus allowing to avoid polar effects of vector insertion. Its integration upstream of a target gene yields conditional, IPTG-dependent, mutants. About 2000 mutants have been constructed and submitted to preliminary characterization, which included growth properties in rich and minimal medium and the pattern of expression. This analysis has shown that the reporter gene activity could be detected for about two thirds of the genes, indicating that a majority of the genes without known function are expressed to a measurable extent under the conditions used. This value is corroborated by the transcript analysis, carried out by Northern hybridization, and covering over 500 kb. By a more sensitive RT PCR method, used to

analyze several chromosomal regions covering about 100 kb, over 90 % of genes appeared to be transcribed. This suggests that a large majority of *B. subtilis* genes may be expressed, albeit weakly, even under ordinary laboratory conditions. However, it should be noticed, that Northern analysis revealed only about a half of the transcripts predicted by sequence analysis, whereas about a half of the transcripts detected were unexpected. This may be due to unknown signals for transcription initiation and termination or to transcript processing, which should possibly be taken into account when considering massive approaches to transcript analysis based on DNA arrays.

About 50 genes could not be interrupted. Strains in which the corresponding genes were put under the control of the IPTG-regulated promoter required IPTG for growth. This indicates that these genes are essential for growth.

Phenotype tests at three levels were carried out systematically for most of the mutants constructed in Europe. The objective of the first level tests, which are sufficiently simple to allow a relatively high throughput, was to assign mutants to broad physiological categories defined as i) Metabolism of small molecules and inorganics; (ii) Macromolecule metabolism; (iii) Cell structures and motility; (iv) Stress and stationary phase; (v) Cell processes. Most mutants constructed in Europe (above 1000) were submitted to the first level tests, and similar tests are in progress in Japan. A phenotype was detected for over 150 of the mutants tested in Europe. This value underestimates the real frequency of detectable phenotypes, since the tests with many mutants have not been completed yet, but indicates that a number of the genes without known function affects significantly *B. subtilis* cells under laboratory conditions. The second level tests, which are more complex and have a lower throughput, are presently being carried out, with the objective of validating gene assignment deduced from the first level tests. The third level tests, which aim to determine gene function, were carried out on some selected genes, and the examples will be presented at the meeting.

In conclusion, the approach to study function of unknown genes in a systematic way by mutant construction and phenotype analysis is efficient and informative in the case of a bacterium such as *B. subtilis*. It has uncovered about 50 new essential genes, and a number of new and interesting genes involved in different aspects of cell metabolism.

# Plenary Session 2: Genome Projects I

**Wren, Brendan W.**
not received in time for publication

## Completion of the Genome Sequence of *Deinococcus radiodurans*

**Owen White**, John Heidelburg, Claire Fraser and J. Craig Venter. The Institute for Genomic Research, Rockville, MD.

*Deinococcus radiodurans* is a non-pathogenic, non-sporulating, red-pigmented Gram+ bacterium. *D. radiodurans* was originally found in radiation sterilized food that under went spoilage. It is remarkable in that it is the most radioresistant organism to have ever been isolated (Moseley, 1983). An important component of this resistance is the ability to repair damage to chromosomal DNA. *D. radiodurans* cultures exposed to 1.5 Mrad of radiation displayed reduction in size of genomic DNA fragments corresponding to approximately 100 double stranded breaks (DSBs) per genome. (Typically, most prokaryotic and eukaryotic organisms cannot tolerate more than 5 double stranded breaks per genome without reduced survival.) Remarkably, within eight to ten hours after exposure, *D. radiodurans* genomic fragment lengths are restored to size ranges seen in non-treated cells. During this repair time, cellular replication of *D. radiodurans* is arrested (Daly *et al.*, 1994); however, after this eight to ten hour interval, the cells display 100% survival with no detectable mutagenesis of their completely restored genomes. The genome sequence of Deinococcus is complete and we have determined the genome is composed of 3 chromosomes and a small plasmid; a number of unique sequence elements have been identified. The content of the genome, along experimental results will be discussed in context of this organism's unique ability to withstand gamma radiation.

## Genomics and Engineering of a Radioresistant Bacterium

**Kenneth W. Minton**, Kira S. Makarova, Michael J. Daly, Eugene V. Koonin, L. Aravind, Hassan Brim, and Ajay Sharma. Uniformed Services University of the Health Sciences, Bethesda, MD.

The eubacterium Deinococcus radiodurans is the most DNA damage-resistant organism discovered to date. It is therefore of intrinsic interest to study its DNA repair mechanisms, and towards this end the full genomic sequence of this organism has recently been obtained by TIGR. We have fully annotated this sequence with special attention to properties that might render this organism radioresistant. Features noted to date include a novel enzyme, combining potential repair domains from three independent repair proteins. This gene is currently being knocked out of the deinococcal genome and properties of the null mutant will be reported. Similarly, desiccation-resistance proteins similar to those seen in plants have also been discovered in the Deinococcus radiodurans genome. This is of particular significance, as there is a known positive correlation between deinococcal desiccation-resistance and radioresistance. The properties of knock out mutants will be reported. Finally, an expansion of several protein families, including phosphatases, proteases, acyl transferases, the mutT family of pyrophosphatases, and thioredoxins have been noted. Deinococcus radiodurans' genome is extraordinarily rich in repeated sequences, suggesting a mechanism of repair that will be presented. In addition, it is the first bacterium to be sequenced that has multiple chromosomes (three). Engineering of this versatile organism for organopollutant degradation in radioactive mixed waste environments and engineering of heavy metal resistance in this organism will be described. Finally, current attempts to acquire large amounts and crystallize Deinococcus' extraordinary and highly toxic RecA protein will be described.

## UNCONTrolled Rearrangements among Borrelia Linear Plasmids?

**Sherwood Casjens**, Wai Mun Huang, Margaret Robertson, Nanette Palmer, Jeremy Peterson[1], Granger Sutton[1] and Claire Fraser[1]. Department of Oncological Sciences, University of Utah School of Medicine, Salt Lake City, UT; and [1]The Institute for Genomic Research, Rockville, MD.

The complete sequence of the genome of the Lyme disease spirochete *Borrelia burgdorferi* strain B31 reveals that it carries 12 linear and 9 circular plasmids. Analysis of the sequence has shown that there are many paralogous families of sequences scattered about the linear plasmids. Analysis of these groups of similar sequences shows that the complex relationships among the plasmids within this strain apparently represent recombinational events of the following types: (i) integration of a circular plasmid into a linear plasmid, (ii) transposition, (iii) insertion, (iv) deletion, (v) inversion, (vi) amplification of short direct repeat tracts, (vii) exchange of telomeres between linear plasmids and between linear plasmids and the linear chromosome, and (viii) homologous recombination between plasmids. Non-homologous and non-reciprocal recombination events appear to have occurred. One result of these apparently uncontrolled rearrangements among the plasmids is the presence of many apparently nonfunctional fragments of genes or pseudogenes. These gene fragments account in part for the observation that the predicted fraction of protein-encoding DNA on these linear plasmids varies from 83% down to 32%, compared to about 90% for the *Borrelia burgdorferi* chromosome and circular plasmids as well as in other completely sequenced bacterial genomes. Of the 845 recognizable putative plasmid "genes", 114 appear to be damaged, 568 appear to be potentially functional, 15 are silent members of a "cassette" diversity generation system for the surface protein VlsE, and 148 are short unique ORFs whose functionalities are often suspect. The damaged genes are found almost exclusively on 9 of the 21 plasmids, and the damage ranges from genes containing a single frameshift to (for example) a gene that has suffered an insertion, 3 major deletions, 7 frameshifts, 6 in-frame stop codons, and one large inversion.

## Chromosome 2 Sequence of the Human Malaria Parasite *Plasmodium falciparum*

**Malcolm J. Gardner**[1], Hervé Tettelin[1], Daniel J. Carucci[2], Leda M. Cummings[1], L. Aravind[3], Eugene V. Koonin[3], Hamilton O. Smith[1], Mark D. Adams[1], J. C. Venter[1], Stephen L. Hoffman[2]. [1]The Institute for Genomic Research, Rockville, MD; [2]Malaria Program, Naval Medical Research Institute, Rockville, MD; and [3]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD.

Chromosome 2 of the malaria parasite *Plasmodium falciparum* was purified from pulsed field gels and sequenced using a random shotgun strategy. The chromosome is 947 kb in length, has a base composition of 80.2% A+T, and contains 210 predicted genes. In comparison to the *Saccharomyces cerevisiae* genome, chromosome 2 has a lower gene density, a greater proportion of genes containing introns, and nearly twice as many proteins containing predicted non-globular domains. Five protein families unique to *Plasmodium* were identified. One of these new families was a group of putative surface proteins, rifins, which are encoded by a gene family comprising up to 7% of the protein-encoding genes in the genome. The rifins exhibit considerable sequence diversity and may play an important role in antigenic variation. Other families include a cluster of 8 tandemly repeated genes encoding putative thiol proteases, and 5 genes encoding DnaJ-type domains. Sixteen genes on chromosome 2 showed definite signs of a plastid or mitochondrial origin. Completion of the chromosome 2 sequence demonstrates that the A+T-rich genome of *P. falciparum* can be sequenced by a shotgun strategy. Sequencing of chromosomes 10, 11, and 14 is underway, and we are developing several strategies for analysis of gene expression.

# Plenary Session 3: Comparative Genomics

## From Genome to T Cell: Searching for Agretopes

**Anne S. De Groot**, Michael Gonzalez, Natasha Chinai, Jean David Barnea and Judy A. George. TB/HIV Research Laboratory, Brown University, Providence, RI.

Pattern analysis of genome sequences facilitates the identification of peptides that bind to MHC molecules (agretopes). Agretopes can promote or impede cell-mediated immunity; therefore, the prospective identification of agretopes may accelerate the pace of development of vaccines and therapeutics. EpiMatrix is a matrix-based algorithm for the selection of agretopes that provides a score (EBP, estimated binding probability) based on match to a pre-determined pattern for a given MHC molecule. *In vitro* studies to test EpiMatrix selection of novel agretopes from HIV sequences have been performed. Binding of predicted peptides ranging from 8 -11 amino acids in length to the surface of MHC allele-specific TAP deficient cell lines was measured by FACS.

RESULTS

| matrix | # binders* | #tested | expected median EBP** | observed %binders |
|---|---|---|---|---|
| A11 | 4 | 7 | 44% | 57% |
| A11 | 6 | 12 | 40% | 38% |
| A2 | 3 | 7 | 87% | 43% |
| B7 | 7 | 7 | 26% | 100% |
| B27 | 7 | 7 | 7% | 100% |

*expected performance for the group of peptides based on median EBP for the sample group. observed performance reports % total group of peptides tested that were observed to bind.

The observed proportion of peptides binding was higher than expected in 5/6 sets of experiments, when compared to the median estimated binding probability (EBP) of the groups of peptides tested. In all cases, control (known agretopes) bound. No binding was observed where it was not predicted, with one peptide exception (out of 33 assayed). In additional studies, 16 novel HIV CTL epitopes based on EpiMatrix predictions for HIV sequences have been identified. EpiMatrix can be used to facilitate research on HIV immunopathogenesis and vaccine development.

## Monitoring Changes in Gene Expression of *Mycobacterium tuberculosis* using DNA Microarrays

**M.A. Wilson**[1], H.H. Kristensen[1], J. DeRisi[2], S. Rane[1], P. Brown[2], G.K. Schoolnik[1]. [1]Division of Infectious Diseases, and [2]Department of Biochemistry at Stanford University School of Medicine, Stanford, CA.

The strategy we are using seeks to identify genes encoding enzymes of particular biosynthetic or metabolic pathways of *Mycobacterium tuberculosis*. The process is based on the availability of a complete *M. tuberculosis* genomic sequence and the use of DNA microarray hybridization analysis to monitor, simultaneously, changes in mRNA expression of nearly all predicted open reading frames of the organism. The approach relies on the ability of microarray comparative hybridization to identify genes whose expressions are induced or repressed when bacteria are exposed to a compound that selectively inhibits an enzymatic step of a multi-enzyme pathway. The results of initial experiments that surveyed responses to several drugs and inhibitors demonstrated that some of the responding genes encode enzymes that participate directly in the affected pathway. Furthermore, the patterns of co-regulated gene networks are beginning to emerge as we compile the various response profiles. Responses to the *M. tuberculosis*-specific drug isoniazid provide examples of this phenomenon and illustrate two ways this approach might be applicable to the discovery of new antimicrobial agents. First, some of the responding genes encode proteins that are appropriate targets for new drug development. Second, genes which specifically respond to a particular inhibitor can be fused to reporter constructs and used to help screen combinatorial libraries for compounds that elicit a similar response.

**Mekalanos, John J.**
not received in time for publication

## A Phylogenomic Analysis of DNA Repair Processes and the Benefits of a Phylogenomic Perspective

**Jonathan A. Eisen** and Philip C. Hanawalt. The Institute for Genomic Research, Rockville, MD.

The ability to recognize and repair abnormal DNA structures is common to all forms of life. Studies in a variety of species have identified an incredible diversity of DNA repair pathways. This diversity is seen with regard to the specificity, complexity, and mechanisms of the different pathways as well as the overlap with other cellular functions. Documenting and characterizing the similarities and differences in repair between species has important value for understanding the origin and evolution of repair pathways as well as for improving our understanding of phenotypes affected by repair (e.g., mutation rates, lifespan, tumorigenesis, survival in extreme environments). Unfortunately, while repair processes have been studied in quite a few species, the ecological and evolutionary diversity of such studies has been limited. Complete genome sequences can provide potential sources of new information about repair in different species. I present a global comparative analysis of DNA repair proteins and processes based upon the analysis of publicly released complete genome sequences. I use a new form of analysis that combines genome sequence information and phylogenetic recreations into one composite phylogenomic analysis. I use this phylogenomic analysis to study the evolution of repair proteins and processes and to predict the repair phenotypes of those species for which we now know the complete genome sequence. In addition, I discuss why combining evolutionary and genomic analysis improves our understanding of DNA repair and can also improve studies of any biological process.

## Categorizing Functions of Gene Products

**Monica Riley**. Bay Paul Center for Molecular and Comparative Evolution, Marine Biological Laboratory, Woods Hole, MA.

Some years ago a simple one-dimensional classification of gene products of E. coli was formulated to see how much we know and do not know about the molecular genetics of Escherichia coli [Microbiological Reviews 57:862-952 (1993) and a later revision in vol. II of "Escherichia coli and Salmonella" (1997). ed. F. C. Neidhardt *et al.*, pp.2118-2202, ASM, Wash., DC]. This classification has served as a point of departure for characterizing gene products of other bacteria whose genomes have been sequenced.

However, there is much room for improvement. The simple one-dimensional classification system is inadequate to express the complexity of the cellular roles of many bacterial genes/gene products. A multidimensional classification is proposed to allow more realistic and more useful designation of the functions of bacterial genes. Dimensions may include:

* metabolic role, more than one metabolic role being allowed for each enzyme (depending on conditions) and including not only the enzymes but the transporters and regulators related to each metabolic function
* regulation, listing regulators connected to the genes regulated, the type of regulation and the conditions invoking global regulators
* information transfer including DNA replication, RNA and protein synthesis
* cell structure, including membrane proteins, cellular appendages, antigenic determinants, etc. that have functions described in other dimensions
* processes such as cell division, DNA repair, motility, chemotaxis, transport, osmotic adaptation, starvation response, sporulation where present
* common ancestry as judged by membership in paralogous groups (Riley & Labedan, J.Mol. Biol. 268, 857-868, 1997).

# Plenary Session 4: Genome Biology II

## Genomic Sequence Comparison of Two Strains of *Helicobacter pylori*

**Richard A. Alm.** Astra Research Center Boston, Cambridge, MA.

This first comparison of the complete genomic sequence of two different isolates of the same bacterial species, in this case *Helicobacter pylori*, has yielded results with important implications for understanding *H. pylori* associated disease and for the development of novel approaches for drug and vaccine therapeutic intervention. *H. pylori* strain J99, isolated from a patient with a duodenal ulcer, has a circular genome of 1,643,831 base pairs and 1,495 predicted, protein coding genes. Comparison of this strain's sequence with the previously published sequence of another strain (26695) showed that these *H. pylori* genomes display some unusual organizational differences and, despite the widely held doctrine of extensive genomic diversity in this species, are remarkably conserved in size and gene number, order, content, and function. Most of the endpoints defining organizational differences between the two strains were anchored within non-coding regions and were associated with mobile insertion elements and/or DNA restriction/ modification genes. Several regions of both chromosomes were of significantly lower (G+C)% content than the remainder of the genome. These included the putative *cag* pathogenicity island and a region of the genome which has been termed "the plasticity zone" because it contains about 45% of the genes unique to each strain and which may represent a novel pathogenicity island. Of the 89 *H. pylori* J99 strain-specific genes, 25 have homology to genes of predicted function, 8 have homology to genes of unknown function and 56 have no significant homology to any genes in public databases. *H. pylori* 26695 has 117 strain-specific genes with a similar distribution of functional assignments. Despite its apparent limited regulatory repertoire, *H. pylori* has been predicted to regulate expression of specific genes by slipped-strand repair. Analysis of these two complete genomic sequences has permitted higher discrimination of which genes may be subject to this phenomenon, including potential colonization factors and outer membrane proteins which may facilitate colonization or play a role in the persistence of infection. Aside from the diversity in the plasticity zone, the extent of conservation between these unrelated *H. pylori* strains suggests that host factors may also play a significant role in the differences in type, severity and outcome of *H. pylori* associated diseases.

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

**Robert D. Fleischmann**[1], Owen White[1], Erin Hickey[1], Rebecca Clayton[1], Robert Dodson[1], Michelle Gwinn[1], John Heidelberg[1], Jeremy Peterson[1], Lisa McDonald[1], Terry Utterback[1], William Bishai[2], Arthur Delcher[3], Claire Fraser[1], and J. Craig Venter[1]. [1]The Institute for Genomic Research, Rockville, MD; [2]The Johns Hopkins University, Baltimore, MD; and [3]Celera Genomics, Rockville, MD.

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 specie 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 170,000 bp), differences in copy number of tandem repeats (~ every 90,000 bp), and the frequency of single nucleotide polymorphisms (~ every 5,000 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.

## Genomic and Functional Analysis of the Hyperthermophilic Archaeon *Pyrococcus furiosus* and Related Species

**Frank T. Robb.** Center of Marine Biotechnology, University of Maryland, Baltimore, MD.

Recently isolated Archaea of the genus *Pyrococcus* consist of 12 species, many of which grow rapidly at temperatures exceeding 100°C. Complete genomic sequences of two members of this group, *P. furiosus* and *P. horikoshii*, have recently been released and the genomic sequence of a third species, *P. abyssi* is completed but not yet available. These genomic sequences provide a unique resource for comparative studies of hyperstable enzymes, and insights into protein structure and function at very high temperatures.

The annotation of the 1.92 mbp circular genomic sequence of *P. furiosus* reveals that the otherwise closely related sequences of *P. horikoshii* and *P. furiosus* have diverged significantly in gene order. *P. furiosus* was isolated from a shoreline solfatara on Vulcano Island in Italy, and *P. horikoshii* was isolated from a vent in the Okinawa Trough at a depth of 1390 m. Since both of the isolates are strict anaerobes growing optimally at 98-100°C, it seems that ongoing genetic exchange is unlikely. The divergence of the genomes includes multiple insertion, deletion and inversion events. For example, the pre-insertion sites of several inteins can be detected and the rapid divergence of homologous inteins is evident. Clustered 37 bp tandem, highly repetitive elements which are conserved in the *Pyrococcus* genomes, have diverged both in copy number and in the position and number of clusters: the *P. furiosus* genome has five clusters whereas *P. horikoshii* has two clusters.

The tryptophan and maltose-trehalose operons of *P. furiosus* are linked, however the operons encoding maltose utilization, histidine biosynthesis and tryptophan biosynthesis have been lost in *P. horikoshii*. The maltose region is of particular interest: A 17.8 kbp region flanked by putative transposons in the *P. furiosus* genome contains an operon encoding maltose/trehalose uptake and catabolism genes resembling those from bacteria. We propose that the

region has been subject to a lateral transfer event. The entire maltose-trehalose region in the *P. furiosus* genome is 100% homologous to a transposon-flanked section in the genome of *Thermococcus litoralis*, a strain growing at 88°C which was isolated from the same site in Italy as *P. furiosus*. We propose that a recent intergeneric lateral transfer event has inserted the maltose region, presumably after the divergence of *P. furiosus* and *P. horikoshii*.

**Gaasterland, Terry**
not received in time for publication

## New Methods for the Prediction of Protein Structure and Function from Sequence

**Jeffrey Skolnick**, Jacqueline Fetrow, Angel R. Ortiz and Andrzej Kolinski. Department of Molecular Biology, The Scripps Research Institute, La Jolla CA.

A novel method for the prediction of protein function based on the sequence-to-structure-to-function paradigm has been developed. First, the tertiary structure of the sequence of interest is predicted from either *ab initio* folding or threading. Then, using a library of three dimensional descriptors of protein active sites, termed "fuzzy functional forms" or FFFs, the resulting structures are screened. If the geometry and residue types in the predicted structure matches an FFF, then the protein is predicted to have the specified molecular function. By way of illustration, predictions of the disulfide oxidoreductase activity of the glutaredoxins and thioredoxins are described. The FFFs correctly identify the active sites in a library of experimental structures as well as in models produced by *ab initio* folding or threading. This shows that low-to-moderate resolution models whose a-carbon root mean square deviations from native range from 3.5-6 Å are sufficient to identify protein active sites. Next, this approach is applied to the screening of the *E. coli*, *H. influenzae* and *M. jannaschii* genomes. For *E. coli*, the method identifies all ten sequences known or proposed to exhibit disulfide oxidoreductase activity. Furthermore, this activity is predicted for two other sequences that have not been previously identified. Six of eight sequences in the other two genomes that are annotated as being glutaredoxins, thioredoxins, or thiol/disulfide interchange proteins are found. Since one of the missed sequences is probably incorrectly annotated, the success ratio is likely to be six of seven. An additional 11 sequences in *H. influenzae* and five sequences in *M. jannaschii* are predicted to have the requisite activity. If these predictions are proven to be correct, then the detection of protein function will have been pushed much further into the twilight zone of sequence identity.

**Eisenberg, David**
not received in time for publication

# Plenary Session 5: Genome Analysis I

## How Good are Deep Phylogenetic Trees?

**Hervé Philippe**. Laboratoire de Biologie Cellulaire, Université Paris-Sud, Orsay Cedex, FRANCE.

The study of deep branching organisms is very important since they can display primitive characters. Unfortunately, the phylogenetic inference of the deep nodes is by essence very difficult because many multiple substitutions have occurred since these ancient cladogenetic events. In addition, a well-known tree reconstruction artifact, the long branch attraction (LBA) phenomenon PROGCOMP ENRf8 (Felsenstein, 1978), appears also to be of prime importance. Because the outgroup species necessarily emerge well before the ingroup species, it represents a long branch. This is particularly obvious for Eubacteria and Eukaryotes in the case of ribosomal RNA, for which the outgroup (for instance, Archaebacteria) is very distantly related. If an ingroup species evolves faster than the others, it will constitute a long branch, which will be attracted by the long branch of the outgroup. In consequence, it will emerge very early. As a result, it is likely that many, if not all, first emerging species are in fact fast evolving species. For example, in mammalian phylogeny based on complete mitochondrial genomes, two groups of rodents (mouse and rat on the one hand and guinea-pig on the other) emerge very early, in a paraphyletic way PROGCOMP ENRf8 (D'Erchia *et al.*, 1996), in marked disagreement with morphological evidence PROGCOMP ENRf8 (Luckett and Hartenberger, 1993). However, the use of paleontological data demonstrates that mouse and rat evolved about 5 times faster than carnivores and that their early emergence is very likely due to an LBA artifact PROGCOMP ENRf8 (Philippe, 1997).

As a result, LBA artifact can seriously mislead molecular phylogenies, even for recent events (less than 100 MY) and for very long sequences (more than 10000 nucleotides). It is thus likely that inference for more ancient events, such as the rooting of the tree of Life or the phylogeny of Eukaryotes, can be seriously confused by LBA artifact. Indeed, rooting the universal tree of Life requires using a paralogous gene PROGCOMP ENRf8 (Schwartz and Dayhoff, 1978), which is the result of a duplication that occurred before the emergence of the three domains. This paralogous gene is very distant and constitutes a very long branch, which is susceptible to attract any fast evolving ingroup. Interestingly, the branch lengths leading to the three domains were quite different according to the universal genes studied (rRNA, ATPase, Ile-tRNA synthetase and elongation factors), meaning that the evolutionary rate varied greatly between genes and between domains. For anciently duplicated genes, on an unrooted tree, the Eubacteria always display the longest branch. As a result, the Eubacteria are attracted by the long branch of the outgroup, resulting in an artifactual rooting in the eubacterial branch PROGCOMP ENRf8 (Philippe and Forterre, in press). To confirm this artifactual nature, we have studied separately the slow- and the fast-evolving positions of the elongation factors PROGCOMP ENRf8 (Lopez *et al.*, in press). As expected, the signal for ancient events has mainly been found in slow evolving positions, except for the eubacterial rooting. In contrast, the eubacterial rooting is rather supported by fast evolving positions, which are the most prone to generate tree reconstruction artifacts. This strongly suggests that the rooting of the universal tree of Life is not confidently determined by molecular phylogenetic analysis and remains

an open question. The recent suggestion of an eukaryotic rooting PROGCOMP ENRf8 (Jeffares *et al.*, 1998; Poole *et al.*, 1998) is a promising hypothesis since it provides the best explanation for the transition from the RNA-world to extant organisms.

References

PROGCOMP ENBib D'Erchia, A., Gissi, C., Pesole, G., Saccone, C., and Arnason, U. (1996). The guinea-pig is not a rodent. Nature 381, 597-600.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27, 401-410.

Jeffares, D. C., Poole, A. M., and Penny, D. (1998). Relics from the RNA world. J Mol Evol 46, 18-36.

Lopez, P., Forterre, P., and Philippe, H. (in press). A method for extracting ancient phylogenetic signal: the rooting of the universal tree of life based on elongation factors. J Mol Evol.

Luckett, W. P., and Hartenberger, J.-L. (1993). Monophyly or polyphyly of the order Rodentia: possible conflict between morphological and molecular interpretations. J. Mam. Evol. 1, 127-147.

Philippe, H. (1997). Rodent monophyly: pitfalls of molecular phylogenies. J Mol Evol 45, 712-5.

Philippe, H., and Forterre, P. (in press). The rooting of the universal tree of life is not reliable. J Mol Evol.

Poole, A. M., Jeffares, D. C., and Penny, D. (1998). The path from the RNA world. J Mol Evol 46, 1-17.

Schwartz, R. M., and Dayhoff, M. O. (1978). Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. Science 199, 395-403.

## Life's Third Domain: An Established Fact or an Endangered Paradigm?

**Radhey S. Gupta.** Department of Biochemistry, McMaster University, Hamilton, CANADA.

The three domain proposal of Woese *et al.* (1990) divides all living organism into three primary groups named Archaea (or archaebacteria), Bacteria (or eubacteria) and Eucarya (or eukaryotes), with Eucarya being specific relatives of Archaea. Although this proposal is currently widely accepted, sequence features and phylogenies derived from many highly conserved proteins are inconsistent with it and point to a close and specific relationship between archaebacteria and Gram-positive bacteria, whereas Gram-negative bacteria are indicated to be phylogenetically distinct. A closer relationship of archaebacteria to Gram-positive bacteria is generally seen for the majority of the available gene/protein sequences. To account for these results, and the fact that both archaebacteria and Gram-positive bacteria are prokaryotes surrounded by a single cell membrane, I have proposed that the primary division within prokaryotes is between Monoderm prokaryotes (surrounded by a single membrane) and Diderm prokaryotes (*i.e.*, all true Gram-negative bacteria containing both an inner cytoplasmic membrane and an outer membrane) (Mol. Microbiol.29,1998,695-707; Microbiol.Mol.Biol.Rev.,62,1998, 1435-1491). This proposal is consistent with both cell morphology and signature sequences in different proteins. Of these two groups, monoderm prokaryotes are indicated to be ancestral. It is suggested that both archaebacteria and diderm prokaryotes evolved from gram-positive bacteria, most likely, in response to the antibiotic selection pressure. Signature sequences in different proteins also provide a means to define a number of different taxa within prokaryotes (namely, low G+C and high G+C gram-positive, Deinococcus-Thermus, cyanobacteria, chlamydia-cytophaga related, and several divisions of Proteobacteria) and to indicate how they evolved from a common ancestor. Phylogenetic information derived from protein sequences also strongly indicate that all eukaryotic cells have received significant gene contributions from both an archaebacterium and a Gram-negative eubacterium. Thus the hypothesis that the ancestral eukaryotic cell directly descended from within the archaebacterial lineage is erroneous. These results call into question the validity of the currently popular three domain proposal and provide evidence for an alternate view of the evolutionary relationship among living organisms.

