Eleventh International
Genome Sequencing and Analysis Conference
Fontainebleau Hilton, Miami Beach, FL
September 18-21, 1999

April 7, 1999

Dr. Jonathan Eisen
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850

Dear Dr. ~~Eisen~~ Jonathan:

On behalf of The Institute for Genomic Research and the other co-chairs, Arthur Caplan and Gerald Rubin, I would like to invite you to speak at the 11th International Genome Sequencing and Analysis Conference to be held September 18-21, 1999, at the Fontainebleau Hilton in Miami Beach, Florida.

We would like you to speak during the **Model Organisms II Plenary Session scheduled for Monday morning, September 20.** Plenary talks will be 25 minutes, followed by 5 minutes of discussion. Should the agenda change, we will notify you. We appreciate, in advance, your flexibility. **This year's emphasis will be on Agriculture, Model Organisms I, Model Organisms II, Mammalian Genome Projects and Hot Button Topics for the Next Millenium.**

All speakers are requested to submit abstracts of their talk by Monday, May 31, 1999. These abstracts will be published as part of the program and abstract book, which will be distributed at the meeting. Abstracts not received by July 23, will not be included in the journal. A preliminary agenda is enclosed.

Bernie Lauro, Vice President of Conferences, Training and Education, would be happy to answer any programmatic questions you may have. You can reach her at 301-838-3561 or by e-mail at "bglauro@tigr.org".

Please contact Jane Ulmer at 301-610-5964 or by e-mail at "jfulmer@tigr.org" by Friday, April 23, to confirm your availability. Please complete the speaker registration form included herein and fax it to the conference office at 301-838-0229. We have enclosed complete logistical information with this letter. Any questions pertaining to logistical matters should be directed to Jane Ulmer at the above number.

We look forward to an exciting conference and hope that you will be able to take part in this memorable occasion.

Sincerely,

J. Craig Venter, Ph.D.
President and Chief Scientific Officer
Celera Genomics Corporation

Enclosures

11th INTERNATIONAL
GENOME SEQUENCING AND ANALYSIS CONFERENCE

Fontainebleau Hilton, Miami Beach, Florida
September 18-21, 1999

SPEAKERS LOGISTICS FACTSHEET

The TIGR Conference Department looks forward to your participation at the 11th International Genome Sequencing and Analysis Conference. The following information is provided 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 Jane Ulmer at 301-610-5964.

ABSTRACTS: As indicated in your speaker invitation letter, **all speakers are asked to submit abstracts by May 31, 1999.** Abstracts will be included in the Liebert journal, *Comparative & Microbial Genomics* that will be distributed at the conference. If abstracts are not received at the conference office by **July 23**, they will not be included in the journal.

Abstract specifications are as follows:

- 3-1/4" wide by 4-1/2" long.
- Heading is 11 point Helvetica; Text is 9 point Times Roman.
- Longer abstracts will be returned to authors for editing.
- Title in initial cap and lower case, lead author in bold followed by co-authors (all authors use first name, middle initial, last name) and affiliations (company, city, state/country) on next line.

You may submit your abstract as a WordPerfect or Microsoft Word (DOS) document on a 3-1/2" floppy disk. Label the disk with the name of the presenting author and indicate the format and originating program (i.e., Word).

You may submit your abstract by email 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. We must have both hard copy (by fax or by mail) and an electronic version (by email or by website or on disk) of all abstracts. Please mail your materials to:

11th International Genome Sequencing and Analysis Conference
TIGR Conferences, Education and Training Department
9712 Medical Center Drive
Rockville, MD 20850-3319

Airfare: Your roundtrip airfare **(coach class only)** to either Miami International Airport, or Ft. Lauderdale/Hollywood International Airport, Florida 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

11th International
Genome Sequencing and Analysis Conference

PRELIMINARY AGENDA

Saturday, September 18, 1999

Noon–9:00 pm	Registration
	Dinner on your own
7:00 – 10:15 pm	**Plenary Session 1: Agriculture**
10:15 pm	Welcome Reception

Sunday, September 19, 1999

7:00 am	Registration re-opens
7:00–8:15 am	Continental Breakfast
8:30 am–Noon	**Plenary Session 2: Model Organisms I**
Noon–2:00 pm	Lunch (Exhibit Hall and French Hall)
1:00–3:00 pm	Poster and Electronic Poster Session I (French Hall)
3:00–5:00 pm	Concurrent Sessions (Fontainebleau Ballroom) Genomics I, Technology I, Bioinformatics I, Agriculture I
	Free Night

Monday, September 20, 1999

7:00 am	Registration re-opens
7:00–8:15 am	Continental Breakfast
8:30 am–Noon	**Plenary Session 3: Model Organisms II**
Noon–2:00 pm	Lunch (Exhibit Hall and French Hall)
1:00–3:00 pm	Posters and Electronic Posters (French Hall)
3:00–5:00 pm	Concurrent Sessions (Fontainebleau Ballroom) Genomics II, Technology II, Bioinformatics II, Agriculture II
5:30-7:00 pm	Dinner (Grand Ballroom)
7:00–10:45 pm	**Plenary Session 4: Mammalian Genome Project**

Tuesday, September 21, 1999

7:00–8:15 am	Continental Breakfast
8:30 am–12:30 pm	**Plenary Session 5: Hot Button Topics for the Next Millennium**
12:00–2:00 pm	Lunch (Exhibit Hall and French Hall)
1:00–3:00 pm	Posters and Electronic Posters (French Hall)
3:00–5:00 pm	Concurrent Sessions (Fontainebleau Ballroom) Genomics III, Technology III, Bioinformatics III, Agriculture III
7:00 pm–Midnight	Beach Blast

4/07/99

11th INTERNATIONAL GENOME SEQUENCING AND ANALYSIS CONFERENCE

Registration Form

Please type or print neatly.

Name Jonathan A. Eisen, Ph.D.

First M.I. Last Degree(s)

Job Title Assistant Investigator

Department Microbial Genomics

Affiliation T.I.G.R.

Address 9712 Medical Center Drive
Rockville, MD 20850

City State ZIP Code

Telephone (301) 838-3507 FAX (301) 838 0208

Email

j	e	i	s	e	n	@	t	i	g	r	.	o	r	g				

@

Abstract title (only lead author should submit title)

Phylogenomic Analysis of Complete Genomes from Closely Related Species Sets

	Postmarked on or by 5/31/99	Postmarked 6/1 - 8/17/99	Postmarked 8/18 - 9/10 then On-Site*
Individual Registration	□ $675	□ $900	□ $1,125
(this fee includes meals provided by the conference)			
Predoctoral Student Registration	□ $375	□ $450	□ $525
(this fee includes meals provided by the conference)			

Registration must be accompanied by a letter from your thesis advisor

***NOTE: No registrations will be accepted from September 10 until on-site registration opens on Saturday, September 18 at noon.**

□ Please check here for vegetarian meals.

Method of payment (Please Note--We Do Not Accept American Express)

Check one: □ VISA □ MasterCard □ Check (Payable to: The Institute for Genomic Research)

Credit Card Number □□□□-□□□□-□□□□-□□□□

Expiration Date □□/□□

Card Holder's Name (please print clearly)________________________________

Card Holder's Signature____
I agree to pay above total charges according to card issuer agreement.

I have enclosed the following:
□ Registration form
□ Payment (check or credit card number or details of wire transaction)
□ Fax of abstract (if you intend to submit an abstract)

I understand that my registration cannot and will not be processed without proper payment.
Please initial _________.

How did you hear about us?
□ Web
□ Colleague/Previous attendee
□ Magazine ad

Send your registration form, abstract and payment to:

11th International Genome Sequencing and Analysis Conference
Department of Conferences, Education and Training
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850-3319

PHONE: 301-610-5959
FAX: 301-838-0229
seqconf@tigr.org

Last modified at Monday, April 19 1999 11:01

CONFERENCES, EDUCATION AND TRAINING

11th International Genome Sequencing and Analysis Conference

Fontainebleau Hilton
Miami Beach, Florida

September 18 - 21, 1999

Abstracts:

Abstracts will be published in the journal, Microbial & Comparative Genomics, which will be distributed at the conference. Because of our publishing deadlines, all abstracts to be included must be received no later than July 23, 1999. THERE ARE NO EXCEPTIONS. Poster presentations, electronic poster presentations and concurrent session speakers will be chosen from abstracts received and approved. Each participant is encouraged to submit one abstract. **No registrant may submit more than one abstract.** Once you have submitted your registration form, payment and hard copy of your abstract (for formatting purposes, either by fax or by mail) you will be notified how to submit your abstract by email, by web, or on disk.

To submit an abstract when you register:
If you have included a copy of your abstract with your faxed or mailed registration form, the access code and email address and instructions for mailing a disk will be automatically provided to you in the registration confirmation that will be mailed to you as soon as your registration and payment have been processed.

To submit an abstract after you have registered:
If, after you have registered, you decide to submit an abstract, you will need to fax a copy of your abstract to 301-838-0229. You will then be contacted and given the access code, email address, and instructions for mailing a disk.

Your abstract must conform to the following format (see sample below):

A Curated Set of Conserved Hypothetical Membrane Proteins from Completely Sequenced Microbial Genomes

Karen A. Ketchum, Gennie I. Fermin, Valentina DiFrancesco, Delwood Richardson, Owen White, Anthony R. Kerlavage and J. Craig Venter.
The Institute for Genomic Research, Rockville, MD.

Hydropathy analysis of coding regions identified in completely sequenced genomes indicates that 33% of the proteome resides in cellular membranes. Many of these proteins (65%), annotated by sequence comparisons, perform essential biological functions like solute transport, signal transduction, energy metabolism, or synthesis and stabilization of the cell envelope. However, the physiological roles of the remaining membrane components are unknown.

We have searched the 7 complete microbial genomes sequenced at The Institute for Genomic Research (TIGR) to identify novel membrane proteins. Gene products have been grouped into protein families and we hypothesize that those sequences found in several organisms, classified as conserved hypothetical proteins, are likely to have central cellular functions. To help direct protein structure discovery and functional characterization of these gene products, a database is being constructed to display multiple sequence alignments, secondary structure alignments, and the predicted topology of each transmembrane protein family. This database will be a visual tool for researchers who are interested in identifying essential amino acid residues and signature motifs characteristic of each class. In addition, these data will facilitate gene annotation as new genome sequences are generated.

Supported by National Science Foundation grant MCB-9723579 to KAK and Department of Energy grant DE-FC02-95ER61962 to JCV

Your abstract must conform to the following format:

- 3-1/4" wide by 4-1/2" long.
- Heading is 11 point Helvetica; Text is 9 point Times Roman.
- Longer abstracts will be returned to authors for editing.
- Title in initial cap and lower case, lead author in bold followed by co-authors (all authors use first name, middle initial, last name) and affiliations (company, city, state/country) on next line.

You may submit your abstract as a WordPerfect or Microsoft Word (Windows or Macintosh) document on a 3-1/2" disk. Label the disk with the name of the presenting author and indicate the format and originating program (i.e., Mac/Word, Win/WordPerfect, etc.).

You will receive an acknowledgment of receipt of your abstract by email or mail once your registration form and fee have been processed. If you do not receive an acknowledgment by July 26, 1999, contact the conference office to confirm that your abstract has reached the conference office.

To summerize:

- No abstract will be accepted or reviewed without paid registration
- A faxed or mailed copy of your abstract must be submitted to TIGR for formatting purposes
- No abstract will be considered if postmarked later than July 23, 1999

Click here to enter your Access Code for abstract submission

Back to G.S.A.C. main page

Last modified at Wednesday, April 21 1999 02:24

11TH International
Genome Sequencing and Analysis Conference
Fontainebleau Hilton Hotel, Miami Beach, FL
September 18 – 21, 1999

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

Name Jonathan A. Eisen

Degree(s) Ph.D.

JobTitle Assistant Investigator

Department Microbial Genomics

Affiliation The Institute for Genomic Research

Address 9712 Medical Center Drive

City Rockville State MD Zip Code 20850

Telephone (301) 838 3507 FAX (301) 838-0208

Email Address: jeisen@tigr.org

Presentation Title: Phylogenomic analysis of complete genomes from closely related species sets: insights in mutation, recombination, and gene transfer.

Audiovisual Requirements:
☒ 35mm projector ☐ overhead projector ☐ LCD ☐ other

Accommodations: We will make your reservation at the Fontainebleau Hilton, Miami Beach.
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: 9/18/99 ~6pm
date estimated time of arrival

I will check out: 9/22/99 ~2pm
date estimated time of departure

Meals: I request: ☐ Vegetarian meals

Please fax this form NO LATER THAN April 23 to: 301-838-0229 or mail to 11th International Genome Sequencing and Analysis Conference at the address below.

TIGR Conference Department ■ 9712 Medical Center Drive ■ Rockville, MD 20850
301-610-5964 ■ FAX: 301-838-0229 ■ seqconf@tigr.org ■ http://www.tigr.org/conf

make your travel reservations, please call **Omega** at **1-800-955-5959 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 the 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.

International travelers are requested to use a U.S. carrier.

Ground transportation: Transportation from either airport to the hotel and return is a reimbursable item. **SUPER SHUTTLE** from Miami International Airport can be reached at **1-800-874-8885** or **305-871-2000, FAX 305-871-8475**. Reservations are required. **AIRPORT EXPRESS** from Ft. Lauderdale/Hollywood International Airport can be reached at **1-800-244-8252** or **954-561-8888, FAX 954-565-7054**. If you rent a car, use a taxi or some other means of transport, you will be reimbursed only up to the equivalent of the authorized carrier expense (**$22 roundtrip from Miami International Airport and $24 roundtrip from Ft. Lauderdale/Hollywood International Airport).** You will be responsible for paying the difference.

You will also be reimbursed by the conference for transportation to and from your originating airport and for parking charges (up to 6 days @ **$10.00** /day) you incur. Automobile mileage is reimbursable at the rate of **$.31**/mile.

Hotel: Your hotel room and tax will be covered by the master account for up to 4 nights, Saturday, September 18 through Tuesday, September 21. If you are travelling from Europe or Asia, you will also be covered for the night of Friday, September 17. **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 directly.** If staff or family need to contact you during the meeting, they should call the Fontainebleau Hilton at 1-800-221-2424 or 305-538-2000 (Fax 305-673-5351). You will not receive a confirmation from the hotel because you are registered under TIGR's room block.

MEALS: Most meals are provided by the conference. You will be provided meal tickets for:

Sunday, September 19	Breakfast, Lunch
Monday, September 20	Breakfast, Lunch, Dinner
Tuesday, September 21	Breakfast, Lunch, Dinner

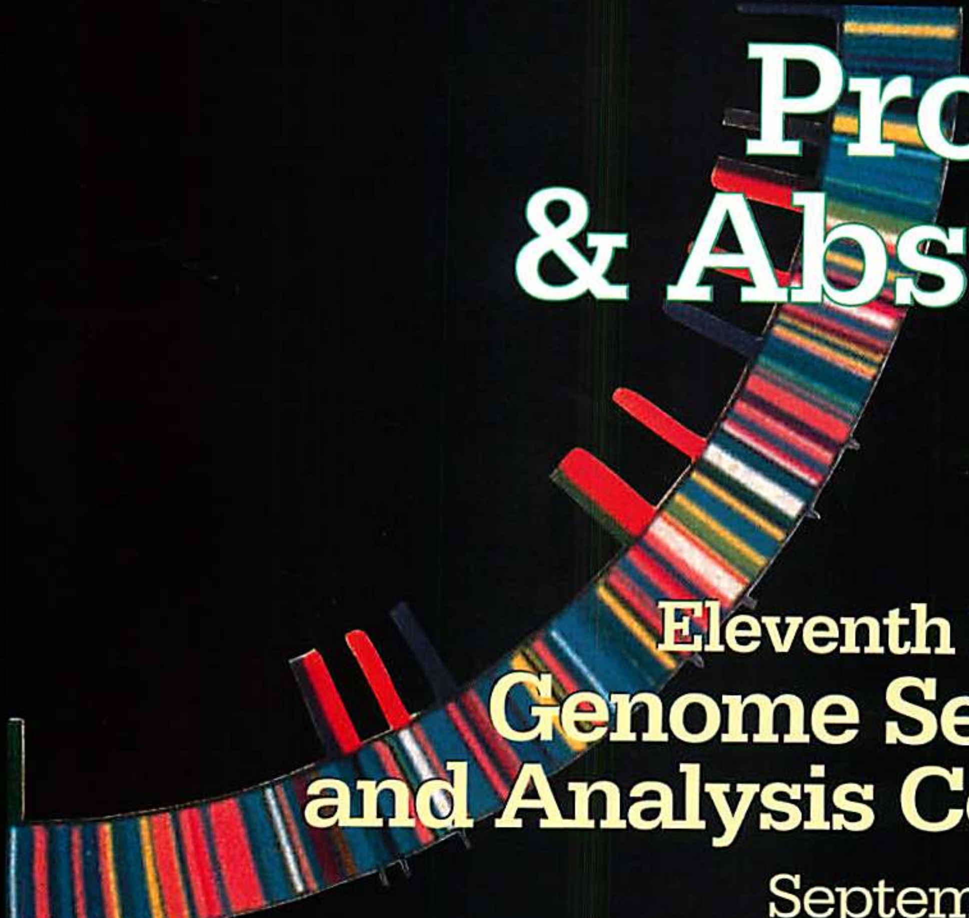
Meals eaten in lieu of those provided by the conference are not a reimbursable expense, with the exception of dinner on Saturday the 18th and dinner on Sunday the 19th, up to $20.00 per meal and does not include any alcohol.

REGISTRATION FORM: Please complete and return the enclosed registration form by **April 23, 1999** to the conference office.

Listed in
Index Medicus and MEDLINE

Volume 4, Number 2, 1999
ISSN 1090-6592

Microbial & Comparative Genomics



Program & Abstracts

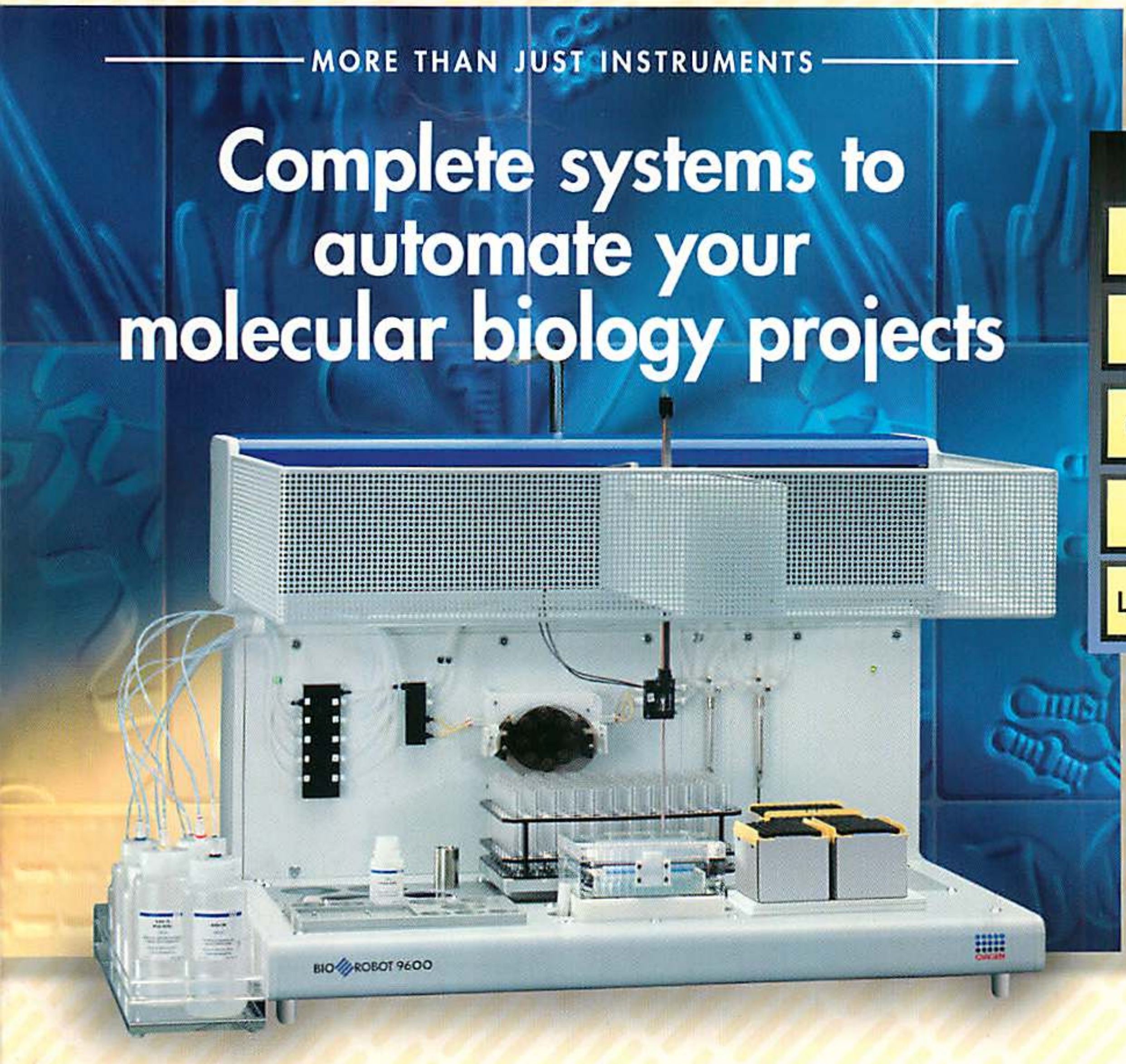
Eleventh International Genome Sequencing and Analysis Conference

September 18 - 21, 1999

Mary Ann Liebert, Inc. publishers

MORE THAN JUST INSTRUMENTS

Complete systems to automate your molecular biology projects



PROJECTS

- Sequencing
- PCR
- Genomic DNA
- RNA
- Liquid handling

QIAGEN® BioRobot™ Systems

Successful automation of molecular biology projects needs more than just an instrument. Our BioRobot™ Systems go further by integrating proven QIAGEN® chemistries, ready-to-run protocols, and dependable instrumentation — for consistent, reliable results.

Versatile and dependable — The BioRobot Systems rapidly purify nucleic acid templates and PCR products, reliably set up enzymatic reactions (PCR, sequencing, and restriction digests), and simplify other liquid-handling procedures such as agarose-gel loading and sample rearray. New protocols and hardware components can easily be added as your needs change.

Complete system support — The knowledgeable QIAGEN support team is always there to help you with all aspects of the system — chemistry, applications, and instrumentation — for fast startup and continued success.

Complete automated systems — from a company that understands molecular biology

Contact QIAGEN today for more information.