## A DNA Structural Atlas of Promoters in Microbial Genomes

**David W. Ussery.** Center for Biological Sequence Analysis, Institute of Biotechnology, The Technical University of Denmark, Lyngby, DENMARK.

We have analysed the complete nucleotide sequence of genomes from more than 20 different organisms, including eubacteria, archae, and eukaryotes. The sequences are examined in terms of a set of 6 different DNA structural parameters: DNase I sensitivity, DNA curvature, helix rigidity (propeller twist), protein-induced deformability, base-stacking energy, and the ability to wrap around protein (position preference). Each individual organism's chromosome sequence is displayed at the genome level as either a circular or linear plot. On a global scale, it is possible to visualise large regions of the genome which could potentially act as organisational determinants of chromosome architecture. At the more local level of individual genes, we have found a preferential localisation of sequences prone towards specific structures within intergenic regions. We find that structural profiles for aligned promoter regions generally tend towards a more rigid structure within about a hundred bp upstream of the transcription start sites, and more deformable are immediately downstream of the +1 sites. We see the same general pattern for all of the genomes in this study. The biological significance of our finding of similar types of DNA structures in promoter regions for such a wide range of organisms is discussed.

## Novel Approaches for the Identification of Microbial Antigens

Michael Buschle, Aaron Hirsh, Walter Schmidt, Max L. Birnstiel and **Alexander von Gabain.** INTERCELL Inc., Vienna Biocenter, Vienna, AUSTRIA.

INTERCELL (IC) is a young biopharmaceutical, science-driven company at the Vienna Biocenter, with an integrated system for the development of chemically defined vaccines for the treatment of human diseases including cancer and infections. IC's portfolio comprises a technique, called CISTEM, to identify cancer antigens and pathogen-derived antigens as well as a vaccine technology, termed TRANSVAX, to induce both strong humoral and cellular immunity against these diseases.

CISTEM is a novel technology which allows the selection and identification of cell clones which "present" protein epitopes at the surface and bind to a certain ligand. At its simplest, a candidate ligand-binding epitope is inserted into a receptor which is a known entry site of a pathogenic agent or cytotoxic drug; e.g., a virus. Ligand and virus are mixed with cells. Binding of the ligand to the epitope prevents adhesion of the virus to the cell, and thereby its destruction. Cells displaying a binding epitope are thus positively selected and the genetic information coding for the displayed protein can be rapidly and easily retrieved using standard molecular techniques. CISTEM far exceeds in sensitivity and specificity other molecular techniques by a very large margin. CISTEM is, therefore, extremely well suited for the detection of protein-ligand interactions, including the discovery of antigens.

TRANSVAX also developed by IC's scientists allows the presentation of disease-specific peptides on the surface of dendritic cells, a special class of antigen-presenting cells of the immune system. This is achieved by the simple procedure of mixing proteins or peptides with the adjuvants polyarginine or polylysine and injecting the mixture subcutaneously. Experimental evidence suggests that the dendritic cells charged with tumor-specific proteins or peptides under the influence of the TRANSVAX adjuvants migrate to draining lymph-nodes where they activate cancer-specific cytotoxic lymphocytes. These then patrol the body and eradicate distant cancer cells. Mice vaccinated with such vaccines resist challenge with live tumor cells and do so equally well as mice vaccinated with genetically altered syngeneic tumor cells. The TRANSVAX technology is not restricted to therapeutic applications, but can also be employed to transload dendritic cells for the identification of antigens.

The discovery of antigens is of pivotal significance in human cancer and diseases caused by viruses, parasites and bacteria. Bacterial genomics provide an unique wealth to "extract" protein epitopes from the genome that are recognized by the immune system. Immunisation of mice by pathogenic microbes leads to the formation of specific antibodies and lymphocytes. We discuss strategies how IC's technologies are combined to identify the cognate peptides recognized by the humoral and cellular arm of the immune system.

# Index to Poster Abstracts

**POSTER SESSION**
**Saturday, January 30, 1999**
**7:30 pm**


| | Poster Number | |
|---|---|---|
| **Alterovitz**, Gil | 01 | Comparing Secondary Structure Prediction with Genomic Structure Datasets from *Methanococcus jannaschii* and *Aquifex aeolicus* |
| **Amano**, Naoki | 02 | Informatical Analysis of Archaeal Genomic DNA Sequences 1: Informatical Basis for ORF Identification and Its Application to Developing an Efficient Algorithm for the Precise ORF Identification |
| **Azuma**, Yoshinao | 03 | Informatical Analysis of Archaeal Genomic DNA Sequences (2): Gene Composition of Archaeal Genomes |
| **Bentley**, Stephen | 04 | Genome Sequencing of *Neisseria meningitidis* Serogroup A Strain Z2491 |
| **Brettin**, Thomas S. | 05 | Nirvana: A Microbial Genome Annotation Tool Kit |
| **Cahill**, Dolores | 06 | Automated High-throughput Technologies in Expression Analysis of *Neisseria meningitidis* and *Neisseria lactamica* |
| **Clayton**, Rebecca A. | 07 | Complete Genome Sequencing of *Vibrio cholerae* and *Shewanella putrefaciens:* An Opportunity for Comparative Genomics |
| **Eroshkin**, Alexey M. | 08 | Automatic Clustering and Domains Dissection in Bacterial and Eukaryotic Proteomes |
| **Gill**, Steven R. | 09 | Genome Analysis of *Staphylococcus aureus* strain COL |
| **Goodner**, Brad W. | 10 | An *Agrobacterium tumefaciens* Genome Project at a Primarily Undergraduate Institution—Current Progress and Future Goals |
| **Govindarajan**, Sridhar | 11 | "The Master Catalog" as a Genome Annotation Tool |
| **Guiliano**, David B. | 12 | The Human Filarial Parasite *Brugia malayi* Contains an a-Proteobacteria Endosymbiont |
| **Harris**, David | 13 | The *Candida albicans* Sequencing Project at The Sanger Centre |
| **Hartsch**, Thomas | 14 | *Thermus thermophilus*: the Genome of an Extremely Thermophilic Bacterium with a Natural Transformation System |
| **Heilig**, Roland | 15 | Complete Genomic Sequence of the Hyperthermophilic Archaeon *Pyrococcus abyssi* |
| **Hickey**, Erin K. | 16 | Automated Data Management for TIGR Microbial Genomes: Small_Genome_Control |
| **Hide**, Winston | 17 | *Mycobacterium tuberculosis* has Eukaryotic Genes |
| **Hihara**, Satoshi | 18 | Micro Array Analysis of Gene-Expression during the Yeast Life Span |

| | | |
|---|---|---|
| **Hinkle**, Gregory | 19 | Comparative Genomics of *Aspergillus nidulans* |
| **Hogue**, Christopher | 20 | A Potential to Predict Protein Thermostability Derived from W.V.the Genome of *Methanococcus jannaschii* |
| **Holtzapple**, Erik | 21 | Genome Sequencing of *Streptococcus pneumoniae* Type 4 |
| **Jacobi**, Carsten | 22 | *Methanosarcina mazei*, the First Genome of the Third Order of Methanogens |
| **Kawarabayasi**, Yutaka | 23 | The First Entire Genomic Sequence of an Aerobic Hyperthermophilic Crenarchaeota, *A. pernix* K1, Comparing with an Anaerobic Hyperthermophilic Euryarchaeota, *P. horikoshii* OT3, and Other Archaea |
| **Ketchum**, Karen A. | 24 | Production of High-density DNA Microarrays for Microbial Genomes at TIGR |
| **Khouri**, Hoda M. | 25 | Strategies for Optimizing the Sequencing of Difficult Genomic Regions |
| **Larimer**, Frank | 26 | Annotation of Microbial Genomes |
| **Lawson**, Daniel | 27 | Protozoan Sequencing at The Sanger Centre |
| **LI**, Guangshan | 28 | Molecular Diversity of the Genes Involved in Anaerobic Energy Metabolism in *Shewanella putrefaciens* MR-1 |
| **Liu**, Jia Yeu | 29 | The Identification of a New Family of Sugar Efflux Pumps in *Escherichia coli* |
| **Machida**, Masayuki | 30 | Pilot Scale Genome Sequencing of *Aspergillus nidulans* by an Internal-labeling Long Read Protocol |
| **Michalickova**, Katerina | 31 | An In-House Integrated Database System |
| **Movahedzadeh**, Farahnaz | 32 | Genes Involved in the Inositol Metabolism of Mycobacteria |
| **Mungall**, Karen | 33 | Genome Sequencing of *Campylobacter jejuni* strain NCTC11168 |
| **Murphy**, Lee | 34 | The *Streptomyces coelicolor* Genome Sequencing Project at The Sanger Centre |
| **Ohfuku**, Yuko | 35 | Positioning of High Repetitive DNA Sequences in Reference to the Direction of Transcription and Replication |
| **Pallen**, Mark J. | 36 | Genome Browsers for Bacterial Pathogens |
| **Parkhill**, Julian | 37 | Microbial Genome Sequencing at the Sanger Centre |
| **Peterson**, Scott | 38 | DNA Microarray Gene Expression Monitoring of the Extreme Radiation Resistant Bacteria *Deinococcus radiodurans* |
| **Pham**, Xuan-Quynh T. | 39 | Sequencing of *Pseudomonas aeruginosa* PA01 |

| | | |
|---|---|---|
| **Qi**, Rong | 40 | Whole Genome Shotgun Optical Mapping of *Deinococcus radiodurans* and *Trypanosoma brucei* |
| **Qin**, Haiying | 41 | Identification of a Novel Site-specific DNA-binding Activity in *Deinococcus radiodurans* |
| **Qiu**, J. | 42 | Direct Fluorescent Labeling and Automated Analysis of Genomic Clones for Mapping |
| **Raleigh**, Elisabeth A. | 43 | On the Specificity of Insertion of MiniTn7 Elements Mediated by TnsABC$^{A227V}$ |
| **Reynolds**, Thomas R. | 44 | Development of a Novel Lane Tracking Reagent for Automated DNA Sequencers |
| **Simons**, Guus | 45 | cDNA AFLPÆ: A Robust High-throughput Transcript Profiling Technology for Microorganisms |
| **Slaven**, Bradley E. | 46 | Distributed Homology Searches using PVM and WU-Blast on a Heterogenous Computer Network |
| **Stoker**, Neil G. | 47 | A DNA Microarray Facility for the Functional Analysis of Bacterial Pathogens |
| **Takami**, Hideto | 48 | Genome Analysis of Facultatively *Alkaliphilic bacillus Halodurans* C-125 |
| **Tatusova**, T. | 49 | Microbial Genomes at NCBI: Complete and Incomplete |
| **Vamathevan**, Jessica J. | 50 | Closure or Ribosomal RNA Repeats in Whole Genomes |
| **Ussery**, David | 51 | Will present his abstract from his speaker presentation |
| **Viari**, A. | 52 | Replication-induced Constraints on Genes and Proteins Composition in Bacteria |
| **Whiteside**, Simon T. | 53 | Functional Analysis of Microbial Genomes Using an Exhaustive Two-hybrid Approach |
| **Williams**, Kerstin | 54 | Random Shot-gun Cloning and Sample Sequencing of the *Francisella tularensis* |
| **Wood**, Todd C. | 55 | Estimating the Extent of Gene Transfer by Simulating Proteome Evolution |
| **Wu**, Thomas.D. | 56 | A Combinatorial Approach to Finding Promoter Elements in Genomic Sequences |
| **Zengel**, Janice M. | 57 | Phylogenetic Analysis of the Regulation of Ribosomal Protein Synthesis |

# Poster Abstracts

## Comparing Secondary Structure Prediction with Genomic Structure Datasets from *Methanococcus jannaschii* and *Aquifex aeolicus*

Gil Alterovitz and Christopher W.V. Hogue. Samuel Lunenfeld Research Institute, Mt. Sinai Hospital, Toronto, CANADA and Department of Biochemistry, University of Toronto, CANADA.

We wished to test an hypothesis regarding the effect of genomic protein composition on secondary structure prediction. Conventionally, secondary structure prediction methods scramble the taxonomic information in their training sets. It has been noted anecdotally that secondary structure prediction with thermophilic sequences is poor. We wondered whether a taxonomically consistent database would improve the performance of secondary structure prediction. Thermophilic organisms are an obvious starting point as the thermostable proteins in their genomes have clear differences in physicochemical properties, and these could be used in comparison with a heterogeneous group of structures gathered from the PDB in a taxonomy-independent fashion. Genomic sequences from *M. jannaschii* and *A. aeolicus* with PDB structural homologues were carefully aligned and assigned the corresponding secondary structure as recorded by the MMDB annotation. These sequence-secondary structure datasets were compiled into query databases separately for corresponding genomic proteins. We tested secondary structure prediction with the GOR method, as it does not use multiple alignment information or neural networks which would have interfered with our hypothesis regarding taxonomy. We varied the quality of alignments in these genome specific databases. Lowering alignment identity cutoff thresholds from 25% to 15% results in larger scoring databases that improved prediction. We used the two genome-based sequence database to provide query proteins, and made predictions using GOR's usual scoring database. The Q3 score of *M. jannaschii* and *A. Aeolicus* were poorer than the GOR method Q3, confirming the anecdotal reports of poor results in secondary structure prediction, and we are pursuing mesophilic genome controls for this data. In particular, beta strand prediction was no better than chance. We then replaced GOR's scoring database with the thermophile genome specific database and repeated the comparison with jackknifing. This showed significant improvement in beta strand predictability.

## Informatical Analysis of Archaeal Genomic DNA Sequences 1: Informatical Basis for ORF Identification and Its Application to Developing an Efficient Algorithm for the Precise ORF Identification

Naoki Amano[1,2], Yuko Ohfuku[1,3,4], Jun Kakinuma[1,5], Hideaki Koike[1], Masaru Tateno[1], Joerg M. Suckow[1], Masashi Suzuki[1,5]. [1]AIST-NIBHT CREST Centre of Structural Biology, Tsukuba, JAPAN; [2]Doctoral Program in Medical Sciences, University of Tsukuba, Tsukuba, JAPAN; [3]Doctoral Program in Agricultural Sciences, University of Tsukuba, Tsukuba, JAPAN; [4]National Institute of Technology and Evaluation, Tokyo, JAPAN; and [5]Graduate School of Human and Environmental Sciences, University of Tokyo, Tokyo, JAPAN.

We have developed an informatical method that can be used in order to identify genes, operon structures and pseudo-genes on the basis of archaeal genomic DNA sequences. Identification is carried out, first by deducing the standard signal sequences for transcription and translation, and physical characteristics that are expected to be shared by all the genes in the genome, and second by analyzing all the candidates informatically by using scoring systems that can evaluate the closeness of given sequences to the standard sequence. The first process is carried out by using a limited number of genes that are essential for the function of the genome and thus can be identified manually. Thus it can be regarded as a knowledge-based approach, while the second step statistical. Gene identification is essentially a process of separating formal candidates that are defined as sections which start with possible start codons and end at possible stop codons, while can code certain numbers of amino acid residues in between, to two groups, real genes and artifacts. It is essential that the success rate of separating the two groups can be estimated, when a certain cut-off value is used for the identification of each type of signal. By monitoring the success rate the cut-off value can be optimized. By combining multiple scoring systems a set of cut-off values can be chosen, so that the number of false negative entries is expected to be close to 0—i.e. no real gene is missing, while the rate of false positive entries is kept smaller than 10%. An algorithm has been developed, which can analyze a whole archaeal genomic DNA sequence semi-automatically in three hours with a set of given cut-off values, and by the use of which the cut-off values can be optimized in less than a week.

## Informatical Analysis of Archaeal Genomic DNA Sequences (2): Gene Composition of Archaeal Genomes

Yoshinao Azuma[1], Yuko Ohfuku[1,2,3], Hideaki Koike[1], Naoki Amano[1,4], Jun Kakinuma[1,5], Masaru Tateno[1], Joerg M. Suckow[1], Masashi Suzuki[1,5]. [1]AIST-NIBHT CREST Centre of Structural Biology, Tsukuba, JAPAN; [2]Doctoral Program in Agricultural Sciences, University of Tsukuba, Tsukuba, JAPAN; [3]National Institute of Technology and Evaluation, Tokyo, JAPAN; [4]Doctoral Program in Medical Sciences, University of Tsukuba, Tsukuba, JAPAN; and [5]Graduate School of Human and Environmental Sciences, University of Tokyo, Tokyo, JAPAN.

By using the algorithm developed by Amano *et al.* for the identification of the transcription and translation units, the

genomic DNA sequence of a hyper-thermophilic archaeon, Pyrococcus sp. OT3 has been analyzed. Among 1788 ORFs identified, 220 are closer to eukaryotic homologues than to eukaryotic homologues (here referred to as eukaryotic type genes), while 545 are closer to eubacterial homologues than to eukaryotic homologues (here referred to as eubacterial type genes).

Genes of the eubacterial type are found to be distributed equally covered the whole genome of P. OT3, while genes of the eukaryotic type cluster within the region that covers approximately 20% of the genome. The border separating the genes of the eukaryotic type and eubacterial type does not lie between two of the major functional categories-i.e. transcription, translation, replication, etc. But each functional category is further divided to the two groups of the eukaryotic and eubacterial type. The category of transcription is composed of the genes of eukaryotic type that is involved in general transcription and the genes of eubacterial type that is involved in specific transcription regulation. The category of replication is composed of the genes of eukaryotic type that is involved in replication initiation and the genes of eubacterial type that is repair.

It suggests that the ancestral eukaryotic genome inserted to the ancestral eubacterial genome to create the ancestral archaebacterial genome and the genes of eukaryotic and eubacterial types differently contribute in archaebacteria. The simple synbiogenesis hypothesis of eukaryotic cells needs to be modified, since the hypothesis in the simplest expectation, could be concluded that eukaryotic genomes are mosaic, while archaebacterial and eubacterial genomes are homogenous.

## Genome Sequencing of *Neisseria meningitidis* Serogroup A Strain Z2491

**Stephen Bentley**, Julian Parkhill and the Pathogen Sequencing Group. The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

The Sanger Centre has nearly completed the sequence of *Neisseria meningitidis* serogroup A strain Z2491. Serogroup A strains of *N. meningitidis* are responsible for epidemics and pandemics of meningitis in the developing world, in contrast to the serogroup B and C strains, which are responsible for endemic disease in Europe.

The genome is being sequenced in collaboration with Brian Spratt at the Centre for the Epidemiology of Infectious Disease, University of Oxford and Mark Achtman at the Max-Planck-Institute for Molecular Genetics. We are following a whole genome shotgun approach, using two pUC libraries with insert sizes of 0.5 to 0.8 kb and 1.0 to 1.5 kb. The relatively small insert size was chosen to avoid problems with the small dispersed intergenic repetitive elements present in the Neisserial genome. To aid the assembly we have also generated end sequences from 10kb clones in a low copy number plasmid vector, and 20kb clones in a lambda-based vector.

At the time of writing, the sequence consists of 68 contigs over 1kb, totalling 2.167 Mb and assembled from 53,360 individual sequences. The assemblies are updated weekly, and are available from the Sanger Centre website at: http://www.sanger.ac.uk/Projects/N_meningitidis/.

The results of some preliminary analysis will be presented.

## Nirvana: A Microbial Genome Annotation Tool Kit

**Thomas S. Brettin**, Mark D. Brettin and Gerald Myers. Theory Division, Los Alamos National Laboratory, NM.

The recent release of several complete bacterial genomes has prompted us to develop a graphical user interface (GUI) to a relational database. The database maintains information that is relevant to a particular bacterial genome and the GUI allows the user simple access to the data. The driving force behind this project is the desire to make simple the organization, compilation, storage, and retrieval of genetic sequence information. The current application allows for the handling of multiple bacterial genomes. In addition to bacterial genomes, the Nirvana package has been successful on plasmid sequences. Nirvana is being used as our in-house GUI providing annotators a facility to access the data behind the publicly accessible STD genomes database (http://www.stdgen.lanl.gov). It is fully dynamic. Newly added information is immediately available without the need to rebuild the database. This application makes the addition of new genomes and uploading of large data sets fast. Automatic updating features exist in the package that help maintain the integrity of dependent data. Release of software packages of this nature will aid the experimentalist in formulating hypotheses for laboratory evaluation. The package is simple to use, which allows the user to focus more time and energy on laboratory discovery. The foundation of the Nirvana package is built on the Apache web server. Navigation through the package is done on an Intranet or on the Internet with a web browser such as Netscape. This work was funded by the Sexually Transmitted Diseases Branch of the Division of Microbiology and Infectious Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda Maryland.

## Automated High-throughput Technologies in Expression Analysis of *Neisseria meningitidis* and *Neisseria lactamica*

**Dolores Cahill**, Brian Cusack, Ellen Sattler, Olivia Wegner, Rampalli Srinivas, John O'Brien, Holger Eickhoff, Mark Achtman and Hans Lehrach. Max-Planck-Institute of Molecular Genetics, Department of Professor Lehrach, Berlin, GERMANY.

A full understanding of the expression profile of a tissue or organism requires the screening of many genetic samples, in parallel, as rapidly as possible. We describe those steps which have been automated and miniaturised in our laboratory to enable a high-throughput and highly parallel hybridisation-based approach to expression profiling of *Neisseria meningitidis* and *Neisseria lactamica*, to facilitate the comparative analysis of these genomes. Also initial steps towards large-scale protein analysis will be described, which will assist the functional characterization of these microorganisms.

We use robot technology to clonally array PCR products from genomic DNA expression libraries of *N. meningitidis* and *N. lactamica* on a whole genome basis. Membrane filters are produced by robot spotting at a density of 27, 648 clones per $22.2cm^2$. Each 5x5 block contains a central salmon sperm guide dot surrounded by twelve clones spotted in duplicate. After hybridisation the guide dots permit automatic grid finding and, subsequent, automatic image analysis. Sequences of spotted clones are identified by hybridisation with specific DNA probes, or alternatively, by oligonucleotide

fingerprinting, a technique which has been developed to characterise DNA libraries. The DNA filters have been used to do subtractive and complex hybridisation experiments to profile differentially expressed gene between *N. meningitidis* and *N. lactamica*.

To apply this technology to proteomics, we have clonally expressed proteins from these arrayed genomic DNA expression libraries. These proteins were arrayed at the same density and in the same numbers as DNA filters, these protein filters were used for interaction screening with specific ligands, results obtained with specific antibodies are presented. These antibodies detected specific protein products. This approach makes translated gene products directly amenable to high-throughput experimentation, allowing a link between expressed protein and sequence information, facilitating the construction of gene product-antibody catalogues for these microorganisms.

## Complete Genome Sequencing of *Vibrio cholerae* and *Shewanella putrefaciens:* An Opportunity for Comparative Genomics

**Rebecca A. Clayton**[1], John Heidelberg[1], John J. Mekalanos[2], Rita R. Colwell[3], Kenneth Nealson[4], J. Craig Venter[1] and Claire M. Fraser[1]. [1]The Institute for Genomic Research, Rockville, MD; [2]Department of Microbiology and Molecular Genetics, Harvard Medical School, Boston, MA; [3]Office of the Director, National Science Foundation, Arlington, VA; and [4]NASA Jet Propulsion Lab, Pasadena, CA.

*Vibrio cholerae*, the etiologic agent of cholera, has been completely sequenced and assembled, and is undergoing the final genome annotation process prior to publication. *Shewanella putrefaciens*, a model organism for bioremediation, is midway through the assembly process. We present preliminary analysis of these two genomes, in addition to whole genome comparisons with other completed microbial genomes. *Vibrio cholerae* and *Shewanella putrefaciens* fall close to one another on the small subunit ribosomal RNA phylogenetic tree, and, along with completely sequenced genomes *Haemophilus influenzae* and *Escherichia coli*, are members of the large and diverse gamma subdivision of the Proteobacteria. We present evidence for a highly conserved set of gamma Proteobacteria genes, with evolutionary divergence between these organisms centering in energy metabolism, transport, and other proteins key to environmental interaction.

## Automatic Clustering and Domains Dissection in Bacterial and Eukaryotic Proteomes

**Alexey M. Eroshkin**. Axys Pharmaceuticals, Inc., La Jolla, CA.

Comparative analysis of protein sequences from complete proteomes is a way to predict function for large set of newly sequenced genes. Domain dissection in multi-domain proteins, that are especially common in eukaryotes, and clustering of related domains are essential steps in such analysis. An automatic clustering and domains dissection of proteins from seven complete genomes (*Escherichia coli, Haemophilus influenzae, Mycoplasma genitalium, Mycoplasma pneumoniae, Synechocystis sp., Methanococcus jannaschii, Saccharomyces cerevisiae*) and *C. elegans* (near complete at the time of analysis) was made using Geanfammer package. First, 32637 proteins were compared by BLAST and clustered to complex single linkage clusters. All complex clusters but largest one (artifactual cluster with 8641 sequences) were broken down by DIVCLUS procedure into simple clusters, that contain only related modules. To solve the biggest cluster, the long sequences (>500 aa) were taken out and the whole process was repeated separately for the short and long sequences. The large sequences were then compared to the clusters produced from the short ones and added to them. Clusters that contain only large sequences were also found. As the result, 16771 single and multi-domain sequences formed 5771 clusters corresponding to single duplication modules. The resulting protein families are between 2 and 258 sequences in size. The duplication modules are between 50 and 4100 residues in size. The biggest protein families are eukaryotic protein kinases (258 proteins), ABC transporters (249), 7TM receptors (202), zinc finger proteins (170). Close to 7300 proteins in obtained clusters have no functional annotation, biochemical function for many of them may be predicted using functional data from other members of the same cluster. Clusters that have proteins from three or more distant phylogenetic lineages correspond well to Clusters of Orthologous Groups (COG), recently developed at NCBI. Extension of the results obtained after the completion of *C.elegans* genome will be discussed.

## Genome Analysis of *Staphylococcus aureus* strain COL

**Steven R. Gill**, Brook Craven, Teresa Utterback, J. Craig Venter and Claire M. Fraser. The Institute for Genomic Research, Rockville, MD.

The gram-positive cocci, *Staphylococcus aureus* is recognized as one of the most common causes of both hospital- and community acquired infections worldwide. Of the ~40 million yearly admissions to acute care hospitals in the United States, ~2-3 million patients (6%) develop nosocomial infections- ~300,000 (10%) of those caused by *S. aureus*. They are members of the indigenous mammalian microbial flora, and in humans, inhabit the skin, upper respiratory and intestinal tracts. As opportunistic pathogens, they are prone to initiate infections when host resistance has been compromised by a skin cut or burn, a previous bacterial or viral infection, chemotherapy, extensive surgical therapy, or introduction of biomaterials such as catheters or prosthetic devices. *S. aureus* infections are often acute and with the evolution of multiply antibiotic resistant strains, including the recent isolation of vancomycin resistant *S. aureus* in the United States, innovative approaches are urgently needed to control this major bacterial pathogen. At TIGR, we are sequencing the *S. aureus* strain, COL; the prototypic strain used by many labs to investigate the genetics and biochemistry of methicillin resistance. COL is also one of the methicillin resistant *S. aureus* (MRSA) source strains used to generate a vancomycin resistant isolate, and may serve as a model for investigating vancomycin resistance.

The 2.8 Mb COL genome is being sequenced by the random whole genome strategy used to complete the sequence of multiple bacterial genomes at TIGR. Random 1.6-2.0 kb and 3.0-4.0 kb libraries were constructed and to date, over 38,000 sequence fragments have been obtained and assembled into contigs representing greater than 2.7 Mb unique sequence. We are currently at ~6X genome coverage and anticipate reaching 8X coverage (~48,000 sequence fragments) before initiating

closure. A summary of the projects current status and a preliminary analysis of gene content will be presented.

*This work is funded by NIH grant RO1AI43567-01.*

## An *Agrobacterium tumefaciens* Genome Project at a Primarily Undergraduate Institution—Current Progress and Future Goals

**Brad W. Goodner**, Brian P. Markelz, M. Casey Flanagan, Chris B. Crowell, Jodi L. Racette, Brittany A. Schilling, J. Scott Mellors and Courtney M. Lappas. Department of Biology, University of Richmond, VA.