QIAGEN BioRobot Systems are not available in all countries; please inquire.
The PCR process is covered by U.S. Patents 4,683,195 and 4,683,202 and foreign equivalents owned by Hoffmann-La Roche AG.

www.qiagen.com

QIAGEN:	**Germany**	**USA**	**Australia**	**Canada**	**France**	**Japan**	**Switzerland**	**UK**
Tel.	02103-892-240	800-426-8157	03-9489-3666	800-572-9613	01-60-920-930	03-5805-7261	061-319-30-31	01293-422-999
Fax	02103-892-255	800-718-2056	03-9489-3888	800-713-5951	01-60-920-925	03-5805-7263	061-319-30-33	01293-422-922

DISTRIBUTORS: Austria/Hungary/Slovenia R. u. P. MARGARITELLA Austria (01) 889 18 19 **Belgium/Luxemburg** Westburg b.v. 0800-1-9815 **Brazil** Labtrade do Brazil (11) 543 1455 or 0800 55 1321 **Central & South America** Labtrade Inc. USA (305) 828-3818 **China** Gene Company Limited (852)2896-6283 **Czech Republic** BIO-CONSULT spol. s.r.o. (02)4447 1239 **Denmark** KEBO Lab A/S 43 86 87 88 **Egypt** Clinilab 525 7212 **Finland** KEBO Lab Oy (09) 804 551 **Greece** BioAnalytica S.A. (01)-640 03 18 **India** Genetix (011)-542 1714 or (011)-515 9346 **Israel** BioLab Ltd. 02. 584 1111 **Italy** GENENCO (M-Medical srl) (055) 500 1871 **Korea** LRS Laboratories, Inc (02) 924 86 97 **Malaysia** Research Biolabs Sdn. Bhd. (03)-7312099 **Mexico** Quimica Valaner S.A. de C.V. (5) 525 57 25 **The Netherlands** Westburg b.v. (033)-4950094 **New Zealand** Biolab Scientific Ltd. (09) 980 6700 or 0800 933 966 **Norway** KEBO Lab AS 22 90 00 00 **Poland** Syngen Biotech (071) 351 41 06 or 0601 70 60 07 **Portugal** IZASA PORTUGAL, LDA (1)-751 6000 **Singapore** Research Biolabs Pte Ltd. 445 7927 **Slovak Republic** BIO-CONSULT Slovakia s.r.o. (07) 50221 336 **South Africa** Southern Cross Biotechnology (Pty) Ltd (021) 671 5166 **Spain** IZASA, S.A. (93) 902.20.30.90 **Sweden** KEBO Lab (08) 621 34 00 **Taiwan** TAIGEN Bioscience Corporation (02) 2880 2913 **Thailand** Theera Trading Co. Ltd. (02) 412-5672 **In other countries contact:** QIAGEN GmbH, Germany

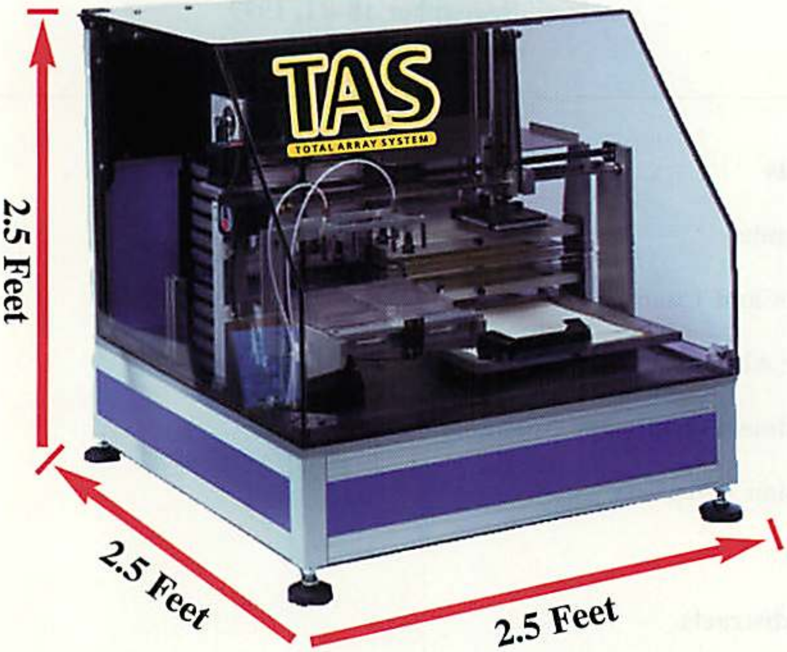
Microbial & Comparative Genomics

VOLUME 4 NUMBER 2 1999

Program and Abstracts
ELEVENTH INTERNATIONAL GENOME SEQUENCING
AND ANALYSIS CONFERENCE
September 18–21, 1999
Miami Beach, Florida

Instructions for Authors can be found at the back of the issue.

Mary Ann Liebert, Inc. publishers
www.liebertpub.com



TAS
TOTAL ARRAY SYSTEM
2.5 Feet
2.5 Feet
2.5 Feet

11th International Genome Sequencing and Analysis Conference

Preliminary Agenda
(as of 8/16/99)*

Saturday, September 18

Noon	**Electronic Poster Set-up** (Champagne Room)
Noon-7:00 pm	**Internet Café** sponsored by Compaq (Brittany Room)
Noon–9:00 pm	**Registration** (Grand Ballroom)
Dinner on your own	
7:00– 9:45 pm	**Plenary Session I: Agriculture (Grand Ballroom)**
	J. Craig Venter – Celera Genomics and The Institute for Genomic Research *Welcoming Comments and Introductions*
	Chair: **Chris Somerville** – Carnegie Institution of Washington *Introduction Overview*
	Xiaoying Lin – The Institute for Genomic Research
	Shauna Somerville – Carnegie Institution of Washington
8:30 pm	**Break** - Sponsored by InforMax, Inc. (Grand Gallerie)
	Michael Bevan – John Innes Centre, Department of Molecular Genetics
	Daphne Preuss - University of Chicago

9:45 - 10:30 pm	**Special Session** with **Dr. Arthur Caplan** University of Pennsylvania Health System
10:30 pm	**Welcome Reception** - Sponsored by Pangea Systems (Fontainebleau Ballroom)

Sunday, September 19

7:00–8:15 am	**Breakfast** (Fontaine Room)
7:00 am-4:00 pm	**Registration** (Grand Gallerie)
7:00 am-6:00 pm	**Internet Café sponsored by Compaq** (Brittany Room)
8:30 am–Noon	**Plenary Session 2: Model Organisms I (Grand Ballroom)**
	Chair: **Gerald Rubin** – University of California, Berkeley
	Mark Adams – Celera Genomics
	Gene Myers – Celera Genomics
	Michael Ashburner – European Bioinformatics Institute
	Gary Karpen– Salk Institute for Biological Studies
10:15 am	**Break** - Sponsored by Molecular Applications Group and TIGR
	Allan Spradling – Carnegie Institution of Washington
10:00-6:00 pm	**Poster and Electronic Poster Session I** - sponsored by SGI (Champagne, Lorraine and Monaco Rooms)
Noon–2:00 pm	**Lunch** (Exhibit Hall and in French Hall)
3:00–5:00 pm	**Concurrent Sessions** (Fontainebleau Ballrooms) Genomics I, Technology I, Bioinformatics I
Free Night	

Monday, September 20

7:00–8:15 am **Breakfast**

7:00 am-7:00 pm **Internet Café** - sponsored by Compaq (Brittany Room)

8:30 am–Noon **Plenary Session 3: Model Organisms II**

Chair: **Claire M. Fraser** – The Institute for Genomic Research

Edward DeLong- University of California, Santa Barbara

Molly Schmid – Microside

10:00-7:00 pm **Posters and Electronic Posters** Sponsored by SGI (Champagne, Lorraine and Monaco Rooms)

10:15 am **Break** – Co-sponsored by Molecular Applications Group and TIGR

Jonathan Eisen – The Institute for Genomic Research

David Eisenberg – UCLA

Paul Herrling – Novartis

Noon–2:00 pm **Lunch** (Exhibit Hall and in French Hall)

Noon–2:00 pm **Roundtable Discussion on Current Technology and Methodology in DNA Sequencing Laboratories** - Sponsored by Association of Biomolecular Resource Facilities (Grand Ballroom)

Scottie Adams, Trudeau Institute - Moderator
What is the ABRF?
Brief overview of the ABRF discussing research groups, electronic communications, scope of technologies covered.

Susan Hardin, University of Houston
How well do sequencing labs do on a standard template?
Presentation and discussion on the 1999 DNA Sequence Research Group. Results on assessing the current state of the art in DNA Sequencing Laboratories when a standard pGEM template is used.

Ted Thannhauser, Cornell University
How well do sequencing labs do on GC rich templates?
Presentation and discussion on the 1999 DNA Sequence Research

Group. Results on assessing the current state of the art in DNA Sequencing Laboratories using 2 different GC rich templates.

Dina Leviten, ICOS and **Duane Bartley**, Johns Hopkins University
The 3700 - Real Results from Core Labs
Presentation and discussion on the new PE/ABD 3700 DNA Sequencer.

George Grills, Albert Einstein College of Medicine
New Methods for High Throughput DNA Sequencing: 1000 Bases at 100% Accuracy
Presentation and discussion on techniques to extend your read length.

OPEN DISCUSSION
Audience participation will be encouraged.

3:00–5:00 pm **Free Time**

5:30–7:00 pm **Dinner** (Grand Ballroom) - Sponsored by PE Biosystems and Celera Genomics

7:00–10:15 pm **Plenary Session 4: Mammalian Genome Projects**

Richard Gibbs – Baylor College of Medicine

Howard J. Jacob – Medical College of Wisconsin

8:30 pm **Break**

Stephen J. O'Brien - National Cancer Institute

Randy Scott - Incyte Pharmaceuticals, Inc.

John Quackenbush – The Institute for Genomic Research

Tuesday, September 21

7:00 am-3:00 pm **Internet Café** - Sponsored by Compaq (Brittany Room)

7:30–8:45 am **Breakfast** (Fontaine Room)

9:00 am–Noon **Plenary Session 5: Hot Button Topics For The Next Millenium** (Grand Ballroom)

Chair: **J. Craig Venter** – Celera Genomics

	Jennifer A. L. Smith – Federal Bureau of Investigations
10:00-4:00 pm	**Posters** (Champagne, Lorraine and Monaco Rooms)
10:15 am	**Break**
	Mark R. Hughes – Wayne State University
	Jeffrey Isner – Tufts University
Noon–2:00 pm	**Lunch** (Exhibit Hall and in French Hall)
3:00–5:00 pm	**Concurrent Sessions** Genomics II, Technology II, Bioinformatics II (Fontainebleau Ballrooms)
4:00 pm	**Poster Breakdown/Exhibit Hall Breakdown**
6:00-10:00 pm	**Beach Blast** - Dinner sponsored by Incyte Entertainment sponsored by TIGR (Great Lawn – Fontainebleau Hilton)
10:00 pm	**Meeting Adjournment**

* Please note: Current and finalized agenda will be available at the conference site in the Pocket Agenda and addenda.

Plenary Speakers and Chairs

Mark D. Adams Ph.D.
Vice President, Genome Programs
Celera Genomics
Gene Research
45 West Gude Drive
Rockville, MD 20850
P: 240-453-3000
F: 240-453-3755
AdamsMD@celera.com

Michael Ashburner Ph.D.
Research Coordinator, Research
European Bioinformatics Institute (EMBL)
Wellcome Trust Genome Campus
Hinxton, Cambs CB10 1SD
UNITED KINGDOM
P: 44-1223-494-648
F: 44-1223-494-470
ashburner@ebi.ac.uk

Michael W. Bevan Ph.D.
Department Head
John Innes Centre
Department of Molecular Genetics
Colney Lane
Norwich, NR4 7UJ
UNITED KINGDOM
P: 44-1603-452835
F: 44-1603-505725
bevan@bbsrc.ac.uk

Arthur L. Caplan Ph. D.
Director
University of Pennsylvania Health System
Center for Bioethics
University City Science Center
3401 Market Street, Suite 320
Philadelphia, PA 19104-3308
P: 215-898-7136
F: 215-573-3035
caplan@mail.med.upenn.edu

Edward F. DeLong Ph.D.
Chair, Science Division
Monterey Bay Aquarium Research Institute
Research and Development
7700 Sandholdt Road,Box 628
Moss Landing, CA 95039
P: 831-775-1843
F: 831-775-1645
delong@mbari.org

Jonathan Eisen Ph.D.
Assistant Investigator
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20870
P: 301-838-3507
F: 301-838-0208
jeisen@tigr.org

David Eisenberg Ph.D.
Director
UCLA
Energy Lab of Biology & Molecular Med.
Box 951570
Los Angeles, CA 90095-1570
P: 310-825-3754
F: 310-206-3914
david@mbi.ucla.edu

Claire M. Fraser
President/CEO
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850
P: 301-838-0200
F: 301-838-0208
cmfraser@tigr.org

Richard A. Gibbs Ph.D.
Director
Baylor College of Medicine
Human Genome Sequencing Center,
Houston, TX 77030
P: 713-798-6539
F: 713-798-5741
agibbs@bcm.tmc.edu

Paul Herrling Ph.D.
Head of Research
Novartis Pharma, Ltd.
WSJ-386.1304
Postfach
Basel, 4002
SWITZERLAND
P: 41 61 324 5284
F: 41 61 324 2141
paul.herrling@pharma.novartis.com

Mark R. Hughes Ph.D.
Wayne State University
Center for Molecular Medicine
5047 Gullen Mall, Room 5107
Detroit, MI 48202
P: 313-993-1353
F: 313-577-6200

Jeffrey Isner Ph.D.
Tufts University School of Medicine
Department of Biomedical Research
St. Elizabeth's Medical Center
736 Cambridge Street
Boston, MA 02135
P: 617-789-2392
F: 617-789-5029

Howard J. Jacob Ph.D.
Associate Professor
Medical College of Wisconsin
Department of Biology
8701 Watertown Road
Milwaukee, WI 53226
P: 414-456-4887
F: 414-456-6516
jacob@mcw.edu

Gary H. Karpen Ph.D.
Associate Professor
The Salk Institute
MBVL
10010 North Torrey Pines Road
La Jolla, CA 92037
P: 619-453-4100
F: 614-622-0417
karpen@salk.edu

Xiaoying Lin Ph.D.
Staff Scientist
The Institute for Genomic Research
Eukaryotic Genomics
9712 Medical Center Drive
Rockville, MD 20850-3319
P: 301-838-3530
F: 301-838-0208
xlin@tigr.org

Gene W. Myers Ph.D.
Senior Director
Celera Genomics
Informatict Research
45 West Gude Drive
Rockville, MD 20850
P: 7240-453-3007
F: 240-453-3324
MyersGW@celera.com

Stephen O'Brien
Chief
National Cancer Institute - FCRDC
Laboratory of Genomic Diversity
Building 560, Room 21-105
Frederick, MD 21702-1201
P: 301-846-1296
F: 301-846-1686
obrien@ncifcrf.gov

Daphne Preuss Ph.D
Assistant Professor
University of Chicago
Molecular Genetics and Cell Biology
1103 E. 57th Street (EBC 304),C
Chicago, IL 60637
P: 773-702-1605
F: 773-702-9270
dpreuss@midway.uchicago.edu

John Quackenbush Ph.D.
Associate Investigator
The Institute for Genomic Research
Department of Eukaryotic Genomics
9712 Medical Center Drive
Rockville, MD 20850-3319
P: 301-838-3528
F: 301-838-0208
johnq@tigr.org

Gerald M. Rubin Ph.D.
Professor of Genetics
University of California, Berkeley
Department of Molecular and Cell Biology
545 Life Sciences Addition
Berkeley, CA 94720-3200
P: 510-643-9945
F: 510-643-9947
gerry@fruitfly.berkeley.edu

Molly B. Schmid Ph.D.
Vice President
Microcide Pharmaceuticals, Inc.
Research Alliances
850 Maude Avenue
Mountain View, CA 94043
P: 650-428-3542
F: 650-428-3534
mschmid@microcide.com

Randy W. Scott Ph.D.
President and Chief Scientific Officer
Incyte Pharmaceuticals Inc.
3174 Porter Drive
Palo Alto, CA 94304
P: 650-845-4533
F: 650-845-4500
khenson@incyte.com

Jenifer A. Smith Ph.D.
Unit Chief
DNA Analysis Unit
J. Edgar Hoover Building, Rm. 3905
935 Pennsylvania Avenue NW
Washington, DC 20535
P: 202-324-5436
F: 202-324-8090

Chris Somerville Ph.D.
Director
Carnegie Institution of Washington
Department Plant Biology
260 Panama Street
Stanford, CA 94305-4101
P: 650-325-1521 X203
F: 650-325-6857
crs@andrewz.stanford.edu

Shauna Somerville Ph.D.
Staff Scientist
Carnegie Institution of Washington
Department of Plant Biology
260 Panama Street
Stanford, CA 94305
P: 650-325-1521
F: 650-325-6857
shauna@andrew2.stanford.edu

Allan C. Spradling Ph. D.
Investigator and Staff Member
Howard Hughes Medical Institute
Department of Embryology
Carnegie Institution of Washington
115 West University Parkway
Baltimore, MD 21210
P: 410-554-1213
F: 410-467-1147
spradling@MAILl.ciwcmb.edu

J. Craig Venter Ph.D.
President and Chief Scientific Officer
Celera Genomics
45 West Gude Drive
Rockville, MD 20850
P: 240-453-3500
F: 240-453-3650
JCVenter@celera.com

Plenary Speaker Abstracts

Plenary Session I: Agriculture
Saturday, September 18
7:00 – 9:45 pm

Somerville, Chris
Abstract not received in time for publication

After Sequencing: Activities toward the Annotation of the Arabidopsis Genome

Xiaoying Lin, The Institute for Genomic, Rockville, MD

The sequencing of the first plant genome of *Arabidopsis thaliana* by the international Arabidopsis Genome Initiative (AGI) is well under way, with the entire sequence expected to be completed within a year. The challenge ahead is how to best utilize this sequence to identify and characterize all of the genes in the genome and their respective biological functions. To efficiently reach such a goal, it is critical that bioinformatic researchers not only identify potential genes in the sequence but also integrate and curate these data with other available sequence and functional data and provide them to the general research community. Current annotation is produced by separate AGI groups with different standards and formats and data must be accessed from different sites. There is a great need for the community to have a centralized database to access all sequence and annotation data in a standardized and uniform format. To address this need, TIGR has created such a relational database, the Arabidopsis Genome Annotation Database (AGAD). Sequence generated by all AGI labs, along with annotation where available, are parsed and loaded to AGAD. Computational analyses (database searches and gene predictions) are performed and manually curated and then displayed on the TIGR web-site. New sequences will be incorporated into this database as they become available. All annotation data can be accessed through the web (http://www.tigr.org/tdb/at/at.html), by way of map position, clone name, and gene name. We are in the process of evaluating the published Arabidopsis annotation using TIGR annotation routines to bring this database up to date. Some interesting findings from our sequence annotation and analysis will be discussed.

Somerville, Shauna
Abstract not received in time for publication

Arabidopsis Genome Sequence and Applications for Crop Plant Improvement

Michael W. Bevan, John Innes Centre, Norwich, UK

The sequence of chromosomes 2 and 4 from the plant Arabidopsis are essentially completed except for detailed characterisation of complex repeat clusters associated with the centromere. Chromosome 5 will also be completed in 4-6 months, and the complete genome is scheduled to be completed in late 2000. An analysis of features revealed by sequencing 22 Mb of chromosome 4 will be presented.

The next challenge is to define the functions of genes systematically, particularly the 40% of genes that appear to be plant-specific. The techniques used for large scale gene function search using forward and reverse screens and progress in scaling up their use, will be described. Initial progress made in characterizing large gene families will be reviewed.
Finally, the approaches taken to apply the wealth of new knowledge revealed by genomics to crop plant improvement, using a strategy for systematically defining the functions of genes in cereals using knowledge gained in Arabidopsis, will be discussed.

This work is supported by grants from the EC, BBSRC, GSF, NSF and USDA.

Preuss, Daphne
Abstract not received in time for publication

Special Session
Saturday, September 18
10:30 am

Caplan, Arthur
Abstract not received in time for publication

Plenary Session II: Model Organisms I
Sunday, September 19
8:30 am-Noon

Rubin, Gerald
Abstract not received in time for publication

Whole Genome Shotgun Sequencing of *Drosophila melanogaster*

Mark D. Adams, Celera Genomics, Rockville, MD

Drosophila melanogaster represents an extraordinary model system for the study of such complex and broadly applicable biological processes as development, neurobiology, and eukaryotic chromosome structure and evolution. At about 120 Mbp, the euchromatic portion of the genome is modest in size compared to the human genome at 3,000 Mbp. Because of its value for biological research and because of the excellent genomic resources already in place, Celera Genomics and the Berkeley *Drosophila* Genome Project are collaborating to accelerate the sequencing of the *Drosophila* genome. The method used is 'whole-genome shotgun sequencing,' which has been successfully applied to bacterial genomes and eukaryotic chromosomes. The combination of a 10-fold coverage of random sequence fragments derived from whole-genome libraries with extensive STS-content and BAC fingerprint maps, extensive BAC end sequences, and minimal shotgun coverage from mapped BACs provides a powerful test system for the efficacy of the whole-genome shotgun method and the supporting information that is necessary to validate assembly from the shotgun sequence data. The current status of the project and lessons learned will be presented.

A Whole Genome Assembler for Drosophila

Granger Sutton, Eric Anson, Art Delcher, Ian Dew, Catherine Jordan, Saul Kravitz, Stefano Lonardi, Clark Mobarry, Knut Reinert, Karin Remington, **Gene Myers**, Celera Genomics, Rockville, MD.

We will be reporting on the overall design of a whole genome shotgun assembler and our initial experience with it on simulated data for C. Elegans and real data from Drosophila. The assembler takes a collection of paired end reads, called mates, sequenced from the ends of a discrete number of sized insert libraries — in the current protocol, 2Kbp, 10Kbp, and BAC-inserts. The assembler is unusual in that it uses quality values primarily for sequence trimming and consensus/SNP evaluation. The key to scaling shotgun to whole genomes is not based on trying to prioritize overlaps based on *local* information, but instead to detect and delimit repetitive regions from a *global* perspective and use mates to navigate around and through them.

Our assembler can compute all overlaps between the 2.4 million reads expected for Drosophila in under 3 hours on a 10-processor Compaq platform, and is capable of the entire assembly in under 12 hours. The assembler correctly identifies all low copy portions of the genome, correctly building contigs for each and ordering them into scaffolds spanning each of the chromosomes. The uibiquitous repeats that lie between these contigs are identified. In a first pass we are at the time of writing able to resolve 95% of the repeats at the sequence level and further efforts should yield a >99% result.

An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila* - The '*Adh*' region.