Active involvement of undergraduate students in "cutting edge" technology is a fantastic teaching opportunity and can lead to productive research progress. This poster describes a genome project involving undergraduate students, both in the classroom and in independent projects. With the exception of a few chromosomal virulence genes, the chromosome(s) of *Agrobacterium tumefaciens* have been relatively understudied in comparison to the Ti (Tumor-inducing) plasmid. Several early genetic mapping studies were consistent with a single circular chromosome. However in 1993, Allardet-Servent *et al.* (J. Bacteriology 175:7869-7874) presented pulse-field gel electrophoresis (PFGE) data showing that *A. tumefaciens* C58 has two chromosomes, a 3.0 Mbp circle and a 2.1 Mbp linear. In order to resolve the differences between these two data sets, we have constructed a combined physical and genetic map of the genome of *A. tumefaciens* A348, a derivative of C58. We collected a large number of auxotrophic and prototrophic transposon insertions using Tn5(pfm) (Wong & McClelland, 1992. J. Bacteriology 174:3807-3811), and physically mapped each Tn5(pfm) insertion using rare restriction sites and PFGE. Our results confirm the results of Allardet-Servent *et al.* and go much further to establish the order of PacI and SwaI restriction sites and various genes on each chromosome. We will present our current chromosome maps, the essential features of the genome, and a possible explanation for why the early genetic mapping experiments missed the linear chromosome. We are now in the genome sequencing phase of the project, focusing on a 710 kbp PacI fragment containing one end of the linear chromosome. We will present our genomic cloning and sequencing strategy, summarize the current state of the project, and look to the future.

References:
1. Allardet-Servent *et al.*, 1993. J. Bact. 175:7869-7874.
2. Cooley & Kado, 1991. Mol. Gen. Genetics 230:24-27.
3. Hooykaas *et al.*, 1982. Mol. Gen. Genetics 188:12-17.
4. Metts *et al.*, 1991. J. Bact. 173:1080-1087.
5. Miller *et al.*, 1986. Mol. Gen. Genetics 205:153-159.
6. Pischl & Farrand, 1984. J. Bact. 159:1-8.
7. Wong & McClelland, 1992. J. Bact. 174:3807-3811.

## "The Master Catalog" as a Genome Annotation Tool

**Sridhar Govindarajan**[1], Stephen G. Chamberlin[1], Lukas J. Knecht[1] and Steven A. Benner[2], [1]Bioinformatics Division, Sulfonics Inc., Alachua, FL; and [2]Department of Chemistry, University of Florida, Gainesville, FL.

The rapidity in which new genomic data can be obtained presents new challenges to the molecular biologist. Tools for gene finding and assignment of function to putative genes are of key importance. We have approached this problem by first constructing a "Master Catalog" from previously published genomic sequence data. The essence of this method is to group fragments into families based on their evolutionary histories. This approach has major benefits: an evolutionary history can be traced from the data and related to each member; the family can be represented as a probabilistic sequence and used to identify other members; most importantly, the observed patterns of variation (at both the protein and DNA level) can be used to infer details of the underlying evolutionary constraints. This is the essence of *ab-initio* annotation—the automated identification of features in genomic sequence data. We demonstrate how our recently reported "Master Catalog" can be used to assist in the annotation process, illustrated with an analysis of the newly sequenced *Rickettsia prowazekii* genome. We show how the evolutionary view provided by the "Master Catalog" can be navigated and used to confirm or deny functional relationships with homologous proteins.

## The Human Filarial Parasite *Brugia malayi* Contains an α-Proteobacteria Endosymbiont

**David B. Guiliano**[1], Barton E. Slatko[2], and The FGP[3]. [1]ICAPB, Ashworth Labs, King's Buildings, University of Edinburgh, Edinburgh, SCOTLAND; [2]New England Biolabs, Inc., Beverly, MA; and [3]Filarial Genome Project.

Filarial nematodes are responsible for lymphatic elephantiasis and cutaneous filiariasis, including African river blindness. These parasites infect over 100 million people worldwide causing widespread morbidity in endemic populations. Transmission election microscopic (TEM) studies of several filarial nematodes revealed the presence of intracellular bodies in the hypodermis of larvae and adults and in the gonads of females. It was speculated that these represented an endosymbiotic bacterial association unique to filarids. In 1992, the Special Programme for Research and Training in Tropical Diseases (TDR) of the World Health Organization sponsored the creation of a Filarial Genome Project. This international collaboration of seven endemic and non-endemic laboratories has successfully implemented a program of gene discovery, genome mapping, and post genomic analysis. cDNA libraries representative of the entire life cycle and large insert genomic DNA libraries have been created and serve as the core of the gene discovery and genome mapping initiatives. These resources have also enabled the positive identification and preliminary characterization of the genome of the endosymbiotic bacteria identified in the TEM studies. The endosymbiont is closely related to Wolbachia, a group of a-proteobacteria which establishes intracellular infections in insects and arthropods. Mapping of the genome of the endosymbiont found in the lymphatic filarial nematode *Brugia malayi* using a BAC (Bacterial Artificial Chromosome) library gridded on a high density filter array is in progress. Several genes useful for

phylogenetic and immunological studies, including *ftsZ* and *GroEL* homologues are also being characterized. Once the physical map is completed, a minimum tiling path of BACs will be used as a template for sequencing of the entire endosymbiont genome, as preliminary studies have indicated that the endosymbiont may be a novel target for drug development.

## The *Candida albicans* Sequencing Project at The Sanger Centre

**David Harris** and The Pathogen Sequencing Unit, The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

*Candida albicans* is the most frequent cause of human fungal infections. This pathogenic yeast has a 15 Mb genome contained in 8 chromosomes that vary in size between 1 Mb (chromosome 7) and 4 Mb (chromosome R). A project to sequence mapped cosmids from strain 1161 is underway at The Sanger Centre in collaboration with Prof. Duncan Shaw (University of Aberdeen, UK). The pilot project is intended to produce regions of finished sequence to examine the feasibility of a clone based strategy to sequence the entire genome. The project has produced 232 kb finished sequence from 6 cosmids. The sequence contains known Candida genes (including RCC1, POL3, CANIK1 and CHS1) and also two complete copies of a large scale repeat which has previously been partially characterised (RPS1). A comparison of homologous proteins showed expected similarities between *C. albicans* and *S. cerevisiae* and regions with with similar gene organisations in the two yeasts have been identified. This work is being funded by The Wellcome Trust.

## *Thermus thermophilus*: the Genome of an Extremely Thermophilic Bacterium with a Natural Transformation System

**Thomas Hartsch**[1], Carsten Jacobi[1], Mechthild Marschall[1], Ulrike Bode[1], Hans-Peter Klenk[1,2], Reinhard Sterner[1], Gerhard Gottschalk[1], and Hans-Joachim Fritz[1]. [1]Goettingen Genomics Laboratory, Institute of Microbiology and Genetics, Goettingen, GERMANY; and [2]EPIDAUROS Biotechnology AG, Bernried, GERMANY.

*Thermus thermophilus* is a Gram-negative, aerobic bacterium which can grow at temperatures from 52 to 85°C. *T. thermophilus* strain HB27 was originally isolated from a natural thermal environment in Japan and is closely related to *T. flavus* and *T. thermophilus* strain HB8. It is less closely related to *T. aquaticus* (the type species of the genus Thermus and the source of Taq polymerase). The genus Thermus is considered to be a deep branch within the domain Bacteria, located within the cluster of the green nonsulfur-bacteria along with the genus Deinococcus. Revaluation of the phylogenetic position of the genus Thermus, based on protein sequence information, is one of the reasons for our genome sequencing project.

The genomes of several extremely thermophilic and hyperthermophilic microbes (archaea and bacteria) have been sequenced during the last three years; however, none of these microbes has a biotechnologically useful transformation system. *T. thermophilus* is the only presently known extremely thermophilic organism that can be (naturally) transformed with high efficiency. This property makes *T. thermophilus* a valuable tool for genetic studies and an attractive target for whole-genome sequencing.

A whole-genome sequencing project (random shotgun approach) of *T. thermophilus* strain HB27 is presently underway in our laboratory; the genomes of *T. thermophilus* strain HB8 and of *T. aquaticus* are currently also in the process of being sequenced in other labs. The relatively high GC-content of *T. thermophilus* (about 70%) and the large number of known IS-elements are the main difficulties for the analysis of the 1.82 Mbp genome and the 0.24 Mbp mega-plasmid pTT27 of *T. thermophilus* strain HB27.

## Complete Genomic Sequence of the Hyperthermophilic Archaeon *Pyrococcus abyssi*

**Roland HEILIG**[1], Valerie BARBE[1], Simone DUPRAT[1], Catherine ROBERT[1], Virginie VICO[1], Jean-Louis PETIT[1], Patrick WINCKER[1], Philippe BROTTIER[1], Odile LECOMPTE[3], Joel QUERELLOU[3], Gael EURAUSO[4], Daniel PRIEUR[4], Olivier POCH[5], Jean-Claude THIERRY[5], Patrick FORTERRE[2], Jean WEISSENBACH[1], Yvan ZIVANOVIC[2] and William SAURIN[1]. [1]GENOSCOPE,Centre National de Sequencage, Evry Cedex, FRANCE; [2]IGM, University Paris-Sud, Orsay Cedex, FRANCE; [3]IFREMER, Plouzané, FRANCE; [4]Laboratoire de Microbiologie Marine, Station Biologique, Roscoff Cedex, FRANCE; and [5]IGBMC, Illkirch Cedex FRANCE.

Hyperthermophilic archaea of the genus Pyrococcus are presently among the most studied hyperthermophiles world wide. *Pyrococcus abyssi*, strain GE5, which was isolated from a deep see hydrothermal vent, contains a plasmid of 3.4 kb, pGT5, currently used as starting material to design genetic tools for hyperthermophilic archaea. This plasmid should also be helpful in studying DNA replication and DNA repair in archaea. *P. abyssi* (GE5 strain) could thus become a model species for hyperthermophilic archaea. In order to complete our knowledge of this organism, sequencing of its 1.8 Mb long chromosome was performed at Genoscope by subjecting the entire molecule to a pairwise ends sequencing strategy. Two types of libraries were constructed: the first one, in a plasmid vector of Bluescript type, with 5ñ6 kb sized inserts, was responsible for the major part of the sequence contigs; the second one, made in a cosmid vector (inserts ranging 35 to 45 kb) was essential for the scaffolding on large distance. The initial shotgun phase was carried out on LI-COR 4200 sequencer (900bp average). Assembly was achieved with PHRED/PHRAP and monitored with CONSED. When the number and size of the gaps, deduced from the scaffold, were judged short enough, the directed walking step has been initiated. This stage was attained with 19300 plasmids and 1635 cosmids end-sequences. Twelve residual cloning gaps, 0.8 to 5.0 kb in size, were recovered by sequencing three independent LA-PCR products. Walking and finishing reactions were analysed on ABI 377 sequencer. Finally, validation of the complete assembly was achieved by comparing the sequence-deduced map to the patterns of four restriction digests from *P. abyssi* total genomic DNA. ORFs spanning at least 100 amino acids were determined. Regions homologous to proteins contained in Swissprot were used to train a hidden Markov model which was then used to assess the coding potential of each ORF. An automatic annotation performed by comparing these potential genes with the COG databank and Swissprot. One cluster of gene encoding ribosomal RNAs was identified by similarity with the RNA subsection of Genbank and 46 tRNAs were identified by using the tRNAscanSE program.

A web site (http://www.genoscope.cns.fr/PyrococcusAbyssi) will disseminate this data. An expert annotation is under way.

## Automated Data Management for TIGR Microbial Genomes: Small_Genome_Control

**Erin K. Hickey**, Jeremy D. Peterson and Owen White. The Institute for Genomic Research, Rockville, MD.

With updates to the annotation of numerous completed microbial genomes and the current annotation of genomes in progress, the management of new data would become overwhelming if attempted in a non-automated fashion. New data can consist of changes to the physical characteristics of an open reading frame (ORF), such as the coordinates, or to information associated with an ORF, such as the best non-redundant-amino-acid database match (olapid). These two types of data are stored in a number of different tables in TIGR's small-genome relational-databases and are edited through TIGRs annotation tools. A change in a single physical or associated characteristic, i.e. a change to one field in one table, may lead to subsequent changes in other fields across one or several tables. For example, a change in the 5' end of an ORF may require updates to the fields containing the nucleotide sequence, amino acid sequence, pairwise alignment (including a new 'similarity' and 'identity'), number of membrane spanning regions, molecular weight and pI, and potential signal peptide. In addition, the multiple FASTA files containing all the ORF's nucleotide and protein sequences would have to be updated. Alternately, a change in olapid would only affect the pairwise alignment and not the physical characteristics of the ORF.

A simple approach to ensure all database fields reflect the current data would be to update all information on a periodic, perhaps nightly, basis. However, with over 15,000 microbial ORFs being maintained at TIGR, this would be a severe waste of computational resources. To solve this problem, the TIGR microbial annotation team uses a system, Small_Genome_Control (SGC), to handle the day-to-day processing of this data. SGC has been developed to only update information on ORFs whose physical or associated characteristics have been changed, thereby limiting the number of ORFs being processed. In addition, SGC only updates those fields related to the initial change, limiting the number of changes per ORF. A description of SGC's single UNIX makefile and the 25+ PERL scripts will be provided.

## *Mycobacterium tuberculosis* has Eukaryotic Genes

**Winston Hide**, Andrey Ptitsyn and Junaid Gamieldien. South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa.

We report evidence of multiple gene transfer between eukaryotes and a pathogenic bacterium. *Mycobacterium tuberculosis* infects one third of all human beings and is the most widespread of lethal human pathogens (1). The bacterium lives inside macrophages, providing opportunity for gene transfer from hosts as a result of selection for intracellular survival. At least 8 genes have been transferred from eukaryotes to *Mycobacterium tuberculosis*. All share functional similarity with those normally located in the mammalian mitochondrial apparatus, providing a clear indication that *M. tuberculosis* is strongly selected by respiratory constraints. Evidence for transfer, and phylogenetic support is provided. Detection of horizontal gene transfers into pathogens may potentially be used as a marker of the major gene classes under selection for virulence.

1. Bloom, B. R., Murray, C. J. L. Tuberculosis: commentary on a re-emergent killer. *Science* 257, 1055-64 (1992).

## Micro Array Analysis of Gene-Expression during the Yeast Life Span

**Satoshi Hihara**[1], Tomoko Yabe[1], Hideaki Hiraki[1], Yasuo Uemura[1], Akio Kanai[1] and Hirofumi Doi[1,2]. [1]Doi Bioasymmetry Project, ERATO, JST; and [2]Fujitsu Labs Ltd., Tsukuba, Japan.

Budding yeast, *Saccharomyces cerevisiae*, has a limited life span and the mean life span varies in a strain-dependent manner. To understand yeast cellular senescence at the molecular level in detail, we analyzed gene expression changes at the whole genome level during the course of aging of the strain CG379 (mean life span=19.1 $B!^ (B6.6) by subtraction between the cDNAs of young cells and those of old cells.

As a first step, we prepared the pure populations of young cells which divided 3 times and old cells which divided 15 times by a combination method of cell separation in sucrose gradients and cell sorting by biotinylation and avidin conjugated magnetic beads. After extraction of total RNAs from these cells, cDNAs were synthesized. DNA fragments of up-regulated genes in young or old cells were equalized and amplified by using PCR-SelectTM cDNA Subtraction Kit (CLONTECH). The resulting PCR products were labeled with 33P and hybridized to Yeast Index Gene Filters (Research Genetics) which contain 6,144 PCR products corresponding to about 97% of all yeast ORFs.

The signal intensities of all spots were quantified and the ratios of young to old of each signal were determined. To better understand the results, the spots of functionally known genes were classified by their function (e.g., DNA synthesis and replication, biogenesis of cytoskeleton, etc.). Processing of these complicated data was simplified by our newly developed computer program.

The results showed that expression level of genes encoding proteins of cytosolic translation complexes, especially ribosomal proteins, glycolysis pathway, and core histones was down regulated during the course of aging. On the contrary, that of amino acid biosynthesis pathway, nucleotide biosynthesis pathway, and some transporters was up regulated. The differentially expressed gene classes in respect to cellular senescence will be discussed.

## Comparative Genomics of *Aspergillus nidulans*

**Gregory Hinkle**, Vicky Gavrias, Yongwei Cao, Craig DeLoughery, Radhika Uppaluri, JaeHyuk Yu, Roger Wiegand, Tom H. Adams and William E. Timberlake. Cereon Genomics, Cambridge, MA.

We have sequenced the 29 Mb genome of the filamentous ascomycete *Aspergillus nidulans* to greater than 3X depth of coverage using a 'shot-gun' sequence and assembly methodology. Homology-based and algorithmic ORF prediction methods trained for the *A. nidulans* genome

recognize >9300 *A. nidulans* ORFs. Comparative genomic analyses of *A. nidulans and S. cerevisiae* have identified the genes unique to *A. nidulans* as well as the fraction with significant sequence similarity to *S. cerevisiae* coding regions. An overview of the annotation of *A. nidulans* including the *in silico* metabolic reconstruction will be presented.

## A Potential to Predict Protein Thermostability Derived from the Genome of *Methanococcus jannaschii*

Howard Feldman, Gil Alterovitz, Nick Plaskos and **Christopher W.V. Hogue.** Samuel Lunenfeld Research Institute, Mt. Sinai Hospital, Toronto, CANADA, and Department of Biochemistry, University of Toronto, CANADA.

The genome of *Methanococcus jannaschii* has been expected to yield important clues for understanding how amino acids are organized in thermophilic proteins. Threading potentials, as derived by Bryant and Lawerence are made from observed 3-D amino acid contacts from nonredundant 3-D structure data sets, counted as pairwise pseudo-beta-carbon contacts with 10 Angstroms (excluding contacts from i-4 to i+4). We undertook to make such a threading potential for this thermophile using homologous sequences from the *M. jannaschii* genome aligned carefully to known 3-D structures. This new potential may be used as a scoring function to suggest mutations leading to the deliberate engineering of thermostable proteins. High-quality sequence-structure alignments were created with position-specific gap penalties in ClustalW. The additional secondary structure information resolved most ambiguous subalignments by placing them into loops, corresponding to observed indels in protein structure. evolution. The effects of errors owing to fragmentary misalignments was very carefully considered and controlled. We determined through testing that a conservative 25% identity in a subsequence comprising a 3-D structural domain (as defined by the VAST algorithm) was a suitable threshold for including a sequence-structure alignment in our dataset. Of 171 alignments, this process left a dataset of 115 alignments containing domains that passed the threshold. Indel residues were masked for the generation of the potential. Two potentials comprising all pairwise interactions stratified into 6 distance bins (0-5,5-6,6-7,7-8,8-9,9-10 Angstroms) were compiled, for the parent PDB structures, and for the "virtual" *M. jannaschii* structures. The differences between these two potentials highlights all the specific amino acid pairwise interactions that are altered in the thermostable proteins of *M. jannaschii*. Several unsuspected amino acid pairwise interactions are observed. Many of these can be explained by trends in optimizing residue packing, electrostatics and by the influence of the unusual osmolytes found in Methanococcus species.

## Genome Sequencing of *Streptococcus pneumoniae* Type 4

**Erik Holtzapple**, Hervé Tettelin, Matthew Lewis, Diana Radune, Hoda Khouri, Teresa Utterback, Brian A. Dougherty, Erin K. Hickey, Owen White, Hamilton O. Smith, J. Craig Venter and Claire M. Fraser. The Institute for Genomic Research, Rockville, MD.

*Streptococcus pneumoniae*, a encapsulated gram-positive coccus, is the most common cause of bacterial pneumonia around the world as well as a common cause of middle ear infection and meningitis. The Institute for Genomic Research has initiated the sequencing of the whole genome of *S. pneumoniae* type 4 by the shotgun method, the random sequencing phase is complete and the genome is in gap closure phase. This closure process has been hampered by several problems, including genomic DNA purity, non-randomness of large insert libraries (lambda clones), presence of many repeated elements, etc. The methods applied to solve these problems: direct genome sequencing, Clontech genome-walker-kit, multiplex-PCR, and confirmation of the repeats assembly will be presented and discussed, together with the preliminary annotation of the sequence content.

## *Methanosarcina mazei*, the First Genome of the Third Order of Methanogens

**Carsten Jacobi**, Hans-Peter Klenk, Hans-Joachim Fritz and Gerhard Gottschalk. Goettingen Genomics Laboratory, Institute of Microbiology and Genetics, University Goettingen, Goettingen, GERMANY.

Biological methanogenesis plays an important role in the carbon cycle on earth. It is the terminal step of anaerobic breakdown of organic matter in many anaerobic habitats. *Methanosarcina mazei* belongs to the order Methanosarcinales within the group of methylotrophic methanogens. These organisms utilize simple C1-compounds such as methanol and methylamines. In contrast, the group of obligate hydrogenotrophic organisms is restricted to the utilization of H2 + CO2 and formate and can be divided into the orders Methanobacteriales, Methanococcales and Methanomicrobiales. As there are already two completed genome sequences of methanogens available (*Methanococcus jannaschii* and *Methanobacterium thermoautotrophicum*) comparisons of these genomes with that of *M. mazei* should reveal new insights in the phylogenetic relationship of the methanogens. Extensive comparison with the genomic information from phylogenetically more distantly related organisms such as the sulfate reducing *Archaeoglobus fulgidus*, the extremely halophilic *Halobacterium salinarium* and *Thermoplasma acidophilum*, and several recently sequenced species of Pyroccocus should generate valuable information for an improved classification of the Euryarcheaota. The complete genomic sequence of *M. mazei* is being determined by a whole-genome-shotgun approach. In the presently performed initial phase of the project random sequences are generated from a small insert library carrying inserts of approximately 2.5 kb length. Sequencing is done on LI-COR 4200 IR2 machines with dye primer chemistry and cycle sequencing. The sequences are assembled with the Staden Package software running under Solaris 2.6 on a SUN Server. Gaps between contigs will be closed by sequencing of a large insert cosmid library, which will also be used as backbone for the control of the assemblies, and direct sequencing of genomic DNA with dye terminator chemistry on ABI Prism 377 DNA sequencers. The genome annotation will be performed using the MAGPIE (Multipurpose Automated Genome Project Investigation Environment) software for comparison with already annotated sequences from public databases.

## The First Entire Genomic Sequence of an Aerobic Hyperthermophilic Crenarchaeota, *A. pernix* K1, Comparing with an Anaerobic Hyperthermophilic Euryarchaeota, *P. horikoshii* OT3, and Other Archaea

**Yutaka Kawarabayasi**[1,*], Yumi Hino[1], Hiroshi Horikawa[1], Syuji Yamazaki[1], Yuji Haikawa[1], Koji Jin-no[1], Mikio Takahashi[1], Mitsuo Sekine[1], Sin-ichi Baba[1], Akiho Ankai[1], Hiroki Kosugi[1], Akira Hosoyama[1], Shigehiro Fukui[1], Yoshimi Nagai[1], Keiko Nishijima[1], Hidekazu Nakazawa[1], Minako Takamiya[1], Sayaka Masuda[1], Tomomichi Funahashi[1], Toshihiro Tanaka[1], Yutaka Kudoh[1], Jun Yamazaki[1], Norihiro Kushida[1], Akio Oguchi[1], Ken-ichi Aoki[1], Kenji Kubota[1], Yoshinobu Nakamura[1], Yoshihiko Sako[3] and Hisasi Kikuchi[1]. [1]Biotechnology Center, Tokyo, JAPAN; [2]NIBH, Ibaraki, JAPAN; [3]Kyoto University, Kyoto, JAPAN

The entire genome analysis of an anaerobic hyperthermophilic Euryarchaeaota, *P. horikoshii* OT3, was already finished and published*. And the entire sequence data and additional other information is released on the Internet homepage (URL, http://www.bio.nite.go.jp).

To compare with the genome of anaerobic microorganism, the next genome project of an aerobic archaeon, *Aeropyrum pernix* K1 isolated from the Kodakara island Japan in 1993, was started at last April. To determine the entire genome sequence of this archaeon, the alternative whole genome shotgun method was selected. Adding to the whole genome long (2 kb insert) and short (1 kb insert) shotgun clones, the short shotgun clones constructed from the restriction fragments of genomic DNA of *A. pernix* K1 were used. The collected 28,000 raw sequencing data were used for assembling by PhredPhrap. The remaining short gaps (maximum 300 bp long) were filled by walking of long shotgun clones. The determination of the entire nucleotide sequence of this archaeon was finished last December, and this is the first entire genome sequence of hyperthermophile in Crenarchaeota.

To identify both the different and common points between the aerobic and anaerobic and between the Euryarchaeota and Crenarchaeota, we have performed the comparative analysis among these two archaea and with other archaea determined entire genome sequence. In this meeting we will discuss results of such comparative analysis.

*Complete Sequence and Gene Organization of the Genome of a Hyperthermophilic Archaebacterium, *Pyrococcus horikoshii* OT3 (1998) DNA Research, 5, 2, 55-76 & 147-155

## Production of High-density DNA Microarrays for Microbial Genomes at TIGR

**Karen A. Ketchum**, Alyson E. Hazen, Scott W. Avery, Linda C. Banerjei, Patee Gesuwan, John Quackenbush, and Claire M. Fraser. The Institute for Genomic Research, Rockville, MD.

High-density DNA microarrays are one of the most promising technologies for parallel analysis of gene expression. This process can be used to document global changes in transcription as microorganisms modify their metabolism and physiology in changing environments. The goal of this project is to develop standard operating procedures to be used at TIGR for high throughput production of microbial microarrays. We have focused on 5 areas for our initial optimization studies.

1- PCR amplification of target DNAs from microbial templates
2- Chemistries for DNA attachment to glass surfaces
3- Fluorescent probe labeling
4- Hybridization conditions
5- Cross talk among paralogous gene sequences

Many factors affect the success of array experiments, including, the humidity of the printing environment, composition of the buffer used to resuspend DNA targets, type of glass selected and the coating that is placed on the surface, ratio of non-labeled to Cy-conjugated nucleotides, and hybridization buffers. This poster will detail our results from each of these areas and highlight critical factors. In addition, we will outline the parameters that we have established for quality control of array construction.

*This research is supported by grants from the United States Department of Energy, Office of Biological and Environmental Research and The Energy Biosciences Program.*

## Strategies for Optimizing the Sequencing of Difficult Genomic Regions

John E. Gill, **Hoda M. Khouri**, Haiying Qin, Jessica J. Vamathevan and Teresa R. Utterback. The Institute for Genomic Research, Rockville, MD.

As TIGR moves towards sequencing organisms of medical and environmental benefit, efforts must be made to streamline sequencing through the difficult region of these complex genomes.

Difficult genomic regions include several types of repeats of varying length and base composition, long GC-rich stretches, areas of DNA secondary structure, sequences with hard stops and unclonable promoters. Standard techniques and chemistries used for sequencing DNA fail to sequence through these areas, thus impeding the process of closure.

Several different techniques were developed for optimizing both PCR and sequencing reactions, which have proven in most cases to allow sequence through these problem areas. These methods include, but are not limited to, the following: the addition of betaine, Q-solution buffer and other enhancers, increasing primer lengths, utilizing a variety of polymerases, increasing the denaturation temperature of PCR reactions, changing the conditions for cycle-sequencing, sub-cloning PCR products for sequencing with Dye-Primer and BigDye chemistries, and the construction of micro-libraries.

Examples of sequence through these difficult regions and the different techniques that were used to resolve them will be presented and discussed.

## Annotation of Microbial Genomes

**Frank Larimer**, Richard Mural, Morey Parang, Manesh Shah, Victor Olman, Inna Vokler, Jay Snoddy, and Edward Uberbacher. Computational Biosciences, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN.

We are implementing integrated display tools for microbial annotation systems within the context of the Genome Channel and the Genome Annotation Consortium. In cooperation with The Institute for Genomic Research, we currently have views of the various complete microbial genomes sequenced by TIGR available in the Genome Channel. Other complete genomes will be added shortly and views of genomes in progress will be developed. Among the features being implemented are:

A visual, integrated contextual browser for viewing genomic relationships from the chromosome to the nucleotide level, within and between genomes;

Improved and consistent gene-calling, with emphasis on accurate prediction of translation start, as well as accurate calling of short (<300nt) genes;

Annotation of structural features, including operon and regulon organization, promoter and ribosome binding site recognition, repressor and activator binding site calling, transcription terminators, and other functional elements;

Linkage and integration of the gene/protein/function catalog to phylogenetic, structural, and metabolic relationships.

The rapidly growing microbial genome database poses significant challenges in both analysis and presentation, particularly in multi-gene, multiple-genome comparisons. The exploration of microbial diversity and an understanding of the nature and origin of evolutionary change requires depth of analysis as well as breadth of sampling. Comprehensive annotation tools will be critical to access the richness of genomic complexity.

*Research sponsored by the Office of Health and Environmental Research, USDOE under contract number DE-AC05-96OR22464 with Lockheed Martin Energy Research Corp.*

## Protozoan Sequencing at The Sanger Centre

**Daniel Lawson**, Sharen Bowman, Mike Quail, Carol Churcher, David Harris, Marie-Adele Rajandream and Bart Barrell. Pathogen Sequencing Unit, The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, UK.