M. Ashburner, S. Misra, J. Roote, S. Lewis,S. Celniker and G. M. Rubin. Department of Genetics, University of Cambridge, Cambridge, England; Berkeley Drosophila Genome Project, Department of Molecular and Cell Biology, University of California, Berkeley; Howard Hughes Medical Institute, Life Sciences Annex, University of California, Berkeley, CA; Lawrence Berkeley National Laboratory, Berkeley, CA.

A contiguous sequence of nearly 3-Mb from the genome of *Drosophila melanogaster* has been sequenced from overlapping P1 and BAC clones. An analysis of the sequence predicts 218 protein coding genes, 11 tRNAs and 17 transposable elemens. At least 38 of the protein coding genes are arranged in clusters from 2 to 6 closely related genes, suggesting extensive tandem duplication. The gene density is one protein coding gene every 13-Kb; the TE density is one element every 171-Kb. Of 73 genes in this region identified by genetic analysis, 53 have been located on the sequence; P-element insertions have been mapped to 43 genes. 95 (44%) of the known and predicted genes match a *Drosophila* EST, and 144 (66%) have clear similarities to proteins in other organisms. Genes known to have mutant phenotypes are more likely to be represented in cDNA libraries, and far more likely to have products similar to proteins of other organisms, than are genes with no known mutant phenotype. Over 650 chromosomes aberration breakpoints map to this chromosome region, their non-random distribution on the genetic map reflects variation in gene spacing on the DNA.

Sequence Analysis of a Drosophila Centromere

Janice Wahlstrom, Hiep Le and **Gary H. Karpen**. The Salk Institute, La Jolla, CA.

Heterochromatin is an important and mysterious region of the genome, but it is excluded from 'whole' genome sequencing efforts due to the presence of repeated DNAs. We need to develop approaches to map and sequence heterochromatin, since it contains centromeres and other elements responsible for nuclear organization and function.

We have investigated the structure and function of heterochromatin using a Drosophila minichromosome, Dp1187. We have localized the fully-functional centromere to a 420 kb region, and determined its molecular structure and composition with strategies that circumvent the problems normally associated with analyzing repeated DNA. A complete restriction map has been generated, and 25% of the centromere has been sequenced to date. Our results reveal striking features of the centromere: it is primarily composed of uniform satellite arrays and single, complete transposable elements. Whole genome analysis reveals that other regions of Drosophila heterochromatin are organized in a similar fashion. Surprisingly, the Dp1187 centromeric satellites and transposable elements are neither unique to centromeres nor present at all centromeres. This work constitutes the first detailed molecular structure of a functional centromere in a multicellular organism. The impact of these results on our understanding of heterochromatin structure, and on the determinants of centromere identity and function, will be discussed.

The BDGP gene disruption project: P element insertions mutating 25% of vital Drosophila genes

Allan C. Spradling, [1] Dianne Stern, [1] Amy Beaton, [2] E. Jay Rhem[2] Todd Laverty, [2]Nicole Mozden, Sima Misra[2] and Gerald M. Rubin[2]. [1]Howard Hughes Medical Institute Research Laboratories, Carnegie Institution of Washington, Baltimore, MD, [2]Howard Hughes Medical Institute Research Laboratories, University of California, Berkeley, CA

A fundamental goal of genetics and functional genomics is to identify and mutate every gene in model organisms such as *Drosophila melanogaster*. The Berkeley Drosophila Genome Project (BDGP) gene disruption project generates single P element insertion strains that each mutate unique genomic open reading frames. So far, 1,052 strains have been produced that disrupt more than 25% of the estimated 3,600 Drosophila genes that are essential for adult viability. Strains in the BDGP collection are available from the Bloomington Stock Center and have already assisted the research community in characterizing over 250 Drosophila genes. Sequences flanking more than 920 insertions have been determined to exactly position them in the genome, and to identify the likely open reading frame mutated in 723 lines (67%). Our results provide a rationale for expanding the collection based entirely on insertion site sequencing. We predict that this approach can bring more than 85% of all Drosophila open reading frames under experimental control.

Plenary Session III: Model Organisms II

Sunday, September 19
8:30 am-Noon

Fraser, Claire M.
Abstract not received in time for publication

Dissecting the genomes of uncultivated microorganisms: prospects, problems, and promise

Edward F. DeLong[1], Oded Beja[1], Ronald Swanson[2], and Robert Feldman[3]. [1]Monterey Bay Aquarium Research Institute, Moss Landing, CA. [2] Diversa Corp., San Diego, CA. [3] Molecular Dynamics, Inc., Sunnyvale, CA.

Cultivation independent surveys of naturally-occurring microbial populations have revealed an astonishing variety of microbial species new to science. A diverse array of novel microbes, undetected by culture-based methods, has been shown to exist in virtually every habitat examined. These novel microbes, evolutionarily distant from known cultivated types, can tell stories that will be not be revealed by study of common laboratory cultivars. Some of the newly discovered microorganisms represent the most abundant microbial types on our planet: their genomes harbor secrets of the inner workings of global biogeochemical cycles. Coupling modern genomic analysis with methods in microbial ecology promises to shed light on the nature and properties of a hugely important, yet still uncharacterized microbial world. Initial attempts in our laboratories have provided a glimpse into the genomes of uncultivated archaea, revealing features of their evolutionary history, and providing reagents to test biochemical and physiological hypotheses. Our current efforts focus on producing higher quality libraries, to better represent uncultivated microbial genomes. Results are promising, and include successful construction of BAC libraries from oceanic picoplankton, yielding BACs with inserts in excess of 100 kb.

Genetic and Genomic approaches to Anti-microbial Drug Discovery

Molly B. Schmid, Ph.D., Vice President, Research Alliances, Microcide Pharmaceuticals, Inc.

With the completion of numerous microbial genome sequences, the discovery of antimicrobial drugs has fully entered the genomic era.

As in other organisms, microbial genomes contain both pharmaceutically relevant and pharmaceutically non-relevant genes. Microcide has developed a "targeted genomics" platform that allows it first to assess pharmaceutical relevance of genes, then rapidly develop screens. While some of the techniques rely on the tools available in microbial systems, many of the strategies for screen development are broadly applicable in other therapeutic areas.

Phylogenomics: The Benefits of an Evolutionary Perspective in Genome Studies and Genome Analysis in Evolutionary Studies.

Jonathan A. Eisen, The Institute for Genomic Research, Rockville, MD

I will discuss the uses and applications of integrating evolutionary reconstructions and genome analysis into a single composite approach, which I refer to as phylogenomics. A phylogenomic approach is useful for three main reasons: a) evolutionary methods can benefit comparative genomic studies; b) genome analysis is incredibly useful in studies of evolution and in particular analysis of complete genomes allows one to address questions never before possible in evolutionary studies; and c) there are feedback loops between genome analysis and evolutionary reconstruction such that they can be combined for a mutual benefit. I will show how phylogenomic analysis can be used to make better predictions of gene functions, to infer evolutionary events such as gene duplication and lateral gene transfer, to test protein structure models, and to scan the genome for genes with unusual histories. I will focus on phylogenomic analysis of current TIGR genome projects including the extremely radiation resistant bacterium Deinococcus radiodurans, the thermophilic bacterium Thermotoga maritima, and chromosome II of the plant *Arabidopsis thaliana.*

PROTEIN FUNCTIONS FROM GENOME SEQUENCES.

David Eisenberg, Edward Marcotte, Matteo Pellegrini, Michael Thompson & Todd Yeates. Box 951570, UCLA, Los Angeles CA

New computational methods have been developed for assigning protein functions and interactions to genome sequences. These methods make use of information from all fully sequenced genomes and are distinct from standard homology methods.

The first method is called the phylogenetic profile method; it examines the correlated inheritance of proteins in different species. It is based on the assumption that proteins that evolve in a correlated fashion are likely to function together in a pathway or structural complex. During evolution, such functionally linked proteins tend to be either all preserved or eliminated in a new species. This property of correlated evolution is described by characterizing each protein by its phylogenetic profile, a string that encodes the presence or absence of a protein in every fully sequenced genome.

Studies show that proteins having matching or similar phylogenetic profiles strongly tend to be functionally linked. Conversely, proteins having similar function tend to have similar phylogenetic profiles. This correlation permits us to assign functions to uncharacterized proteins that have phylogenetic profiles similar to proteins of known function.

The second method examines correlated domains in proteins and it's termed the Rosetta Stone method. It is based on the observation that some pairs of interacting proteins have homologs in another organism fused into a single protein chain. Many members of these pairs are confirmed as functionally related. Some proteins have links to several other proteins; these coupled links appear to represent functional interactions such as complexes or pathways.

Combining these methods has permitted us to carry out a genome-wide prediction of protein function for *Saccharomyces cervisiae*, using information from some 20 fully sequenced genomes. For the 6,217 proteins of yeast, over 98,000 pairwise links between functionally-related yeast proteins have been discovered. Links between characterized and uncharacterized proteins allow a general function to be assigned to more than half of the 2,557 previously uncharacterized yeast proteins. Of the pairwise links, we find some 21,000 from phylogenetic profiles, and 50,000 from Rosetta Stone sequences. From these links, we designate 4,152 as being of the "highest confidence."

In short, the new methods lead to functional assignments for a large fraction of previously uncharacterized proteins.

A Transgenic Model for Alzheimer's Disease

Paul Herrling and Christine Sturchler-Pierrat, Novartis Pharma AG, CH4002 Basle Switzerland

Transgenic mouse models provide an important genomics tool for evaluation of function of candidate disease genes. Such models will contribute to a better understanding of the molecular basis of disease onset and progression, and provide *in vivo models* which focus discovery efforts on new causal therapeutics to treat major unmet medical needs such as Alzheimer's Disease (AD).

Several transgenic mouse lines have been generated which express the human amyloid precursor protein (APP) with different AD-linked mutations. The APP protein is critical for the formation of amyloid plaques, one of the major pathological hallmarks of AD, as well as other possible events that lead to disease development. A transgenic mouse model, APP 23 overexpressing human APP751 cDNA carrying the Swedish double mutation expresses the exogenous human APP mRNA 7-fold over the endogenous APP mRNA. The mice develop amyloid deposits by 6 months, and deposits increase in size and number with age and occupy a substantial area of the cortex and hippocampus in 24-month-old mice. The APP 23 transgenic mouse line also mimics other aspects of human AD pathology and displays deficits in a test of spatial memory. This transgenic animal model provides a tool to assess therapeutic leads that affect amyloid plaque formation and reverse disease progression. Hence such a model offers possibilities for experimental approaches to test novel therapeutic strategies.

Sturchler-Pierrat *et al.* (1997), PNAS, 94, 13287 - 13292

Calhoun *et. al*, Nature (1998) 395, 755-756

Plenary Session IV:
Mammalian Genome Projects

Sunday, September 19
7:00 – 10:15 pm

Sequencing Mammalian DNA At The Baylor College Of Medicine Human Genome Sequencing Center

R. A. Gibbs and the staff of the Baylor College of Medicine Human Genome Sequencing Center, Houston TX

The (BCM-HGSC) is sequencing DNA from the human genome in addition to the mouse, fly, Dictyostelium and cDNAs. The sequencing strategy we have developed is based upon the analysis of individual BACs using M13 shotgun methods. Distinctive technical aspects of the BCM-HGSC process include the use of BODIPY dye primers, and custom built automatic workstations for different steps in the sequencing pipeline. In the month of June, 1999 we carried out 140,000 reactions, with an overall greater than 80% success rate (Q20>200). Approximately 30 Megabases of finished human DNA sequence has been generated. We aim to continue ramping to produce more than 350,000 reactions per month by next April.

The sequence reads are currently being used to generate a draft of the human genome, as part of a multi-laboratory international consortium. The Drosophila work also involves a mapping component that is a collaboration between the BCM-HGSC and the Berkeley Drosophila group. The Dictyostelium sequencing is being carried out in collaboration with several different groups, and the initial target is Chromosome 6. cDNA studies involve workers in three other laboratories who are generating full length cDNA clone libraries.

The Rat: An Ideal Model for Functional Genomics

Howard J. Jacob, Masahide Shiozawa, Anne Kwitek-Black. Laboratory for Genetics Research, Department of Physiology, Medical College of Wisconsin, Milwaukee, WI.

The laboratory rat has long been a favorite model for biochemists, pharmacologists and physiologists; since 1966 more than 500,000 articles have been published, most involving a model of human disease. The Rat Genome Project has been building the infrastructure required for the application of molecular genetic tools to this important physiological model system. The Table below illustrates that these genomic resources are in place.

Resource	US Project	International	Total
Genetic map:	>5,500 markers	>3,500	>9,000
Mapping cross:	>1,000 meioses	>1,000	
YAC libraries:		~10X coverage	~10X
coverage ~20X			
PAC library:	~10X coverage	~10X	
BAC library:	~10X coverage	~10X	
RH map:	>10,000 markers	>5,000 markers	>10,000
Normalized librs:	12 different tissues ?		>12 libraries
cDNA/EST project:	>100,000		

We have taken these resources one step further by developing detailed comparative maps between rat, human and mouse. We are now able to integrate the strengths of all three systems to address the genetic basis of complex diseases including diabetes, hypertension, and renal failure. Our development of a functional genomics platform ranging from bioinformatics, consomic rats, genomics and integrative physiology will be discussed.

O'Brien, Stephen J.
Abstract not received in time for publication

Scott, Randy
Abstract not received in time for publication

Whole Genome Functional Analysis in Humans Using cDNA Microarrays

John Quackenbush, The Institute for Genomic Research, Rockville, MD

A goal of the Human Genome Project is identification of the complete set of human genes and the role played by these genes in development and disease. Microarrays provide the opportunity to study gene expression patterns on a genomic scale. Thousands of cDNA clones are arrayed on a microscope slide and relative expression levels of the genes they encode are determined by measuring fluorescence intensity of labeled mRNA hybridized to the arrays. The data provided by microarray analysis promises functional information on a genomic scale, allowing a significant fraction of the genes in any organism to be assayed in a single experiment. Further, they provide a means of identifying candidate genes that may play a role in development and progression of human disease.

We have begun to realize that promise. We have assembled a collection of cDNA clones representing more than 40,000 distinct genes, developed laboratory hardware and protocols, and created databases and data analysis tools necessary to analyze differential expression. High-density cDNA microarrays containing more than 19,200 PCR amplified clones have been used to study differential expression patterns in colon tumor metastasis, using cell lines of low metastatic (KM12C) and high metastatic (KM12L4A, KM12SM) potential as a model. This analysis has allowed the identification of genes, many of unknown function, which may play a role in tumor metastasis. Known genes within this set present a well-defined picture of the metastatic process and the additional genes provide candidates for further analysis.

Plenary Session V:
Hot Button Topics for the Next Millenium

Tuesday, September 21
9:00 am – Noon

Venter, J. Craig
Abstract not received in time for publication

Forensic Applications of DNA Analysis in the FBI Laboratory

Jenifer A.L. Smith, Unit Chief of the DNA Analysis Unit I, FBI Laboratory, Washington, DC

In December of 1988, the FBI Laboratory began conducting DNA analysis on items of evidence stained with body fluids such blood or semen. Since the initial implementation of the first DNA typing procedure, dramatic changes have occurred to improve and expand DNA testing. The first DNA typing procedure implemented by the FBI Laboratory involved analyzing Restriction Fragment Length Polymorphisms (RFLP). Additional DNA typing techniques using the Polymerase Chain Reaction (PCR) have subsequently been implemented. Using these various technologies, the DNA Analysis Unit I (DNAU I) of the FBI Laboratory has generated results on thousands of cases. Technical advances were made due to the combined efforts of scientists from public forensic laboratories, academia and industry. Battles waged in the courtroom were won by proponents of the powerful technology so that today, results from DNA analyses are commonly given in testimony in courts throughout the United States and the international law enforcement community. In October of 1997, the DNA Analysis Unit I (DNAU I) of the FBI Laboratory issued a policy allowing examiners to identify the source of an evidentiary body fluid stain in FBI Laboratory reports. Such a conclusion is reached when a match between a known and questioned sample is determined and the probability of the DNA profile obtained is sufficiently rare. Individualization of the source of the evidentiary DNA, to a reasonable degree of scientific certainty, can be assessed from the probability of the DNA profiles obtained. Analysis using both the RFLP and PCR-based STR procedures may provide sufficiently rare profiles to establish individualization. Additionally, DNA typing is now being used to assist investigators in solving crimes in which a suspect has not been readily identified. The Combined DNA Index System (CODIS) is a collection of DNA databases from forensic laboratories around the United States. DNA profiles of individuals previously convicted of serious crimes such as rapes and homicides are maintained in computer files and compared to DNA profiles collected from cases with no suspects. All states have passed legislation that allows them to collect DNA samples from convicted offenders. These profiles are compared to DNA profiles of unsolved crimes, creating an extremely valuable investigative tool. This presentation will give an historical overview of the changes in DNA technology, the battle in the courts, the CODIS system and the current status of DNA testing as experienced by the forensic examiners of the FBI Laboratory.

Hughes, Mark R.
Abstract not received in time for publication

Isner, Jeffrey
Abstract not received in time for publication

Concurrent Sessions Preliminary Agenda

Bioinformatics Concurrent Session I

Sunday, September 19
3:00 – 5:00pm

Automated Domain Identification in Protein Classification
Kimmen Sjölander, Molecular Applications Group, Palo Alto, CA

Life on the Edge: using genome-scale *in silico* models of microorganisms to interpret and predict metabolic phenotypes
Bernhard O. Palsson, Department of Bioengineering, University of California-San Diego

Function Prediction Using the Sequence→Structure→Function Paradigm: Analysis of Disulfide Oxidoreductase Activity in Eight Genomes
Jacquelyn S. Fetrow, GeneFormatics, Inc., San Diego, CA

Computational Structural Genomics
Steven E. Brenner, Department of Structural Biology, Stanford University, Stanford, CA

Genomics Concurrent Session I

Sunday, September 19
3:00 – 5:00pm

Whole genome sequencing of Vibrio cholerae, the etiologic agent of cholera
J.F. Heidelberg, The Institute for Genomic Research, Rockville, MD

PAC physical mapping for rice genome sequencing in RGP
Satoshi Katagiri, Rice Genome Research Program (RGP), National Institute of Agrobiological Resources / Institute of the Society for Techno-innovation of Agriculture, Forestry and Fisheries, Tsukuba, Ibaraki, Japan

Tomato EST Database: A Genomics Approach To Plant Research
C.M. Ronning, The Institute for Genomic Research

Detection of GEM Cross-Reactivity within Gene Families
Elisabeth Evertsz, Incyte Microarray Systems, Fremont, CA

Technology Concurrent Session I

Sunday, September 19
3:00 – 5:00pm

High Throughput Genomic Sequencing Using the MegaBACE 1000 Capillary Sequencer
Helene Jones, Oxagen Ltd, Oxford, UK

Genome-Wide Gene Expression Profiling by cDNA Microarray with Colorimetric Detection
Konan Peck, Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan

Effects of Template Quality on the Performance of Capillary Electrophoresis
Helmut Hilbert, QIAGEN GmbH, Max-Volmer-Straße 4, D-40724 Hilden, Germany

Performance of the ABI 3700 using BODIPY Dye-Primer Chemistry
Donna M. Muzny, Baylor College of Medicine, Human Genome Sequencing Center, Department of Molecular and Human Genetics, Houston, TX

JGI-LANL Sequencing Process Enhancement: R&D Results
Owatha L. Tatum, Center for Human Genome Studies – Joint Genome Institute, Los Alamos National

Bioinformatics Concurrent Session II
Tuesday, September 21
3:00 – 5:00pm

Defining the Pathway Structure of Metabolic Genotypes through Convex Analysis
Christophe H. Schilling, Bioengineering, University of California-San Diego

A Database Of Functionally Conserved Subfamilies Speeds Genome Annotation
Brendan Loftus, The Institute for Genomic Research, Rockville, MD

Integrating Pathway Information with Gene Expression and Sequence Analysis
Paul Thomas, Molecular Applications Group, Inc. Palo Alto, CA

Confidence Scoring of DNA Base Calls with the Automated Trace Quality Assessment (ATQA) System
Max A. Karlovitz, Malvern, PA

Finding the Information in DNA Chip Expression Data
Yixin Wang, Department of Molecular Biology, Department of Pathology and Experimental Toxicology, Parke-Davis Pharmaceutical Research, Warner-Lambert Company, Ann Arbor, MI

Genomics Concurrent Session II
Tuesday, September 21
3:00 – 5:00pm

Sequencing of a Plant Chromosome: The *Arabidopsis thaliana* Chromosome II Project
Kaul, S., The Institute for Genomic Research, Rockville, MD

Genotyping and Identification of Human Neurological Disorder Genes
Jay Ji, Biotech Research Labs, Boston Biomedica Inc., Gaithersburg, MD

Use of the Rat Gene Index to examine gene expression patterns from Src-transformed rat fibroblasts
Renae L. Malek, The Institute for Genomic Research, Rockville, MD

Mapping And Sequencing The ~200-Kb Segment Of Chromosome 17p13 Containing The Nephropathic Cystinosis Gene
Jeff W. Touchman, NIH Intramural Sequencing Center, Bethesda, MD

Global profiling of gene expression during maturation of human antigen presenting cells.
Akbar S. Khan, Army Medical Research Institute of Infectious Diseases, Frederick, MD

Human BAC Ends
Shaying Zhao, The Institute for Genomic Research, Rockville, MD

Technology Concurrent Session II
Tuesday, September 21
3:00 – 5:00pm

Three-Dimensional Arrays of Microtransponders Derivatized with Oligonucleotides
Wlodek Mandecki, PharmaSeq, Inc., Monmouth Junction, NJ

Microarrays in Three Easy Steps
Priti Hegde, The Institute for Genomic Research, Rockville, MD

A Single-tube Primer Extension Reaction for SNP Interrogation
Martin Johnson, PE Applied Biosystems, Foster City

High-throughput, multiplex analysis of SNPs implicated in disease using a novel luciferase-based analysis tool.
Daniel D. Kephart, Hartford Hospital, Hartford, CT

Use of FlexJet(tm) inkjet-based technology to construct 50k human gene microarrays
Stephen Friend, Rosetta Inpharmatics, Kirkland, WA

Concurrent Session Abstracts

Bioinformatics Concurrent Session I

Sunday, September 19
3:00 – 5:00pm

Automated Domain Identification in Protein Classification

Kimmen Sjölander, Paul Thomas, Brian Karlak Molecular Applications Group, Palo Alto, CA

One of the pitfalls of high-throughput sequence classification stems from the mix-and-match domain structure of proteins. While the function of a protein is derived from the sum of its parts, annotations associated with a protein are seldom decomposed accordingly. Instead, the protein as a whole is assigned a function, and matches of unknown sequences to any part of the protein are used to assign the function of the protein as a whole to the query. This is an obvious problem in automated classification of protein sequences when the region of alignment is local.

To address this problem, we developed an automated method to identify protein domains, and construct hidden Markov models for these domains. The resulting HMMs give us increased specificity of function assignment in high-throughput classification, without loss of sensitivity.

This method is demonstrated on a set of ion channels, where we found some surprising local homologies between a number of TNF-alpha-induced proteins and the tetramerization domain at the N-terminus of certain potassium channels, and kinases containing C-AMP binding domains and a region at the C-terminus of other potassium channels.

Life on the Edge: using genome-scale *in silico* models of microorganisms to interpret and predict metabolic phenotypes

Bernhard O. Palsson, Department of Bioengineering, University of California-San Diego

Small genome sequencing and annotation are leading to the definition of metabolic genotypes in an increasing number of organisms. We show how *in silico* metabolic genotypes are formulated based on genomic, biochemical, and microbiological data. Such metabolic genotypes have been formulated for *E. coli*, *H. influenzae*, and *H. pylori*. The *in silico* models are based on the philosophy of using applicable physico-chemical (such as stoichiometric structure) and capacity (maximum fluxes) constraints on the integrated functioning of the metabolic networks. Given these constraints, optimal phenotypes can be computed and compared to experimental data. They are found on the edge of the allowable solution spaces – a space that basically represents the reaction norm of the defined genotype – where the governing constraint on cellular functions can be identified. For *E. coli*, this process leads to quantitative prediction of growth and metabolic by-product secretion data in batch, fed-batch, and continuous cultures, and to the accurate prediction of the metabolic capabilities of 73 of 80 mutants examined. Furthermore, we present mathematical methods that allow for the analysis, interpretation, prediction, and engineering of the metabolic genotype-phenotype relationship, and for the interpretation of expression array data.

Function Prediction Using the Sequence→Structure→Function Paradigm: Analysis of Disulfide Oxidoreductase Activity in Eight Genomes

Jacquelyn S. Fetrow, Naomi Siew, and Jeffrey Skolnick, GeneFormatics, Inc., San Diego, CA, The Scripps Research Institute, La Jolla, CA

The practical exploitation of the vast numbers of sequences in the genome sequence databases is crucially dependent on the ability to identify the function of each sequence. Unfortunately, current methods, including global sequence alignment and local sequence motif identification, are limited by the extent of sequence similarity between sequences of unknown and known function; these methods increasingly fail as the sequence identity diverges into and beyond the twilight zone of sequence identity.

To address this problem, a novel identification method for protein function based on the sequence→structure→function paradigm is described. Descriptors of enzyme active sites, termed "functional forms" or FFs, are created based on the geometry and conformation of catalytic active sites. We show that the geometric descriptor created for disulfide oxidoreductase activity in the glutaredoxin/thioredoxin family is specific for the high-resolution atomic structures that are known in this family. We then make the FFs "fuzzy", to create "fuzzy functional forms" or FFFs and show that these descriptors can be applied to inexact protein models created by current fold prediction algorithms. The method is applied to all sequences in 8 different genomes, creating a curated set of all proteins with predicted disulfide oxidoreductase activity. The active site residues are identified in each predicted protein structure. For comparison, a similar analysis using the Blocks motifs for glutaredoxins and thioredoxins was applied to the same 8 genomes.

Computational Structural Genomics

Steven E. Brenner and Michael Levitt, Department of Structural Biology, Stanford University, Stanford, CA

Structural genomics aims to provide a good experimental structure or computational model of every tractable protein in a complete genome. Underlying this goal is the immense value of protein structure, especially for recognition of distant evolutionary relationships for proteins whose sequence analysis has failed to find any significant homolog. The solved structures will be similarly useful for elucidating the biochemical or biophysical role of proteins that have been previously ascribed only phenotypic functions. More generally, knowledge of an increasingly complete repertoire of protein structures will aid structure prediction methods, improve understanding of protein structure, and ultimately lend insight into molecular interactions and pathways.

We use computational methods to select families whose structures cannot be predicted and which are likely to be amenable to experimental characterization. A critical component is consultation of the PRESAGE database, which records the community's experimental work underway and computational predictions. The protein families are ranked according to several criteria including taxonomic diversity and known functional information and proteins are selected from these families as targets for structure determination. The solved structures are examined for structural similarity to other proteins of known structure and homologs are modeled.

Genomics Concurrent Session I

Sunday, September 19
3:00 – 5:00pm

Whole genome sequencing *of Vibrio cholerae*, the etiologic agent of cholera

J.F. Heidelberg[1], W. Nelson[1], T.D. Read[1], D.H. Haft[1], E.L. Hickey[1], M.L Gwinn[1], R.J. Dodson[1], R. Clayton[1], R.R. Colwell[2], J. J. Mekalanos[3], and C. M. Fraser[1], [1]The Institute for Genomic Research, Rockville, MD [2]Biotechnology Institute, University of Maryland, MD, [3]Department of Microbiology and Molecular Genetics, Harvard Medical School, Boston, MA

The complete genome sequence of *Vibrio cholerae* serotype O1, Biotype El Tor, strain N16961 was determined to be 4,024,233 base pairs containing 4,575 predicted coding sequences (open reading frames). The genome consists of 2 circular chromosomes, 1,072,915 and 2,951,318 base pairs with 1,302 and 3,273 ORFs respectively. Virtually all of the genes encoding DNA replication and repair, transcription, translation, cell wall biosynthesis, and a variety of central catabolic and biosynthetic pathways have been identified. However, the vast majority of recognizable genes for these essential biological functions are located on the large chromosome. Similarly, genes known to be essential for bacterial pathogenicity (i.e., those encoding the toxin co-regulated pilus, cholera toxin, lipopolysaccharide, and the extracellular secretion machinery) are also located on the large chromosome of N16961. In contrast, the small chromosome contains a larger fraction (66%) of hypothetical genes compared to the large chromosome (53%). Thus, the small chromosome might be involved in unique but undefined biological properties of the genus Vibrio. On such property is the capacity to form resting cells; this viable but nonculturable state (VBNC) can be induced starvation. In this regard, the genomic sequence provides a qualitative assessment of the regulatory and nutritional capacities of this organism. A total of 462 paralogous gene families containing a total of 1935 ORFs (42.3%) were identified in *V. cholerae*. The largest families are the ABC transporter family with 63 members, and the response regulator receiver domain with 61 members. A total of 225 paralogous gene families have only 2 members, and 119 families contain 382 genes that have no assigned biological role. The integron island of *V. cholerae* is also located on the small chromosome. This region contains 147 *V. cholerae* repeats (VCR) and spans 147,193 base pairs. Genes within this region have been implicated in virulence and other apparently encode antibiotic resistance genes; however, like the rest of the small chromosome, a large percent of the ORFs in this region have no assigned biological function. The integron island has a lower GC content than the overall chromosome (42.3% compared with 47.5% for the small chromosome) and has a significantly different Chi2 value for the tri-mers, suggesting many of these ORFs were acquired by lateral transfer.

PAC physical mapping for rice genome sequencing in RGP

Satoshi Katagiri,Tomoya Baba, Shoko Saji, Masao Hamada, Marina Nakashima, Masako Okamoto, Yoshino Chiden, Mika Hayashi, Ryoichi Tanaka, Kazuhiro Koike, Jianzhong Wu, Takashi Matsumoto Takuji Sasaki, Rice Genome Research Program (RGP), National Institute of Agrobiological Resources/Institute of the Society for Techno-innovation of Agriculture, Forestry and Fisheries, Tsukuba, Ibaraki, Japan

As the core of rice genome sequencing, RGP has constructed a PAC (P1-derived artificial chromosome) genomic library from Oryza sativa ssp. japonica cultivar, Nipponbare, in order to establish a sequence-ready physical map. This PAC genomic library consists of about 71,000 clones with average insert size of 112 kb and 16-fold genome coverage. To construct PAC contigs, we screen our PAC library by PCR using STS primers of mapped DNA markers as well as ESTs. At present we concentrate our efforts on ordering of PACs on rice chromosomes 1, 5, 6 and 10. So far, a total of 245 DNA markers and 565 ESTs mapped on these chromosomes have been used for screening. As a result, we have selected 3,790 PACs aligned in 257 contigs with a total physical length of about 35 Mb, corresponding to approximately 30% of these chromosomes. We have also started the complete genomic sequencing of some of these ordered PACs. Initial results of physical mapping with PACs as well as genomic sequencing have shown the uneven distribution of genes in the rice genome. Thus in order to identify the most number of genes at the outset of our sequencing efforts, we will focus our PAC physical mapping strategy on gene-rich regions of the genome.

Tomato EST Database: A Genomics Approach To Plant Research

C.M. Ronning[1], A. L. Matern[2], T. Vision[2], R. van der Hoeven[2], M.B. Craven[1], C.L. Bowman[1], M. D'Ascenzo[2], X. He[2], J. Lyman[2], J. Alcala[3], J. Vrebalov[3], R. White[3], S.D. Tanksley[2], J.J. Giovannoni[3], G.B. Martin[2], W. Nierman[1], C.M. Fraser[1], J. C. Venter[1]. [1]The Institute for Genomic Research; [2]Cornell University/Boyce Thompson Institute; [3]Texas A&M University

The tomato (*Lycopersicon esculentum L.*) has been the model organism in several aspects of plant biology, including pathogen response and fruit development. In order to identify genes involved in such processes, cDNA libraries constructed from five core tissues (shoot, root, seed, callus, and carpel) are being sequenced to generate Expressed Sequence Tags (ESTs), a very cost effective way of identifying a large number of genes expressed in a tissue. In addition, clones will be sequenced from libraries derived from susceptible and from resistant tomato lines challenged with *Pseudomonas syringae* pv. tomato, from tissues exposed to mixed elicitors, and from early- and late-ripening fruit. The 90,000 ESTs thus generated will be used to construct a nonredundant Tomato Gene Index (LGI; http://www.tigr.org/tdb/tdb.html), that will be annotated to identify possible functions of each gene. The source of each EST will also be part of this database, and therefore it will be possible to identify tissue-specific genes, developmental stage-specific genes, and genes switched on in response to attacking pathogens. This data is also being used to create the SYNTOM database at Cornell, which will be a very useful source of information for the Solanaceae as well as other plant species. These sequences will be a valuable resource for a large number of experimental approaches including mapping, expression studies, and genome annotation. It is expected that the sequences and the clones from this project will become part of the collection of tools used by tomato researchers, and indeed researchers working on any number of other species.

Detection of GEM Cross-Reactivity within Gene Families

Elisabeth Evertsz, Janice Au-Young, Mike Ruvolo, Steven Daniel, Ai Ching Lim, Tom Theriault, and Mark Reynolds, Incyte Microarray Systems, 6519 Dumbarton Circle, Fremont, CA

Gene Expression Microarrays (GEM) are powerful tools for the elucidation of differential gene expression among two different MRNA samples. The objective of this project was to determine the level of hybridization cross-reactivity within gene families using standardized stringency conditions. We chose representative members of several pharmaceutically relevant gene families: chemokines, cytochrome P-450 isozymes, G proteins, and proteases. The percent identity within the 6 families ranged between 55 -100% at the nucleotide level, as determined by BLASTN. Targets were amplified from cDNA clones by PCR using vector specific primers, arrayed onto a glass surface and chemically bonded to enable parallel hybridization experiments. RNA transcripts were generated from a designated parent clone member of each family by adding a T7 promoter site to the PCR primer. These transcripts were then used to generate fluorescent cDNA probes using standard protocols. As expected, hybridization signals were highest at the spotted elements corresponding to the designated parent clone members of each clone family. The signal intensities for non-parent clones were significantly decreased relative to the parent. Relative intensities correlated with the percent identity between probe and target. These results suggest that the effects of cross-reactive probes on differential expression data can be interpreted based upon the degree of sequence similarity between targets.

Technology Concurrent Session I

Sunday, September 19
3:00 – 5:00pm

High Throughput Genomic Sequencing Using the MegaBACE 1000 Capillary Sequencer

Helene Jones, Gayle Vincent, Jo Wynne, Alisoun Carey. Oxagen Ltd, Oxford, UK

Oxagen is carrying out large scale family studies to identify genes involved in common diseases such as IBD and asthma. In order to identify polymorphic markers and novel genes in regions of linkage or association, we have developed a high throughput sequencing capability.

High quality DNA is prepared from BACs containing genomic inserts from regions of interest. The DNA is mechanically sheared using a Hydroshear to produce random fragments which are then subcloned into pUc19. DNA from individually picked clones is prepared using a semi-automated method developed at Oxagen. This is a modified alkaline lysis which utilises the Packard Multiprobe II liquid handling robot. Sequencing reactions are performed using ET Terminators and then analysed using the MegaBACE 1000 Capillary Sequencer. The quality of sequence from a 150 min run is comparable to that obtained from an ABI 377 with 48cm WTR plates during a 10hr run. Average read lengths obtained are 650bp after quality and vector clipping which is again comparable to those seen on an ABI377.

The methods we have established have provided Oxagen with a high-throughput sequencing capacity capable of generating approximately 7Mb sequence (6x coverage) per year with minimal staffing requirements.

Genome-Wide Gene Expression Profiling by cDNA Microarray with Colorimetric Detection

Konan Peck, Yuh-Pyng Sher, Yue-Zung Lee, Jeremy J.W. Chen, Meng-Hsuan Han, and Wei-Chen Kao. Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan 115, R.O.C.

A high capacity system based on cDNA microarray approach and colorimetric detection is developed to simultaneously monitor the expression of many genes in different tissues or cell lines at different physiological conditions in a microarray format on nylon membranes. A UniClone database based on NCBI Unigene clustering and an indexed library containing 60,000 non-redundant gene clones have been constructed for genome-wide gene expression studies. To array 60,000 non-redundant cDNA fragments on nylon membranes, a high capacity arraying machine capable of holding 31,000 PCR products at a time and depositing 31,000 spots on a piece of 1.8 cm by 2.7 cm nylon membrane have been developed. The current detection limit of the system allows gene expression profiling by using RNA samples derived from a few hundred cells. The characterization of the gene screening and quantification system will be presented together with examples of gene expression profiles of human tissues and cell lines.

Effects of Template Quality on the Performance of Capillary Electrophoresis

Holger Wedler, Helmut Hilbert, Ralf Himmelreich, Dörte Möstl, Erika Wedler, and Andreas Düsterhöft QIAGEN GmbH, Max-Volmer-Straße 4, D-40724 Hilden, Germany

The recent introduction of capillary electrophoresis into molecular biology has fueled the hope to finish large scale genome sequencing projects much faster than previously anticipated. However, fluorescent sequencing on these machines has to be optimized and integrated into each lab's individual workflow. Critical factors include template purification method, set-up of sequencing reactions and subsequent clean-up, and ionic strength of the loading buffer.

We have investigated in these factors and present data that demonstrate that both the success rate and the reading length on the 3700 can be increased to 95% and >500 bases. Data from various genome projects demonstrate that the quality of data from the 3700 capillary sequencer are comparable to 48 cm WTR from the 377 instrument.

Performance of the ABI 3700 using BODIPY Dye-Primer Chemistry

Donna M. Muzny, Benedict Bodota, John Bouck, James H. Gorrell and Richard A. Gibbs, Baylor College of Medicine, Human Genome Sequencing Center, Department of Molecular and Human Genetics, Houston, TX

Central to our production model at the Baylor College of Medicine, Human Genome Sequencing Center (BCM-HGSC) has been the implementation of the BODIPY Dye-Primer Chemistry. We previously reported the use of the BODIPY fluorophores for automated sequencing showing improved spectral and electrophoretic properties over fluorescein/rhodamine dyes. Our goal in implementing the ABI 3700 capillary sequencers has been to optimize the 3700 run parameters for the BODIPY Dye-Primer chemistry as well as the BIG DYE terminator chemistry. In order to use the BODIPY Dye-Primer chemistry on the ABI 3700, a new matrix of discrete fragment lengths was generated by PCR with the BODIPY Dye-Primer Dye set. A spectral calibration on the 3700 was performed using standard procedures. The mobility file for the BODIPY Dye-Primer set was provided by PE-ABI. Run parameters were optimized to 40 C cuvette temperature, 50 C run temperature, 45 sec injection time and 2000 V run voltage.

Parallel experiments were performed comparing the performance of BODIPY Dye Primer chemistry and Big Dye terminator chemistry on the ABI 3700 using the program PHRED to assess data quality and accuracy of the called bases. Phred analysis has shown that the BODIPY Dye-Primer reactions yielded higher quality data in the longer read lengths with an average of 715bp of PHRED 20 quality data as compared to 600 bp for the BIG Dye terminators. To expand these studies we have now compared the quality and read lengths of the BODIPY Dye-Primers, Big Dye terminators and Energy Transfer terminators on additional sequencing platforms including the Molecular Dynamics MegaBACE 1000 and the ABI 377XL-96. Performance of the BODIPY Dye-Primer sets in completed and draft assemblies have also been evaluated. The results of these comparisons will be reported.

JGI-LANL Sequencing Process Enhancement: R&D Results

Owatha L. Tatum and P. Scott White, Center for Human Genome Studies – Joint Genome Institute, Los Alamos National Laboratory, Mail Stop M888, Los Alamos, NM 87545

The Center for Human Genome Studies at Los Alamos National Laboratory (LANL) as a member of the Joint Genome Institute (JGI) has an active sequencing research and development group. With recent dramatic increases in the Joint Genome Institute's sequencing effort, the need to improve efficiency and reduce costs while maintaining high quality standards has become of utmost importance. Aspects of Los Alamos sequencing R&D goals include improvements in sequencing chemistry, both in the form of modifications to existing methods and investigation into the development of new sequencing technologies and process automation systems.

Work is ongoing to improve each step of our sequencing process ranging from DNA template preparation to sequencing chemistry. The most aggressive effort has gone into modifications of commercially available sequencing chemistries. The results of this work have been significant reductions in overall cost with an improvement in read length and sequence quality as reflected by both a reduced IUB ambiguity and increase Q20 score. The protocols resulting from these R&D efforts have been implemented in the LANL production and finished sequencing efforts with great success. Data obtained from difficult templates (i.e. BAC DNA) have improved dramatically as a result of chemistry R&D as well. The LANL R&D group is currently working on novel methods for high throughput BAC end sequencing that is fully automatable and scalable from clone DNA purification to sequencing. While improvements in chemistry have had the most immediate impact on cost, LANL has also focused on quality control and automation issues to further streamline the sequencing process. Commercially available automation equipment has been integrated into the production process line with a considerable reduction in technician hands-on time. In addition to time and cost reduction, high throughput automated systems have also been implemented to improve quality control early in the sequencing process. R&D is ongoing to ensure that LANL will continue to take full advantage of its current instrumentation capacity while adapting the sequencing process to even further throughput increases, and currently includes an ABI 3700 capillary machine. Our sequencing capacity is increasing from approximately 5 Mb of high-quality finished sequence/year to more than double that for the coming year. Examples of improvements in data quality and process enhancement will be presented for each aspect discussed above with emphasis on methods unique to LANL.

Bioinformatics Concurrent Session II

Tuesday, September 21
3:00 – 5:00pm

Defining the Pathway Structure of Metabolic Genotypes through Convex Analysis

Christophe H. Schilling and Bernhard O. Palsson
Bioengineering, University of California-San Diego

Small genome sequencing and annotation are leading to the complete definition of metabolic genotypes in an increasing number of organisms. Proteomics is beginning to give insights into the use of the metabolic genotype under given growth conditions. These data sets provide the basis for systematically studying the genotype-phenotype relationship through *in silico* biology. We present a complete theory for the functional definition of biochemical pathways that describes the complete capabilities of metabolic networks. From principles of convex analysis the unique set of extreme pathways for a metabolic network can be determined. These pathways are stoichiometrically balanced pathways that are systemically independent. Together they can be combined in a non-negative manner to achieve every possible steady state flux distribution and hence metabolic phenotype that the network is capable of generating. Furthermore, we illustrate how to determine the pathway structure of a complete metabolic genotype and combine the pathway analysis with other quantitative analysis techniques for the study of metabolic networks. Through this conceptual framework a general classification scheme of pathways is developed based on systemic function as opposed to historical discovery. Systems science modeling efforts such as this will be important in defining, characterizing, and studying the genotype-phenotype relationship in the post-genomic era.

A Database Of Functionally Conserved Subfamilies Speeds Genome Annotation

Brendan Loftus, Delwood L. Richardson, Daniel Haft and Owen White. The Institute for Genomic Research (TIGR). 9712 Medical Center Drive, Rockville, MD 20850

We have developed a collection of hidden Markov Models (HMMs) for protein families predicted to share both function and evolutionary origin. Most of these families are hypothesized each to represent a set of functionally conserved subfamilies. These statistical models, based on curated alignments for clusters of proteins from completed genomes, are being used to discriminate functionally conserved subfamilies from other homologs and to achieve more sensitive and less error-prone protein identification than is possible by pairwise sequence comparison.

This database currently contains 500 families and is growing rapidly. Of these, 92% represent functionally conserved subfamilies, 4% domains, 2% superfamilies, and 3% paralog and subfamilies. The families are based upon a collection of completed genomes that was clustered based upon fasta3 and wu-blast searches. These clusters of related proteins are then analyzed based upon conservation of sequence similarity and the currently available annotation for each of the members. Clusters are split into groups that appear to have conserved functionality and representation among several genomes. We have done comparisons with the PFAM collection of HMMs and this will be discussed. A variety of web-based interactive-tools have been developed to speed construction of new subfamilies. The subfamilies will represent a protein family based entry into the Comprehensive Microbial Resource (CMR), a new database of all completed genomes that is being developed at TIGR. In addition, the entire dataset of HMMs will be searched against predicted proteins of newly sequenced genomes, allowing rapid recognition of new members of each group therein. A two stage score cutoff contains a trusted cutoff that can be used for automatic annotation and a noise cutoff that excludes spurious hits to other proteins nearby in sequence space. Recognition by an HMM facilitates assignment of EC number, biological role and curated names to new members of each subfamily.

Integrating Pathway Information with Gene Expression and Sequence Analysis

Michael Campbell, Fang-Fang Cai, Peter Covitz, Michael Mueller, and **Paul Thomas**, Molecular Applications Group, Inc. Palo Alto, CA

We have developed a statistical approach that integrates large scale gene expression analysis with biochemical and physiological pathway prediction. Our approach leverages a highly curated database of assignments of human genes to biological pathways that we created for this purpose. Using this method on microarray gene expression data from human cell lines we were able to predict – with statistical confidence measures - the pathway function of groups of transcriptionally co-regulated genes. These predictions were combined with exhaustive sequence analysis to produce testable hypotheses about gene function. We have automated several steps in our approach with the goal of facilitating rapid identification of novel functions for previously uncharacterized genes.

Confidence Scoring of DNA Base Calls with the Automated Trace Quality Assessment (ATQA) System

Max A. Karlovitz and Michael L. Catalano-Johnson, Daniel H. Wagner, Associates, Malvern, PA.