The Sanger Centre is currently involved in a number of protozoan genome sequencing projects, namely *Plasmodium falciparum*, *Leishmania major*, *Trypanosoma brucei* and *Dictyostelium discoideum*. All these species are eukaryotic with small genomes (approximately 30 Mb) organised into multiple chromosomes. Projects cover extremes of DNA composition, ranging from an [A+T] of 0.35 for *L.major* to 0.82 for *P. falciparum*.

The apicomplexan sporozoa *P.falciparum* is the predominant causative agent of human malaria. The Sanger Centre is sequencing approximately half of this malaria genome: chromosomes 1, 3-9, and 13. The chosen sequencing strategy involves a chromosomal shotgun with a light 'skim' of mapped Yeast Artificial Chromosomes (YACs).

The kinetoplastid flagellate *L. major* is the causative agent of leishmaniasis. The Sanger Centre is involved in two projects whose aim is to sequence 4.3 Mb of *L. major* DNA from 7 chromosomes. The chosen sequencing strategy involves sequencing a minimal tiling path of overlapping cosmid clones.

The kinetoplastid flagellate *T. brucei* is the causative agent of sleeping sickness. The Sanger Centre is sequencing chromosome 1 (1.15 Mb) as a pilot project. The chosen sequencing strategy involves a chromosomal shotgun

The social amoeba *D. discoideum* is the model organism of choice for developmental biologists. The Sanger Centre is involved in the sequencing of 1.5 Mb from chromosome 6 as part of an international consortium. The DNA composition of Dictyostelium, [A+T] = 0.75, is reminiscent of malarial DNA and a similar sequencing strategy has been employed involving the physical YAC map.

Here we summarise the goals for each of the projects and give a progress update. More information can be found at The Sanger Centre web site (URLs given below).

URLs for the Protozoan projects:
P.falciparum HYPERLINK
http://www.sanger.ac.uk/Projects/P_falciparum/
http://www.sanger.ac.uk/Projects/P_falciparum/
L.major HYPERLINK
http://www.sanger.ac.uk/Projects/L_major/
http://www.sanger.ac.uk/Projects/L_major/
T.brucei HYPERLINK
http://www.sanger.ac.uk/Projects/T_brucei/
http://www.sanger.ac.uk/Projects/T_brucei/
D.discoideum HYPERLINK
http://www.sanger.ac.uk/Projects/D_discoiduem/
http://www.sanger.ac.uk/Projects/D_discoiduem/

## Molecular Diversity of the Genes Involved in Anaerobic Energy Metabolism in *Shewanella putrefaciens* MR-1

**Guangshan Li**[1], Jizhong Zhou[1], Douglas Lies[2], Rebecca Clayton[3], Kenneth H. Nealson[2], Claire Fraser[3] and James M. Tiedje[4]. [1]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN; [2]Department of Geology and Planetary Sciences, Jet Propulsion Laboratory and California Institute of Technology, Pasadena, CA; [3]The Institute for Genomic Research, Rockville, MD; [4]Center for Microbial Ecology, Michigan State University, East Lansing, MI.

*Shewanella putrefaciens* MR-1 is a facultative bacterium in the gamma Proteobacteria and able to reduce a variety of compounds including dimethyl sulfoxide, fumarate, nitrate, nitrite, thiosulfate, Fe(III), and Mn(IV). Ninety-nine percent of the MR-1 genome sequence is available. To understand molecular diversity of anaerobic metabolic pathways, the amino acid sequences of genes involved in anaerobic electron-accepting processes from other species were searched against the available genome sequence of MR-1. Three ORFs in MR-1 had 68-100 % identities to the periplasmic nitrate reductase and two ORFs had 57-66 % identities to the formate-dependent nitrite reductases. However no genes that are involved in known nitric and nitrous oxide respiration were detected. Five genes, which

may be involved in Fe(III) and Mn(IV) reduction in MR-1, showed 20-30% homology to the current database sequences from other species. Also there are three ORFs which were 40-70% identical to the genes involved in fumarate reduction. Genes similar to those involved in thiosulfate and dimethyl sulfoxide reduction were detected. In addition, One ORF had 54-75% identities to that of FNR family, which is a pleiotropic activator of genes involved in anaerobic respiration. These results suggest that the genes involved in anaerobic energy metabolism in MR-1 are quite divergent from those in other species.

## The Identification of a New Family of Sugar Efflux Pumps in *Escherichia coli*

**Jia Yeu Liu**, Paul F. Miller[1], Mark Gosink[2], and Eric Olson. Infectious Diseases and [2]Molecular Biology Department, Parke-Davis Pharmaceutical Research, Division of Warner-Lambert Company, Ann Arbor, MI; and [1]Present address: Infectious Diseases Department, Pfizer Central Research, Groton, CT

Using a functional cloning strategy with an *Escherichia coli* genomic plasmid library, we have identified a new family of sugar efflux proteins with three highly homologous members in the *Escherichia coli* genome. In addition, two open reading frames, one present in *Yersinia pestis* and the other in *Deinococcus radiodurans*, appear to encode closely related proteins. An *in vitro* transport assay using inside-out membrane vesicles prepared from overproducing strains demonstrated that members of this new family can efflux [$^{14}$C]-lactose and [$^{14}$C]-glucose. Substrate specificity was further explored using both the inhibition of the *in vitro* [$^{14}$C]-lactose transport assay and a cell based assay. The results indicate that one member of this family (SetA) prefers glycosides with alkyl or aryl substituents as both isopropyl-b-D-thiogalactoside (IPTG) and *o*-nitrophenyl-b-D-thiogalactoside are substrates for efflux. As sugar efflux phenomena have been reported previously in several bacterial species including *E. coli*, the identification of a new family of sugar efflux proteins may help reveal the physiological role of sugar efflux in metabolism. It is proposed that the *E. coli* members of this family, whose functions were previously unknown, be given the gene family designation *set* for *s*ugar *e*fflux *t*ransporter.

## Pilot Scale Genome Sequencing of *Aspergillus nidulans* by an Internal-labeling Long Read Protocol

**Machida, Masayuki**[1], Nakagawa, Mari[1], Kunihiro, Sumiko[1] and Takei, Shigeo[2]. [1]Molecular Biology Department, National Institute of Bioscience and Human-Technology, AIST, Tsukuba, JAPAN; and [2]Analytical Instrument Section, Radiation and Biology Engineering Department, Aloka Co., Ltd., Tokyo, JAPAN.

In the course of world wide effort to complete the genome sequence of *Aspergilllus nidulans*, we started the pilot scale sequencing of a part of the ordered cosmids library (1). We picked the cosmids locating in the middle part of chromosome VIII and neighboring to the cosmid which has been sequenced by Prade *et al.* (2). We have tentatively sequenced at both ends of the cosmids and analyzed the overlap among the cosmid clones by LD-PCR. The sequencing was done mainly by the primer walking method using the internal-labeling protocol and analyzed by LI-COR model 4200L DNA sequencer. Since longer than 800 nucleotide sequence could be analyzed in a single run, approximately 160 reactions were expected to complete the sequence of both strands of cosmid's inserts. However, the internal-labeling cycle sequencing protocol applied to LI-COR DNA sequencer was limited to as long as 10 kbp DNA template. We examined the condition for the internal-labeling protocol applicable for cosmid sequencing and found that the successful long-read sequencing depended on higher concentration of IRD-labeled dATP, optimization of cosmid amount and the higher temperature for denaturation step. The *A. nidulans* cosmid sequencing is under progress.

1. Prade, R. A., Griffith, J., Kochut, K., Arnold, J. and Timberlake, W. E., Proc. Natl. Acad. Sci. 94, 14564-14569 (1997)2. Kupfer, D. M., Reece, C. A., Clifton, S. W. and Roe, B. A., Fung. Genet. Biol. 21, 364-372 (1997)

## An In-House Integrated Database System

**Katerina Michalickova** and Christopher W.V. Hogue. Samuel Lunenfeld Research Institute, Mt. Sinai Hospital, Toronto, CANADA and Department of Biochemistry, University of Toronto, CANADA.

We are interested in a number of Bioinformatics approaches to studying complete genomes and their relationships with 3-D structures and compact protein domains. While previously we had advocated the use of NCBI's Entrez interface for remote querying of the Entrez databases, we have found that these methods of access are unreliable for automated query methods for a variety of reasons. We require a reliable in-house database containing the same information as found in Entrez. While Entrez is not presented by NCBI in an exportable format, the NCBI ftp site has a number of datasets that can be used to reconstruct a functional in-house database resource that behaves like Entrez, but currently without the neighbor information. We use the embedded information in ASN.1 sequence and structure files for generating link information and store it together with the original ASN.1 binary data in CodeBase databases (Sequiter Software Inc., Alberta). The CodeBase database system is easily linked with the NCBI toolkit on a number of platforms and provides a high-speed and streamlined database subsystem capable of holding all of GenBank, MMDB and the associated link databases. The database is updated daily from NCBI ftp site. At the present stage, a web interface facilitates database searches for sequences based on unique identifiers (GI), GenBank accession numbers, NCBI taxonomy identifiers, and Medline identifiers. Nucleic acid and protein sequences can be displayed in several formats. The query also triggers a search for linked nucleic acid or proteins. Taxonomy and Medline searches provide linking identifiers to the Entrez and Pubmed data at NCBI. In addition we support linking identifiers to other sequence and structure databases. We are developing our own client-server interface which retrieves data from our database through http protocol and a CGI-based server. The new interface will perform some operations which are not addressed in the Entrez API such as obtaining nonredundant set of sequences from organisms or complete genomes.

## Genes Involved in the Inositol Metabolism of Mycobacteria

**Farahnaz Movahedzadeh**, Tanya Parish and Neil G. Stoker. London School of Hygiene and Tropical Medicine, London, UK.

Tuberculosis remains the infectious disease which causes the greatest mortality world-wide. The genome sequence for *Mycobacterium tuberculosis* has recently been completed and this provides an excellent source for analysis of gene function.

We have isolated a transposon mutant of the fast-growing non-pathogen *Mycobacterium smegmatis*, which has an altered cell envelope. Sequence analysis showed the insert lies in an inositol monophosphatase gene (*impA*).

Mycobacteria are unusual in that, unlike most bacteria, inositol is an important molecule. It plays a key role in the structure of lipoarabinomannan in the cell wall, and in the major thiol, mycothiol. We have now identified the homologue of *impA* in *M. tuberculosis*, as well as several other genes which may be involved in inositol metabolism. The aim of this project is to investigate the functions of these genes in order to understand the importance of these inositol-containing molecules in the growth and survival of *M. tuberculosis.*

## Genome Sequencing of *Campylobacter jejuni* strain NCTC11168

**Karen Mungall**, Julian Parkhill and the Pathogen Sequencing Group. The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

The Sanger Centre has recently finished the genome sequence of *Campylobacter jejuni* strain NCTC11168 in collaboration with Brendan Wren of the Department of Medical Microbiology, St Bartholomew's and the Royal London School of Medicine and Dentistry and Julian Ketley of the Department of Genetics, University of Leicester. *Campylobacter jejuni* is the leading cause of food-borne gastroenteritis in developed countries such as the UK and the USA. In addition, the disease can have debilitating long term sequelae including Guillain-Barre syndrome, reactive arthritis and other autoimmune conditions.

The sequence is 1,641,480 bp in length and was generated from 33,824 sequencing reads. At present we are in the process of systematic annotation of the genomic sequence. We have attempted to make as much information as possible publicly available while the annotation is in progress. To this end we have released an initial gene prediction, along with further information for each CDS such as matches to the Pfam database of protein families, and regularly updated comparisons with the protein databases. All of this information is available from the *C. jejuni* website at: http://www.sanger.ac.uk/Projects/C_jejuni/.

One of the most striking results of the genome sequence so far is the presence of hypervariable homopolymeric tracts within the clonal DNA, suggesting an extremely rapid rate of slip-strand mispairing during replication.

## The *Streptomyces coelicolor* Genome Sequencing Project at The Sanger Centre

**Lee Murphy**, Stephen Bentley and The Pathogen Sequencing Unit, The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

The Streptomyces are Gram-positive bacteria similar to mycobacteria in having high G+C DNA. They are unusual procaryotes in that they undergo a complex cycle of morphological development and also in their production of a large number of antibiotics and other secondary metabolites. *Streptomyces coelicolor* A3(2) has an 8 Mb linear genome and a project to sequence the entire chromosome is underway at The Sanger Centre. The clone-based strategy uses a set of 320 overlapping cosmids produced and mapped at The John Innes Centre, Norwich, UK. Strategies have been developed to solve the difficulties resulting from the very high G+C content of the DNA (>70%) and from the presence in the genome of very many small inverted repeats. The project has produced >2 Mb of finished sequence and a further 0.5 Mb unfinished sequence is available. The sequences have been obtained from several regions of the genome including two of the 4 known antibiotic gene clusters. The high gene density typical of bacterial genomes seen in all sequenced cosmids indicates that the DNA is likely to contain 8000 genes (more than are found in *Saccharomyces cerevisiae*). The sequencing project has suggested possible reasons for the apparent high number of genes. This work is being funded jointly by the UK Biotechnology and Biological Sciences Research Council and The Wellcome Trust.

## Positioning of High Repetitive DNA Sequences in Reference to the Direction of Transcription and Replication

**Yuko Ohfuku**[1,2,3], Hideaki Koik[1], Naoki Amano[1,4], Jun Kakinuma[1,5], Masaru Tateno[1], Joerg M. Suckow[1], and Masashi Suzuki[1,5]. [1]AIST-NIBHT CREST Centre of Structural Biology, Tsukuba, JAPAN; [2]Doctoral Program in Agricultural Sciences, University of Tsukuba, Tsukuba, JAPAN; [3]National Institute of Technology and Evaluation, Tokyo, JAPAN; [4]Doctoral Program in Medical Sciences, University of Tsukuba, Tsukuba, JAPAN; and [5]Graduate School of Human and Environmental Sciences, University of Tokyo, Tokyo, JAPAN.

It is reported here that when the circular genomic DNA molecule is displayed by a 12-hr-clock, transcription of most of the identified ORFs in the genome of Pyrococcus sp. OT3 proceeds in the two directions -i.e. clockwise and anti-clockwise, starting from the position of 12 o'clock. The one direction goes clockwise and ends at 7 o'clock, while the anti-clockwise direction is found in the range of 7 to 12 o'clock. Other researchers have proposed that in some eubacterial species the direction of transcription is likely to be coupled with the direction of replication. An archaeon, P. OT3, might be another of such examples. A 29mer sequence repeating more than hundred times has been identified in the genome of P. OT3. Two neighbouring units in the repeat are separated by 68 base pairs on average according to each start position. The repeating DNA units are clustered into six regions. Five clusters are organized in the direction of replication, while one not. Some notable correlation appears to be found between the positioning of the repeating sequences and the direction of transcription. The genomic DNA sequences of three other archaeal species also contain similar repetitive DNA sequences. In the genome of *M. thermoautotrophicum*

only two clusters are found. The sequence consensus to the two clusters is very close to the sequence consensus of all sequenced archaebacterial species. In contrast, in the genome of *M. jannaschii*, repeating units are divided to a number of clusters. The sequence consensus to the clusters is diverged from each other and from the archaeal consensus sequence. Thus, it appears that there is a correlation among the divergence of the consensus repeat sequence of each species from the archaeal consensus repeat sequence, the divergence of consensus sequences of the clusters from the species specific consensus sequence and the division of the repeating units into clusters. Namely, it suggests that they are traces of evolution.

## Genome Browsers for Bacterial Pathogens

Nicholas J. Loman and **Mark J. Pallen.** Department of Medical Microbiology, St Bartholomew's and the Royal London School of Medicine and Dentistry, London, UK.

Bacterial genome sequencing projects exemplify the Pareto principle: around 80% of the work is done in 20% of the time. A large amount of sequence data becomes available during shot-gun sequencing. However, the definitive annotated genome sequence may not be released for many months, as the finishing and annotation stages can take just as long as shot-gun sequencing, if not longer. While, in most cases, laboratory-based researchers can query the incomplete sequence through cut and paste web interfaces, this approach is too onerous if one is interested in more than a few sequences or wishes to track the appearance of sequences in a pathway (e.g. aromatic amino acid biosynthesis) or protein complex (e.g. type-III secretion systems) while genome sequencing progresses.

To help bacteriologists mine genomic data while sequencing is in progress, we have built several web-based "genome browsers" (http://www.medmicro.mds.qmw.ac.uk/genomes). TBLASTN searches with proteins sequences from other completed genomes are performed against incomplete genome sequence data and are automatically repeated each time a new batch of sequence data is made public. Results are presented through an attractive web interface, that includes a hierarchical classification of protein function. In addition to using protein sets from completed genomes we allow users , through a "private browser", to set up their own lists of proteins sequences to be tracked as the sequencing project progresses. Sequences of interest can be entered individually in FASTA format or en masse as Entrez UIDs.

As TBLASTN searches identify paralogues as well as orthologues, we provide a simple web-based confirmatatory system, where the regions of predicted protein sequence identified by TBLASTN can be searched by BLASTP against the non-redundant protein database. In this way, one can, for example, quickly distinguish between hits to type-III secreted proteins that represent components of the flagellar biosynthesis system from genuine type-III secreted proteins.

A user-friendly facility for retrieving information from unassembled sequences or from contigs sequence means that users can explore the local context of any protein-coding sequence by performing BLASTX search with the relevant nucleotide sequence and its flanking regions.

## Microbial Genome Sequencing at the Sanger Centre

**Julian Parkhill,** Bart Barrell and the Pathogen Sequencing Group. The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

The Sanger Centre has an active and expanding microbial genome sequencing group. In addition to the completion and publication of the Mycobacterium tuberculosis genome last year, we have also recently completed the sequence of *Campylobacter jejuni*, and are in the process of the systematic annotation of this genome. The shotgun sequences of three further organisms; *Neisseria meningitidis*, *Yersinia pestis* and *Bordetella pertussis*, are complete, and these are now in various stages of the finishing process. In addition, we are in the process of shotgun sequencing *Salmonella typhi* and *Bordetella bronchiseptica*, and are testing libraries of *Clostridium difficile* and *Bordetella parapertussis* in preparation for shotgun sequencing.

In addition to these whole genome shotguns, we are also sequencing two organisms, *Streptomyces coelicolor* and *Mycobacterium leprae*, using a cosmid-based clone-by-clone approach.

Further information, and access to search engines and sequences, can be obtained from the microbial genome sequencing web page at: http://www.sanger.ac.uk/Projects/Microbes/.

## DNA Microarray Gene Expression Monitoring of the Extreme Radiation Resistant Bacteria *Deinococcus radiodurans*

**Scott Peterson,** Robin Cline, John Heidelberg and Owen White. The Institute for Genomic Research, Rockville, MD.

*Deinococcus radiodurans* is best known for its extreme resistance to ionizing radiation. It has been documented that a *D. radiodurans* cell has the ability to repair 150 double-stranded DNA breaks without mutagenesis or DNA rearrangement. This enormous potential to survive X-irradiation is known to be dependent on *rec*A, *pol*I and *uvr*A. Beyond this, the underlying mechanism of X-irradiation resistance is not well understood. *D. radiodurans* also has elevated resistance to U.V. irradiation, dessication and oxidative stress. Mutational analysis suggests that there is significant overlap in the proteins and activities required for resistance to these insults and stresses. However, some mutations have been uncovered that appear to dissect the pathways. An example is that of *kat*A, which has a role in oxidative stress resistance but not ionizing radiation resistance. We have initiated an analysis of *Deinococcus radiodurans* RNA expression profiles using DNA microarray technology, by comparing expression patterns of cells after high doses of X-irradiation (1.75 Mrads) and conditions of oxidative stress induced by hydrogen peroxide.

## Sequencing of *Pseudomonas aeruginosa* PA01

**Xuan-Quynh T. Pham**[1], Karen Avery-Phelps[1], Karen Blochlinger[1], Kerry Bubb[1], Jina Chan[1], Cindy Demarais[1], David Gordon[1], Robert Hubley[1], William Hughes[1], Shawn Iadonato[1], Don Jewett[1], Arnie Kas[1], Regina Lim[1], Lloyd Lytle[1], Charles Magness[1], Mortimer Petralia[1], Bethany Richards[1], Channakhone Saenphimmachak[1], Elizabeth Simms[1], Kim Smith[1], David Spencer[1], Chris Weinmann[1], Gane Ka-Shu Wong[1], Lettie Goltry[2], Wendy Hufnagle[2], Micki Lagrou[2], Ernie Tolentino[2], Scott Mizoguchi[2], Gita Bangera[3], Stephen Lory[3], Bob Hancock4, Phil Green[1], Maynard Olson[1], Ken Stover[2], Rick Garber[2]. [1]The Human Genome Center, Department of Medicine, University of Washington, Seattle, WA; [2]Pathogenesis Corporation, Seattle, WA; [3]Department of Microbiology, University of Washington, Seattle, WA; 4Department of Microbiology, University of British Columbia, Vancouver, BC.

*Pseudomonas aeruginosa* is an opportunistic pathogen which primarily infects immunocompromised patients, such as people who suffer from Cystic Fibrosis. Two years ago, a collaborative sequencing project between the University of Washington Genome Center and the Pathogenesis Corporation was initiated to sequence a clinically virulent strain, PA01.

The majority of the sequencing and finishing was carried out at the UWGC. The PA01 strain is 6Mbp, with no known plasmids. Our initial goal was to cover the genome at 7.5X redundancy with 90,000 reads. Realizing the greater probability of compressions in this GC rich genome (>60%), we decided to implement precautionary measures. We used ABI 377 sequencers because of their higher running temperatures (52C). Half of the M13 shotgun reads were dye primer and the other half were dye terminator. Shotgun sequencing began in June of 1997 and ended in August of 1998.

Finishing began after 93,973 shotgun reads were generated and assembled into 97 contigs. We have divided the finishing effort into four components: contig joining, strengthening weak regions, cosmid end sequencing, and MCD mapping. We are using an automatic finishing program called 'Autofinisher', which is a feature of Consed. It designs finishing reactions that will resolve the most errors at the lowest cost. The first round of automated finishing produced 1042 reads and reduced the assembly to 12 major contigs. These contigs have a total size of 6.2Mbp, and 6.1Mbp of this is in 40Kbp (or larger) regions with an integrated error rate of less than 0.0001. With ~3X cosmid coverage of the genome we have made 8 contig joins in the MCD map resulting in 4 large contigs and 1 possible 11kbp misassembly. The largest gap is ~1Kbp. We will confirm these MCD map joins with cosmid end sequencing and then close these gaps using the cosmids as templates for directed sequencing. There will be a continued effort to improve the quality of the sequencing data and to join all remaining contigs into one contig. The annotation of this sequence is being shared between the UWGC, Pathogenesis, Lory Lab at UW, and Hancock Lab at UBC.

## Whole Genome Shotgun Optical Mapping of *Deinococcus radiodurans* and *Trypanosoma brucei*

**Rong Qi**, Jieyi Lin, Zhongwu Lai, Virginia Clarke, Stephanie Delobette, Christopher Aston, Junping Jing, Thomas Anantharaman[1], Bud Mishra[1], Najib M. El-Sayed[2], Owen White[2], and David C. Schwartz. W.M.Keck Lab for Biomolecular Imaging, Department of Chemistry, [1]Department of Computer Science, New York University, NY, NY; and [2]The Institute for Genomic Research, Rockville, MD.

Optical Mapping is a proven approach for the rapid generation of ordered restriction maps across whole microbial genomes. Maps are constructed by imaging restriction endonucleases cutting of single DNA molecules with fluorescence microscopy. We have used shotgun Optical Mapping to generate physical maps of the *Deinococcus radiodurans* genome. The final *Nhe* I map ( average fragment size of 29 kb ) was assembled without gaps at an average depth of 35x. We calculated the genome to be 2.7 mb by optical means and also discovered the presence of a second chromosome, 415 kb in size. Our maps have greatly facilitated sequence assembly efforts at TIGR. Construction of high resolution restriction maps for the *Trypanosoma brucei* genome has been recently initiated. The haploid genome of *T. brucei* is approximately 37 mb in size. Three size classes of *T.brucei* chromosomes have been defined: the minichromosomes of 50~150 kb, the intermediate chromosomes of 200~900 kb, and we are currently mapping the 11 pairs of diploid chromosomes of 1.0~6.0 Mb. We evaluated a number of restriction enzyme candidates by determining cut site frequency from BAC and P1 clone end sequence information in the TIGR database. Restriction enzyme *Xba* I gives fragments of an optimum size (14 kb) for Optical Mapping and has been chosen to map *T. brucei* genomic DNA molecules. *T. brucei* DNA was prepared from agarose inserts which were provided by TIGR. Our maps will serve as a scaffold for assembly and verification of the large-scale sequencing. We aim to link the *T.brucei* optical map with existing sequence information and physical map resources such as pulse field electrophoresis-generated karyotypes and EST markers mapped to the P1 genomic DNA library. Optical Mapping will also enable detailed whole genome analysis for related African trypanosome species.

## Identification of a Novel Site-specific DNA-binding Activity in *Deinococcus radiodurans*

**Haiying Qin**, Kelly S. Moffat, William C. Nelson, John F. Heidelberg and Owen White. The Institute for Genomic Research, Rockville, MD.

*Deinococcus radiodurans* is a Eubacterium that is resistant to large doses of gamma-irradiation. Bioinformatic analysis of whole-genome sequence data from this organism has revealed a variety of repeat elements found throughout the three chromosomes that make up the genome. Fifty-eight copies of one such repeat, (poridge, 151 base pairs), are evenly spaced throughout the genome at intervals of, on average, 70 kilobasepairs. Small circles of chromosomal DNA have been identified that are hypothesized to be recombinational intermediates across these repeats. Using an electrophoretic mobility shift assay, we have identified a DNA-binding activity from a soluble cell extract that is specific for the poridge repeat. Sequence analysis of poridge revealed an A/T-rich inverted repeat at one end. A truncated poridge probe lacking this inverted repeat failed to be bound by the cellular

extract in a mobility shift assay, thereby tentatively identifying the recognition site. Cellular extracts were made from *D. radiodurans* cells harvested at various time points during recovery following gamma-irradiation. A decrease in binding activity is observed at two hours post-irradiation, and recovery of activity to pre-irradiation levels is achieved by six hours. Mobility shift assays using cell extract from a recA mutant strain of *D. radiodurans* still demonstrated binding activity. Isolation and identification of the protein responsible for this activity is in progress.\

## Direct Fluorescent Labeling and Automated Analysis of Genomic Clones for Mapping

M. Esser, A. Oommen, H. Osterman and J. Qiu. LI-COR, Inc., Biotechnology Division, Lincoln, NE.

Direct labeling and detection of nucleic acids has many advantages in DNA fragment analysis for genomic mapping. The "extendase" (dA addition) activity of some DNA polymerases, which catalyzes the addition of single nucleotides to the 3' terminus of blunt-ended PCR products, has been previously employed in various cloning strategies. We demonstrate the same activity can be used to label DNA fragments generated by either restriction enzyme digestion or PCR. Furthermore, a single-step reaction protocol has been successfully developed in which restriction digestion and labeling are performed simultaneously and differentially in a single tube. Specifically, DNA is digested and labeled with IRD-700 or IRD-800 dNTPs in a single, 30-minute reaction. The IRD-labeled restriction fragments are then detected and analyzed within a few hours on a horizontal agarose gel using a LI-COR IR2 automated DNA sequencer. Several different DNA polymerases and templates, including Human BAC clones, are compared for labeling efficiency. This protocol can be a valuable tool for efficient mapping of genomic clones for high throughput laboratories.

## On the Specificity of Insertion of MiniTn7 Elements Mediated by TnsABC$^{A227V}$

**Elisabeth A. Raleigh**[1], Ann S. Stellwagen[2], Matthew C. Biery[2], Barton Slatko[1], and Nancy Craig[2]. [1]New England Biolabs, Inc, Beverly, MA; and [2]Howard-Hughes Medical Institute/Johns Hopkins School of Medicine, Baltimore, MD.