We describe the development of algorithms and software that assign *confidence scores* to DNA base calls by evaluating the quality of the fluorescent trace evidence from which the base call is derived. A confidence score corresponds to the probability that the associated base call is correct. In fact, our algorithms assign three distinct confidence scores to each base call corresponding to the probabilities of substitution, insertion, and deletion errors, respectively.

We present performance results for the ATQA system on sequence data provided by the Association of Biomolecular Resource Facilities (ABRF). We investigate the performance of the model as a function of the broad range of sequencing conditions that are represented in this data set.

We also present preliminary results on extensions of the ATQA system to the identification of heterozygosity and to consensus base call confidence scoring.

Supported by grants from NIST (50-DKNB-5-00118 and 50-DKNB-6-90158), from the NSF (DMI -9612376), and from NIH (1 R43 HG01848-01).

Finding the Information in DNA Chip Expression Data

Yixin Wang, Niam Saiti, ·Maggie Johns, Steve Madore, Steve Bulera, Jeffrey Thomas, Department of Molecular Biology, Department of Pathology and Experimental Toxicology, Parke-Davis Pharmaceutical Research, Warner-Lambert Company, 2800 Plymouth Road, Ann Arbor, MI

DNA chip technology has provided a chance to simultaneously monitor expression of thousands of genes. However, bioinformatics tools capable of extracting information from the data and revealing patterns relevant to biological functions are yet to be developed. We applied two statistical methods to analyze DNA chip expression data from rat liver samples treated with vehicle or several known hepatotoxicants. The selective expression test identifies genes whose expression changes are peculiar to the treatment of one compound. The clustering method uses hierarchical clustering algorithms to arrange genes according to similarity in pattern of gene expression and group genes responsive to the treatment of the specific compound. Our results indicate that the selected gene markers and/or the expression patterns observed in the experiments can potentially be used as indications of toxicity. Furthermore, clustering genes of known function with uncharacterized genes may provide leads to new gene functions that have not been extensively studied. The approach affords a new way to determine patterns of gene expression induced by well-characterized toxicants and use these patterns to study the compounds with unknown mechanisms of toxicity.

Genomics Concurrent Session II

Tuesday, September 21
3:00 – 5:00pm

Sequencing of a Plant Chromosome: The *Arabidopsis thaliana* Chromosome II Project

Kaul, S[1]., Lin, X[1]., Rounsley, S[1]., Shea, T.P[1]., Fujii, C.Y[1]., Mason, T[1]., Bowman, C.L[1], Barnstead, M. [1], Adams, M[1]., Feldblyum[1], T., Koo[1], H., Moffat[1], K., Cronin, L., Shen, M., Pai, G. [1], Van Aken, S. [1], Umayam, L. [1], Tallon, L. [1], Gill, J. [1], Ketchum, K.A. [1], Ronning,C.M[1]., Benito, M-I. [1], Carrera, A.J. [1], Creasy, T.H. [1], Buell, C.R. [1], Town, C.D. [1], Goodman, H.M,[2], Somerville, C.R. [3], Nierman, W.C. [1], White, O. [1], Eisen, J.A. [1], Salzberg, S. [1], Fraser,C.M. [1], Venter, J.C. [1], [1]The Instutute for Genomic Research, Rockville, MD, [2] Massachusetts General Hospital, Boston, MA, [3] Carnegie Insitution of Washington, Stanford, CA

TIGR has recently completed sequencing chromosome II of *Arabidopsis thaliana*. The chromosome is nearly 20 Mb in length, excluding the 4 Mb region of ribosomal DNA repeats at the north end, making it 40% longer than the size estimate derived from the YAC based physical map. The top arm is nearly 4 Mb in length and runs from the nucleolar organizing region to the northern boundary of the centromere, whereas the bottom arm is approximately 16 Mb in length and runs from the southern boundary of the centromere all the way to our southernmost clone.

Initially, several seed clones distributed along the chromosome were chosen for sequencing. Upon sequencing these seed clones, we used the BAC end sequence strategy to build our contigs. After the chromosome was nearly one-third complete, we built contigs using the fingerprint data made available by Washington University. By combining both BAC end sequences with the fingerprint data, we were able to build tiling paths for the entire chromosome and greatly increase our production.

Towards the latter part of the sequencing phase, we utilized the Perkin Elmer 3700 machines. Data from these machines accounted for nearly 15% of the entire chromosome sequence. We also made a number of technological advances in informatics which helped in data management, process control, and closure of BACs. In fact, by implementing these advances, we were able to decrease the average amount of time a clone stayed in finishing from 2 months to 8 days.

We have sequenced over 2 Mb of sequence which falls within the boundaries of the functional centromere. Though as expected, most of the DNA in this region consists of retroelements, we have identified a number of genes in this region. We have also confirmed an interesting evolutionary event. Within the centromeric region of chromosome II, we have discovered a large insertion of the mitochondrial genome. These, along with other exciting discoveries, will be discussed in our presentation.

Genotyping and Identification of Human Neurological Disorder Genes

Michael Mullokandov and **Jay Ji**, Biotech Research Labs, Boston Biomedica Inc., Gaithersburg, MD

A systematic approach has been established in our laboratory in attempt to localize the genes responsible for several inherited human neurological disorders. Several nuclear family pedigrees with various neurological disorder have been examined for the past two years. Experimental conditions for genome-wide mapping at 10-cM density were optimized with ABI microsatellite markers by adjusting PCR conditions and multiplexing amplification reactions. This approach allows us to reduce the total numbers of PCR required to produce a complete genome scan by four- folds. The amount of DNA needed for 10 cM genomtyping can be as little as 5 ug. Two-point and multi-point parametric, as well as non-parametric (NPL) linkage analysis, were performed for the resulting 20,000 genotyping data points using LINKAGE and GENEHUNTER programs. Putative candidate gene locations within 20 cM have been identified for one nuclear family, with a maximum LOD score value of 2.40 and NPL score of 3.5. Multi-point linkage analysis produced a maximum LOD score of 3.78 and NPL score of 8.81 at the same chromosomal position. Haplotype analysis provided further evidence for the disease gene location. In order to confirm and refine the gene location, 23 markers at 1 cM density around the region is under investigation.

Genomic libraries of interested genes have been isolated and sub-clones were constructed. A nested deletion method was used to provide downstream sequence analysis. This approach resulted in a minimal interference from complex genomic sequence, such as micro-satellite and mini-satellite repeats. The selected subclones from the nested deletion library were sequenced progressively further into the large fragment in steps of several hundred nucleotides with the same universal primer. Thus far over 20 nested deletion libraries have been constructed in the laboratory in the course of study. Our systematic efforts of genotyping, library screening and sequencing have significantly facilitated the progress in localization, identification and characterization of genes involved in neurological disorders.

Use of the Rat Gene Index to examine gene expression patterns from Src-transformed rat fibroblasts

Renae L. Malek[1], Qingbin Guo[2], Mauro Ruffy[2], Edison T. Liu[2], Ingeborg Holt[1], Ishwar Chandra[1], Feng Liang[1], Jonathan Upton[1], John Quackenbush[1], Richard Jove[3], Timothy J. Yeatman[3] and Norman H. Lee[1] [1]The Institute for Genomic Research, Rockville, MD 208502, [2]National Cancer Institute, Bethesda, MD 20892, [3]H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL 33612

We are developing the Rat Gene Index as a non-redundant gene resource which will contain information on gene identity, function, expression patterns as well as links to mapping sites and to orthologous genes in other species. The non-redundant transcript database was generated by assembling cleaned EST (Expressed Sequence Tags) and non-redundant ETs (Expressed Transcript) into TCs (Tentative Rat Consensus sequence). TCs were searched against a non-redundant amino acid database and assigned a putative identity where possible. ESTs not contained in a TC were searched individually. The current RGI Release, Version 2.0, contains almost 90,000 ESTs, which assemble into over 26,000 singletons and almost 14,000 TCs. Role categories were assigned to estimate the number of genes represented by a variety of cellular and organismal functions. We identified several hundred TCs with no significant homology matches. We have used the Index to identify tissue specific TCs and have confirmed the tissue specific expression of several hypothetical and unknown proteins by Northern analysis. We compared gene expression patterns between Src-transformed rat fibroblasts using a glass array containing cDNAs from over 5000 clones selected from the Rat Gene Index.

Mapping And Sequencing The ~200-Kb Segment Of Chromosome 17p13 Containing The Nephropathic Cystinosis Gene

Jeff W. Touchman,[1] Yair Anikster,[2] Valerie V. Braden,[3] Nicole L. Dietrich,[1] Gerry G. Bouffard,[1] Steve M. Beckstrom-Sternberg,[1] William A. Gahl,[2] and Eric D. Green[1,3] [1]NIH Intramural Sequencing Center, [2]National Institute for Child Health and Development, and [3]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892

Nephropathic cystinosis is an autosomal recessive disorder caused by the defective lysosomal transport of cystine. The causative gene (*CTNS*), which resides on human chromosome 17p13, was recently identified by a positional cloning strategy. *CTNS* encodes an integral membrane protein called cystinosin. Analysis of affected patients has revealed that many of the responsible mutations are deletions, including one >55 kb in size that represents the most common cystinosis allele encountered to date.

In an effort to determine the precise genomic organization of *CTNS* and to gain sequence-based insight about the DNA within and flanking cystinosis-associated deletions, we sought to map and sequence the region of human chromosome 17p13 encompassing *CTNS*. Specifically, a bacterial artificial chromosome (BAC)-based physical map spanning ~1 Mb that includes *CTNS* was constructed by sequence-tagged site (STS)-content mapping. The resulting BAC contig provides order information for 43 STSs. Two overlapping BACs that together contain all of the *CTNS* exons as well as extensive amounts of flanking DNA were selected and subjected to systematic shotgun sequencing.

A total of ~200 kb of contiguous, highly accurate sequence was generated. Analysis of the resulting data has allowed the intron-exon organization of *CTNS* to be established and provided precise measurements of *CTNS* intron sizes. The sequence has also provided valuable information about the genomic segments commonly deleted in cystinosis as well as information about five genes that neighbor *CTNS*. One of these genes, encoding a novel carbohydrate kinase (*CARKL*), resides within the region most commonly deleted in cystinosis. The latter finding is particularly interesting, since it indicates that there is a second protein (the novel carbohydrate kinase) that is completely absent in 44% of cystinosis patients.

Global profiling of gene expression during maturation of human antigen presenting cells.

Akbar S. Khan, Mechelle D. Lesher, Kamal U. Saikh and Robert G. Ulrich. Laboratory of Molecular Immunology, Army Medical Research Institute of Infectious Diseases, 1425 Porter Street, Frederick, MD 21702.

Dendritic cells degrade and present antigenic fragments of pathogens to T lymphocytes in response to primary immune assaults. These antigen-presenting cells differentiate from precursor cells that are activated by inflammatory stimuli. We analyzed the changes in gene transcription that occur during maturation of these cells, by using gene array profiling. Comparisons were made between precursors and dendritic cells that were induced by treatment with granulocyte-macrophage colony-stimulating factor (GM-CSF). In addition, the antigen presentation competency of these dendritic cells was greatly enhanced by treatment with tumor necrosis factor-a, (TNF-a), and diminished by interleukin-10 (IL-10) treatment. TNF-a enhanced antigen presentation by stabilizing cell-surface expression of HLA-DR, while IL-10 diminished class II expression at transcriptional level. Further analysis with a cDNA array of known genes confirmed that a complex profile of gene expression occurred in response to each of these cytokine signals. An 11,366 EST-component, non-redundant array was next used to measure global transcription levels after each cytokine treatment. IL-10 induced a dramatic increase in expression of several factors associated with apoptosis, such as caspase 9 and 10. Furthermore, this analysis identified differential expression of multiple genes of signaling pathways, transcriptional factors, cell adhesion molecules, as well as genes of unknown function. These results record the complex nature of the molecular events that occurred during the transformation of a quiescent precursor cell to a new functional phenotype.

Human BAC Ends

Shaying Zhao, Joel Malek, Gregory Mahairas, Lily Fu, William Nierman, J. Craig Venter, and Mark D. Adams The Institute for Genomic Research, Rockville, MD, and Department of Molecular Biotechnology, University of Washington, Seattle, WA

End sequences from Bacterial Artificial Chromosomes (BACs) have been playing critical roles in large scale genomic sequencing projects: 1) selecting clones for sequencing, 2) validating, joining and ordering contigs, and 3) building genome assembly scaffolds. To date, we have generated >300,000 end sequences from >180,000 human BAC clones with an average read length 460 bp for a total of 140 MB covering ~4.6% of the genome. Over 60% of the clones have BAC end sequences (BESs) from both ends representing >5.5X coverage of the genome by the paired ends clones. Our quality assessments and sequence analyses indicate that BESs from human BAC libraries developed at California Institute of Technology (CalTech) and Roswell Park Cancer Institute (RPCI) have similar properties. The analyses also indicate that BESs generated by The Institute for Genomic Research (TIGR) and the University of Washington (UW) are sufficiently accurate for use in both building the minimum tiling path of sequence-ready clones across the genome and building genome assembly scaffolds. The annotation results of BESs for the contents of available genomic sequences, sequence tagged sites (STSs), expressed sequence tags (ESTs), protein encoding regions and repeats indicate that this resource will be valuable in many areas of genome research.

Technology Concurrent Session II

Tuesday, September 21

3:00 – 5:00pm

Three-Dimensional Arrays of Microtransponders Derivatized with Oligonucleotides

Wlodek Mandecki, Steven Gray and Eileen Ernst. PharmaSeq, Inc., Monmouth Junction, NJ

In the assay, DNA molecules immobilized on the microtransponders hybridize with target DNA, and each target DNA molecule is labeled with a fluorophore. A custom detection system allows us to monitor the binding of the target DNA to the probe on the solid phase microtransponder by measuring emitted fluorescence.

The targetted dimensions of the transponder are 250 x 250 x 250 microns. These tiny chips are manufactured as elements of a silicon wafer. The basic function of the transponder is to encode the serial number that identifies the DNA molecule attached chemically to the surface of the transponder. The transponder is activated by light, which can come from a laser or another light source. The light serves two purposes. First, it activates the electronic circuits within the transponder. This causes the transponder to emit radio waves carrying the serial numbers of the transponder and the sequence of the probe. The light also activates fluorogenic molecules on the surface of the transponder, and the detector measures the fluorescence emanating from the surface of the transponder. These measurements are done in a flow chamber of a flow fluorometer, at a significant rate of speed, and a series of measurements of the serial numbers and fluorescence is made automatically.

The technology is ideal for assays in which screening for several genes, gene fragments or mutations is necessary. Presently PharmaSeq is working to optimize the hardware and the assay methodologies.

Earle-Hughes, Julie	P-023	Analysis Of Cell Cycle-Regulated Gene Expression Patterns In A Human Neuroblastoma Cell Line Using High Density cDNA Microarrays.
Rodriguez-Tomé, Patricia	P-024	Corba Servers And Services At EBI
FitzGerald, Micheal G.	P-025	A Sequence Variation Discovery Platform
George, Reed A.	P-026	Technologies For Finishing The Drosophila Genome Sequence
Gharizadeh, Baback	P-027	Screening A Selected cDNA Library By Pyrosequencing
Goldsborough, Mindy D.	P-028	Room Temperature Archiving Of Plasmid Clones In An Automatable 96-Well Format
Goldsborough, Mindy D.	P-029	A Paper-Based System For The Collection, Archiving, Purification And Analysis Of DNA Samples
Goo, Young Ah	P-030	Analysis On Gene Expression Patterns In Halobacterium Halobium Using Of DNA Samples
Cowles, S.	P-031	Integrated Monitoring And Control Of A 7x24 High Throughput Sequencing Operation
Guettler, Robert D.	P-032	High-Capacity Oxygenated Microwell Plate Shaker
Guettler, Robert D.	P-033	High-Throughput 96-Channel DNA Synthesis
Hahnenberger, Karen M.	P-034	Integrated Genomic/ Proteomic Analysis Of Adenoviral Gene Therapy Vectors
Burkhart-Schultz, K.	P-035	Progress Towards Sequencing The Genome Of The Nitrifying Bacterium, N. Europaea
Hartsch, Thomas	P-036	Sequencing The Whole Genome Of Thermus Thermophilus HB27
Gay, Cheryl	P-037	Identification Of Genes Involved In Colon Cancer Metastasis Using CDNA Microarrays.,
Hellwig, Randolph J.	P-038	Development And Validation Of Automated High Throughput Plasmid Purification Systems
Holmberg, Anders	P-039	A Novel Automated Workstation For Large-Scale Sequence Reaction Cleanup
Hunicke-Smith, Scott	P-040	Performance Of The Genemachines Omnigrid Microarrayer For Printing DNA Chips For Gene Expression Analysis.
Iisley, Diane	P-041	Optimizing Fluorescent Dye-Nucleotide Incorporation Into DNA Probes For CDNA Microarrays
Jiang, Lingxia	P-042	Sequencing Through Large Gaps By Utilizing Micro-Libraries And Transposon-Mediated Libraries
Fischer, H.P.	P-043	Identification Of Antibacterial Targets Using Large Scale Genome Comparisons
Joubert, F.	P-044	The Use Of Homology-Modeled Malaria Proteins For Ligand Discovery
Ju, Jingyue	P-045	Novel Fluorescent Reagents And Solid Phase Sequencing Chemistry For Genetic Analysis
Kanamori, Hiroyuki	P-046	Rice Genome Sequencing Project: Sequencing Of Rice Genomic DNA
Koo, Hean L.	P-047	Application Of New Strategies To Increase Efficiency Of BAC Closure In Arabidopsis Thaliana Chromosome II
Kozyavkin, S.	P-048	Chemically Modified Primers For Advanced DNA Sequencing
Lachenmeirer, Eric	P-049	A Fully Integrated cDNA Sequencing, Clone Management, And Microarray Fabrication Process

Index to Posters

Wier, Lawrence	P-001	Hybrigel Purification: A Novel Technique For Accelerated Preparation Of DNA Sequence Products For Capillary Electrophoresis And Multiplexing
Adams, Pamela S.	P-002	ABRF Roundtable Discussion On Current Technology And Methodology In DNA Sequencing Laboratories
Ahmadi, Ali R.	P-003	The Industrialization Of Genomics: Current Status And Future Potential
Ahmadian, Afshin	P-004	Single Nucleotide Polymorphism Analysis By Pyroseqencing
Alderborn, Anders	P-005	Reliable SNP Determination Using Real-Time Pyrosequencing
Ansorge, W.	P-006	5000 Bases In A Single Sequencing Reaction On ARAKIS, Towards 1 Mega Base DNA Sequencer With Automated Loading
Aoki, Hiroyoshi	P-007	The Rice Genome Sequencing Project: Construction Of Shotgun Library
Batchelder, Keith A.	P-008	A New Price/Performance Point For Plaque Pickers
Benito, Maria -Ines	P-009	Analysis Of Transposons And Other Repeat Elements In The Complete Sequence Of Arabidopsis Thaliana, Chromosome II
Bevirt, JoeBen	P-010	Automating The Genomics Revolution
Bjerke, Michael P.	P-011	Automated Fluorescent Sequencing Reaction Cleanup Using Magnesil Paramagnetic Particles
Brasch, Michael Dr.	P-012	GATEWAY Cloning Technology: A High-Throughput Gene Transfer Technology For Functional Analysis And Protein Expression
Buice, Carl U.	P-013	A High-Throughput, Plasmid Template Preparation Machine At $0.05 Per Prep.
Carpenter, Liane	P-014	Sequencing Of The M. Tuberculosis Genome; Comparison Of A Recent Clinical Isolate With The Laboratory Strain
Ch'ang, Lan-Yang	P-015	Oncodevelopmentel Genes Of Human Hepatocellular Carcinoma Identified By The Complexity Reduction Analysis
Chen, F.	P-016	Sequencing In The Proximal Half Of The Q-Arm Of Human Chromosome 22, And Syntenic Regions Of Mouse Chromosomes 6, 10 And 16
Clancy, Yevette C.	P-017	Octamer Sequencing In A Core DNA Sequencing Facility
Clark, Doni J.	P-018	Increased DNA Sequencing Throughput On The CEQ 2000 DNA Analysis System
Cline, Robin T.	P-019	Analysis Of Competence Genes From S. Pneumoniae Using DNA Microarrays
Conaway, Concetta A.	P-020	Automated, High-Throughput Analysis Of A Rice Genomic SNP Correlated With Amylose Content And Processing Quality
Dogget, Norman A.	P-021	Draft And Finished Sequencing Of Human Chromsome 16
Durrant, Ian	P-022	Recent Technical Advances Leading To Improved Performance Of Microarry Based Expression Analysis

Earle-Hughes, Julie	P-023	Analysis Of Cell Cycle-Regulated Gene Expression Patterns In A Human Neuroblastoma Cell Line Using High Density cDNA Microarrays.
Rodriguez-Tomé, Patricia	P-024	Corba Servers And Services At EBI
FitzGerald, Micheal G.	P-025	A Sequence Variation Discovery Platform
George, Reed A.	P-026	Technologies For Finishing The Drosophila Genome Sequence
Gharizadeh, Baback	P-027	Screening A Selected cDNA Library By Pyrosequencing
Goldsborough, Mindy D.	P-028	Room Temperature Archiving Of Plasmid Clones In An Automatable 96-Well Format
Goldsborough, Mindy D.	P-029	A Paper-Based System For The Collection, Archiving, Purification And Analysis Of DNA Samples
Goo, Young Ah	P-030	Analysis On Gene Expression Patterns In Halobacterium Halobium Using Of DNA Samples
Cowles, S.	P-031	Integrated Monitoring And Control Of A 7x24 High Throughput Sequencing Operation
Guettler, Robert D.	P-032	High-Capacity Oxygenated Microwell Plate Shaker
Guettler, Robert D.	P-033	High-Throughput 96-Channel DNA Synthesis
Hahnenberger, Karen M.	P-034	Integrated Genomic/ Proteomic Analysis Of Adenoviral Gene Therapy Vectors
Burkhart-Schultz, K.	P-035	Progress Towards Sequencing The Genome Of The Nitrifying Bacterium, N. Europaea
Hartsch, Thomas	P-036	Sequencing The Whole Genome Of Thermus Thermophilus HB27
Gay, Cheryl	P-037	Identification Of Genes Involved In Colon Cancer Metastasis Using CDNA Microarrays.,
Hellwig, Randolph J.	P-038	Development And Validation Of Automated High Throughput Plasmid Purification Systems
Holmberg, Anders	P-039	A Novel Automated Workstation For Large-Scale Sequence Reaction Cleanup
Hunicke-Smith, Scott	P-040	Performance Of The Genemachines Omnigrid Microarrayer For Printing DNA Chips For Gene Expression Analysis.
Iisley, Diane	P-041	Optimizing Fluorescent Dye-Nucleotide Incorporation Into DNA Probes For CDNA Microarrays
Jiang, Lingxia	P-042	Sequencing Through Large Gaps By Utilizing Micro-Libraries And Transposon-Mediated Libraries
Fischer, H.P.	P-043	Identification Of Antibacterial Targets Using Large Scale Genome Comparisons
Joubert, F.	P-044	The Use Of Homology-Modeled Malaria Proteins For Ligand Discovery
Ju, Jingyue	P-045	Novel Fluorescent Reagents And Solid Phase Sequencing Chemistry For Genetic Analysis
Kanamori, Hiroyuki	P-046	Rice Genome Sequencing Project: Sequencing Of Rice Genomic DNA
Koo, Hean L.	P-047	Application Of New Strategies To Increase Efficiency Of BAC Closure In Arabidopsis Thaliana Chromosome II
Kozyavkin, S.	P-048	Chemically Modified Primers For Advanced DNA Sequencing
Lachenmeirer, Eric	P-049	A Fully Integrated cDNA Sequencing, Clone Management, And Microarray Fabrication Process