An *in vitro* transposition system with desirable properties has been developed for use in sequencing applications. Desirable characteristics include sequence-nonspecific insertion, predominance of single simple insertions, high insertion efficiency and possession of useful sites for rare-cutting restriction enzymes. The properties of this Tn7-based system depend on a mutation in the target-acquisition component of the four-subunit transposase. This mutation (TnsC$^{A225V}$) makes the target-recognition protein (alternate requirements are TnsD and TnsE) dispensable, without materially changing the reaction efficiency or the transposition-immunity properties of the enzyme. The resulting three-component enzyme catalyzes insertion of mini-Tn7 elements into a highly random distribution of sites. The global specificity of insertion of transposition mediated by TnsABC* was examined in two ways: by analyzing the distribution of 63 insertions along the length of a target sequence (pER183, a pACYC184 derivative carrying the *mcrBC* genes of *E. coli* K-12; 9.9 kb), and by analysis of the sequences within the 5 bp of target sequence that is duplicated during the insertion process. The local specificity within a smaller target sequence sample was examined by a physical assay: PCR amplification of pooled insertions using one primer binding to a fixed site in the target molecule and another primer specifically binding to one end of the miniTn7 transposon. Multiple insertions are rarely obtained; the most comment aberrant product can be avoided by ensuring a monomeric state of the donor plasmid.

## Development of a Novel Lane Tracking Reagent for Automated DNA Sequencers

**Thomas R. Reynolds** and Gregory A. Meyers. Commonwealth Biotechnologies, Inc. Richmond, VA.

Recently, we have developed a reagent (ACCUTRAC™) which provides a marking tool for sample tracking in Automated DNA sequencing applications. Specifically, inclusion of AccuTrac in the sample loading dye "marks" the lane of the acrlyamide gel into which each sample is individually loaded and electrophoresis is applied. Marking each lane helps to manually identify each lane during the tracking process and ensures that the instrument correctly reads the corresponding data from each lane and accurately applies the data to the appropriate sample information. The reagent incorporates a fluorescent dye which is recognized by most, if not all, currently available DNA sequencers. AccuTrac is incorporated into a DNA sequencing loading buffer through a formulation which is **PROPRIETARY** to CBI. The reagent is detected by an automated sequencer without interfering with the data generated by the instrument.

In the work described in this poster, we show the application of the new reagent technology to lane/sample tracking during automated DNA sequence. Side by side comparison of gels run with and with out AccuTrac is shown. Both gels contain the same samples loaded into the same lane number. Direct comparison of these gels clearly shows that in the gel containing AccuTrac, lanes can be easily tracked using the dye marker which appears at the bottom of the gel as (a) blue band(s). The gel without AccuTrac present in the loading is more ambiguous in terms of lane identification. The AccuTrac bands are present before any "real" data is resolved in the lane.

The dye front of the AccuTrac gel facilitates efficient lane identification. This effect plays an important role when 96 lanes are run and technical people are required to scan through gels at a rapid rate to check tracking or to manually track gels. Using AccuTrac in the gel loading buffer has aided our facility by less time spent manually tracking gels. The use of AccuTrac at CBI has virtually eliminated mistakes made in tracking lanes.

## cDNA AFLPÆ: A Robust High-throughput Transcript Profiling Technology for Microorganisms

**Guus Simons**, Martin Reijans, Antoinette Oude Groeneger, Rutger van Rooyen* and Pieter Vos. KeyGene N.V., Wageningen, The Netherlands; and *Gist-Brocades, Delft, The Netherlands.

We present an adaptation of the AFLPÆ technology to monitor genome-wide expression in microorganisms. cDNA-AFLP fingerprinting has overcome the limitations associated with other techniques and is able to fully evaluate differential gene expression. Advantages of cDNA-AFLP are: 1) template preparation requires very low amounts of starting material to

yield high quality fingerprints, 2) the use of restriction enzymes, adapter ligation and amplification with selective nucleotides enables quantitative profiling of transcripts, 3) low abundance transcripts can be detected easily by applying additional selective nucleotides, 4) full analysis of the transcriptome can be achieved by using a limited set of enzyme combinations, 5) the technique is relatively insensitive to amplification conditions resulting in a high level of reproducibility, 6) cDNA-AFLP fragments can be easily purified and sequenced, and are derived from various parts of the transcript and not predominantly from the 3' untranslated region, which increases coding potential and function prediction, 7) nonannotated ORF (NORF) genes and up to now uncharacterized genes can be detected and 8) no prior sequence information is required.

Combining cDNA-AFLP with bioinformatics expertise provides a versatile tool for high-throughput differential expression analysis comparing gene expression patterns of related strains or of identical strains under different growth conditions.

Examples in *Saccharomyces cerevisiae*, *Penicillium chrysogenum* and *Aspergillus niger* will be presented. (Note: AFLPÆ is a registered trademark of KeyGene N.V.)

## Distributed Homology Searches using PVM and WU-Blast on a Heterogenous Computer Network

**Bradley E. Slaven**, Hanif Khalak, Daniel Kosack and Dellwood Richardson. The Institute for Genomic Research, Rockville, MD.

The advent of large sequence datasets presents serious computational requirements for nucleotide and protein searches on a genomic scale. The emergence of inexpensive processor technologies have lead to the development of supercomputer class capabilities on desktop platforms. We leveraged off the shelf computers and publicly available software products to conduct searches on large scale genomic class datasets. To this end, we networked a cluster of personal computers running the Linux Operating System (Red Hat Release 5.1) using Parallel Virtual Machine (PVM) software (Oak Ridge National Laboratory). We have implemented a system that searches each sequence in a FASTA file against a database utilizing WU-Blast (Warren Gish, unpublished). This approach incorporates existing native program executables for searching (WU-Blast) and output parsing (btab) within a PVM framework.

Several DNA and polypeptide datasets were queried using WU-Blast using this system against 16 completed genomes. Results are presented including cpu, memory, network, and disk usage statistics as compared with runs with single platform implementations.

## A DNA Microarray Facility for the Functional Analysis of Bacterial Pathogens

**Neil G. Stoker**[1], Philip D. Butcher[2], Joseph A. Mangan[2], Robert M. Davies, Julian Parkhill[3], Bart G. Barrell[3] and Brendan W. Wren[4]. [1]London School of Hygiene and Tropical Medicine, London, UK; [2]St George's Hospital Medical School, London, UK; [3]Sanger Centre, Hinxton, Cambridge, UK; and [4]St Bartholomew's and Royal London Hospital School of Medicine & Dentistry, London UK.

We are currently developing a microarray facility to exploit the sequencing of bacterial pathogens being carried out at the Sanger Centre and elsewhere. The microarrays will be produced using PCR products generated from defined plasmids used in the sequencing projects, or made with ORF-specific primers. These will be arrayed on a standard glass microscope slides at a density of 500 / $cm^2$ using a BioRobotics microgridding robot. The array will be hybridized with fluorescently-labelled cDNA made from bacteria grown in different environmental conditions. Alternatively, genomic DNA will be used as the probe for comparative genomic studies. Dual fluorescence will be used, allowing simultaneous detection of two RNA or DNA populations. The output of the arrays will be read using a confocal laser scanner. Different binding strategies, and methods to assess the concentration of DNA bound to the array have been investigated. Initial work is being carried out using *Campylobacter jejuni* and *Mycobacterium tuberculosis* whole genome arrays. A tiling programme has been designed to select 1700 coding sequences from the 8-10-fold redundant pUC18 library from the *C. jejuni* genome project. Using a single pair of vector primers, amplicons from the miniset of plasmids will form the basis of a complete DNA microarray from *C. jejuni*.

## Genome Analysis of Facultatively *Alkaliphilic bacillus Halodurans* C-125

**Hideto Takami**, Kaoru Nakasone and Koki Horikoshi. Japan Marine Science and Technology Center/Deep-sea Microorganisms Research Group, Yokosuka, JAPAN.

Generally, *alkaliphilic Bacillus* strains cannot grow or grow poorly under neutral pH conditions, but grow well at pH higher than 9.5. Since 1969, we have isolated a great number of *alkaliphilic Bacillus* strains from various environments and have purified many alkaline enzymes. Over the past two decades, our studies have focused on the enzymology, physiology, and molecular genetics of alkaliphilic microorganisms to elucidate their mechanisms of adaptation to alkaline environments. Industrial applications of these microbes have also been investigated and some commercial enzymes from *alkaliphilic Bacillus* strains have brought great advantages to industry. Thus, it is clear that *alkaliphilic Bacillus* strains are quite important and interesting not only academically but also industrially. Recently, whole genome analysis of *B. subtilis*, which is taxonomically related to *alkaliphilic B. halodurans* strain C-125, except for the alkaliphilic phenotype, has been completed. Knowledge of the complete nucleotide sequence of the *B. subtilis* genome will help us in analysis of *B. halodurans* C-125 genome. From this research background, we initiated genome analysis of *alkaliphilic B. halodurans* C-125 and have completed a physical map of the genome which has a size of 4.25 Mb. The oriC region of the chromosome was identified and a 18.5 kb DNA fragment containing the oriC region of the chromosome was obtained by means of PCR. In addition, a lambda phage library of the chromosomal DNA was constructed and three independent DNA inserts (15-20 kb) were sequenced and

analyzed to determine their genetic features. The ORFs identified in these fragments were found to be similar to genes in the *B. subtilis* chromosome but were organized in a different order. These results will be helpful to consider the necessity of gene organization in the chromosome. Systematic sequencing of the whole genome of *Bacillus halodurans* C-125 has been routinized since last April and 80% of genome sequencing has been done so far.

## Microbial Genomes at NCBI: Complete and Incomplete

**T. Tatusova**, I. Mizrachi and J. Ostell. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD.

The Genomes division of Entrez currently contains 774 entries that represent 659 completely sequenced genomes, including 425 viruses, 17 bacteria, and *Saccharomyces cerevisiae*, complete sequences of *Leishmania major* chromosome 1 and *Plasmodium falciparum* chromosome 2, and a number of plasmids and organelles. The entries are divided in six large taxonomy groups: Archaea, Bacteria, Eukaryotae, Viroids, Viruses and Plasmids. The list of organisms in each group can be viewed in alphabetic order or as a taxonomic tree generated using the information from NCBI Taxonomy database.

WWW Entrez Genomes(http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html) provides flexible views, precomputed relationships, and immediate access to analytical tools for all complete sequence records. Genomes records not exceeding 350 limit for a single record (as agreed upon by International Sequence Database Collaboration) can be downloaded in several formats from the Entrez Nucleotide division. Large records including 17 complete microbial genomes, *Saccharomyces cerevisiae* and *Caenorhabditis elegans* can be downloaded from the NCBI FTP site (ncbi.nlm.nih.gov/genbank/genomes) in a variety of formats.

In addition to the completely sequenced genomes, Entrez Genomes contains mapping data and contiguous sequence islands for some eukaryotic genomes whose sequencing is in progress. The organisms included in this group are *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Oryza sativa* and *Zea mays*.

Preliminary sequence data from microbial genome sequencing projects in progress are accessible through a specialized BLAST Web page. Databases of assembled contigs are submitted to NCBI for BLAST similarity searches prior to the sequences being formally deposited into GenBank. The search employs Gapped BLAST 2.0 program and supports TBLASTN, BLASTN, and TBLASTX search modes. The sequences can be retrieved from the submitting sequencing centers.

## Closure or Ribosomal RNA Repeats in Whole Genomes

**Jessica J. Vamathevan**, Haiying Qin, Teresa R. Utterback, John F. Heidelberg, and Rebecca Clayton. The Institute for Genomic Research, Rockville MD.

The resolution of repeated sequences is one of the most challenging steps during the closure of microbial genomes. This is particularly evident for repeated elements which are larger than the size of the small insert library. One such repeat is the ribosomal RNA operon, which often exists in multiple copies in a genome. In many bacteria, this operon consists of the 16S rRNA promoter, the 16S rRNA gene, a 16S-23S intergenic spacer region (variable between operons), the 23S rRNA gene, and the 5S rRNA gene. The large size and high degree of sequence homology between operons make this a difficult problem for closure. Further complications in this repeat include the high amount of secondary structure which can make PCR across difficult, and the fact that the rRNA promoter can be toxic to *E. coli* , and therefore the DNA sequence flanking the 16S rRNA gene are not represented in the sequence generated from random sequencing phase.

To contend with multiple rRNA operons, TIGR has developed the following strategy: 1) determine the number of operons by Southern hybridization, 2) identify the 16S and 5S flank sequence in the assemblies, 3) if the 16S promoter is toxic to *E. coli* , identify the 16S rRNA flanks by use of the Clonetech Universal Genome Walker kit or direct genome sequencing, 4) perform combinatorial PCR 16S and 5S rRNA unique flank primers, 5) obtain the complete sequence of the PCR product. This method has proven useful because no step is contingent on the random phase of sequencing being completed, consequently we are able to resolve one of the most complicated repeats prior to the formal closure phase of a genome project.

**Ussery, David**
Will present his abstract from his speaker presentation

## Replication-induced Constraints on Genes and Proteins Composition in Bacteria

**Viari A.**[1], Rocha E.P.C[1,2], Médigue C.[1,2] and DanchinA.[2]. [1]Atelier de BioInformatique - Université Paris, Paris, FRANCE; and [2]Institut Pasteur, Paris Cedex 15, FRANCE.

Bacterial chromosome replication usually starts at a single origin and two replication forks propagate in opposite directions up to termination signals. As the replication mechanism differs for the two strands of the duplex DNA, this process may, in principle, give rise to compositional asymmetries between the leading (continuously replicated) strand and the lagging strand. Indeed, numerous evidences of these asymmetries have recently been given, essentially reflecting an excess of G over C in the leading strand. We introduce the use of Linear Discriminant Analysis in order to elucidate the presence of biases at the gene level, i.e. by considering more variables than the four nucleotide frequencies. We first show that strong compositional asymmetries between the leading and lagging strands do exist at the level of nucleotides, codons and, very surprisingly, amino acids. For some species, this bias is so high that the sole knowledge of a protein sequence allows to predict with very high confidence (96%) whether the gene is transcribed from the leading or the lagging strand. Furthermore, the examination of 16 complete genomes shows that these biases are not species specific but appear to be ubiquitous across the species. For instance, there is a general over-representation of Valine in the proteins corresponding to genes transcribed from the leading strand. The existence of these constraints have important consequences in our understanding of fundamental biological processes in bacteria such as replication fidelity, codon usage in genes and amino acids usage in proteins. They should therefore be taken into account in the course of a bacterial genome analysis.

## Functional Analysis of Microbial Genomes Using an Exhaustive Two-hybrid Approach

**Simon T. Whiteside**, Jean-Christophe Rain, Luc Selig, Céline Reverdy, Stéphane Simon, Véronique Battaglia, Miklos Fejes, Pierre Legrain, Hybrigenics S.A., Paris, FRANCE.

Large scale DNA-sequencing leads to the prediction of the structure of the proteome. New tools are needed to link these linear data to well-defined biological functions. Protein-protein interactions are key determinants of the life cycle of the cell. We have developed a highly selective and standardized procedure that allows us to constitute networks of protein-protein interactions - Protein Interaction Maps, or PIMs. This patented method, when applied to the yeast genome, has already revealed new factors involved in a given pathway as well as suggested the existence of novel complexes and ascribed novel biological functions to proteins[1]. Hybrigenics, a functional proteomics company, has developed high-throughput screening technology and is currently performing an exhaustive analysis on the ulcer-provoking bacterium *Helicobacter pylori*. The company's first products are protein-protein interaction databases, to be followed by detailed analysis of interacting domains - Selected Interaction Domains, SIDs - and isolation of compounds capable of modulating specific protein-protein interactions.

[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.

## Random Shot-gun Cloning and Sample Sequencing of the *Francisella tularensis*

Richard W. Titball[1], **Kerstin Williams**[2], Kerri A. Mack[1], Petra C. F. Oyston[1], Richard Prior[1], Nicola Chatwell[1], Mark Pallen[2], Nick Loman[2], Karin Hjalmarsson[3], Thomas Svensson[3], Luther Lindler[4], Siv G. E. Andersson[5], Gunnar Sandstrom[3] and Brendan W. Wren[2]. [1]Defence Evaluation and Research Agency, Salisbury, UK; [2]Department of Medical Microbiology, St. Bartholomew's and the Royal London School of Medicine and Dentistry, London, UK; [3]Department of Microbiology, National Defence Research Establishment, Umea, Sweden; [4]Department of Bacterial Diseases, Walter Reed Army Institute for Research, Washington, DC; and [5]Department of Molecular Biology, University of Uppsala, Uppsala, SWEDEN.

The human pathogen *Francisella tularensis* is the causative agent of tularaemia, a disease which occurs almost exclusively in the northern hemisphere. Despite the medical significance of tularaemia, little is known about the organism or how it causes disease. An international consortium of five laboratories has begun to sequence the 2 Mb genome of the highly virulent Biovar strain Schu4 of *F. tularensis*. A random shot-gun library has been constructed of 1-1.5 kb fragments of *F. tularensis* Schu4 DNA in pUC18. To date, approximately 8000 clones (~4-fold coverage) have been sequenced and are currently being analysed. Both known and novel *Francisella tularensis* genes have been identified. Updates of preliminary sequence data, BLAST search results and comparative genomics of Francisella DNA sequence data searched against gene products from several other genomes can be viewed via the genome browser (www.medmicro.mds.qmw.ac.uk/ft). Determination and analysis of the complete genome sequence will provide insight into the pathogenicity and the taxonomic position of this unusual bacterium.

## Estimating the Extent of Gene Transfer by Simulating Proteome Evolution

**Todd C. Wood** and William R. Pearson. Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA.

We have simulated proteome evolution to estimate the fraction of the eukaryotic nucleus that originated by horizontal transfer from the eubacteria. Random "mutations" were introduced into each protein sequence of the *M. genitalium* proteome using different distributions of PAM distances. The distribution of protein sequence similarities for these simulated proteomes matches closely the distribution of similarities between modern eubacterial and archaebacterial proteomes. Using a distribution of distances derived from comparisons of real proteomes, we simulated the origin of the three domains of life, including various amounts of gene transfer from the eubacteria to the eukaryotes at 2.0 billion years ago (Ga). Our results show that if the divergence of the eukaryotes and archaebacteria occurred at 3.5 Ga, the amount of gene transfer necessary to account for the similarities of real eubacterial and eukaryotic proteomes is less than 15% of the modern eukaryotic proteome. Alternatively, if the eukaryotes and archaebacteria diverged 2.25 Ga, 60% of the modern eukaryotic proteome must have been acquired through gene transfer from the eubacteria to account for the distribution of similarities seen in modern proteomes. To distinguish these two possibilities (ancient and recent divergence), we constructed trees using both simulated and real protein sequences from all three domains. Using the sequences of 46 arbitrarily-chosen proteins from our simulations, we find that the frequency of correct trees for the ancient divergence/low transfer phylogeny is 78.3%. For the recent divergence/high transfer phylogeny, the frequency of correct trees is only 37.0%. Using sequences of 93 different real proteins from all three domains, we find that the frequency of correct trees is 80.6%, which is only consistent with an 3.5 Ga origin of the eukaryotes and a small fraction genes acquired from the eubacteria.

## A Combinatorial Approach to Finding Promoter Elements in Genomic Sequences

**Thomas D. Wu.** Department of Biochemistry, Stanford University School of Medicine, Stanford, CA.

I describe a method for identifying potential promoter elements in a set of unaligned genomic sequences. This method uses a strategy of combinatorial alignment and refinement to extract statistically significant patterns in the input sequences. The method depends on precise probability calculations involving finite state automata that account for Markov dependencies among neighboring nucleotides. The probability calculations are based on Markov statistics from the input sequences, in order to increase the sensitivity and applicability of the algorithm to a variety of different organisms.

In contrast with stochastic methods for finding patterns, such as Gibbs sampling, the combinatorial method presented here attempts to identify the complete set of significant patterns in the given input sequences, by systematically searching a space of initial seed patterns. The method identifies seed

# Exhibit Descriptions

**National Center for Biotechnology Information (NCBI)** provides integrated access to DNA and protein sequence data, associated mapping data, protein structures, and MEDLINE. Demonstrations of the GenBank database, the Entrez retrieval system, PubMed for MEDLINE searching, BLAST and VAST similarity searches for sequence and structures, and the BankIt and Sequin sequence submission software will be provided.

**Contact person:**
Barbara Rapp
Operations Research Analyst
National Center for Biotechnology Information (NCBI)
National Library of Medicine
National Institutes of Health
8600 Rockville Pike
Bldg. 38A, Rm. 8N803
Bethesda, MD 20894
301-496-2475
301-480-9241 FAX
info@ncbi.nlm.nih.gov

**New England Biolabs** is the leading supplier of restriction enzymes in the USA, offering a full range of reagents for Molecular Biology and cell signaling analysis. Products include nine 8-base cutters; enzymes for DNA sequencing; phosphospecific antibodies; and innovative kits e.g.GPS-1 (Genome Priming System) for generating sequencing templates in vitro, IMPACT T7 (Intein based protein expression).

**Contact person:**
Dr. Barton Slatko
Director
DNA Sequencing Group
New England Biolabs, Inc.
32 Tozer Road
Beverly, MA 01915
978-927-5054
978-921-1350 FAX
dnaseq@neb.com

# Conference on Microbial Genomes

Westfields Marriott
Chantilly, VA
January 29-February 1, 1999

## Registrants
### (through January 26, 1999)

**Richard Alm**
Research Scientist and Genomics Project Leader
Molecular Biology and Microbiology
Astra Research Center Boston
128 Sidney Street
Cambridge, MA 02139-4239
617-234-2540
FAX: 617-576-3030
richard.alm@arcb.us.astra.com

**Gil Alterovitz**
Fullbright Scholar
Institute of Biomedical Engineering
University of Toronto
SLRI, Mt. Sinai Hospital
600 University Avenue
Toronto, Ontario M5G 1X5
CANADA
416-586-4800 X 2790
FAX: 416-586-8869
bp837@cleveland.freenet.edu

**Naoki Amano**
Student
AIST-NIBHT CREST
Center of Structural Biology
University of Teukuba
Higasi 1-1 Teukuba
Ibaraki, 305-0538
JAPAN
81-298-54-6407
FAX: 81-298-54-6534
nmano@nibh.go.jp

**Beth Andrews**
Senior Research Associate
Astra Research Center Boston
128 Sidney Street
Cambridge, MA 02139
617-234-2564
FAX: 617-576-3030
beth.andrews@arcb.us.astra.com

**Paul Appeddu**
Project Leader
Product Development
Genosys Biotechnologies
1442 Lake Front Circle
The Woodlands, TX 77380
281-363-3693
FAX: 281-363-5909
appeddu@genosys.com

**Jaspreet Arora**
Field Application Specialist
Automation
QIAGEN, Inc.
28159 Avenue Stanford
Valencia, CA 91355
800-426-8157x358
FAX: 301-309-6369
jaspreet@earthlink.net

**Scott Avery**
Research Assistant II
Functional Genomics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-0200
FAX: 301-838-0208
savery@tigr.org

**Yoshinao Azuma**
Researcher
Centre of Structural Biology
National Institute of Bioscience and Human Technology
Agency of Industrial Science and Tech.
Tsukuba
Ibaraki, 305-8566
JAPAN
0298-54-6535
FAX: 0298-54-6534

**Linda Banerjei**
Postdoctoral Fellow
Functional Genomics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319
301-315-2524
FAX: 301-838-0208
banerjei@tigr.org

**Valerie Barbe**
Research and Development
Genoscope, CNS
2 rue Gaston Cremieux, BP191
Evry Cedex, 91006
FRANCE
33-01-60-87-25-56
FAX: 33-01-60-87-25-89
vbarbe@genoscope.cns.fr

**Mary Barnstead**
Team Leader
Library Construction-Seqcore
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319
301-838-0200
FAX: 301-838-0208
maryb@tigr.org

**Brigitte Bathe**
Project Coordinator Genomics
Department of Genetics
University of Bielefeld
PO Box 100131
Bielefeld, 33501
GERMANY
49-521-1065602
FAX: 49-521-1065626
brigitte.bathe@genetik.uni-bielefeld.de

**Michael Beach**
Virtek Vision, Inc.
300 Wildwood Avenue
Woburn, MA 01801
781-933-3456
FAX: 781-933-3461

**Stephen Bentley**
Senior Computer Biologist
Pathogen Sequencing Unit
The Sanger Centre
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SA
UNITED KINGDOM
44-122-383-4244
FAX: 44-122-349-4919
sdb@sanger.ac.uk

**Eric Blair**
Research Associate I
Seqcore
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-0200
FAX: 301-838-0208
eblair@tigr.org

**Craig Bloch**
Assistant Research Scientist
Biological Chemistry and Pediatrics
University of Michigan
MSRB I, Room A500
1150 West Medical Center Drive
Ann Arbor, MI 48109-0656
734-703-2005
FAX: 734-647-9703
cbloch@umich.edu

**Jeffrey Bock**
Research Associate
Genomics
Pharmacia and Upjohn, Inc.
7242-267-510
301 Henrietta Street
Kalamazoo, MI 49007
616-833-1305
FAX: 616-833-2599

**Sharen Bowman**
Research Associate
Pathogen Sequencing Unit
The Sanger Centre
Hinxton Hall
Hinxton, Cambridge CB10 1SA
UNITED KINGDOM
44-122-383-4244
FAX: 44-122-349-4919
sharen@saner.ac.uk

**Thomas Brettin**
Computer Technician
Theoretical Biology and Biophysics
Los Alamos National Laboratory
Group T-10, MS K710
Los Alamos, NM 87545
505-667-2307
FAX: 505-665-3493
brettin@lanl.gov

**Peter Bretting**
National Program Leader
Plant Germplasm and Genomes
USDA/ARS
Building 005 BARC-W
10300 Baltimore Avenue
Beltsville, MD 20705
*301-504-5541*
FAX: 301-504-5467
pkb@ars.usda.gov

**Joan Brooks**
Vice President
Strategic Development
Proteome, Inc.
100 Cummings Center, Suite 435M
Beverly, MA 01915
978-922-1643
FAX: 978-922-3970
jeb@Proteome.com

**Robert Bruccoleri**
President
Department of Macromolecular Structure
Congenomics, Inc.
114 W. Franklin Avenue, Suite K1,10
PO Box 314
Pennington, NJ 08534
609-737-6383
FAX: 609-737-7528
bruc@acm.org

**Martin Burnham**
Assistant Director
Anti-infectives
SmithKline Beecham Pharmaceuticals
1250 South Collegeville Road
Collegeville, PA 19426
610-917-6349
FAX: 610-917-7901
Martin-K-Burnham@sbphrd.com

**Dolores Cahill**
Group Leader
Dept. of Prof. Hans Lehrach
Max-Planck Institute of Molecular Genetics
Ihnestasse 73
Berlin - Dahlem, D-14195
GERMANY
49-0-30-8413-1222
FAX: 49-0-30-8413-1380
cahill@mpimg-berlin-dahlem.mpg.de

**Weiguo Cao**
Instructor
Microbiology and Immunology
Cornell University Weill Medical College
1300 York Avenue, Box 62
New York, NY 10021
212-746-6517
FAX: 212-746-8587
wgc@mail.med.cornell.edu

**Stephane Carbonell**
MSC
Eurogentec
Parc Scientifique du sart
Seraing, 4102
BELGIUM
32 4 366 01 50
FAX: 32 4 365 16 04
d.allaer@eurogentec.be

**Liane Carpenter**
Research Associate
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319
301-838-0200
FAX: 301-838-0208
liane@tigr.org

**Sherwood Casjens**
Professor
Department of Oncological Sciences
University of Utah
University of Utah Medical Center
Salt Lake City, UT 84132-0001
801-581-5980
FAX: 801-581-3607
sherwood.casjens@
qmserver.genetics.utah.edu

**Stephen Chamberlin**
Bioformatics Group Leader
Sulfonics, Inc.
12085 Research Drive
Alachua, FL 32615
904-462-2000
FAX: 904-418-1680
sgc@biotech.ufl.org

**Ludmila Chistoserdova**
Research Assistant Professor
Department of Chemical Engineering
University of Washington
PO Box 351750
Seattle, WA 98195
206-616-6954
FAX: 206-543-8297
milachis@u.washington.edu

**George Church**
Professor
Department of Genetics
Harvard Medical School
Warren Alpert Building, Room 513
200 Longwood Avenue
Boston, MA 02115
617-432-7562
FAX: 617-432-7266
Church@arep.med.harvard.edu

**Henry Cittone**
Research Associate II
Sequencing Facility
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319
301-838-0200
FAX: 301-838-0208
hcittone@tigr.org

**Rebecca Clayton**
Assistant Investigator
Microbial Genomics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-3540
FAX: 301-838-0208
rclayton@tigr.org

**Sandra Clifton**
Small Projects Coordinator
Department of Genetics -
Genome Sequencing Center
Washington University School of Medicine
4444 Forest Park Boulevard
St. Louis, MO 63108
314-286-1467
FAX: 314-286-1810
sclifton@watson.wustl.edu

**Stephen Clougherty**
Marketing Assistant
Marketing Department
MJ Research, Inc.
149 Grove Street
Watertown, MA 02472
617-972-8143
FAX: 617-924-2148
stevec@mjr.com

**Jerry Coffey**
President
Sooner Scientific, Inc.
PO Box 180
Garvin, OK 74736
580-286-9408
FAX: 580-286-7047
sonrsci@ionet.net

**John Coleman**
TIGR Board of Trustees
9712 Medical Center Drive
Rockville, MD 20850
301-838-0200
FAX: 301-838-0208

**Daniel Davison**
Principal Scientist
Bioinformatics
Bristol-Myers Squibb
5 Research Parkway
Department 853
Wallingford, CT 06492-7660
203-677-7958
FAX: 203-677-6771
davisond@bms.com

**Anne DeGroot**
TB/HIV Research Laboratory
Brown University
International Health Institute
Box G/B 473
Providence, RI 02912-4799
401-863-1374
FAX: 410-863-1243
Anne_DeGroot@Brown.edu

**Herman Deweerd**
Vice President
Virtek Vision, Inc.
300 Wildwood Avenue
Woburn, MA 01801
781-933-3456
FAX: 781-933-3461

**Floyd Dewhirst**
Senior Member of Staff
Molecular Genetics Department
Forsyth Dental Center
140 Fenway
Boston, MA 02115
617-262-5200 x 298
FAX: 617-262-4021
FDewhirst@forsyth.org

**Robert Dodson**
Bioinformatics Analyst
Bioinformatics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319
301-838-0200
FAX: 301-838-0208
rjdodson@tigr.org

**Hirofumi Doi**
Project Director
Doi Bioasymmetry Project, ERATO
Japan Science and Technology Corporation
WBG Marive East 12F
Nakase 2-6, Mihama-ku
Chiba, 261-7112
JAPAN
81-43-299-1351
FAX: 81-43-297-7530
doi@bioa.jst.go.jp

**Brian Dougherty**
Senior Research Investigator II
Applied Genomics
Bristol-Myers Squibb
5 Research Parkway
Wallingford, CT 06492
203-677-7604
FAX: 203-677-6771
dougherb@bms.com

**Darrell Doyle**
Senior Investigator
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-3520
FAX: 301-838-0208
djdoyle@tigr.org

**Andreas Duesterhoeft**
Business Unit Manager
QIAGEN GmbH
Max-Volmer Strasse 4
Hilden, D-40724
GERMANY
49-2103-892-307
FAX: 49-2103-892-308
a.duesterhoeft@de.qiagen.com

**Bernard Dujon**
Biotechnologies
Institut Pasteur
25 Rue du Dr Roux
Paris Cedex 15, 75724
FRANCE
33-1-45-68-84-82
FAX: 33-1-40-61-34-56
bdujon@pasteur.fr

**Margaret Duncan**
Associate Member of the Staff
Department of Molecular Genetics
Forsyth Dental Center
140 Fenway
Boston, MA 02115
617-262-5200 X 344
FAX: 617-262-4021
mduncan@forsyth.org

**Susan Eddins**
Next Generation Product Manager
Genetic Analysis Marketing
PE Biosystems
850 Lincoln Centre Drive
Foster City, CA 94404
650-638-5470
FAX: 650-638-6666
eddinsSK@pe-bio.com

**Charles Edmonds**
Physical Scientist
Division of Medical Science
Office of Biological and Environmental Research
US Department of Energy
19901 Germantown Road, ER-73, B-119
Germantown, MD 20874-1290
301-903-0042
FAX: 301-903-0567
charles.edmonds@science.doe.gov

**Marti Edwards**
Conference Coordinator
Conferences, Education and Training
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-3509
FAX: 301-838-0229
medwards@tigr.org

**Dusko Ehrlich**
Laboratoire de Genetique Microbienne
Institut National de la Recherche Agronomique
Jouy en Josas Cedex, 78352
FRANCE
33-1-34 65 25 11
FAX: 33-1-34 65 25 21
ehrlich@biotec.jouy.inra.fr

**Ingvar Eidhammer**
Associate Professor
Informatics
University of Bergen
HIB
Bergen, N5020
NORWAY
47-55-58-41-64
FAX: 47-55-58-41-99
ingvar@ii.uib.no

**Jonathan Eisen**
Assistant Investigator
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20870
301-652-5844
FAX: 301-838-0208
jeisen@tigr.org

**David Eisenberg**
Director
Department of Energy Lab of Structural Biology and Molecular Medicine
UCLA
Box 951570
Los Angeles, CA 90095-1570
310-825-3754
FAX: 310-206-3914
david@mbi.ucla.edu

**Gregory Endress**
Staff Scientist
Research and Development
Life Technologies, Inc.
9800 Medical Center Drive
Rockville, MD 20850
301-610-8327
FAX: 301-610-8371
gendress@lifetech.com

**Alexey Eroshkin**
Senior Scientist
Bioinformatics
Axys Pharmaceuticals
11099 North Torrey Pines Road, Suite 160
La Jolla, CA 92037
619-646-8378
FAX: 619-452-6653
alexey@axyspharm.com

**Allison Errett-Golden**
Research Assistant
Biology
University of New Mexico
Biology Building -167 Castetter Hall
Albuquerque, NM 87131-1091
505-277-9339
FAX: 505-277-2103
errett@unm.edu

**Shawn Estrem**
Post Doctoral Scientist
Infectious Disease
Eli Lilly and Company
Lilly Corporate Center
Drop Code 0428
Indianapolis, IN 46285
317-433-3040
FAX: 317-227-0778
stestrem@lilly.com

**Tamara Feldblyum**
Director
DNA Sequencing Facility
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-610-5945
FAX: 301-838-0208
tamaraf@tigr.org

**Stephen Fey**
Center for Proteome Analysis
Odense University
International Science Park Odense
Forskerparken 10B
Odense M, DK-5230
DENMARK
011-45-63-15-7239
FAX: 011-45-63-15-7240
sjf@cpa.spo.dk

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

**Claire Fraser**
President
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-0200
FAX: 301-838-0208
cmfraser@tigr.org

**Marvin Frazier**
Director of Health Effects and Life Sciences Division
Office of Biological and Enviornmental Research
US Department of Energy
19901 Germantown Road
Germantown, MD 20874-1290
301-903-5368
FAX: 301-903-8521
marvin.frazier@science.doe.gov

**Joyce Fuhrmann**
Technical Training Specialist
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-610-5950
FAX: 301-838-0208
fuhrmann@tigr.org

**Dean Gaalaas**
Edge BioSystems
19208 Orbit Drive
Gaithersburg, MD 20879-4149
301-990-2685
FAX: 301-990-0881

**Terry Gaasterland**
Laboratory for Computational Genomics
The Rockefeller University
1230 York Avenue, Box 250
New York, NY 10021-6399
212-327-7755
FAX: 212-327-8544
gaasterland@rockefeller.edu

**Emilio Garcia**
Senior Scientist
Biology and Biotechnology
Lawrence Livermore National Laboratory
7000 East Avenue, L-452
Livermore, CA 94550
925-422-8002
FAX: 925-422-2282
garcia12@llnl.gov

**Malcolm Gardner**
Assistant Investigator
Eukaryotic Genomics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319
301-838-3519
FAX: 301-838-0208
gardner@tigr.org

**James Garrels**
President and CEO
Proteome, Inc.
100 Cummings Center, Suite 435M
Beverly, MA 01915
978-922-1643
FAX: 978-922-3971
jg@proteome.com

**Patee Gesuwan**
Research Associate
Functinal Genomics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319
301-838-0200
FAX: 301-838-0208
pgesuwan@tigr.org

**Beverly Giammara**
Marketing Development Manager
Virtek Vision, Inc.
300 Wildwood Avenue
Woburn, MA 01801
781-933-3456
FAX: 781-933-3461
bgiammara@compuserve.com

**Steven Gill**
Assistant Investigator
Functional Genomics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-315-2521
FAX: 301-838-0208
srgill@tigr.org

**John Gill**
Research Associate II
Molecular Biology/Closure
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-0200
FAX: 301-838-0208
jgill@tigr.org

**John Glass**
Senior Biologist
Department of Infectious Diseases Research
Lilly Research Laboratories
Eli Lilly and Company
Lilly Corporate Center
Indianapolis, IN 46285-0438
317-277-0143
FAX: 317-276-1743
glass_john_i@lilly.com

**Andre Goffeau**
Professor
FYSA
Universite Catholique de Louvain
Place Croix du Sud 2-20
Louvain la Neuve, B-1348
BELGIUM
32-10-47-36-14
FAX: 32-10-47-38-72
goffeau@fysa.ucl.ac.be

**Brad Goodner**
Assistant Professor of Biology
Biology
University of Richmond
Science Center E-105
Richmond, VA 23173
804-289-8661
FAX: 804-289-8233
bgoodner@richmond.edu

**Michael Gottlieb**
Parasitology and International
Programs Branch

Solar Building, Room 3A12
Bethesda, MD 20892-7630
301-496-2544
FAX: 301-402-0659
mgottlieb@
mercury.niaid.nih.gov

**Gerhard Gottschalk**
Professor of Microbiology
Microbiology and Genetics
Georg-August Universitat
Grisebach Str. 8
Gottingen, 37077
GERMANY
551-39-3781
FAX: 551-39-3808
ggottsc@gwdg.de

**Sridhar Govindarajan**
Scientist/Systems Architect
Bioinformatics
Sulfonics, Inc.
12085 Research Drive
Alachua, FL 32615
904-462-2000
FAX: 904-418-1680
sridharg@biotech.ufl.org

**Jessica Gray**
Research Associate I
Seqcore
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-0200
FAX: 301-838-0208
jbeach@tigr.org

**David Guilliano**
Ph.D. Student
ICAPB
University of Edinburgh
NEB Inc.
32 Tozer Road
Beverly, MA 01915
978-927-5054 x 327
FAX: 978-921-1350
dguilian@holyrood.ed.ac.uk

**Radhey Gupta**
Department of Biochemistry
McMaster University
Hamilton, Ontario L8N 3Z5
CANADA
905-525-9140 x 22639
FAX: 905-522-9033
gupta@fhs.mcmaster.ca

**Michelle Gwinn**
Staff Scientist
Bioinformatics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319
301-315-2536
FAX: 301-838-0208
migwinn@tigr.org

**Daniel Haft**
Bioinformatics Analyst
Informatics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-610-5952
FAX: 301-838-0208
Haft@tigr.org

**Jean Hani**
Research and Bioinformatics
Biomax Informatics GmbH
Lochhamerstr.11
Martinstied, D-82152
GERMANY
49 89 89590434
FAX: 49 89 89590725
Jean.Hani@biomax.de

**Kimberly Hansen**
Product Specialist
Department of Genetic Analysis
PE Biosystems
850 Lincoln Centre Drive, 800 Building
Foster City, CA 94404
650-638-5202
FAX: 650-638-6200
hansenkj@pebio.com

**Thomas Hansen**
Research Associate I
Seqcore
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-0200
FAX: 301-838-0208
thansen@tigr.org

**Philip Harriman**
Program Director
Microbial Genetics
National Science Foundation
4201 Wilson Boulevard, MCB, Room 655
Arlington, VA 22230
703-306-1439
FAX: 703-306-0355
pharrima@nsf.gov

**David Harris**
Team Leader
Pathogen Sequencing Unit
The Sanger Centre
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SA
UNITED KINGDOM
44-122-383-4244
FAX: 44-122-349-4919
deh@sanger.ac.uk

**Barbara Harris**
Team Leaderr
The Sanger Centre
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SA
UNITED KINGDOM
44-122-383-4244
FAX: 44-122-349-4919
brh@sanger.ac.uk

**Thomas Hartsch**
Gottingen Genomics Laboratory
Institute of Microbiology and Genetics
Grisebachstrasse 8
Goettingen, Lower Saxony 37077
GERMANY
49-551-393841
FAX: 49-551-394195
thartsc@Uni-MolGen.gwdg.de

**Florence Haseltine**
Director
Center for Population Research
National Institute of Child Health
and Human Development
National Institutes of Health
6100 Executive Boulevard
Room 8B-07 - MSC-7510
Bethesda, MD 20892
301-496-1101
FAX: 301-496-0962

**Shin-ichi Hashimoto**
Researcher
Tokyo Research Laboratories
Kyowa Hakko Kogyo Company, Ltd.
3-6-6 Asahi-machi
Machida-shi
Tokyo, 194-8533
JAPAN
81-427-25-2555
FAX: 81-427-26-8330
shashimoto@kyowa.co.jp

**Judith Healy**
Senior Scientist
Drug Discovery
Scriptgen Pharmaceuticals, Inc.
610 Lincoln Street
Waltham, MA 02451
781-768-3403
FAX: 781-768-0040
healy@scriptgen.com

**John Heidelberg**
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319
301-838-0200
FAX: 301-838-0208
jheidel@tigr.org

**Roland Heilig**
Researcher
Research and Development
Genoscope, CNS
2 rue Gaston Cremieux, BP191
Evry Cedex, 91006
FRANCE
33-1-69-47-28-00
FAX: 33-1-60-77-86-98
heilig@genscope.cns.fr

**Josef Hermanns**
Senior Manager
Genomics
Lion Bioscience Ag
Im Neuenheimer Feld 515
Heidelberg, 69120
GERMANY
49-6221-4038127
FAX: 49-6221-4038101
hermanns@lion-ag.de

**Erin Hickey**
Microbial Annotation Coordinator
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319
301-838-3563
FAX: 301-838-0208
ehickey@tigr.org

**Paula Hicks**
Senior Scientist
Biotechnology Central Research
Cargill, Inc.
PO Box 5702
Minneapolis, MN 55440-5702
612-742-3000
FAX: 612-742-3010

**Winston Hide**
Director
South African National Bioinformatics
Institute
University of the Western Cape
Private Bag X17
Bellville, Capetown 7535
SOUTH AFRICA
27-21-959-3645
FAX: 27-21-959-2266
winhide@sanbi.ac.za

**Satoshi Hihara**
Researcher
Doi Bioasymetry Project. ERATO
Japan Science and Technology Corporation
Tsukuba Research Consortium
5-9-9 Tokadai
Tsukuba, Ibaraki 300-2635
JAPAN
81 298 48 1515
FAX: 81-298 47 8901
hiha@bioa.jst.go.jp

**Helmut Hilbert**
Laboratory Manager
QIAGEN GmbH
Max-Volmer Strasse 4
Hilden, D-40724
GERMANY
49-2103-892-307
FAX: 49-2103-892-308
h.hilbert@de.qiagen.com

**Gregory Hinkle**
Computational Biologist
Bioinformatics
Ceeon Genomics
One Kendall Square, Building 300
Cambridge, MA 02139
617-577-3504
FAX: 617-577-3562
gregory.hinkle@cereon.com

**Christopher Hogue**
Scientist
Samuel Lunenfeld Reearch Insititut
Mt. Sinai Hospital
600 University Avenue
Toronto, Ontario M5G1X5
CANADA
416-586-4800
FAX: 416-586-8857
hogue@mshri.on.ca

**Erik Holtzapple**
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319
301-838-0200
FAX: 301-838-0208
eholtz@tigr.org

**John Houghton**
Office of Biological
and Environmental Research
US Department of Energy, ER-74
19901 Germantown Road
Germantown, MD 20874-1290
301-903-8288
FAX: 301-903-8519
john.houghton@science.doe.gov

**Ping Hu**
Bioinformatics Investigator
Bioinformatics
diaDexus
3303 Octavius Drive
Santa Clara, CA 94588
408-330-5064
FAX: 408-330-4926
phu@diadexus.com

**Jennie Hunter-Cevera**
Head-Center for Environmental
Biotechnology
Life Sciences
Lawrence Berkeley National Laboratory
1 Cyclotron Road, Building 70A-3317
Berkeley, CA 94720
510-486-7359
FAX: 510-486-7152
jchunter-cevera@lbl.gov

**Ana Hurtado**
Postdoctoral Positions
Microbiology
Universidad Miguel Hernandez
Facultad de Medicina
Campus de San Juan, Facultad de Medicina
Crta. Valencia Km 87-Apdo. 18
San Juan, 03550
SPAIN
34-96-5919313
FAX: 34-96-5919457

**Carston Jacobi**
Goettingen Genomic Laboratory
Institute of Microbiology and Genetics
Grisebachstrasse 8
Goettingen, Lower Saxony 37077
GERMANY
49-551-393-841
FAX: 49-551-394-195
cjacobi@gwdg.de

**Inge Jonassen**
Post Doctorial
Department of Informatics
University of Bergen
HIB
Bergen, N5020
NORWAY
47 55 58 4713
FAX: 47 55 58 4199
inge@ii.uib.no

**Louis Jones**
Bioinformatics Scientist
Service Informatique Scientifique
Institut Pasteur
28, rue du Docteur Roux
Paris, 75724
FRANCE
33-1 40 61 37 25
FAX: 33-1 40 61 30 80
lmj@pasteur.fr

**Jan Karlsson**
Researcher
Microbiology
FOA NBC Defense
Umea, SE-901 82
SWEDEN
46-90-106600
FAX: 46-90-106806
jan.karlsson@ume.foa.se

**Hiroyuki Kato**
Senior Scientist
Eisai Research Institute of Boston, Inc.
100 Research Drive
Wilmington, MA 01887
978-661-7204
FAX: 978-657-7715
hiroyuki_kato@eri.eisai.com

**Arthur Katz**
Physical Scientist
Office of Biological and Environmental
Research
US Department of Energy
ER 72/GTN
19901 Germantown Road
Germantown, MD 20874-1290
301-903-4932
FAX: 301-903-8521
arthur.katz@science.doe.gov

**Eric Kaufman**
Program Manager
Office of Nonpreliferation of National
Security
US Department of Energy, NN-20
1000 Independence Avenue
Washington, DC 20585
202-586-7770
FAX: 202-586-0485
eric.kaufman@hq.doe.gov

**Yutaka Kawarabayasi**
Biotechnology Center
National Institute of Technology
and Evaluation
2-49-10 Nishibara
Shibuya-ku
Tokyo, 151-0066
JAPAN
81-3-3481-8951
FAX: 81-3-3481-8424
kyutaka@kazusa.or.jp

**Louise Kelleher**
Marketing Services
Genpack Limited
Science Park Square, Falmer
Brighton, BN1 9SB
UNITED KINGDOM
0 12 72 704470
FAX: 0 12 73 626213
info@genpackdna.com

**Karen Ketchum**
Associate Investigator
Functional Genomics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-3531
FAX: 301-838-0208
kketchum@tigr.org

**Hanif Khalak**
Software Engineer
Bioinformatics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319
301-838-2520
FAX: 301-838-0208
hkhalak@tigr.org

**Hoda Khouri**
Molecular Biologist
Closure
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-4050
FAX: 301-838-0208
hkhouri@tigr.org

**Julia Kozlovsky**
Software Engineer
Department of Bioinformatics
Genome Therapeutics Corporation
100 Beaver Street
Waltham, MA 02453-8443
781-398-2666
FAX: 781-398-2470
julia.kozlovsky@genomecorp.com

**Frank Kunst**
Director of Research
Departement des Biotechnologies
Unite de Biochemie Microbienne
Institut Pasteur
25, rue du Docteur Roux
Paris Cedex 15, 75724
FRANCE
33 1 45 68 88 69
FAX: Fax (lab): 33 1 45 68 87 86
Fax: (dept): 33 1 45 68 87 90

**Yoko Kuwabara**
Applied Microbiology Laboratory
Ajinomoto Company, Inc.
Building 1, 1-1 Susuki-cho
Kawasaki-ku
Kawasaki City, Kanagawa 210-8681
JAPAN
81-044-244-7102
FAX: 81-044-245-1548
bld_kuwabara@to2.ajinomoto.co.jp

**Frank Larimer**
Staff Scientist
Computational Biosciences Section
Oak Ridge National Laboratory
1060 Commerce Park Drive
Oak Ridge, TN 37831-0793
423-574-1253
FAX: 423-241-1965
FWL@ornl.gov

**Bernie Lauro**
Conference Director
Conferences, Education and Training
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-3561
FAX: 301-838-0229
bglauro@tigr.org

**Daniel Lawson**
Computer Biologist
Informatics
The Sanger Centre
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SA
UNITED KINGDOM
44-122-383-4200
FAX: 44-122-349-4919
dl1@sanger.ac.uk

**Terrance Leighton**
Professor of Biochemistry
and Molecular Biology
Biochemistry and Molecular Biology
University of California
MBM, 401 Barker Hall
Berkeley, CA 94720-3203
510-642-1620
FAX: 510-643-5035
leighton@socrates.Berkeley.edu

**Nicola Lennard**
Senior Finisher
The Sanger Centre
Wellcome Trust Genome Campus
Hinkton, Cambridge CB10 1SA
UNITED KINGDOM
44-122-383-4244
FAX: 44-122-349-4919

**Matthew Lewis**
Research Associate II
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319
301-838-0200
FAX: 301-838-0208
mrlewis@tigr.org

**Guangshan Li**
Postdoctoral Associate
Environmental Sciences Division
Oak Ridge National Laboratory
Bethal Valley Road
Oak Ridge, TN 37831-6038
423-576-2766
FAX: 423-576-8646
iqu@ornl.gov

**Jia Yeu Liu**
Post-Doctoral Fellow
Infectious Diseases
Parke-Davis Pharmaceutical Research
2800 Plymouth Road
Ann Arbour, MI 48105
734-622-5727
FAX: 734-622-7158
jia.liu@WL.com

**Jennifer Lodge**
Assistant Professor
Biochemistry and Molecular Biology
St. Louis University School of Medicine
1402 South Grand Boulevard
St. Louis, MO 63104
314-577-8143
FAX: 314-577-8156
lodgejk@wpogate.slu.edu

**Brendan Loftus**
Data Specialist
Bioinformatics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319
301-838-3543
FAX: 301-838-0208
bjloftus@tigr.org

**Nicholas Loman**
Medical Student
St. Bartholomew's Hospital
The Royal London School of Medicine and Dentistry
West Smithfield
London, ECIA 7BE
UNITED KINGDOM
44-171-601-8414
FAX: 44-171-601-8409
n.j.loman@mds.qmw.ac.uk

**James Lyons-Weiler**
Sloan Postdoctoral Fellow
Department of Biology
Institute of Molecular Evolutionary Genetics
Pennsylvania State University
328 Mueller Lab
University Park, PA 16802
814-865-1034
FAX: 814-863-7336
jfl8@psu.edu

**Masayuki Machida**
Senior Researcher
Molecular Biology
National Institute of Bioscience & Human-Technology
Higashi 1-1
Tsukuba, Ibaraki 305-8566
JAPAN
81-298-54-6214
FAX: 81-298-54-6240
machida@nibh.go.jp

**Matthias Mack**
Scientist
BAST-LYNX Bioscience AG
Im Neuenheimer Feld 519
Heidelberg, 69120
GERMANY
49-6221-454-733
FAX: 49-6221-454-778
mack@basf-lynx.de

**Kira Makarova**
Postdoctoral Fellow
Pathology
Uniformed Services University of the Health Sciences
4301 Jones Bridge Road
Bethesda, MD 20816
301- 435-5912
FAX: 301-295-1640
makarova@ncbi.nlm.nih.gov

**Dennis Mangan**
Director, Infectious Diseases Program
National Institute of Dental and Craniofacial Research
National Institute of Health
Building 45, Room 4AN-32F
Bethesda, MD 20892-6402
301-594-2421
FAX: 301-480-8318
dennis.mangan@nih.gov

**Joe Mangan**
Research Fellow
Department of Medical Microbiology
St. George's Hospital Medical School
Cranmer Terrace
Tooling
London, SW 17 ORE
ENGLAND
44-181-725-5740
FAX: 44 181 672 0234
jmangan@sghms.ac.uk

**Gerard Manning**
Research Scientist
Molecular Applications Group
607 Hanson Way
Building One
Palo Alto, CA 94304
650-846-3499
FAX: 650-846-3595
ger@mag.com

**Betty Mansfield**
Managing Editor, Human Genome News
Human Genome Management Information System
Oak Ridge National Laboratory
1060 Commerce Park, MS-6480
Oak Ridge, TN 37830
423-576-6669
FAX: 423-574-9888
mansfieldbk@ornl.gov

**Peter Mansfield**
Vice President
Sales and Marketing
Compugen, Inc.
3 Baldwin Green Common, #201
Woburn, MA 01801
781-932-1155
FAX: 781-932-7409
peter@cgen.com

**Tanya Mason**
Research Associate III
Seqcore
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319
301-838-0200
FAX: 301-838-0208
tmbadger@tigr.org

**Paula McCready**
DOE Joint Genome Institute
2800 Mitchell Drive, B100
Walnut Creek, CA 94588
925-296-5635
FAX: 925-296-5710
mccready2@llnl.gov

**Scott McGinnis**
Technical Information Specialist
National Center for Biotechnology Information
8600 Rockville Pike
Building 38A, Room 8N803
Bethesda, MD 20894
301-496-2475
FAX: 301-480-9241
info@ncbi.nlm.nih.gov

**Claudine Medigue**
Researcher
Science de la Vie
Centre National de la Recherche Scientifique
Institut Pasteur
28 rue du Docteur Roux
Paris Cedex 15, 75005
FRANCE
33 01 45 63 84 41
FAX:

**John Mekalanos**
Professor and Chair
Department of Microbiology and Molecular Genetics
Harvard Medical School
200 Longwood Avenue
Building D1 - Room 421
Boston, MA 02115
617-432-1935
FAX: 617-738-7664
jmekalan@warren.med.harvard.edu

**Gregory Meyers**
Advanced Lab Specialist
Commonwealth Biotechnologies, Inc.
601 Biotech Drive
Richmond, VA 23235
804-648-3820
FAX: 804-648-2641
gmeyers@cbi.biotech.com

**Katerina Michalickova**
Graduate Student
Biochemistry
University of Toronto
600 University Avenue
Toronto, Ontario M5G 1X5
CANADA
416-92-586-4800 x 2863
FAX: 416-586-8869
katerina@mshri.on.ca

**Bill Miller**
Director
Marketing
Edge BioSystems
19212 Orbit Drive
Gaithersburg, MD 20879-4911
301-990-2685
FAX: 301-990-0881
bmiller@edgebio.com

**Kenneth Minton**
Professor
Department of Pathology
Uniformed Services University of the Health Sciences
4301 Jones Bridge Road
Bethesda, MD 20814-4799
301-295-3476
FAX: 301-295-1640
kminton@usuhs.mil

**C. Garrett Miyada**
Manager
Product Development
Affymetrix
3380 Central Expressway
Santa Clara, CA 95051
408-731-5534
FAX: 408-481-0435
garry_miyada@affymetrix.com

**Ilene Mizrachi**
Staff Scientit
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
8600 Rockville Pike
Bethesda, MD 20894
301-435-5929
FAX: 301-435-7794
mizrachi@ncbi.nlm.nih.gov

**Kelly Moffat**
Research Assistant II
Seqcore
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319
301-838-0200
FAX: 301-838-0208
kmoffat@tigr.org

**Donald Moir**
Research Director
Pathogen Genetics
Genome Therapeutics Corporation
100 Beaver Street
Waltham, MA 02453-8443
781-398-2365
FAX: 781-398-2476
don.moir@genomecorp.com

**Anne Morgat**
Scientist
Antibacterial Programme
Rhone-Paulenc Rorer
13 Quai Jules Guesde
Central Research, Magendie Building
Vitry-Sur-Seine, Cedex 94403
FRANCE
33-1-55713829
FAX: 33-1-55713960
anne.morgat@rp-rorer.fr

**Subhash Morzaria**
Principal Scientist
Parasitology
International Livestock Research Institute
PO Bo x 3079
Nairobi,
KENYA
254-2-630743
FAX: 254-2-631499
s.morzaria@cgiar

**Farahnaz Movahedzadeh**
Research Fellow
Infectious and Tropical Diseases
London School of Hygiene and Tropical Medicine
Keppel Street
London, WCAIE 7HT
UNITED KINGDOM
171-637-4314
FAX: 171-637-4314
f.movahedzadeh@lshtm.ac.uk

**Jacob Mueller**
Research Associate I
BAC Ends
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-0200
FAX: 301-838-0208
jmueller@tigr.org

**Amanda Mullens**
Technical Representative
Sooner Scientific, Inc.
PO Box 180
Garvin, OK 74736
580-286-9408
FAX: 580-286-7047
sonrsci@ionet.net

**Karen Mungall**
Group Leader
Pathogen Sequencing Unit
The Sanger Centre
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SA
UNITED KINGDOM
44-122-383-6244
FAX: 44-122-349-4919
klb@snager.ac.uk

**Richard Mural**
Senior Staff Scientist
Department of Lifesciences
Computational Biosciences Section
Oak Ridge National Laboratory
1060 Commerce Park Drive
Oak Ridge, TN 37831-6480
423-576-2938
FAX: 423-241-1965
Muralrj@ornl.gov

**Lee Murphy**
Senior FInisher
Pathogen Sequencing Unit
The Sanger Centre
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SA
UNITED KINGDOM
44-122-383-4244
FAX: 44-122-349-4919
ldm@sanger.ac.uk

**Kaoru Nakasone**
Research Associate
Deep-sea Microorganism Research Group
Japan Marine Science and
Technology Center
2-15 Natsushima
Yokosuka, Kanagawa 237-0061
JAPAN
81-468-67-3895
FAX: 81-468-66-6364
nakasone@jamstec.go.jp

**Teruko Nakazawa**
Professor
Department of Microbiology
Yamaguchi University School of Medicine
Yamaguchi
Ube, 755
JAPAN
81-836-22-2226
FAX: 81-836-22-2227
nakazawa@po.cc.yamaguchi-u.ac.jp

**Karen Nelson**
Collaborative Investigator
Prokaryotic Genomics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-3565
FAX: 301-838-0208
kenelson@tigr.org

**William Nelson**
Bioinformatics Analyst
Microbial Annotation
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-315-2522
FAX: 301-838-0208
wnelson@tigr.org.