Index to Posters

Law, Scott M.	P-050	Using The Invader Squared FRET Assay On A Panel Of Familial Genomic Samples To Detect Single Base Changes At Multiple Loci
LeBlanc, James F.	P-051	A New Rapid Method For Proteomic Profiling Using SELDI Proteinchip Arrays Demonstrated On Bacterial Lysates And Cultured Human Glial Cells.
Lou, Yunian	P-052	Basecalling Software Capcall Inside DNA Sequencing Assembly Pipeline At JGI/LBNL
Lukacs, Morten F,	P-053	Automated Method For Purification Of Dye Terminator Sequencing Products
Lundeberg, Joakim	P-054	A Large-Scale Strategy For Multiplex Sequencing
Malek, Joel A.	P-055	Developments In High Throughput BAC Ends Sequencing
Malykh, A.	P-056	Solutions For Sequencing Through Difficult Regions In BAC DNA Templates. Thermofidelase II For High- Throughput Sequencing
Mamone, Anthony J.	P-057	High Throughput Sequencing With The Megabace 1000 System Today
McArdle, Bernard F.	P-058	The Megabace 1000 Methods Development Projects: Analysis Of The Factors Affecting Success And Read Length
Marziali, A	P-059	T2: An Automated, Sub-Microliter Thermal Cycler For High Throughput DNA Sequencing
Matsumoto, Takashi	P-060	Current Status Of Rice EST Mapping
Gill, John	P-061	Utilization Of The ABI 3700 DNA Analyzer For The Sequencing Of Multiple Genomes
Moeckel, Bettina	P-062	Genome Sequencing Of Industrial Microrganisms: The Corynebacterium Glutamicum ATCC 13032 Genome Project
Moffat, Kelley S.	P-063	The Effects Of Oxidative DNA Damage On A Novel Repeat-Specific DNA-Binding Activity In Deinococcus Radiodurans
Monforte, Joseph	P-064	High-Throughput, Multiplexed SNP Genotyping By Mass Spectrometry
Nomura, Nobura	P-065	Complete Sequence Analysis Of 1,500 Human CDNA Clones Harboring Long And Nearly Full-Length Inserts
Paolella, Mary Jane	P-066	Introducing Automated DNA Sequencing Into The American High School Biology Laboratory
Pence, William H.	P-067	Introduction Of The ACAPELLA-5K Automated Fluid Sample Handling System
Qin, Haiying	P-068	Novel Approuches To Closing Repeated And Difficult Regions Of Microbial Genomes
Quinn, Thomas	P-069	The Genomic Species Concept
Reynolds, Mark A.	P-070	DNA Binding Metrics For Microarray Applications
Rhodes, Richard B	P-071	Automated, Multiplex Detection Of Genetically Modified Organisms (Gmos) On Your Grocery Shelves
Roach, Jared C.	P-072	Quantitation Of The Inefficiencies Of Closing Gaps By Walking
Roe, Bruce A.	P-073	The Sequence Of Human Chromosome 22
Russo, James J.	P-074	Mapping, Sequencing And Analysis Of Legionella Pneumophila Genome
Scholler, Patrik	P-075	Directed Minimal Sequencing For Total Genome With Minimized Sequencing Efforts
Schubert, Frank	P-076	Dye Terminator Removal And Sequencing Reaction Cleanup - Utilization Of Superparamagnetic Particles

Poster Session Abstracts

Posters on display
September 19 – 21
1:00 pm – 3:00 pm
Monaco and Lorraine Rooms

P-001
Hybrigel Purification: A Novel Technique for Accelerated Preparation of DNA Sequence Products for Capillary Electrophoresis and Multiplexing

Lawrence Weir, Rahul K. Dhanda, T Christian Boles, Christopher P Adams, Mosaic Technologies, Boston, MA

Current protocols for DNA sequencing require substantial purification of sequence products prior to automated analysis. To obtain high quality data by capillary electrophoresis, sequencing products are usually ethanol precipitated to remove unincorporated dNTPs and salt. We have developed a new electrophoretic technique for purifying the products of DNA sequencing reactions based on Mosaic's Acrydite™ technology. Oligonucleotide probes modified with 5'-acrylamide groups are copolymerized within polyacrylamide electrophoresis gels. During electrophoresis, complementary targets are themselves immobilized whereas non-complementary molecules migrate through the gel unimpeded. Capture probes were designed whose sequences were complementary to products of Mp18 sequencing reactions. These probes efficiently capture sequencing products in miniature Acrydite™ purification devices. The captured material may then be released at higher voltages and retained in a collection chamber bound by a semi-permeable membrane. Template DNA remains trapped in the gel while primers pass into the buffer reservoir. This procedure also enables a multiplexing strategy whereby multiple sequencing reactions performed in the same vessel may be easily separated using specific capture probes. The procedure is ideally suited for automation using the microtitre format.

P-002
ABRF Roundtable Discussion on Current Technology And Methodology in DNA Sequencing Laboratories

Pamela S. Adams[1], Susan Hardin[2], Theodore Thannhauser[3], Dina Leviten , Duane Bartley , and George Grills ., Trudeau Institute, Saranac Lake, NY; University of Houston, Houston, TX; Cornell University, Ithaca, NY; ICOS Corporation, Bothell, WA; Johns Hopkins University, Baltimore, MD; Albert Einstein College of Medicine, Bronx, NY; Association of Biomolecular Resource Facilities.

The resources and experiences that are shared among core resource laboratories are the basis for the Association of Biomolecular Resource Facilities (ABRF). Examples of this cooperation, including the current study on the effects of different instrumentation and methods on the quality of results from sequencing a common standard template and two difficult GC-rich templates will be discussed. Both common and new technologies were analyzed on the basis of accuracy and quality. The role of the new 3700 DNA Analyzer in core labs and the challenges of sample throughput, data handling and software issues of this new technology will be discussed within the limits allowed by its recent introduction. Although new technologies seem to improve data generation, methods for manipulating protocols and reagents using the ABI 377 which allow read lengths of 1000 bases at near 100% accuracy will be presented. A brief overview of the ABRF will be included.

P-003
The Industrialization of Genomics: Current Status and Future Potential

Ali R. Ahmadi, The Automation Partnership, Royston, Herts, UK

Burgeoning demands on existing genomics facilities have led genomics to enter an "industrial" age. An industrial approach is needed for sample storage and processing, data tracking and operations management. A step-change in the application of automation from bench to industrial scale is required to facilitate this. This step-change has already been implemented in other areas of drug discovery, for example high throughput screening. However genomics has so far largely failed to import the existing expertise from other disciplines.

We will demonstrate the differences between conventional and industrial-scale research. Using case studies from our collaborations with genomics companies and institutes, we will compare alternative methods for high throughput genotyping, hybridization, sequencing and protein analysis. Finally we will discuss the feasibility of a fully-industrialized genomics research facility.

P-004
Single nucleotide polymorphism analysis by pyrosequencing

Afshin Ahmadian, Baback Gharizadeh, Anna Gustafsson, Fredrik Sterky, Mathias Uhlén and Joakim Lundeberg, Department of Biotechnology, The Royal Institute of Technology (KTH), SE-100 44 Stockholm, Sweden

As the sequencing of the human genome becomes complete, characterization of genetic variation will be increasingly important. Genetic variation is the basis of human diversity, and efforts aim to correlate this variability with human genetic disease. The most common type of genetic variation is single nucleotide polymorphisms (SNPs). We have investigated the possibility of typing SNPs by using a recently developed technique called pyrosequencing. Pyrosequencing is a sequencing-by-synthesis method in which a cascade of enzymatic reactions yields detectable light which is proportional to incorporated nucleotides. One feature of typing SNPs with pyrosequencing is that each allelic variant will give a unique sequence compared to the two other variants. These variants can easily be distinguished by a simple pattern recognition software. The software displays the allelic alternatives and allows for direct comparison with the pyrosequencing raw-data. For optimal determination of SNPs, various protocols of nucleotide dispensing order was investigated. Here, we demonstrate that typing of SNPs can easily be performed by pyrosequencing. In addition, an automated system for parallel processing of 96 samples in 5 minutes exists suiteable for large scale screening and typing of SNPs. The application of SNP pyrosequencing on a microarray platform will be described.

Index to Posters

Law, Scott M.	P-050	Using The Invader Squared FRET Assay On A Panel Of Familial Genomic Samples To Detect Single Base Changes At Multiple Loci
LeBlanc, James F.	P-051	A New Rapid Method For Proteomic Profiling Using SELDI Proteinchip Arrays Demonstrated On Bacterial Lysates And Cultured Human Glial Cells.
Lou, Yunian	P-052	Basecalling Software Capcall Inside DNA Sequencing Assembly Pipeline At JGI/LBNL
Lukacs, Morten F,	P-053	Automated Method For Purification Of Dye Terminator Sequencing Products
Lundeberg, Joakim	P-054	A Large-Scale Strategy For Multiplex Sequencing
Malek, Joel A.	P-055	Developments In High Throughput BAC Ends Sequencing
Malykh, A.	P-056	Solutions For Sequencing Through Difficult Regions In BAC DNA Templates. Thermofidelase II For High- Throughput Sequencing
Mamone, Anthony J.	P-057	High Throughput Sequencing With The Megabace 1000 System Today
McArdle, Bernard F.	P-058	The Megabace 1000 Methods Development Projects: Analysis Of The Factors Affecting Success And Read Length
Marziali, A	P-059	T2: An Automated, Sub-Microliter Thermal Cycler For High Throughput DNA Sequencing
Matsumoto, Takashi	P-060	Current Status Of Rice EST Mapping
Gill, John	P-061	Utilization Of The ABI 3700 DNA Analyzer For The Sequencing Of Multiple Genomes
Moeckel, Bettina	P-062	Genome Sequencing Of Industrial Microrganisms: The Corynebacterium Glutamicum ATCC 13032 Genome Project
Moffat, Kelley S.	P-063	The Effects Of Oxidative DNA Damage On A Novel Repeat-Specific DNA-Binding Activity In Deinococcus Radiodurans
Monforte, Joseph	P-064	High-Throughput, Multiplexed SNP Genotyping By Mass Spectrometry
Nomura, Nobura	P-065	Complete Sequence Analysis Of 1,500 Human CDNA Clones Harboring Long And Nearly Full-Length Inserts
Paolella, Mary Jane	P-066	Introducing Automated DNA Sequencing Into The American High School Biology Laboratory
Pence, William H.	P-067	Introduction Of The ACAPELLA-5K Automated Fluid Sample Handling System
Qin, Haiying	P-068	Novel Approuches To Closing Repeated And Difficult Regions Of Microbial Genomes
Quinn, Thomas	P-069	The Genomic Species Concept
Reynolds, Mark A.	P-070	DNA Binding Metrics For Microarray Applications
Rhodes, Richard B	P-071	Automated, Multiplex Detection Of Genetically Modified Organisms (Gmos) On Your Grocery Shelves
Roach, Jared C.	P-072	Quantitation Of The Inefficiencies Of Closing Gaps By Walking
Roe, Bruce A.	P-073	The Sequence Of Human Chromosome 22
Russo, James J.	P-074	Mapping, Sequencing And Analysis Of Legionella Pneumophila Genome
Scholler, Patrik	P-075	Directed Minimal Sequencing For Total Genome With Minimized Sequencing Efforts
Schubert, Frank	P-076	Dye Terminator Removal And Sequencing Reaction Cleanup - Utilization Of Superparamagnetic Particles

Sengoku, Eiichi	P-077	An Effcient Method For BAC, PAC And P1 DNA Purification
Shah, Nila	P-078	Identification Of Thousands Of Single Nucleotide Polymorphisms (SNPS) In The Human Genome
Shaw-Bruha, Carla M.	P-079	Ion-Pair Reversed-Phase HPLC (IP-RP-HPLC) Approach For Cloning Of PCR Products And Colony Screening
Shultz, John W.	P-080	Specific DNA Measurement Using A New, Enzyme-Based System, The READIT System
Silk, Chris	P-081	Further Upgrades To The ABD 377 Sequencer For High Throughput DNA Sequencing And Fragment Analysis
Sinibaldi, Ralph	P-082	Covalent Attachment Of Sequence-Optimized PCR Products For DNA Microarrays.
Slatko, Barton E.	P-083	The Filarial Genome Project: Brugia Malayi Contains An A-Proteobacteria Endosymbiont
Smith, Rick	P-084	Changing Gel Based SNP Assays To Non-Gel Based, Digital Read Out Assays
Springer, Amy L.	P-085	A Novel Chemical Affinity System For Cleanup Of Cycle Sequencing Reactions
Shang, J.	P-086	Sequencing Human DNA At The Stanford Human Genome Center
Tan, Ruoying	P-087	Sequential Double-Priming; A Highly Efficient Method To Generate 5'-Biased cDNA Libraries
Thannhauser, Theodore	P-088	Analysis Of The Effects Of Different DNA Sequencing Methods On Accuracy And Quality And Expansion Of A Web-Based Sequencing Resource: Results Of The ABRF DNA Sequencing Group 1999 Study.
Wakamatsu, Ai	P-089	Analysis Of cDNA Clones From Oligo-Cap Libraries And The Improved Oligo-Capping Method For Cloning Full Length cDNA Clones
Vamathevan, Jessica J.	P-090	Complete Sequencing Of The Ribosomal RNAS In Small Genomes
Volckaert, Guido	P-091	Trouble-Shooting In Genome Sequencing Projects, And The Construction Of Temperature-Sensitive Mutant Alleles For Functional Analyses Of The Saccharomyces Cerevisiae Genome.
Wang, Bruce	P-092	DNA Microarray Probe Analysis And Quality Control
Wang, Haiyang	P-093	Semi-Automated Solid-Phase Sequencing Reaction Purification
Wang, Yiwen	P-094	ABI PRISM 3700 DNA Analyzer: A Fully Automated And High-Throughput System For Genescan Applications
Wheeler, David I.	P-095	Optimized Polyacrylamide Gel Matrix For High Throughput DNA Sequencing
Wiemann, Stefan	P-096	Sequencing And Analysis Of Full Length cDNAs In The Course Of The German Genome Project
Wiest, Debra	P-097	Development Of A cDNA Microarry Manufacturing Platform Utilizing A Thermal Ink Jet Deposition System
Wolber, Paul K.	P-098	Iterative Optimization Of Oligonucleotide Arrays That Measure Gene Expression In Saccharomyces Cerevissiae
Womack, Andrew W.	P-099	Contamination Of Human Genomic Libraries Used For Large Scale Sequencing By E. Coli IS186 Insertion Elements.
Wong, Lily Y.	P-100	HPLC Purification Of Differentially Expressed Gene Fragments
Xu, Lisha	P-101	Reaction Additive Improves Sequencing Through Difficult Templates

Index to Posters

Poster Session Abstracts

Posters on display
September 19 – 21
1:00 pm – 3:00 pm
Monaco and Lorraine Rooms

P-001
Hybrigel Purification: A Novel Technique for Accelerated Preparation of DNA Sequence Products for Capillary Electrophoresis and Multiplexing

Lawrence Weir, Rahul K. Dhanda, T Christian Boles, Christopher P Adams, Mosaic Technologies, Boston, MA

Current protocols for DNA sequencing require substantial purification of sequence products prior to automated analysis. To obtain high quality data by capillary electrophoresis, sequencing products are usually ethanol precipitated to remove unincorporated dNTPs and salt. We have developed a new electrophoretic technique for purifying the products of DNA sequencing reactions based on Mosaic's Acrydite™ technology. Oligonucleotide probes modified with 5'-acrylamide groups are copolymerized within polyacrylamide electrophoresis gels. During electrophoresis, complementary targets are themselves immobilized whereas non-complementary molecules migrate through the gel unimpeded. Capture probes were designed whose sequences were complementary to products of Mp18 sequencing reactions. These probes efficiently capture sequencing products in miniature Acrydite™ purification devices. The captured material may then be released at higher voltages and retained in a collection chamber bound by a semi-permeable membrane. Template DNA remains trapped in the gel while primers pass into the buffer reservoir. This procedure also enables a multiplexing strategy whereby multiple sequencing reactions performed in the same vessel may be easily separated using specific capture probes. The procedure is ideally suited for automation using the microtitre format.

P-002
ABRF Roundtable Discussion on Current Technology And Methodology in DNA Sequencing Laboratories

Pamela S. Adams[1], Susan Hardin[2], Theodore Thannhauser[3], Dina Leviten , Duane Bartley , and George Grills ., Trudeau Institute, Saranac Lake, NY; University of Houston, Houston, TX; Cornell University, Ithaca, NY; ICOS Corporation, Bothell, WA; Johns Hopkins University, Baltimore, MD; Albert Einstein College of Medicine, Bronx, NY; Association of Biomolecular Resource Facilities.

The resources and experiences that are shared among core resource laboratories are the basis for the Association of Biomolecular Resource Facilities (ABRF). Examples of this cooperation, including the current study on the effects of different instrumentation and methods on the quality of results from sequencing a common standard template and two difficult GC-rich templates will be discussed. Both common and new technologies were analyzed on the basis of accuracy and quality. The role of the new 3700 DNA Analyzer in core labs and the challenges of sample throughput, data handling and software issues of this new technology will be discussed within the limits allowed by its recent introduction. Although new technologies seem to improve data generation, methods for manipulating protocols and reagents using the ABI 377 which allow read lengths of 1000 bases at near 100% accuracy will be presented. A brief overview of the ABRF will be included.

P-003
The Industrialization of Genomics: Current Status and Future Potential

Ali R. Ahmadi, The Automation Partnership, Royston, Herts, UK

Burgeoning demands on existing genomics facilities have led genomics to enter an "industrial" age. An industrial approach is needed for sample storage and processing, data tracking and operations management. A step-change in the application of automation from bench to industrial scale is required to facilitate this. This step-change has already been implemented in other areas of drug discovery, for example high throughput screening. However genomics has so far largely failed to import the existing expertise from other disciplines.

We will demonstrate the differences between conventional and industrial-scale research. Using case studies from our collaborations with genomics companies and institutes, we will compare alternative methods for high throughput genotyping, hybridization, sequencing and protein analysis. Finally we will discuss the feasibility of a fully-industrialized genomics research facility.

P-004
Single nucleotide polymorphism analysis by pyrosequencing

Afshin Ahmadian, Baback Gharizadeh, Anna Gustafsson, Fredrik Sterky, Mathias Uhlén and Joakim Lundeberg, Department of Biotechnology, The Royal Institute of Technology (KTH), SE-100 44 Stockholm, Sweden

As the sequencing of the human genome becomes complete, characterization of genetic variation will be increasingly important. Genetic variation is the basis of human diversity, and efforts aim to correlate this variability with human genetic disease. The most common type of genetic variation is single nucleotide polymorphisms (SNPs). We have investigated the possibility of typing SNPs by using a recently developed technique called pyrosequencing. Pyrosequencing is a sequencing-by-synthesis method in which a cascade of enzymatic reactions yields detectable light which is proportional to incorporated nucleotides. One feature of typing SNPs with pyrosequencing is that each allelic variant will give a unique sequence compared to the two other variants. These variants can easily be distinguished by a simple pattern recognition software. The software displays the allelic alternatives and allows for direct comparison with the pyrosequencing raw-data. For optimal determination of SNPs, various protocols of nucleotide dispensing order was investigated. Here, we demonstrate that typing of SNPs can easily be performed by pyrosequencing. In addition, an automated system for parallel processing of 96 samples in 5 minutes exists suiteable for large scale screening and typing of SNPs. The application of SNP pyrosequencing on a microarray platform will be described.

P-005
Reliable SNP determination using real-time pyrosequencing

Anders Alderborn and Ulf Hammerling, PyroSequencing AB, Vallongatan 1, S-752 28 Uppsala, Sweden, Eurona Medical AB, 751 06 Uppsala, Sweden.

The feasibility of real-time pyrosequencing for characterization of pre-defined SNPs was investigated. This method is based on indirect luminometric quantification of released pyrophosphate, upon nucleotide incorporation on an amplified template. The employed technical platform comprises a highly automated sequencing instrument that allows the analysis of 96 samples within 10 minutes.

To test the capability of this method to discriminate between various allelic configurations, 11 SNPs within the Renin-Angiotensin-Aldosterone System (RAAS) were studied. In addition to alternative bases of each studied polymorphic position, 5 – 10 downstream bases were sequenced. Evaluation of pyrogram data was accomplished by comparison of peak heights, which are proportional to the number of incorporated nucleotides. Analysis of pyrograms, resulting from alternate allelic configurations for each addressed SNP, revealed a highly discriminating pattern. Homozygous samples produced clear-cut single base peaks in their respective expected position, whereas heterozygous counterparts were characterized by distinct half-height sized peaks in both allelic positions. Whenever one or several SNP-succeeding bases were identical to either of the allelic nucleotides, their signals were added to those of the SNP. This feature, however, did not influence SNP readability.

Despite the highly diversified nature of the selected SNPs within the test panel, all proved amenable to unambiguous determination using the real-time pyrosequencing approach.

P-006
5000 Bases In A Single Sequencing Reaction On ARAKIS, Towards 1 Mega Base DNA Sequencer With Automated Loading

J.Stegemann, R.Ventzki, H.Erfle, C. Schwager, K.Faulstich, V.Benes, J.Zimmermann, Y.Li and **W.Ansorge**, Functional Genomics and Proteomics, European Molecular Biology Laboratory,, Meyerhofstrasse 1, D-69117 Heidelberg, Germany

High throughput and low error rate are essential for fast and accurate DNA sequence determination in large-scale genome projects. The Automated DNA Sequencing System ARAKIS was upgraded to simultaneous sequencing of five different DNA templates, yielding an output of 5000 bases in a single sequencing reaction. Primers employed in the reaction are labelled template-specific with five different fluorescent dyes. Continuous excitation and detection is achieved by five lasers entered into the slab gel through a side window and five array detector rows placed behind the gel, continuously recording the emitted fluorescence. Spectral crosstalk between the five differently labelled samples is below detection limit (dyes, lasers and filters were selected for minimum spectral overlap). A data evaluation and analysis software package (LiTracker) was developed. Derived trace data are transferred to standard base-calling and analysis programs. The device can accommodate up to 80 cm long gels; on 60 cm long gels the evaluated resolution is up to 1300 bp. Simultaneous sample gel loading was developed using porous filter materials and robotics. Collaboration on the commercialisation of the ARAKIS system has been initiated with MWG Biotech.

The system has been applied to EU genome sequencing projects (Human, Arabidopsis, Drosophila), ESTs and cDNA German genome project, EURO-Image EST and cDNA project and to gene expression studies by SAGE.

The ARAKIS system is suitable for all common sequencing applications, random and directed strategies, primer walking in a finishing phase of a project, bi-directional Doublex sequencing (forward/reverse). Due to its large capacity and high accuracy, it is particularly efficient for full-length cDNA and SAGE projects.

Sequencing protocols have been developed to prepare reactions on five different templates simultaneously in one tube. A particularly advantageous way is to perform these reactions as doublex reactions on two double stranded DNA templates simultaneously, resulting in lower price per base.

New array detectors will allow the sequencing of up to 96 clones per dye and laser line, increasing the total capacity to 480 clones run simultaneously. Run twice a day, the potential raw data output of 1 Mega base per device is very competitive with capillary sequencers, because of the long reads, low error rates, resulting low redundancy and significantly lower cost per finished base pair.

The ARAKIS system was developed with a financial support from the European Union.

P-007
The Rice Genome Sequencing Project : Construction of Shotgun Library

Hiroyoshi Aoki, Hiroko Yamane, Masumi Iijima, Tomoko Ito, Kimihiro Terasawa, Yuichi Katayose, Yamamoto Kimiko, Takashi Matsumoto, Takuji Sasaki, Rice Genome Research Program, National Institute of Agrobiological Resources / Institute of the Society for Techno-innovation of Agriculture, Forestry and Fisheries, Tsukuba, Ibaraki, Japan

One of the major goals of the second phase of the Rice Genome Research Program (RGP) is to sequence the whole rice genome. At present, we concentrate our sequencing efforts on chromosomes 1 and 6 as part of the International Rice Genome Sequencing Project (IRGSP). In order to accelerate the sequencing process and generate high quality sequence data we developed a simple and efficient system for construction of shotgun libraries and preparation of sequencing templates.

A PAC clone ordered in chromosome 1 or 6 is purified by ultracentrifugation. Then the purified PAC DNA is sonicated and size-fractionated into 1.5-2 kb and 5.5-7 kb fractions by agarose gel electrophoresis for construction of 2 kb- and 5 kb-library, respectively. Size-selected DNA fragments are ligated to pUC vector and transformed into E. coli DH10B by electroporation. After checking the quality of the library, the shotgun clones are automatically transferred into 96-well microtiter plates and stored at -80C. The 2kb- and 5kb libraries are subjected to plasmid purification by the alkaline-SDS method. We prepare about 4,000 clones from the 2kb library and 1,000 clones from the 5kb library for sequencing using forward and reverse primers. Simple and cost-effective methods of preparing sequencing template will be explored further.

P-008
A New Price/Performance Point for Plaque Pickers

Keith A. Batchelder and Scott Hunicke-Smith, GeneMachines®, San Carlos, CA

The GeneMachines Gel-2-Well™ offers a new level of price versus performance ratio for plaque/colony picking by utilizing a unique approach of parallel processing picking tasks. The twenty picking needles of the G2W are arrayed circumferentially around a rotating platform. This approach allows picking, placing, washing and sterilization to occur simultaneously at each location. Innoculation is the rate-limiting step and typically requires about 1 second, encompassing 5 innoculation strokes. Overall picking speeds are dependent on plate types used. For 384 well output plates and Nunc Omni-Trays as input plates, picking rates of 2000 picks/hour have been achieved. The G2W has a total plate capacity of 84 input and output plates.

P-009
Analysis Of Transposons And Other Repeat Elements In The Complete Sequence Of *Arabidopsis Thaliana*, Chromosome II

Maria-Ines Benito, Xiaoying Lin, Todd H. Creasy, Ana J. Carrera, J. Craig Venter, Claire M. Fraser, The Institute for Genomic Research, Rockville, MD

As a member of the multinational *Arabidopsis* Genome Initiative, The Institute for Genomic Research (TIGR) has been responsible for the complete sequencing and annotation of chromosome II. The *Arabidopsis* genome like that of other higher eukaryotes consists of repetitive sequences intermixed with genes. Unlike many plants, *Arabidopsis* is considered to have a low repetitive DNA content (~25%). We have been able to assemble the complete sequence of chromosome II into two large contiguous segments of DNA, a feat that may not be possible in genomes with more repetitive DNA.

Intergenic sequences such as repetitive sequences not only provide a considerable technical challenge to sequencing teams, but they are equally challenging to annotation groups. Genome annotation efforts for most genome projects have focused on gene prediction and the assignment of protein function to coding sequence. As part of the chromosome II annotation effort we have begun to develop a classification system for transposable and repeat elements in the *Arabidopsis* genome. An analysis of the complete sequence of chromosome II allows us to assess the presence, distribution, and diversity of specific repetitive elements on a single chromosome. Data from this analysis will be presented.

P-010
Automating the Genomics Revolution

JoeBen Bevirt, G. Noah Brinton, Teresa Carabeo, Eric Goldfarb, Georgina Hodgson, Brad Krueger, Brett Krueger, Eric Lachenmeier, Robert Nail, Thorsten Obermaier, Eric Rollins, Jon Wagner, Kathy Warne, Incyte Pharmaceuticals, Palo Alto, CA

We stand at the threshold of a new millennium and the dawn of the genomics age. In our quest to understand the molecular basis of all life we are faced with the daunting and exciting task of processing billions of samples and trillions of data points. To facilitate processing this data better, faster, and cheaper than ever before, we have designed and built a set of ultra-high throughput automated tools. These tools give our manufacturing operations the capacity to sequence 20 million samples/year and to print 100 billion microarray elements per year.

Fully automated modular work cells were designed to perform all of the error prone tasks in a given process. Advanced robots were developed to automate the transport of plates and slides from one station to another. Modular plate stackers and slide cassettes were designed to provide a secure, effective, and accurate method for humans to transport samples to and from the automated systems. Integral barcode reading and application was implemented to ensure robust sample tracking through every step of the process. Customized pipetting equipment was developed to allow small volumes of liquid reagents to be handled accurately. These systems are part of our continuing quest at Incyte to increase throughput and data quality while reducing cost and thus to provide more comprehensive and powerful data and tools to our customers.

P-011
Automated Fluorescent Sequencing Reaction Cleanup using MagneSil™ Paramagnetic Particles

Michael P. Bjerke, Craig E. Smith and Paul Otto, Promega Corporation, Madison, WI

Novel MagneSil™ Paramagnetic Particles exhibit superior capacity for purifying DNA over conventional silica resins and membranes, resulting in increased quality and yield. Here we describe the use of MagneSil to purify fluorescent sequencing reactions.

Two commonly used methods to purify fluorescent sequencing reactions are ethanol precipitation and column filtration. Both of these methods involve centrifugation steps, which are difficult to automate. We have developed an rapid method to purify dye terminator (including BigDye®) and dye primer fluorescent sequencing reactions using MagneSil. Compared to ethanol precipitation and Princeton Centri-Sep Separation Spin Columns, this purification method provides comparable sequence quality and accuracy, yet reduces processing time by half. The use of MagneSil allows low cost, fully automated systems to be developed on common robotic platforms such as the Beckman Biomek.

P-012
GATEWAY Cloning Technology: A High-Throughput Gene Transfer Technology for Functional Analysis and Protein Expression

Dr. Michael Brasch, Senior Scientist, Dept of Gene Expression and Protein Analysis Life Technologies Inc., Rockville, MD

As a result of numerous ongoing genome sequencing projects, large numbers of candidate open reading frames are being identified, many of which have no known function. The analysis of these genes typically involves transfer of various DNA segments into a variety of vector backgrounds for protein expression or functional analysis. We describe a method called Recombinational Cloning (RC) that uses *in vitro* site-specific recombination to promote the transfer of DNA segments between vector backbones. This approach can also be applied to the efficient, directional cloning of PCR products. Such cloned

PCR products or other DNA segments flanked by recombination sites, can be "automatically" transferred into new vector backgrounds by simply adding the desired "Destination" vector and recombinase. By incorporating appropriate selections, the desired subclones are recovered at high efficiency (typically >90%) following introduction into *E. coli*. The method is fast, convenient, and automatable, allowing numerous DNA segments to be transferred in parallel into many different vector backgrounds. The resulting subclones maintain reading frame register, providing for the generation of amino and carboxy translation fusions. Approaches for optimization of protein expression, rapid functional analysis, and the integration of numerous technology platforms will be discussed.

P-013
A High-throughput, Plasmid Template Preparation Machine at $0.05 per Prep.

Carl U. Buice and Nancy Bergsteinsson. GeneMachines®, San Carlos, CA

Based on the array microcentrifuge technology developed at Stanford University, we have developed a machine capable of fully-automated isolation of DNA from plasmids for less than $0.05 per sample at a rate of two 96-well plates an hour. The key to the cost savings is the replacement of expensive filtration steps common to most automated DNA isolation machines with inexpensive centrifugation steps.

The array microcentrifuge consists of 96 individuals rotors arrayed in the standard 96 well format and spacing, allowing completely automated handling of samples by our 96 channel pipetter. To achieve reduced centrifugation times, each rotor spins at 60,000rpm, subjecting each sample to forces of nearly 14,000gs.

The Revolution PrepMachine™ (RevPrep™) integrates the array microcentrifuge into a robotic workstation including a 96 channel pipetter, a bulk reagent dispenser, an 8 position rotating deck and robotic server arm with access to a total of 40 plates. The combination of reduced centrifugation times, low reagent and disposables cost, and long unattended run times makes the RevPrep™ a superior technology for high-throughput, low cost template prep.

P-014
Sequencing of the M. tuberculosis Genome; Comparison of a Recent Clinical Isolate with the Laboratory Strain

Liane Carpenter[1], Amy Mikula[1], Arthur Delcher[2], Jeremy Peterson[1], David Alland[3], Claire Fraser[1], and Robert Fleischmann[1]. [1]The Insitute for Genomic Research, Rockville, Maryland, [2]Celera Genomics, Rockville, Maryland, and [3]Albert Einstein College of Medicine, NY

The completion of both the H37Rv and CDC1551 M. tuberculosis genomes provides the first opportunity for a complete comparison between two closely related organisms of the same bacterial species and the chance to correlate differences in phenotype with genome content and organization. In order to comprehensively identify the genomic differences between the two M. tuberculosis strains,we have developed an approach to providing whole genome alignments of closedly related DNA sequence. Our method combines suffix trees, the longest common subsequence, and Smith-Waterman alignment. The alignment built upon this information allows us to compare a variety of biological features including the number and distribution of the IS6110 element (four in CDC1551 vs sixteen 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.

P-015
Oncodevelopmental Genes of Human Hepatocellular Carcinoma Identified by the Complexity Reduction Analysis

Lan-Yang Ch'ang, Mandy Ting, Biyu Lin, Jeou-Yuan Chen, Wen-Chang Lin, Chin-Wen Chi and Char-Yang Chau. Institute of Biomedical Sciences, Academia Sinica, and Veterans General Hospital-Taipei, Taiwan

We have developed a complexity reduction (CR) approach for the analysis of genes expressed in the fetal and adult livers. The principle of this approach is based on sorting the restricted cDNA fragments into compartments with unique binary codes. Expressed gene fragments in each compartment of a virtual pyramid can be selectively amplified and fingerprinted with CR primers addressing different binary codes. The identities of fingerprinted fragments are determined by sequencing analysis.

Expression analysis of 21 human sequences by RT-PCR showed that five of which were expressed predominantly in the fetal organ and considered as developmentally regulated genes (RDGs). Subsequent examination of liver cancer cell lines suggests that 3 of the 5 RDGs analyzed are likely to be the oncodevelopmental genes (ODGs). Clinical studies performed on tumor specimens surgically resected from five HCC patients verified that the expression of these ODGs, ESF142-144, was at a level higher in human hepatomas than in adult livers. When compared to the expression of the •-fetoprotein (AFP) gene found only in two tumors, ESF142-144 were expressed in all five HCC patients at significantly heightened levels. As molecular markers for liver cancers, these ODGs are of great clinical values for the diagnosis and/or prognosis of human liver cancers. In addition, these unknown genes may serve as potential targets for the development of new therapeutics.

P-016
Sequencing In The Proximal Half Of The Q-Arm Of Human Chromosome 22, And Syntenic Regions Of Mouse Chromosomes 6, 10 And 16

F. Chen, A. Do, T. Do, S. Deschamps, A. Dorman, Y. Fu, F. Fang, P. Hu, X. Hu, S. Kenton, D. Kupfer, A. Hua, H-S. Lai, L. Lane, V. Lao, J. Lewis, S. Lewis, S-P. Lin, E. Malaj, F.Z. Najar, H.Q. Pan, Y. Qian, L. Ray, Q. Ren, J. Tian, R. Tian, J. Tilahun, Z. Yao, Q. Yang, J.D. White, H. Wright, M. Zhan and B.A. Roe. Department of Chemistry and Biochemistry, University of Oklahoma, Norman, Oklahoma

We have completed the sequence of almost 10 million bases of human genomic DNA from Cosmid, P1, PAC, Fosmid and BAC clones from several human chromosomes with almost 8 Mbp being from the proximal half of the q-arm on human chromosome 22 and at least 2 Mb being comparative sequence from the syntenic regions of mouse chromosomes 6, 10 and 16. Our double-stranded shotgun cloning sequencing strategy entails sub -cloning nebulized target large insert cloned DNA into pUC vectors, followed by robotic template isolation, and Taq-FS polymerase catalyzed cycle sequencing with non-fluorescent universal forward and reverse primers and fluorescent-labeled Big Dye terminators. The data are assembled at high stringency (minmatch=30,minscore=55) with Phred/Phrap via a series of perl scripts. Semi-automated closure and proofreading are accomplished using primers with a Phrap concensus quality >30, a Tm of 60degC +/-1degC, and fewer than 12 complementary bases per 20-mer primer to any known sequences in the target database. Primers are synthesized in a 96 well format on a modified MerMade oligonucleotide synthesizer, followed by Taq-FS cycle sequencing directly off the target large insert clones or PCR gap-scanning templates with the standard Big Dye-labeled terminators and either the dITP or dGTP,mixes, or the dRhodamine terminators with or without DMSO. The results of this work will be discussed along with the detailed techniques employed. Our most recent data with Consed error rates and GenBank links, as well as our latest protocols are available at URL: http://www.genome.ou.edu

We acknowledge our collaborators B. Emanuel and M. Budarf at the Children's Hospital of Philadelphia, I. Dunhan and J. Sulston at the Sanger Center, B. Morrow at Albert Einstein Univ., E. Zwarthoff at Erasmus Univ., J. Dumanski at the Karolinska Inst., R. Reeves at John Hopkins Univ., M. Meisler at U. of Michigan, M. Simon at Cal. Tech. and support from NCHGR, NIH-AI, NIH-DI and NSF EPSCoR.

P-017
Octamer Sequencing in a Core DNA Sequencing Facility

Yevette C. Clancy[1], Lisa S. Hayes[1], Melissa T. Cronan[1], Alison E. Maurice[1], Susan H. Hardin[2] and Suzanne P. Williams[1]. [1]Pfizer, Inc, Groton, CT. [2]Department of Biology and Biochemistry, University of Houston, Houston, TX.

Octamer Sequencing Technology (OST), a method of DNA Sequencing that uses single octamer oligonucleotides to prime cycle sequencing reactions, may reduce the time and cost of DNA Sequencing. Turnaround is decreased because a universal set of primers is on hand, eliminating delays associated with design and synthesis of gene specific primers. Cost is reduced because one synthesis of an oligonucleotide produces enough material to prime hundreds of reactions. To test the feasibility of using OST in a Core DNA Sequencing Lab, 75 octamers from an optimized 768 member library were selected to sequence portions of eight distinct cDNA clones, ranging in size from 1.6 to 8.0 kilobases and containing between 39 to 55% GC content. This report describes design of the Optimized Octamer Library, selection of the Octamers for specific clones, Octamer Reaction conditions using ABI BigDye Terminators on an ABI 373XL Automated Sequencer, and Octamer Sequencing results.

P-018
Increased DNA Sequencing Throughput On The CEQ™ 2000 DNA Analysis System

Doni J. Clark, Paul K. Cartier, Barry K. Hanamoto, Scott K. Boyer, Mark D. Dobbs, and Keith W. Roby, Genetic Analysis Development Center. Beckman Coulter, Inc. Fullerton, CA

The CEQ™ 2000 DNA analysis system is a capillary-based automated DNA sequencer capable of unattended denaturation, injection, separation, and data analysis for a full plate of 96 sequencing samples. Breakthroughs in DNA capillary arrays, voltage parameters, and temperature control systems have significantly increased the maximum sequencing read length and shortened the overall separation time. Under ideal conditions, a read length of approximately 800 bases may now be achieved in 60 minutes. We have also found several alternative DNA preparation methods that improve both the efficiency of upstream sample preparation and the quality of DNA sequencing results. When these new methods are implemented, the maximum throughput of the CEQ 2000 system is more than doubled, from 50,000 to 130,000 raw bases per day.

P-019
Analysis of competence genes from S. pneumoniae using DNA microarrays

Robin T. Cline, Donald A. Morrison, Carissa Horst and Scott N. Peterson, The Institute For Genomic Research

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

P-20
Automated, High-Throughput Analysis of a Rice Genomic SNP Correlated with Amylose Content and Processing Quality

Concetta A. Conaway[1] and William D. Park[1], Richard B. Rhodes[2] and Daniel D. Kephart[2], [1]Crop Biotechnology Center, Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX; [2]Promega Corporation, Madison, WI

Amylose content is a key determinant of the cooking and processing quality of rice. The Waxy gene encodes the granule-bound starch synthase responsible for the synthesis of amylose in rice. A single nucleotide polymorphism (SNP) at the 5' leader splice site of the Waxy gene has been shown to correlate strongly with amylose content of rice.

We used Promega's READIT™ technology to generate a screen for the G-T SNP. The screen was used to correctly determine the G-T content at the SNP of a set of 96 genomic DNA samples. Further, the READIT™ system was used to accurately measure the G-T content at the SNP of several complex heterozygotes. The technology is amenable to high-throughput screening in that a set of 96 samples can be assayed in a single day. The READIT™ system offers sensitivity and time savings in the screening of genetic polymorphisms.

P-021
Draft and Finished Sequencing of Human Chromsome 16

Norman A. Doggett, Mark O. Mundt, David C. Bruce, P. Scott White, Jon L. Longmire, and Larry L. Deaven. DOE Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, NM

We are pursuing a plasmid-based approach toward the large scale sequencing of human chromosome 16. Our strategy for sequencing involves nebulization to randomly break DNA, size selection of 3 kb and larger fragments, double adapter cloning into plasmid, and sequencing of both ends up to 4.5X random sequencing coverage. Assembly of sequence contigs is assisted with JAVA tools which exploit the inherent relationship of the paired-end sequences. Closure and finishing is achieved by a combination of primer walking, shatter libraries, longer reads, and alternate chemistry reactions. We have recently introduced the use of the Mermade 96 channel oligonucleotide synthesizer which enables us to reduce the level of shotgun in favor of targeted primer walks. We have also begun double-end plasmid draft sequencing to 4.5X coverage in Phred Q20 bases, producing ordered and oriented contigs. Drafting accelerates the sequencing process, and we are now completing about 1.5 Mb in this format per month. To date we have submitted more than 6.7 Mb of finished sequence, including a 3.0 Mb contig from 16p13.3 spanning four disease genes-- TSC2, PKD1, MEFV and CREBP. Seouence comparison analysis is semi-automated with use of the SCAN program (developed at LANL). This program launches a suite of sequence similarity searches and gene prediction algorithms and summarizes the results into a single integrated annotation report.

Supported by the US DOE, OBER.

P-22
Recent Technical Advances Leading To Improved Performance Of Microarray Based Expression Analysis

Ian Durrant, Alex McDougall, Roland Howe, Allan Hayle, Doug Hurd, Litao Fu, Emma Shipstone and Yvonne Burrus, Amersham Pharmacia Biotech, Amerham UK

The APBiotech microarray system consists of a 12 pen, 36 slide spotter, high performance scanner and analysis software. The application is driven by optimized chemistry components including modified glass surfaces, labeling reactions using CyDye nucleotides and hybridization materials. The combination of hardware, software and chemistry has been integrated for the reliable study of comparative gene expression on more than 4000 duplicate genes per hybridization event.

Recent work has concentrated on a series of chemistry based improvements that include

Metal coated glass surfaces for spotting cDNA for improved hybridized; CyDye labelled probe signal CyDye labelled ribonucleotides for higher probe yields; use of total RNA as a source of probe; hybridization buffer optimization; automated systems to standardise the hybridization event

The combination of these advances in a fully integrated system for microarray expression analysis have led to significant improvements in sensitivity, reproducibility and reliability.

P-023
Analysis of Cell Cycle-Regulated Gene Expression Patterns in a Human Neuroblastoma Cell Line using High Density cDNA Microarrays

Julie Earle-Hughes, Priti Hegde, Cheryl Gay, Sonia Dharap, Renee Gaspard, Alexander Saeed, Vasily Sharov and John Quackenbush, The Institute for Genomic Research, Rockville, MD

The regulation of cell division and differentiation is a complex process coordinated by both internal and external signals. Transcriptional controls of cell cycle related mRNA populations are crucial for proper progression through the cell division cycle. We have conducted a genome-wide expression profiling of cell cycle-regulated transcripts containing more than 19,000 PCR amplified cDNA clones. The human neuroblastoma cell line SK-N-SH was synchronized independently using physical and chemical methods and RNA was extracted at appropriate intervals during each cell cycle phase. Expression ratios were determined by comparing RNA populations from synchronized cells to unsynchronized cells using a two-color (Cy3/Cy5) fluorescence labeling and detection scheme. Identification of genes involved in cell cycle progression on a genome-wide scale provides integrated information into the regulation of the process of cell division.

P-024
Corba Servers And Services At EBI

Patricia Rodriguez-Tomé, EMBL Outstation-EBI, Wellcome Trust Genome Campus, United Kingdom.