**Vish Nene**
Senior Scientist
Molecular Biology
International Livestock
Research Institute
PO Box 30709
Nairobi,
KENYA
254-2-630743
FAX: 254-2-631499
v.nene@cgiar.org

**William Nierman**
Vice President for Research
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319
301-838-3559
FAX: 301-838-0218
wneirman@tigr.org

**Cary O'Donnell**
Senior Research Bioscientist
Bioinformatics
Zeneca Pharmaceuticals
Mereside, Alderley Park
Cheshire
Macclesfield, SK10-4TG
UNITED KINGDOM
44-1625-515573
FAX: 44 1625 586900
Cary.ODonnell@alderley.zeneca.com

**Naotake Ogasawara**
Professor
Graduate School of Biological Sciences
Nara Institute of Science and Technology
8916-5 Takayama
Ikoma
Nara, 630-01
JAPAN
81-743-72-5430
FAX: 81-743-72-5439
nogasawa@bs.aist-nara.ac.jp

**Yuko Ohfuku**
Researcher
Molecular Biology
National Institute of Bioscience and
Human Technology
Higushi I-1. Tsukuba
Ibaraki, 305-6538
JAPAN
81-298-54-6407
FAX: 81-298-54-4534
ohfukku@nibh.go.jp

**Akira Ohyama**
Bioscience Systems
Mitsui Knowledge Industry Co. Ltd.
2-7-14 Higashi-Nakano
Nakano-Ku
Tokyo, 164-8555
JAPAN
81-3-3227-5724
FAX: 81-3-3366-6780
akr@hydra.mki.co.jp

**Onesmo Ole-MoiYoi**
Director
Institute for Molecular and
Cell Biology - Africa
PO Box 30709
Nairobi,
KENYA
254-2-630743
FAX: 254-2-631499
ole-moiyoi@

**Mark Pallen**
Senior Lecturer
Medical Microbiology
St. Bartholomew's Hospital
The Royal London Schools of
Medicines & Denistry
West Smithfield
London, ECIA 7BE
UNITED KINGDOM
44-171-594-5254
FAX: 44-171-594-5255

**Anna Palmisano**
Office of Biological
and Enviromental Research
US Department of Energy
19901 Germantown Road
Germantown, MD 20874
301-903-9963
FAX:

**Julian Parkhill**
Computer Biologist
Pathogen Sequencing Unit
The Sanger Centre
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SA
UNITED KINGDOM
44-122-383-4244
FAX: 44-122-349-4919
parkhill@sanger.ac.uk

**Jeremy Peterson**
Bioinformatics Engineer
Microbial
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-3526
FAX: 301-838-0208
peterson@tigr.org

**Scott Peterson**
Assistant Investigator
Functional Genomic
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-3539
FAX: 301-838-0208
scottp@tigr.org

**Xaun-Quynh Pham**
Research Coordinator
Medicine
University of Washington Genome Center
Fluke Hall - Room 225, Mason Road
Seattle, WA 98195
206-616-3859
FAX: 206-685-7344
xuan@u.Washington.edu

**William Phares**
Lab Head
Genetics
Novartis Forschungsinstitut
Brunnerstr. 59
Vienna, A-1235
AUSTRIA
43 1 86634 9067
FAX: 43 1 86634 727
william.phares@pharma.novartis.com

**Herve Philippe**
Laboratoire de Biologie cellulaire
Universite Paris-Sud
Bat 444
Orsay Cedex, 91405
FRANCE
33-1-69-156481
FAX: 33-1-69-156803
hp@bio4.bc4.u-psud.fr

**Mary Pingitore**
Sales Representative
PE Biosystems
850 Lincoln Center Drive
Foster City, CA 94404
800-248-0281 x 7281
FAX: 650-572-2743
pingitmc@pebio.com

**Rong Qi**
Graduate Student
Chemistry Department
New York University
PO Box 173
31 Washington Place
New York, NY 10003
212-998-8465
FAX: 212-995-4681
qi@carbon.chem.nyu.edu

**Haiying Qin**
Research Associate II
Seqcore
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-0200
FAX: 301-838-0208
hqin@tigr.org

**Jiansheng Qiu**
Scientist
Biotechnology Division
LI-COR, Inc.
4308 Progressive Avenue
Lincoln, NE 68504
402-467-0702
FAX: 402-467-6851
jqiu@bio.licor.com

**Michael Quail**
Senior Research Assistant
Pathogen Sequencing Unit
The Sanger Centre
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SA
UNITED KINGDOM
44-122-383-4244
FAX: 44-122-349-4919
mq1@sanger.ac.uk

**Anne Marie Quinn**
Student
Epidemiology and Public Health
Yale University
38 Marshall Avenue
Guilford, CT 06437
203-737-5011
FAX:
anne.marie.quinn@yale.edu

**Elisabeth Raleigh**
Senior Scientist
New England Biolabs, Inc.
32 Tozer Road
Beverly, MA 01915
978-927-5054
FAX: 978-921-1350
raleigh@neb.com

**Barry Ratzkin**
Associate Director
New Products and Technologies
Amgen Inc.
1 Amgen Center Drive
Thousand Oaks, CA 91360
805-447-3008
FAX: 805-499-9315
bratzkin@amgen.com

**Timothy Read**
Coll. Investigator
Microbial Genomics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-3554
FAX: 301-838-0208
tread@tigr.org

**Rajendra Redkar**
Research Scientist
Institute of Molecular Biology and Medicine
University of Scranton
IMBM Building, Room 303
Scranton, PA 18510
717-941-6353
FAX: 717-941-6229
redkarr1@taol.com

**Karl Reich**
Senior Research Molecular Biologist
Genomics & Molecular Biology
Abbott Laboratories
Dept. of MB, Building AP52
200 Abbott Park Road
Abbott Park, IL 60064-3403
847-938-2635
FAX: 847-938-3403
karl.reich@abbott.com

**Isabel Resende**
Research Associate I
BAC Ends
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-0200
FAX: 301-838-0208
iresende@tigr.org

**Andrew Revel**
R & D Specialist
Seqcore
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-383-0200
FAX: 301-838-0208
atrevel@tigr.org

**Thomas Reynolds**
Executive Vice President
Commonwealth Biotechnologies, Inc.
601 Biotech Drive
Richmond, VA 23235
804-648-3820
800-735-9224
FAX: 804-648-2641
treynolds@cbi-biotech.com

**Del Richardson**
Staff Scientist
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-3570
FAX: 301-838-0208
delwood@tigr.org

**Monica Riley**
Senior Scientist
Molecular Evolution
Marine Biological Laboratory
7 MBL ST
Woods Hole, MA 02543
508-289-7612
FAX: 508-540-6902
mriley@mbl.edu

**Frank Robb**
Professor
University of Maryland Center
for Marine Biology
701 East Pratt Street, Suite 236
Baltimore , MD 21202
410-234-8870
FAX: 410-234-8896
ROBB@mbimail.umd.edu

**Eduardo Rocha**
Ph.D. Student
Unite de Regulation de
l'Expression Genetique
Institut Pasteur
28, rue du Dr Roux
Paris Cedex 15, 75724
FRANCE
33-1-45-68-84-42
FAX: 33-1-43-06-98-35
erocha@pasteur.fr

**John Rosamond**
Senior Scientist
CIRD
Zeneca Pharmaceuticals
14G20X Mereside
Alderley Park, Macclesfield
Cheshire, SK10 4TG
UNITED KINGDOM
44-1625-513086
FAX: 44-1625-517436
john.rosamond@alder

**L. David Rothman**
Scientist
Computing, Modeling and Information
Sciences
Dow Chemical Company
1776 Building
Midland, MI 48674
517-636-3872
FAX:
ldrothman@dow.com

**Paul Roy**
Professor
Department of Biochemistry
Faculty of Sciences
Laval University
Pavillon Vachon
Quebec City, Quebec G1K 7P4
CANADA
418-654-2705
FAX: 418-654-2715
proy@rsvs.ulaval.ca

**James Russo**
Research Scientist
Columbia Genome Center
Columbia University
BB 1605
650 West 168th Street
New York, NY 10032
212-305-3955
FAX: 212-305-1191
jjr4@columbia.edu

**Annmarie Schramm**
Research Technician
NMRI/BDRP
Naval Medical Research Institute
8901 Wisconsin Avenue
Bethesda, MD 20889
301-295-2021
FAX: 301-295-2023
schrammA@nmripo.nmri.nnmc.navy.mil

**Tsutomu Sekizaki**
Head of Laboratory
Department of Bacterilogy and
Parasitology
National Institute of Animal Health
Kannondai 3-1-1
Tsukuba, Ibaraki 305-0856
JAPAN
81-298-38-7743
FAX: 81-298-38-7907
sekizaki@niah.affrc.go.jp

**James Selkirk**
Special Assistant to the Scientific
Director, NIEHS
Dept. of Health and Human Services,
Public Health Service, NIH
National Institute of Environmental
Health Sciences
PO Box 12233 (MD A2-09)
Research Triangle Park, NC 27709
919-541-2548
FAX: 919-541-5002

**Sejal Shah**
Product Coordinator
Product Development
Incyte Pharmaceuticals, Inc.
Microarray Systems
6519 Dumbarton Circle
Fremont, CA 94555
510-739-2177
FAX: 510-739-2200
sshah@synteni.com

**Gideon Shapiro**
Chief Scientific Officer
Sulfonics, Inc.
12085 Research Drive
Alachua, FL 32615
904-462-2000
FAX: 904-418-1680
gshapiro@biotech.ufl.org

**Joe Shaw**
Research Scientist
Institute of Molecular Biology
University of Scranton
IMBM Building, Room 207
Scranton, PA 18510
717-941-6353
FAX: 717-941-6229
joeshaw1953@aol.com

**Rob Shipman**
Project Manager
Gene Foundry
Visible Genetics Inc.
700 Bay Street, Suite 1000
PO Box 333
Toronto, Ontario M561Z6
CANADA
416-813-3240 X 4848
FAX: 416-813-3250

**Guus Simons**
Senior Scientist
Keygene N.V.
Agro Business Park go, PO Box 216
Wageningen, 6700AE
THE NETHERLANDS
31-317-466866
FAX: 31-317-424939
keygene@euronet.nl

**Jeffrey Skolnick**
Professor
Department of Molecular Biology
The Scripps Research Institute
10550 North Torrey Pines Road (TPC 5)
La Jolla, CA 92037
619-784-8821
FAX: 619-784-8895
skolnick@scripps.edu

**Barton Slatko**
DNA Sequencing Group
New England Biolabs, Inc.
32 Tozer Road
Beverly, MA 01915
978-927-5054 x 327
FAX: 978-921-1350
dnaseq@neb.com

**Bradley Slaven**
Software Engineer
Bioinformatics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-3521
FAX: 301-838-0208
bslaven@tigr.org

**Hamilton Smith**
Director of DNA Resources
Celera Genomics Corporation
45 West Gude Drive
Rockville, MD 20850
240-453-3000
FAX: 240-453-3755
smithho@celera.com

**Erik Snesrud**
Research Associate I
Functional Genomics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-0200
FAX: 301-838-0208
esnesrud@tigr.org

**Sylvia Spengler**
Director
Human Genome Program Field Operations
Lawrence Berkeley National Laboratory
One Cyclotron Road
Mail Stop 84-171
Berkeley, CA 94720
510-486-4879
FAX: 510-486-5717
SJSpengler@lbl.gov

**John Spieth**
Research Assistant Professor of Genetics
Genome Sequencing Center
Department of Genomics
Washington University School of Medicine
4444 Forest Park Boulevard
St. Louis, MO 63108
314-286-1840
FAX: 314-286-1810
jspieth@watsonwusfl.edu

**Ashley Stewart**
Research Associate II
Sequencing Core
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319
301-838-0200
FAX: 301-838-0229
astewart@tigr.org

**Fiona Stewart**
New England Biolabs, Inc.
32 Tozer Road
Beverly, MA 01915
978-927-5054
FAX: 978-921-1350
stewart@neb.com

**Marvin Stodolsky**
Molecular Biologist
Office of Biological and Environmental Research
US Department of Energy
19901 Germantown Road, ER-72
Germantown, MD 20874-1290
301-903-4475
FAX: 301-903-8521
marvin.stodolsky@science.doe.gov

**Neil Stoker**
Senior Lecturer
Infectious and Tropical Diseases
London School of Hygiene and
Tropical Medicine
Keppel Street
London, WCIE 7HT
UNITED KINGDOM
171-927-2425
FAX: 171-657-4314
n.stoker@lshtm.ac.uk

**Erik Stokhof**
Eurogentec
Parc Scientifique du Sart Tilman
Seraing, 4102
BELGIUM
32-4-365 51 03
FAX: 32-4-366-01-50
d.allaer@eurogentec.be

**Joey Storer**
Research Leader
Computing, Modeling, and Information
Sciences
The Dow Chemcial Company
1707 Building
Midland, MI 48674
517-636-7883
FAX: 517-636-5406
jstorer@dow.com

**Nancy Stover**
Product Manager
Genome Sequencing
QIAGEN, Inc.
28159 Avenue Stanford
Valencia, CA 91355
800-426-8157
FAX: 800-718-2056

**Richard Sucgang**
Postdoctoral Research Fellow
Biochemistry
Baylor College of Medicine
I Baylor Plaze, Room T328
Houston, TX 77030
713-798-5476
FAX: 713-796-9438
rsucgang@bcm.tmc.edu

**Alexander Szabo**
Affymetrix
3380 Central Expressway
Santa Clara, CA 95051
408-731-5086
FAX:
alex_szabo@affymetrix.com

**Waclaw Szybalski**
Professor of Oncology
Editor-in-Chief, Gene
McArdle Laboratory for Cancer Research
University of Wisconsin Medical School
1400 University Avenue
Madison, WI 53706-1599
608-262-1259
FAX: 608-262-2824
szybalsk@macc.wisc.edu

**Hideto Takami**
Research Associate
Deep-sea Microorganisms Research Group
Japan Marine Science and
Technology Center
2-15 Natsushima
Yokosuka, Kanagawa 237-0061
JAPAN
81-468-67-3895
FAX: 81-468-66-6364
takamih@jamstec.go.jp

**George Tarasenko**
Vice President of Development
GENEPRO, Inc.
451 Bishop Street, Suite B, NW
Atlanta, GA 30318
404-875-4130
FAX: 404-875-9698
george@tarasoftinc.com

**Tatiana Tatusova**
Visiting Scientist
National Center for Biotechnology
Information
National Library of Medicine
National Institutes of Health
8600 Rockville Pike
Bethesda, MD 20894
301-435-5756
FAX: 301-480-9241
tatiana@ncbi.nlm.nih.gov

**Herve Tettelin**
Collaborative Investigator
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319
301-838-3542
FAX: 301-838-0208
tettelin@tigr.org

**Jean-Francois Tomb**
Senior Research Associate
Life Sciences
DuPont
Experimental Station, E328-240
Wilmington, DE 19880-0328
302-695-7651
FAX: 302-695-1829
JEAN-FRANCOIS.TOMB@usa.dupont.com

**Bao Tran**
Research Assistant I
Sequencing Lab.
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-0200
FAX: 301-838-0208
btran@tigr.org

**Jane Ulmer**
Office Assistant II
Conferences, Education and Training
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-610-5964
FAX: 301-838-0229
jfulmer@tigr.org

**Jonathan Upton**
Research Associate I
Sequencing
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-0200
FAX: 301-838-0208
jupton@tigr.org

**David Ussery**
Associate Professor
Center for Biological Sequence Analysis
Institute of Biotechnology
Danish Technical University
Building 208
Lyngby, DK-2800
DENMARK
45 4525 2488
FAX: 45 4593 1585
Dave@CBS.dtu.dk

**Terry Utterback**
Team Leader Research Associate IV
Sequencing Core
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-0200
FAX: 301-838-0208
utterb@tigr.org

**Roy Vallie**
Vice President
Sales and Marketing
5 Prime-3 Prime, Inc.
5603 Arapahoe Avenue
Boulder, CO 80303
800-533-5703/303-440-3705
FAX: 303-440-0835
rvallie@5prime.com

**Jessica Vamathevan**
Research Associate II
Seqcore
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-0318
FAX: 301-838-0208
jvamat@tigr.org

**J. Craig Venter**
President and Chief Scientific Officer
Celera Genomics Corporation
45 West Gude Drive
Rockville, MD 20850
240-453-3502
FAX: 240-453-3650
JCVenter@celera.com

**Alain Viari**
Researcher
Biology
Centre National de la Recherche Scientifique
Institut Pasteur
Atelier de BioInformatique-Universite Paris 6
12, reu Cuvier
Paris, 75005
FRANCE
33-01-44-27-65-36
FAX: 33-01-44-27-63-12

**Alexander von Gabain**
CEO
Microbiology and Genetics
Vienna Biocenter
Dr. Bohrgasse 9
Vienna, 1030
AUSTRIA
43-1-4277-54601
FAX: 43-1-4277-9546
alex@gem.univie.ac.at

**Hartmut Voss**
Vice President
Department of Genomics
Lion Bioscience AG
Im Neuenheimer Feld 517
Heidelberg, 69120
GERMANY
49-622-140-3822
FAX: 49-622-140-3835
voss@lion-ag.de

**Joe Wang**
Associate Research Scientist
Bioinformatics
Pasteur Merieux Connaught
1755 Steeles Avenue W.
Toronto, Ontario M2R 3T4
CANADA
416-667-2700
FAX: 416-661-7960
jwang@ca.pmc-vacc.com

**Richard Warren**
Assistant Director
Genetic Technologies
SmithKline Beecham Pharmaceuticals
1250 Collegeville Road, UP 1340
Collegeville, PA 19426
610-917-7157
FAX: 610-917-7901
richard_warren-1@sbphrd.com

**Janice Weidman**
Research Associate III
Seqcore
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319
301-838-4032
FAX: 301-838-0208
jweidman@tigr.org

**Berthold Weinstein**
Deputy Associate Director
Biology and Biotecnology Laboratory
Lawrence Livermore National Laboratory
PO Box 808, L-452
Livermore, CA 94550
925-422-5352
FAX: 925-423-3110
bweinstein@llnl.gov

**Keith Weinstock**
Senior Scientist
Department of Genomics and Technology
Development
Genome Therapeutics Corporation
100 Beaver Street
Waltham, MA 02453-8443
617-398-2300
FAX: 617-893-9535
kgweinst@genomecorp.com

**Margret Werner-Washburne**
Program Director
Genetics
National Science Foundation
4201 Wilson Boulevard
Arlington, VA 22230
703-306-1439
FAX: 703-306-0355
mwernerw@nsf.gov.

**Owen White**
Deputy Director of Informatics
Informatics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-3534
FAX: 301-838-0208
owhite@tigr.org

**Simon Whiteside**
Progect Leader
R and D
Hybrigenics S.A.
Campus Pasteur
25 Rue Du Dr. Roux
Paris, 75015
FRANCE
33-14-061-3786
FAX: 33-14-061-3798
swhitesi@hybrigenics.fr

**Kerstin Williams**
Post-Doctoral Research Fellow
Medical Microbiology
St. Bartholomew's Hospital
West Smithfield
London, ECIA 7BE
UNITED KINGDOM
0171-601-8410
FAX: 0171-601-8409
kerstin.j.williams@mds.qmw.ac.uk

**Michael Wilson**
Graduate Student
Microbiology/Immunobiology
Stanford University
B239 Beckman Center, S.U.M.C.
300 Pasteur Lane
Stanford, CA 94305
650-723-7026
FAX: 650-723-1399
wilson@cmgm.stanford.edu

**Cindy Winder**
Publications Coordinator
Conferences, Education and Training
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-838-3515
FAX: 301-838-0229
cwinder@tigr.org

**Valerie Wood**
Computer Biologist
Pathogen Sequencing Unit
The Sanger Centre
Wellcome Trust Genome Campus
Hinxton, *Cambridge* CB10 1SA
UNITED KINGDOM
44-122-282-6244
FAX: 44-122-349-4919
val@sanger.ac.uk

**Todd Wood**
Postdoctoral Fellow
Biochemistry and Molecular Genetics
University of Virginia
Box 440, Jordan Hall
Charlottesville, VA 22908
804-924-2821
FAX: 804-924-5069
tcw3r@virginia.edu

**John Wooley**
Deputy Associate Director
Office of Biological
and Environmental Research
US Department of Energy
19901 Germantown Road
Germantown, MD 20874-5051
301-903-3153
FAX: 301-903-5051
john.wooley@science.doe.gov

**Brendan Wren**
Director of Research
Department of Medical Microbiology
St. Bartholomew's Hospital
West Smithfield
London, EC1A 7BE
UNITED KINGDOM
44-1716-018411
FAX: 44-1716-018409
b.w.wren@mds.qmw.ac.uk

**Thomas Wu**
Biochemistry Department
Stanford University School of Medicine
Beckman Center B400
Stanford, CA 94305-5976
650-723-5976
FAX: 650-725-6044
thomas.wu@stanford.edu

**Jennifer Yeager**
DNA Sequencing Application Scientist
Department of Genetic Analysis
PE Biosystems
850 Lincoln Centre Drive, Bldg. 400
Foster City, CA 94404
650-638-6481
FAX: 650-638-6514
yeagerjm@pebio.com

**Haruhiko Yokoi**
Senior Researcher
Applied Genomics
Tokyo Research Laboratories
Kyowa Hakko Kogyo Co., LTD.
3-6-6 Asahi-machi
Machida, Tokyo 194-8533
JAPAN
81-427-25-0781
FAX: 81-427-26-8330
hyokoi@kyowa.co.jp

**Robert Zagursky**
Senior Principal Researcher
Molecular Biology
Wyeth Lederic Vaccines
211 Bailey Road
West Henrietta, NY 14586
716-273-7619
FAX: 716-273-7515
zajurobm@war.wyeth.com

**Janice Zengel**
Senior Research Scientist
Department of Biological Sciences
University of Maryland at Baltimore
1000 Hilltop Circle
Baltimore, MD 21250
410-455-2876
FAX: 410-455-3875
zengel@umbc.edu

**Christina Zschoche**
Technical Training Specialist
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
301-610-5947
FAX: 301-838-0208
czschoche@tigr.org

# *1999 MICROBIAL ON-SITE REGISTRATION*
## *January 29, 1999 - February 1, 1999*

Linda S. Avey
Senior Sales
**PE Biosystems**
850 Lincoln Centre Drive
Foster City, CA 94404
800-248-0281
aveyls@pebio.com

David J. Bacon
Postdoctoral Fellow
Infectious Diseases
**Naval Medical Research Center**
Bethesda, MD 20818
301-295-1769
301-295-6171
bacond@nmripo.nmri.nnmc.navy.m.l

Robin Buell
Technical Support Manager
**MJ Research, Inc.**
149 Grove Street
Watertown, MA 02472
617-972-8064
617-924-2148
robinb@mjr.com

Vishnu Chaturvedi
Director
Mycology Laboratory
Division of Infectious Diseases
**Wadsworth Center**
120 New Scotland Avenue
Albany, NY 12201-0509
518-474-4177
518-486-7971
vishnu@wadsworth.org

Hui-Hsien Chou
Assistant Professor
Zoology/Genetics and Computer Science
**Iowa State University**
339 Science II
Ames, IA 50011
515-294-9246
515-294-8457
hhchou@iastate.edu

David M. Cook
Scientific Sales Representative
**MJ Research, Inc.**
149 Grove Street
Watertown, MA 02472
617-972-8143
617-924-8080
davecook@mjr.com

Tararaj Dharakul
Associate Professor
Immunology Department
**Mahidol University**
Faculty of Medicine Siriraj Hospital
2 Prannok Road
Bangkok, 10700 THAILAND
66-2-4197066
66-2-4181636
sitdr@mahidol.ac.th

Brad W. Goodner
Assistant Professor of Biology
**University of Richmond**
Science Center E-105
Richmond, VA 23173
804-289-8661
804-289-8233
bgoodner@richmond.edu

Patricia Guerry
Research Microbiologist
Enteric Diseases
**Naval Medical Research Center**
12300 Washington Avenue
Rockville, MD 20852
301-295-1514
301-295-6171
guerryp@nmri.nnmcpo.nmri.nnmc.navy.mi

Clyde Hutchison
Kenan Professor
Department of Microbiology and Immunology
**University of North Carolina at Chapel Hill**
C.B. 7290,
Mary Ellen Jones Building
Chapel Hill, NC 27599-7290
919-966-4503
919-962-8103
clyde@email.unc.edu

Joanna Klapacz
Ph.D. Candidate
Biochemistry
**Wayne State University**
471 Chemistry Boulevard
Detroit, MI 48202
iklapacz@chem.wayne.edu

# *1999 MICROBIAL ON-SITE REGISTRATION*
## *January 29, 1999 - February 1, 1999*

Przemek Kowal
Graduate Student
Department of Chemistry
**Wayne State University**
471 Chemistry
Detroit, MI 48202
313-577-7036
pkowal@chem.wayne.edu

Michael Megginson
Sales Representative
**General Scanning Inc.**
500 Arsenal Street
Watertown, MA 02472
P: 617-924-1010 ext. 168
F: 617-926-4093
mmegginson@genscan.com

Stacy O. Montgomery
Senior Scientist
Department of Molecular Biology
**PE Biosystems**
850 Lincoln Centre Drive
Foster City, CA 94404
650-638-6782
650-638-6333
montgoso@pebio.com

Anne Morgat
Scientist
Antibacterial Programme
**Rhone-Paulenc Rorer**
13 Quai Jules Guesde
Central Research, Magendie Building
Vitry-Sur-Seine
Cedex, 94403 FRANCE
33-1-55713829
33-1-55713960
anne.morgat@rp-rorer.fr

Gary B. Nunn
Senior Scientist
Department of Applied Markets
**PE Biosystems**
850 Lincoln Centre Drive
Foster City, CA 94404
650-638-5783
650-638-6333
nunngb@pebio.com

Richard T. Okinaka
Staff Scientist
DNA Damage and Repair Group
Life Sciences Division
**Los Alamos National Laboratory**
M-888
Los Alamos, NM 87545
505-667-2743
505-665-3024
okinaka@telomere.lanl.gov

Garrick J. Peters
Sequencing Sales Manager
Sales and Marketing
**Incyte Pharmaceuticals, Inc.**
3174 Porter Drive
Palo Alto, CA 94304
650-849-8835
650-845-4144
garrick@incyte.com

John Pugh
Director
Research and Development
**MICROBIOTEST, Inc.**
105B Carpenter Drive
Sterling, VA 20164
703-925-0100
703-925-9366
pugh@microbiotest.com

Mike Raum
Instrument Sales Specialist
**QIAGEN**
607 Tayman Drive
Annapolis, MD 21403
410-626-0186
410-626-0186

Margaret E. Riehman
Seq. Tech.
**MIDI Labs Inc.**
125 Sandy Drive
Newark, DE 19713
302-737-4297
302-737-7781
mriehman@midilabs.com

## *1999 MICROBIAL ON-SITE REGISTRATION*
## *January 29, 1999 - February 1, 1999*

Steven Salzberg
Director of Bioinformatics
**The Institute for Genomic Research**
9712 Medical Center Drive
Rockville, MD 20850
301-838-2537
301-838-0218
salzberg@tigr.org

Edgar H. Schreiber
Senior Scientist
Department of Microbial Detection and Identification Group
**PE Biosystems**
850 Lincoln Center Drive
Foster City, CA 94404
650-631-6176
650-638-6333
schreied@pebio.com

Sirirurg Songsivilai
Associate Professor
**Immunology Department**
**Siriraj Hospital**
Faculty of Medicine Siriraj Hospital
2 Prannok Road
Bangkok, 10700 THAILAND
66-2-4197065
66-2-4181636
sissv@mahidol.ac.th

Christine M. Szymanski
Postdoctoral Fellow
Infectious Diseases
**Naval Medical Research Institute**
12300 Washington Avenue
Rockville, MD 20851
301-295-1769
301-295-6171
szymanskic@nmripo.nmri.nnmc.navy.mil

Chris Upton
Assistant Professor
Department of Biochemistry and Microbiology
**University of Victoria**
Petch Building, Room 150
Victoria, BC, V8W 3P6 CANADA
250-721-6507
250-721-8855
cupton@uvic.ca

Peter Wiesner
Vice President
Business Development
**LION Bioscience AS**
Im Nenenheimer Feld 515
Heidelberg, D-69120 GERMANY
49 6221 403 8164
49 6221 403 8101
wiesner@lion-as.de

Jadwiga Wild
Scientist
Department of Oncology
**University of Wisconsin-Madison**
1400 University Avenue
Madison, WI 53706-1599
608-262-2047
608-262-2824
wild@oncology.wisc.edu

**UPDATED ADDRESS AND E:MAIL INFORMATION**

Louise M. Kelleher
Sales & Marketing Executive
**GENPAK Limited**
P: 44 1 273 704 470
F: 44 1 273 626 213
**e-mail: info@genpakdna.com.uk**

Robert Zagursky, Ph.D.
Senior Prinicpal Researcher
**Wyeth-Ayerst Pharmaceuticals**
**e-mail: zagurobm@war.wyeth.com**

Third Annual

# Conference on Microbial Genomes

**Westfields Marriott**
**Chantilly, VA**
**January 29 - February 1, 1999**

October 21, 1998

Jonathan Eisen
Department of Biological Sciences
Stanford University
Stanford, CA 94305-5020

Dear Dr. Eisen: Jonathan

On behalf of the The Institute for Genomic Research (TIGR), Foundation and my cochair, Andre Louis Goffeau of the Universite de Louvain, I would like to invite you to speak at the Third Annual Conference on Microbial Genomes to be held January 29 - February 1, 1999, at the Westfields Marriott in Chantilly, Virginia.

Topics to be covered during this year's conference include: recently completed genome sequencing projects; genome-based biology; functional genomics; molecular evolution; bioinformatics; and databases.

We would like you to speak during the **Genome Biology II** Plenary Session on **Sunday afternoon, January 31**, on the topic of "DNA Repair Gene in Archaea and Bacteria". Plenary talks will be 25 minutes, including a few minutes of discussion.

All speakers are requested to submit abstracts of their talk by November 16. These abstracts will be distributed as part of a program and abstract book. A preliminary agenda for the meeting is enclosed.

I would be happy to answer any programmatic questions you may have. You can reach me at 301-838-3500 or by email at cmfraser@tigr.org.

Bernie Lauro, conference director for TIGR, has enclosed complete logistical information with this letter. Any additional logistical questions should be directed to her at 301-838-3561 or by email at bglauro@tigr.org.

Please let us know of your availability and complete the speaker registration form included by **November 6, 1998**. You may fax this registration form to the conference office at 301-838-0229.

**The Institute for Genomic Research ■ 9712 Medical Center Drive ■ Rockville, MD 20850**
**301-610-5959 ■ FAX: 301-838-0229 ■ genome@tigr.org**

We look forward to an exciting conference and hope that you will be able to participate. Please check the conference website (http://www.tigr.org/conf) for updated agenda and conference information.

Sincerely,

Claire M. Fraser

Claire M. Fraser, Ph.D.
Conference Cochair

Enclosures

Third Annual

# Conference on Microbial Genomes

**Westfields Marriott**
**Chantilly, VA**

**January 29 - February 1, 1999**

**Speaker and Session Chair Registration Form**
(Please type or print neatly)

Name Jonathan A, Eisen

Degree(s) Ph.D. Stanford University, 1998

JobTitle Assistant Investigator, I

Department

Affiliation The Institute for Genomic Research

Address 7206 Honeywell Lane

City Bethesda State MD Zip Code 20814

Telephone (301) 652 5844 FAX ( )

Email Address jeisen@leland.stanford.edu

Presentation Title: A Phylogenomic Analysis of DNA Repair Processes and the benefits of a phylogenomic perspective.

**Audiovisual Requirements:**
☒ 35mm projector ☐ overhead projector

**Accommodations:** We will make your reservation at the Westfields Marriott. **Please do not contact the hotel directly to make or change your own reservation.**

I prefer: ☒ single room ☐ double room ☐ non-smoking

**I will arrive:** 1/29/99 (date) (estimated time of arrival)

**I will check out:** 2/1/99 (date) (estimated time of departure)

**Meals:** I require: ☐ Vegetarian meals ☐ Kosher Meals

**Please fax this form NO LATER THAN November 6, to: 301-838-0229 or mail to Third Annual Conference on Microbial Genomes at the address below.**

**The Institute for Genomic Research • 9712 Medical Center Drive • Rockville, MD 20850**
**301-610-5959 • FAX: 301-838-0229 • genome@tigr.org • http://www.tigr.org**

# THIRD ANNUAL CONFERENCE ON MICROBIAL GENOMES

**January 29 - February 1, 1999**
**Westfields Marriott**
**Chantilly, Virginia**

## WORKING AGENDA

### Friday, January 29, 1999

| | |
|---|---|
| 3:00pm | Registration opens |
| 6:00pm | Dinner |
| 8:00pm | Welcoming Reception |

### Saturday, January 30, 1999

| | |
|---|---|
| 7:00am | Breakfast |
| 8:30am-12:15pm | Plenary Session 1: Genome Biology I |
| 10:15am | Break |
| 12:30pm | Lunch |
| 2:00pm-5:30pm | Plenary Session 2: Genome Projects I |
| 6:00pm | Dinner |
| 7:30pm | Poster Session |

### Sunday, January 31, 1999

| | |
|---|---|
| 7:30am | Breakfast |
| 8:30am-12Noon | Plenary Session 3: Comparative Genomics |
| 10:00am | Break |
| Noon | Lunch |
| 2:00-5:00pm | Plenary Session 4: Genome Biology II |
| 6:00pm | Dinner |

### Monday, February 1, 1999

| | |
|---|---|
| 7:00am | Breakfast |
| 9:00am-11:00am | Plenary Session 5: Genome Analysis I |
| 11:00am | Discussion on the Future Funding for Microbial Genome Studies |
| 12:30pm | Meeting Adjourns |

Third Annual
## Conference on Microbial Genomes

**Westfields Marriott**
**Chantilly, VA**

**January 29 - February 1, 1999**

# Speakers Logistics Factsheet

The Institute for Genomic Research (TIGR), looks forward to your participation at the Third Annual Conference on Microbial Genomes. The following information is included to help you with your planning and to apprise you of reimbursement guidelines for the conference. If you have any questions, please do not hesitate to contact Bernie Lauro at 301-838-3561.

**ABSTRACTS:** As indicated in your speaker invitation letter, **all speakers are asked to submit abstracts by November 16, 1998.** Abstracts will be published in the Journal, *Microbial & Comparative Genomics* and distributed at the conference. Abstracts not received by the deadline will not be included in the journal, but will instead be copied and distributed as an addenda.

Abstract specifications are as follows:

- 5-1/2" wide by 7" long
- Helvetica 12 point
- Title in title case and bold
- Lead author's name in bold
- 1" from top and bottom
- 1.25" from left and right sides
- .5" from edge

You may submit your abstract by email (ASCII only to cwinder@tigr.org) or in Microsoft Word or WordPerfect (DOS or Macintosh) format, on a 3-1/2" disk. Label the disk with your name, computer type and program used to create the abstract. Send all abstracts (hard and disk copy) to:

Third Annual Conference on Microbial Genomics
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319

**AIRFARE:** Your roundtrip airfare **(coach class only)** to either Dulles International Airport, or Ronald Reagan National Airport, will be reimbursed, as follows, by the conference. We have arranged for all speakers to be ticketed through our authorized conference travel agency, Omega Travel Agency. Your ticket will be charged to our master account and sent directly to you. To make your travel reservations, please call Omega at 1-800-955-5949 or 301-330-9155 between 9:00am and 5:30pm Eastern time Monday through Friday, and refer to the TIGR account.

**Please Note:** If you book through a different travel agent, you will be reimbursed at the rate that Omega quotes as the lowest 14 day fare. If you use Omega and book your ticket more than 14 days out from your departure date, your ticket cost will not be

questioned, unless you choose an itinerary that is significantly higher than the lowest fare available. If you book less than 14 days out, or if you use a different travel agent, you will be reimbursed only the amount of the 14 - day fare quoted by Omega. Omega will not charge any tickets to the master account less than two weeks out from conference. TIGR is provided with quotes for all invited speakers, whether they choose to book with Omega or not. It is to your advantage to book with them and charge your ticket to our master account. Otherwise, you run the risk of non- reimbursable expenses.

**GROUND TRANSPORTATION:** Westfields Marriott provides complimentary shuttle service to and from Dulles International Airport. **Please make reservations with the Bell Captain at the Westfields Marriott at 800-635-5666 at least 48 hours in advance of your arrival.** The shuttle will meet you outside of baggage claim on the lower level at Curb 2H. For shuttle service from Ronald Reagan National Airport proceed to the Washington Flyers Super Shuttle booth in the terminal. Their shuttle is $19 one way inclusive of tax. For those conferees who plan to drive, ample free parking is available for all guests. **Taxis taken instead of the shuttle will not be reimbursed.**

You will also be reimbursed by the conference for transportation to and from your originating airport and for parking charges you incur. Automobile mileage is reimbursable at the rate of $0.315 per mile. Parking is reimbursable up to $10 per day, which should allow for parking in remote lots.

**HOTEL:** Your hotel room and tax will be covered by the master account for up to three nights, Friday, January 29 through Sunday, January 31, 1999. If you are traveling from Europe or Asia, you will also be covered the night of Monday, February 1. **You will be personally responsible for paying for extra nights and all incidental room charges.** Please let us know on the enclosed registration form what day you plan to arrive and depart. If you need to change your plans prior to the conference, please inform us so we can notify the hotel directly. **Please do not contact the hotel yourself.** If staff or family need to contact you during the meeting, they should call the Westfields Marriott at 800-635-5666.

**MEALS:** Most meals are provided by the conference. You will be provided meal tickets for:

| | |
|---|---|
| Friday, January 29 | dinner |
| Saturday, January 30 | breakfast, lunch, dinner |
| Sunday, January 31 | breakfast, lunch, dinner |
| Monday, February 1 | breakfast |

Meals eaten in lieu of those provided by the conference are not a reimbursable expense.. Meals eaten en route to or from the conference are not reimbursable.

REGISTRATION FORM: **Please fax this form by November 6, 1999** to 301-838-0229 or mail to the Third Annual Conference on Microbial Genomes at the letterhead address.

C Fraser

- 20+ published

A Goffeau

- yeast as a model system
- Molecular Physiology

  CHIPS

  2D ELECTROPHORESIS

  IN SILICO ANALYSIS

B. Dujon

- real challenge of post genomics is to characterize f(x) of uncharacterized genes

- 3 categories of genes
  - experimentally determined
  - sequence suggests f(x)
  - sequence does not suggest f(x)

- Evolutionary conservation and gene redundancy
  - 3000+ genes in yeast are in families ~75% of these have homologs
  - 2800+ genes are single gen ~½ of these have homologs

KNOW A LOT ABOUT WHAT IS COMMON TO A SPECIES BUT LITTLE ABOUT WHAT IS UNIQUE.

- 3700 genes have homologs in other spec - most have determined f(x)
- 2800 genes don't - most don't have experimentally determined f(x)

How study

Gene disruption

Gene cloning + overexpression

Gene tagging

Systematic transcript analysis

Gene family studies -

Essential genes (conditional overexpression) (13-15% of all genes)

Genes that are conserved are not more likely to be essential

Why gene families?

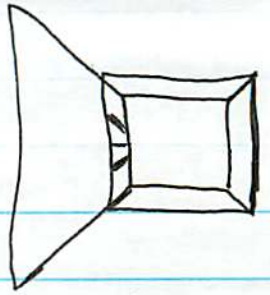
① Subtelomeric regions are copied from 1 chromosome to next

② ~~Signals~~ Single gene duplications

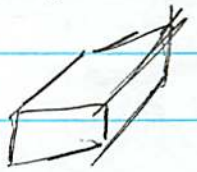
③ tandem gene duplications

④ chromosome duplications

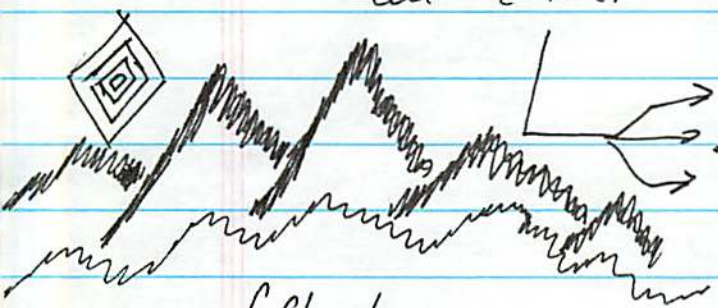
Yeast Proteome Database



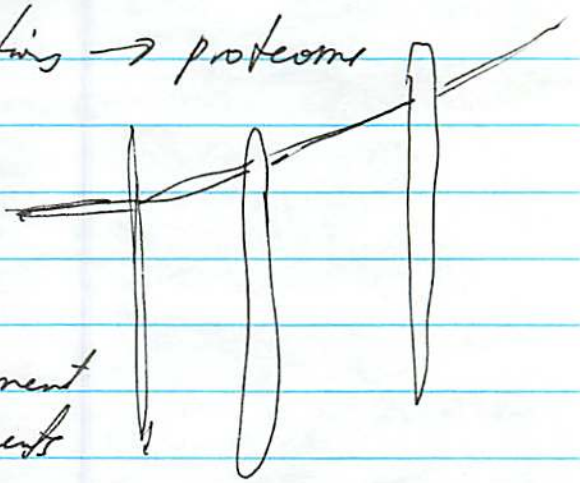
Proteome analysis



cell → extract → Proteome



→ subcellular fractions → proteome



G. Church

- Ecoli + Yeast regulatory networks
- Eco l. KO's ... need to do inframe replacement
- multiplex competitive growth experiments

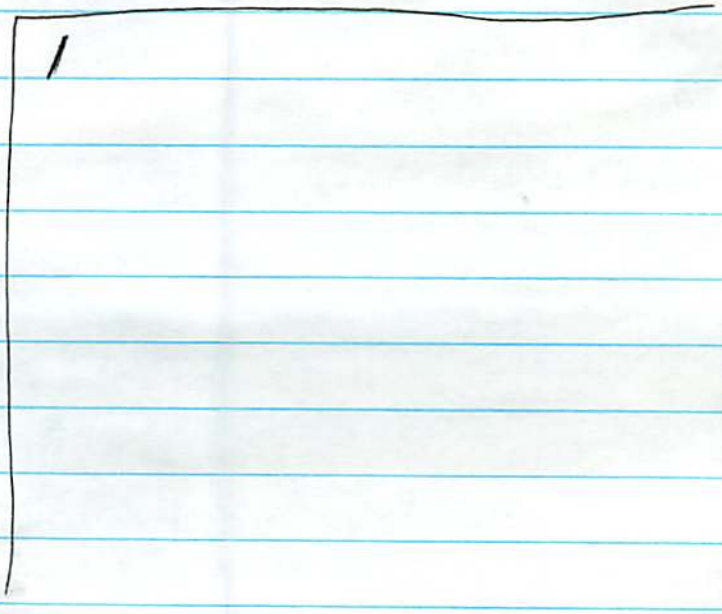
B. subtilis

| | |
|---|---|
| Total ORFs | 4000 |
| Experim. | 1000 |
| Predicted fcn | 1400 (more biochemical predictions than physiological) |
| Weak.. no predictn | 1600 |

Syst KO of D. radiodurans

In 90kb

44 fx predicted
59 fx detected
25 were predicted
34 not predicted



C. jejuni

Epsilon?

- 1st cultured in 1970's
- causes majority of diarrhea cases
- why the increase?
- different presentation in developed vs developing world
- also leading antecedent to Guillen Barre syndrome (probably due to molecular mimickry)

What want to learn from genom

polysacs + GB
cause diarrhea
adherence
environment
host pathogen interaction

① shotgun library
② checked random
③ 8-10x
④ .

- 1731 CDS
- GC = 30.62%
- no plasmids

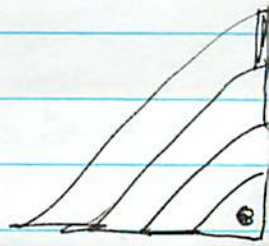
Features

① 25 coding regions w/ frameshift mutations (SSM)
CONTINGENCY many polymorphism
② polysaccharides? -- 3 neuB homologs
③ 3 neuB's are contingency genes
④ no obvious virulence determinants

H. pylori vs C. jejuni

- HP Has CagA locus
- Ureas operon
- $Ni^{-}$ transport
- HpaA not in CJ
- VapA not in CJ

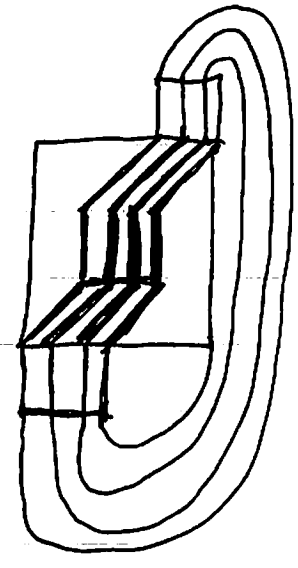
} What's in HP not CJ

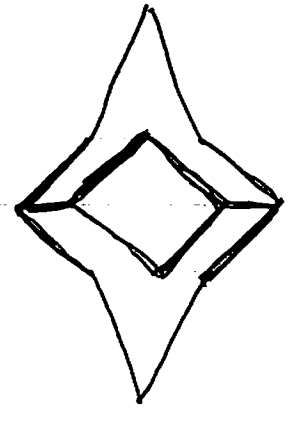


Gwen
- DSBR + repeats
- origin of mama, baby, papa...
- if DNA is held in place

RNA editing
why pseudogenes kept
substitutions

FRAMESHIFT?
RNA EDITING?
→ TRANSPOSASE
- but suggs. not b/c other species don't have frameshifts
- How MANY HAVE HOMOLOGS IN GENOME.





## Genomes to T-cells

### Vaccines

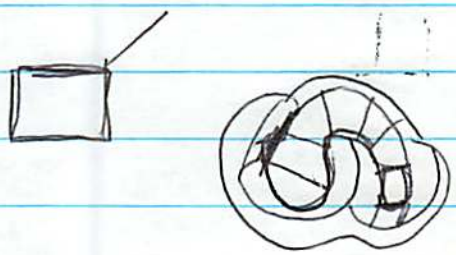
- - how to predict epitopes from genome sequence
- - can ID motifs that make up epitopes by sequencing peptides that bind MHC
- - used motifs to analyze sequences ... found some regions w/ more than 1 MHC presentation motif

You might expect these to be underrepresented in genomes.

- these were not perfect predictors
- - so used a MATRIX for the motifs instead

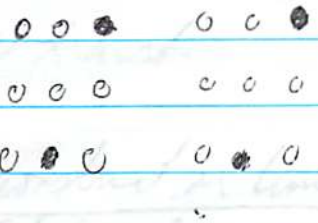
Are these conserved both species.

### Mtub



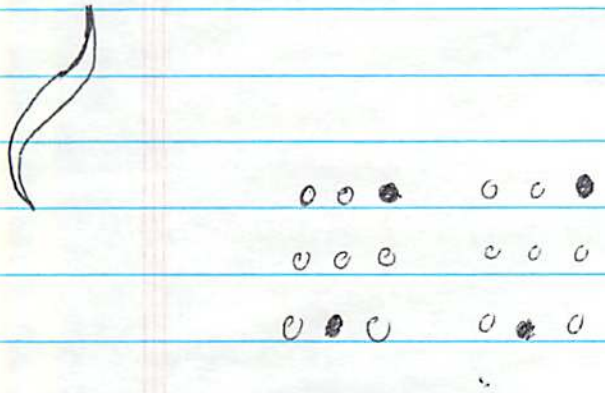
Monica Riley - AKA SLOW, SLOWER, SLOWEST

MULTIPLE FUNCTIONS ATTRIBUTIONS



Monica Riley - AKA SLOW, SLOWER, SLOWEST

MULTIPLE FUNCTIONS ATTRIBUTIONS



H. Phillipe

- many trees contradict rRNA tree

Causes of conflicting trees
- Tree Reconstruction Limits
- Lateral transfers

Tree Reconstruction Limits
GC bias - (can try to correct w/ non-stat models)
Among site variation

Covarion Model
- diff. rates in diff parts of tree

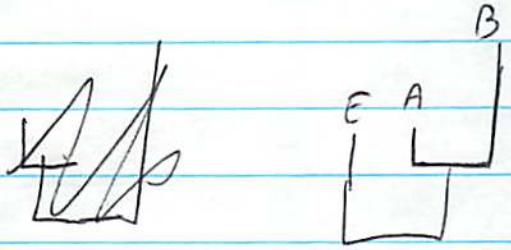
If don't use ~~contents~~ good model get artifacts

Artifacts
① long/short branch attraction
② Outgroups = long branch

Two examples
① root of universal tree
- branch lengths of diff genes v. different
so there is no universal clock
- if root of tree is in MIDDLE of long branches --> usually a long branch artifact

Trees using only slowly evolving positions

Suggests that bacteria have longer branches and so will attract to outgroup




If this is correct then you will get wrong tree a lot of time

Simulation

if outgroup is close to ingroup then get clustering of long branches within ingroup

if outgroup is far from ingroup - get long ingroups at base

Conflicting trees

if rates differ among genes (in which species are long branches) then you will get long branch attraction causing different trees

Asymmetric bases of trees are largely due to long branch attraction. Symmetric tops are more likely to be correct.

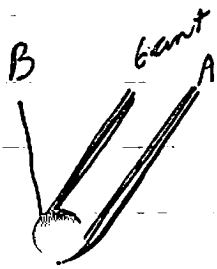
G⁺ G⁻
R Y

## Gupta

- must include all biological information (e.g. fossils)
- ancestor should be prokaryotic b/c of fossils

## Conserved INDELS as phylogenetic markers

- use flanking regions to make sure alignment is good




### HSP70

all G⁺ and A share insert
- many other proteins

### Brown + Doolittle

>2/3 of proteins A + G⁺ show close relationship

### Say TG is G⁺

### What mean:

#### Membranes

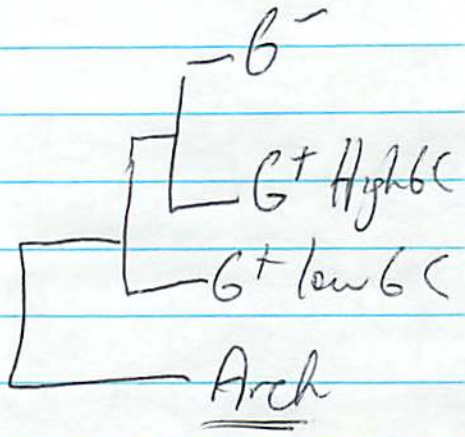
A + G⁺ → one membrane, cell wall not reqd
G⁻ → two membranes

cells w/ one membrane = ancestral

Other inserts

A + low GC $G^+$ = insert

high $GC^+$, $G^-$ = no insert




D/T is transition betn. $G^+$ + $G^-$

## $G^+$ + Archaea

- many trees show A as non-monophyletic
- so how resolve this with genes that show A as monophyletic

↓

Transfer

↙ ↘ ? alternatives

~~A evolved for $G^+$ cell~~

A lat trans betn A+$G^+$ — A+$G^+$ share common ancestor

How distinguish ... categories of genes ... Ab act one interrupts transfer + maybe $G^+$ which make Ab

antibiotic

Suggests ... $G^-$ evolved to become resistant to Ab

Plasmid

# of DNA domains

fit shows only those consistent w/ model.

mutagenesis correlated w/ supercoils

## Phil Heseman

- major funding goes to plants
- Crop genome initiative
- some for photosyn bacteria
- some supp. AW for reg. grants to fund genome
- Life in extreme environments ALSO FROM NASA
  - can receive prop. for sequencing

- New program

BIOCOMPLEXITY - environmental background
- How organisms interact w/ environment
- Next year - 50 million
- several new post-doctoral programs - INFORMATICS
- Microbiology POST-DOC

## MARK FRASER -DOE

Microbial Genome Program - Phylogenetic

Criteria

- energy/envir relevance
- ORF obtainable
- genome size
- genetically manipulable
- nonpathogen
- scientifically interesting

- biosensors
- biomonitors
- fuel production
- disease/drought resistance
- bioprocessing
- photosynthesis

uncultivable

Climate change

Carbon cycle

$CH_3$ + Hz producers

Energy Biosciences Division

Fermentation biology
Extremophiles
Plant biology

Michael Cotterlich - pathogens

- NAIAD -

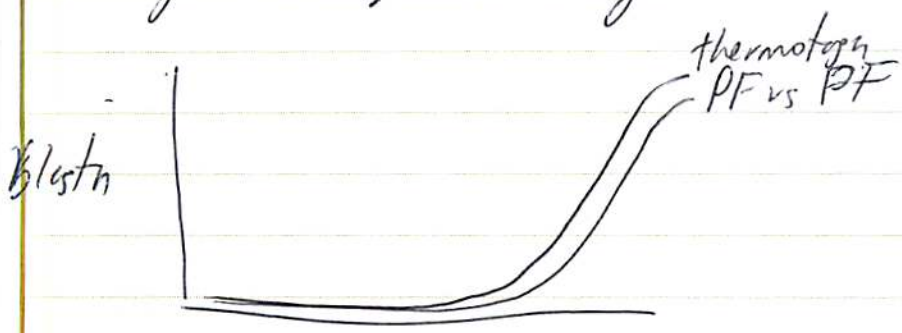
Web access
sprot 36

Don't want ancestral - want true weighted average

java plug in

Pyrococcus

- growth T° ≠ GC of genome




Codon bias

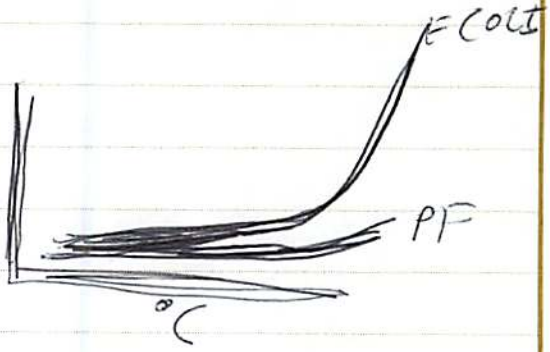
genetic code conserved

GC skewed underabundant

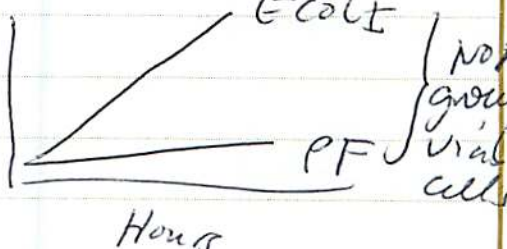
~50% hypothetical

>100°

- dsr ss breaks accumulate
  but looks like it protects some




- V. radioresistant
- 4 hrs ... 1st chromosome assembly 100°C
- 2 copies of chromosome / cell




Introns

RR

MJ vs PF

- no linkage

PF vs PH

- some linkage

R. Alm - HP vs HP99

H. pylori - several different libraries
- many gastric diseases assoc. w/ HP
- conservative annotation

- several diff. libraries to incr. depth
- 26695 isolated in humans... 2yrs of passage
- 99 - recent isolate

- why study?
  1. lab vs. clinical
  2. genomic diversity
  3. are all HP strains equal

| | HP99 | HP |
|---|---|---|
| Class f(x) | 874 | 895 |
| conserved, unknown | 275 | 294 |
| HP specific | 346 | 367 |
| Strain specific | 89 | 117 |

- Length : - v. similar
- % coding - v. similar
- orfs - differs by ~50
- GC skew - some differences
- "Origin" - none ID'd
- two unlinked rRNA both
- large # of RE's both but some separate
- lack some DNA pol III subunits...

Of the strain specific ORFs (of predictable f(x)) most are RE's

Genome Diversity

old studies suggest lots of diversity

- 17 locations where 99 hpn strain specific genes, HP2635 hpn strain-specific gene

Gene order

lack of sy. shuffling indicates low evolut divergence

Plasticity zone

- low GC
- continuous in J99 - 30Kb (26385 - has inversion)
- most genes are HP specific
- most of the strain specific genes are hpn too

- mono + dinucleotid repeat variation
  - 26 genes
  - 62% in same on/off position
  - monomers - don't at (22%)

Rob