The EBI is actively participating in the Life Science Research Task Force of the Object Management Group, whose goal is to provide a forum for the creation of standards for the field of life science. We have developed CORBA interfaces to the EBI databases and services. The most often used are:- EMBL: there are 5 servers providing access to the Nucleotide Sequence Database (sequence, taxonomy, references, geneticCode and metaData). Clients are developed in collaboration with various groups in Europe.- JESAM: the EST alignment and cluster databases servers with associated client to browse, query and analyse the data (GenomeBuilder, ClusterBrowser).- RHalloc which aims is to provide a database for mapping consortiums to prevent duplication of efforts.- RHdb the radiation hybrid database.- MappedServer to answer the question "Has this EST been Mapped". This server itself acts as a client of JESAM and RHdb.- AppLab which allows interfacing command line applications.

This work was supported by a European Uni on (DGXII) grant BIO-4-CT-960346.

P-025
A Sequence Variation Discovery Platform

Michael G. FitzGerald, Kristen Wall, Ulrich Thomann and Lynn Doucette-Stamm, Genome Therapeutics Corporation, Waltham, MA

The GTC Sequencing Center at Genome Therapeutics provides commercial, academic and governmental scientists with access to industrial scale DNA sequencing and sequence variation detection. The Sequence Variation Discovery Platform utilizes this solid infrastructure for rapid gene analysis.

Single Nucleotide Polymorphisms are genetic markers of increasing importance due to their wide distribution and binary nature. Analysis of these markers requires two systems, one for discovery and a second method for detection of known markers. Discovery methods can range from those that scan the entire genome or target specific candidate regions. At Genome Therapeutics we have developed a robust SNP discovery system based on our high-throughput sequencing platform. This system has been designed with the candidate region approach in mind and is capable of discovering all SNP and small INDEL sites for the locus of interest. Our goal is discovery of functional SNPs for use in association studies. The system is based on direct sequence analysis of PCR generated templates. The sequence data are analyzed using our proprietary data analysis pipeline, which has incorporated polyphred (U. Washington, Seattle). Data will be presented demonstrating the robustness and efficiency of this system.

P-026
Technologies for Finishing the *Drosophila* Genome Sequence

Reed A. George[1], Susan Celniker[1], Robert Blazej[1], Clare Doyle[1], Richard Galle[1], Kathym Houston[1], Roger Hoskins[1], Robert Svirskas[2], Ken Wan[1], Evan Baxter[1], Stephen Richards[1], Mark Champe[1], Suzanna Lewis[3], Gerald M. Rubin[3], [1]Lawrence Berkeley National Laboratory, Berkeley, CA, [2]Motorola, Inc. Schaumburg, Ill, [3]Department of Molecular and Cell Biology and Howard Hughes Medical Institute, University of California, Berkeley. (http://www.fruitfly.org)

The Berkeley *Drosophila* Genome Project (BDGP) is a consortium of groups at UC Berkeley, LBNL, Baylor College of Medicine, and the Carnegie Institution. The BDGP and Celera Genomics have signed an agreement to work together to complete sequencing of the *D. melanogaster* genome in 1999. This prompted a new strategy requiring a complete physical map and subclone libraries of BACs spanning the genome. These are being used to produce a "scaffold" sequence for verifying assemblies and selecting templates for directed finishing of whole genome shotgun data. Toward these goals, BDGP has implemented: 1) a high-throughput process, used to produce an STS map of the autosomes (100 Mb), 2) a 4X ramp in subclone library production, 3) optimization of Licor slab-gel and ABI 3700 capillary sequencers, 4) PCR product sequencing, reducing template cost by >50%, 5) minimization of Big Dye terminator reagents, resulting in a 67% cost reduction, and 6) a 192-well oligonucleotide synthesizer for production of custom sequencing primers at > 50% savings over external vendors.

P-027
Screening a Selected cDNA Library by Pyrosequencing

Baback Gharizadeh1, Tommy Nordström1, Nader Pourmand2, Pal Nyrén1, Mostafa Ronaghi1, 1Department of Biotechnology, Royal Institute of Technology, Stockholm, Sweden, 2Department of Medicine, Karolinska Institutet, Stockholm, Sweden.

Pyrosequencing is a non-electrophoretic, bioluminometric DNA sequencing method based on sequencing-by-synthesis. The method employs a four-enzyme mixture to sequentially determine the sequence of a target DNA in real-time. Here, we report on sequencing of sixty-four colonies of a selected cDNA library, each between 100 to 800 bases in length. Pyrosequencing produced accurate sequence data for 30 bases. The data obtained from the first 30 bases with pyrosequencing were in 100% agreement with Sanger DNA sequencing, indicating high accuracy of the system for short reads. The average sequence read was 36 bases and over 23% were over 40 bases. Although, we aimed for a read-length of approximately 30 bases to identify a unique cDNA from the GenBank, we examined a few of the cDNA templates for longer reads. For one cDNA an accurate readings until 68 bases could be obtained. Some ambiguities appeared after this length, although a significant difference between incorporation signals and background signals could be observed even after 90 bases. By using the recently developed fully automated pyrosequencer machine 96 cDNA can be tag-sequenced by this system in less than 45 minutes.

P-028
Room Temperature Archiving of Plasmid Clones in an Automatable 96-Well Format

Mindy D. Goldsborough, Pamela Hansen and Arubala Reddy, Life Technologies, Inc., Rockville, MD

Genome projects generate millions of plasmid clones. While it is desirable to archive all clones, the space and costs for storage in glycerol at -80°C is tremendous.

FTA is a paper-based matrix impregnated with chemicals that lyse cells and protect DNA from degradation, thus allowing room temperature storage. Recently, we have expanded the archiving applications of FTA to plasmid DNA. Plasmids are recovered from storage by transformation of competent cells with a small "punch" of the paper. Plasmid DNA can be applied to FTA in many forms: overnight cultures, resuspended colonies, glycerol stocks, in DNA purification "solution 1" and as purified DNA. Studies will be presented on the successful archival and recovery of plasmids ranging from 2 Kb to > 200 Kb (BACs). Plasmids archived on FTA show no loss in transformation potential over time. In addition to recovery of plasmids from FTA, the DNA on FTA punches can be used directly in PCR and sequencing reactions.

A new 96 well format of FTA has been specifically designed for use with robotic workstations. Additional improvements to the FTA technology with this format include an indicator dye incorporated into the paper that changes color when sample is applied. FTA offers a reliable as well as a space and cost effective method for archiving large numbers of plasmid clones. In addition, FTA can be used as a room temperature distribution method for clones.

P-029
A Paper-Based System for the Collection, Archiving, Purification and Analysis of DNA Samples

Dr. Mindy Goldsborough, Ph.D., Senior Scientist, Genome Analysis Department Life Technologies-Rockville, MD

The FTA Gene Guard System is a revolutionary method for the collection and room-temperature storage of DNA samples from biological specimens. In addition to the ability to collect and archive DNA, FTA enables users to purify the immobilized DNA very rapidly using non-organic reagents. Data from blood, tissue, microorganisms and plant material collected and archived on FTA will be shown.

P-030
Analysis on gene expression patterns in *Halobacterium Halobium* using DNA microarray technology

Young Ah Goo[1], Shiladitya DasSarma[2], WaiLap V. Ng[1], and Leroy Hood[1] Dept. of Molecular Biotechnology, Univ. of Washington, Seattle, WA.[1] ; Dept. of Microbiology, Univ. of Massachusetts, Amherst, MA.[2]

Halophilic Archaea live in a unique habitat where they are exposed to several physiological challenges including extreme salinity, low oxygen concentration, high UV light intensity, and fluctuating temperatures.

Using microarray technology, we are investigating how the genes of the halophilic Archaeon, *Halobacterium halobium*, response to such environmental stresses. In order to optimize the DNA array hybridization conditions, we chose the recently sequenced *H. halobium* megaplasmid, pNRC 100, as a model system. ORFs on pNRC 100 were PCR amplified and arrayed on glass slides. Total RNA isolated from *H. halobium* grown under varying conditions was fluorescently labeled and hybridized with the DNA arrayed on the glass.
The *H. halobium* genome will be completely sequenced in our laboratory in this year. Upon completion of the whole genome sequence, we will investigate genome-wide gene expression patterns using DNA array technology.

Supported by NSF grant DEB 9900497 and NIH grant 5RO1HG01713

P-031
Integrated Monitoring and Control of a 7x24 High Throughput Sequencing Operation

S. Cowles, R. Linton, M. Keating, S. Murphy, G. Kutty, T. Ricker, W. Swagerty, K. Mueller, L. Stuvé, K. Warne, P. Hess, F. Chauveau, D. Sleeter, E. Vershen, J. Wilkinson, S. Saywell, R. Naughton, P. Short, and R. Cathcart, Incyte Pharmaceuticals, Palo Alto, CA

In phases, we are implementing an integrated control and monitoring system for a 7x24 high throughput sequencing facility. This system specifies several functional groups: primary computer and networking hardware, operating systems, vendor-supplied applications, web services, database services, and sequencing services. Our requirements for these systems are that they be highly available with multiple points of redundancy, highly flexible, extensible, compatible with heterogeneous systems, security in user authentication and traffic encryption, very fast, and very cost effective.
Servers run in highly available clusters separated by function. Online storage is mirrored; nearline storage is closely integrated with transient cached online storage for minimum recovery times. File systems and databases are built on top of highly flexible middleware. Database design is aimed at exploiting the flexibility of metadata while simultaneously taking advantage of the speed benefits of an automatically synchronized data warehouse. Web services are tightly coupled to the databases for speed and design flexibility while ensuring secure transactions. Decision support, realtime status reporting, device control and automated data reduction serve sequencing needs.

P-032
High-Capacity Oxygenated Microwell Plate Shaker

Robert D. Guettler, Sara E. Polgar, Christine S. Chang, Robert P. Blunt, GeneMachines®, San Carlos, CA.

In todays genomics laboratories space and processing time are both at a premium. The newly redesigned GeneMachines' HiGro shaker provides relief for both. In one small, 24"x24" footprint, HiGro can incubate 48 microwell plates and directly replace floor shakers that use deep-well plates. Higher density and better quality cell growth are achieved with rotational speeds and orbits that are specifically tuned for the 96-well format. The HiGro is a superior growth environment for both plasmids and phage. The cassettes used to house the microwell plates are directly interchangeable with the GeneMachines' Gel-2-Well for upstream innoculation and the GeneMachines' RevPrep for downstream processing. Growth results from the HiGro and conventional floor shakers will be presented for both plasmid and phage.

P-033
High-Throughput 96-Channel DNA Synthesis

Robert D. Guettler, Jimmy Koh, and Damien Luk, GeneMachines®, San Carlos, CA

The accelerating pace of genome sequencing, functional analysis, and PCR-based techniques is creating a growing need for high-throughput synthesis of oligonucleotides. This need has led GeneMachines to develop a novel, high-throughput oligonucleotide synthesizer. Our PolyPlex™ (pat. pending) instrument features 96-channel parallel synthesis that generates a plate of 20mers in less than four hours. Reagent volumes and hence costs are kept to a minimum by parallel reagent dispensing. The 96-well output format is amenable to downstream high-throughput processing and handling. The instrument features flexible yet easy-to-use software that accepts sequences in a simple text-file format. Synthesis progress for all 96 channels can be monitored by trityl collection in conjunction with off-line colorimetric measurements. Detailed results regarding synthesis yield, purity, and biological activity will be presented.

P-034
Integrated Genomic/Proteomic Analysis of Adenoviral Gene Therapy Vectors

Karen M. Hahnenberger[1], Alex Apffel[1] John A. Chakel[1], Gargi Choudhary[1], William Hancock[1], Joseph A. Traina[2] and Erno Pungor[2]., [1]Hewlett Packard Laboratories, Palo Alto, CA, [2]Berlex Laboratories, Inc., Richmond, CA

An approach for gene therapy of coronary disease involves the transient expression of a growth factor gene, delivered by an adenoviral vector, which will increase blood flow by stimulating the growth of new blood vessels in the heart. Rigorous quality control procedures are required to ensure that the viral vector preparations are functional and safe. For this purpose, we have investigated the potential of using a common analytical platform to obtain correlative genomic and proteomic information for adenoviral gene therapy vectors.

The integrity of the insert in adenoviral vectors containing the FGF4 gene, and the absence of viral replication sequences, was verified by restriction digest, RFLP mapping, and heteroduplex analysis performed using a combination of HPLC, CE, and rapid separations on microfluidic chips. The protein complement of the viral particles was examined by HPLC ESI/MS and MALDI-TOF/MS and suggests that quantitation of specific viral proteins by this method can be used as a measure of the amount of active viral particles.

P-035
Progress towards sequencing the genome of the nitrifying bacterium, *N. europaea*

K. Burkhart-Schultz[1], A. Arellano[1], S. Stilwagen[1], A. Erler[1], T. Attix[1], L. Do[1], W. Regala[1], G. Sakaldasis[1], L. Marieiro[1], B. Winkleblech[1], M. Concepcion[1], S. Duarte[1], A. Kobayashi[1], F. Larimer[1], M. Shah[1], D. J. Arp[2], A. B. Hooper[3], and J. E. Lamerdin[1]. [1]Joint Genome Institute, Lawrence Livermore National Laboratory, Livermore, CA., [2]Botany and Plant Pathology Department, Oregon State University, Corvallis, OR, [3]Department of Biochemistry, University of Minnesota, St. Paul, MN, [4]Oak Ridge National Laboratory, Oak Ridge, TN

As part of the DOE initiative to understand the role of microorganisms in global carbon sequestration, the JGI is sequencing *Nitrosomonas europaea*. This organism is an autotrophic nitrifying bacterium that plays an important role agriculturally and environmentally by oxidizing ammonia to nitrite. *N. europaea* is also capable of degrading a variety of halogenated organic compounds, making it an attractive organism for bioremediation purposes. In order to complete the sequence of the 2.2 Mb genome of *N. europaea*, we have chosen to use a whole genome shotgun strategy, supplemented with a scaffold of fosmid end sequences. Fingerprinting of a minimal spanning path of fosmids will be used to aid verification of the final assembly. We will present progress towards completion of this genome and analysis of its unique gene content.

This work was performed under the auspices of the U.S. DOE by LLNL under contract no. W-7405-ENG-48.

P-036
Sequencing the Whole Genome of *Thermus thermophilus* HB27

Thomas Hartsch, Silke Blume, Mechthild Bömeke, Ulrike Bode, Carsten Jacobi, Hans-Peter Klenk, Gerhard Gottschalk and Hans-Joachim Fritz Göttingen Genomics Laboratory, University of Göttingen, Germany

Thermus thermophilus strain HB27 is a Gram-negative bacterium growing at temperatures ranging from 52 °C to 85 °C. Its genome of 1.82 Mbp (excluding the megaplasmid pTT27 of 0.24 Mbp) has a GC-content of approximately 69%. 16S rRNA sequence firmly establishes *Thermus thermophi-lus* in the cluster of green non-sulfur bacteria. Until present, three additional species within the genus *Thermus* have been identified and named as follows: *Thermus aquaticus*, *Thermus ruber*, *Thermus filiformis*.

Due to its unique natural transformation competence, *Thermus thermophilus* is amenable to genetic studies. In addition to problems of basic research, such as the structural basis of protein thermostability and the accuracy of macromolecular biosynthesis at high temperature, possible technical applications of temperature-stable enzymes of *Thermus thermophilus* contribute to the interest in that organism. The whole-genome sequencing project is carried out by a random shotgun approach. Adressed gap closure was initiated after 25000 random sequences.

P-037
Identification of Genes Involved in Colon Cancer Metastasis using cDNA Microarrays.

Priti Hegde[1], **Cheryl Gay**[1], Julie Earle-Hughes[1], Kristie Abernathy[1], Sonia Dharap, Alexander I. Saeed [1], Vasily Sharov, Norman H. Lee[1], Timothy Yeatman[2], and John Quackenbush[1]. [1]The Institute for Genomic Research, Rockville, MD 20850 and [2]The H. Lee Moffitt Cancer Center, Tampa, FL

Colon cancer is the second most common cause of deaths related to cancer in the United States. Despite significant advances in therapy, poor prognosis of colon cancer metastasis results in high mortality. Unlike tumorigenesis, which requires presence of mutated genes, metastasis is mostly a result of altered expression of genes. cDNA microarrays provide an ideal tool for the high-throughput analysis of gene function on a genome-wide scale. We have assembled a collection of cDNA clones representing more than 40,000 distinct genes, developed laboratory hardware and protocols, and created databases and data analysis tools necessary for analyzing differential expression. High-density cDNA microarrays containing more than 19,200 PCR amplified clones have been used to study differential expression patterns between less metastatic (KM12C) and highly metastatic (KM12L4A, KM12SM) cellular phenotypes in human colon carcinoma cell lines. Statistical analysis of measured expression ratios suggests genes that may be of prognostic or diagnostic value and provide a more complete understanding of gene function and regulation with respect to cancer metastasis.

P-038
Development and Validation of Automated High Throughput Plasmid Purification Systems

Randolph J. Hellwig, Joseph A. Hensley, Scott T. Wathen, Sean S. H. Han, Kathleen Brown-Steinke, and Angela M. Pasquith Eppendorf – 5 Prime, Inc. Boulder, CO.

Automation of plasmid purification in a high throughput format requires the combination of robust, reliable chemistry, hardware platforms, and processes. Automated multitasking systems for the unattended processing of up to 50 96-well bacterial culture plates to ready-to-use plasmid have been developed and recently validated for use with our unique PERFECTprep-96VAC "lyse-bind-wash-elute" plasmid purification technology. A 96-well-plate format is utilized for *E. coli* culture through plasmid elution. Reagents, filter bottom plates, collection plates, and culture plates containing drained bacterial pellets are loaded onto the platform and the program started. Alkaline lysates are prepared, transferred via the 96-well pipettor head to a filter bottom (FB) plate, and the lysate is vacuum filtered directly into a second FB plate. Silica matrix and chaotrope are added to the cleared lysate, crude plasmid is bound, washed free of contaminants, and the purified plasmid is vacuum eluted with water. The automated systems yield high quality, immediately useable plasmid DNA in approximately 0.5 hour or less per 96-well plate. Although varying with copy number, host, and growout, yields of pUC-based plasmids range from 1-10 µg DNA per mL of culture, comparable to manual systems. Plasmids are consistently suitable for automated fluorescent DNA cycle sequencing (typically 650 base reads, or better), PCR, and restriction enzyme digestion.

P-039
A Novel Automated Workstation For Large-Scale Sequence Reaction Cleanup

Anders Holmberg[1], Kimimichi Obata[2,] and Mathias Uhlén[1,]
[1]Royal Institute of Technology, Dep. of Biotechnology, Sweden,
[2] Precision Systems Science, Chiba, Japan

Here we describe two novel automated workstations with flexible software. Using solid-phase technology, the setup has been adapted to automated purification of Sanger cycle sequencing reactions and preparation for PyroSequencing. The multiplex sequencing technique involves cycle sequencing of multiple targets in a single reaction and then, in an iterative manner, capturing and purification of the individual sequencing reactions. The PyroSequencing technique utilises the release of pyrophosphate (PPi) on stepwise incorporation of nucleotides and the subsequent conversion of PPi into ATP and finally light by sulfurylase and luciferase. Two robotic workstations has been developed to facilitate (1) an automated sequence product clean-up with a throughput of 384 reactions (obtained from 96 quatroplex reactions) within two hours, and (2) an automated PCR clean-up and single-strand generation with a throughput of 96 samples within 40 minutes. Examples will be given for both DyeTerminator and DyePrimer reactions as well as PyroSequencing runs.

P-040
Performance of the GeneMachines® OmniGrid™ Microarrayer for Printing DNA Chips for Gene Expression Analysis

Scott Hunicke-Smith, Jaques Fayet-Faber and Poonam .S. Medberry GeneMachines, San Carlos, CA.

Functional genomics and gene expression profiling have recently evolved from the synthesis of oligonucleotides on a chip to a wide range of DNA chips produced by microarraying spotters. As more and more researchers encounter the need to create their own arrays, the demand increases for flexibility, low production costs and high throughput in a microarraying system.

The OmniGrid offers a robust, high-speed and fully integrated solution for making DNA arrays. Data on the instrument's precision, accuracy, flexibility and throughput are discussed. As an example of these parameters, the OmniGrid is capable of printing more than 20,000 spots/slide in 10 hours onto 100 slides using 32 pins. Additionally tip technology, cross-contamination and alignment issues are addressed. Results demonstrate that the OmniGrid achieves high-throughput while maintaining a high quality of DNA chips with spot sizes averaging 150µm and contamination levels below 0.5%.

P-041
Optimizing Fluorescent Dye-Nucleotide Incorporation into DNA Probes for cDNA Microarrays

Diane Ilsley, Peter Tsang, Debra Wiest, and Laurakay Bruhn, Hewlett Packard Company, Palo Alto, CA

Fluorescently labeled DNA probes for cDNA microarrays can be generated using fluorescently labeled primers or by enzymatic incorporation of dye-dNTP analogs. Enzymatic methods are hindered by the fact that dye-dNTPs are often poor substrates for DNA polymerases. We have developed a simple model system to kinetically characterize dye-dNTP incorporation. Using synthetic templates of defined length and sequence, we have quantitatively measured incorporation of a panel of dye-dNTP analogs and subsequent polymerization onto a 3'-dye-dNMP terminus by several DNA polymerases. We found a strong correlation between the kinetic data from the model system and data obtained from two color competitive hybridizations on cDNA microarrays. We found that Cy3-dCTP and Cy5-dCTP were incorporated with similar efficiencies (k_{cat}/K_M), and that the rates of polymerization onto a 3'-dye-dCMP terminus were also comparable. These data were consistent with Cy5 and Cy3 fluorescent signals on hybridized cDNA microarrays. We have applied the findings from these studies to optimize fluorescent dye-dNTP labeling efficiencies of cDNA probes resulting in increased sensitivity in cDNA microarray hybridizations.

P-042
Sequencing Through Large Gaps by Utilizing Micro-Libraries and Transposon-Mediated Libraries

Lingxia Jiang, Haiying Qin, John Gill, and Steven R. Gill,
The Institute for Genomic Research, Rockville, MD

Closure of large physical and sequence gaps is one of the most time consuming and challenging processes in a genome-sequencing project. We present two approaches currently being used at TIGR to accelerate the closure of large gaps in the genome of *Staphylococcus aureus*: Micro-Libraries and Transposon-Mediated Libraries (TML).

Micro-Libraries resembles the traditional shotgun cloned libraries used for whole genome sequencing. However, when coupled with advanced techniques, easily automated sequencing processes and existing software, it becomes a powerful tool for large gap closure. The Micro-Libraries are constructed from sonicated PCR products spanning genomic gaps. Micro-Library inserts are amplified by PCR, sequenced and assembled using TIGR Assembler.

Transposon-Mediated Libraries are generated by random *in vitro* insertion of a transposon (GPS-1, New England Biolabs) into cloned PCR products that span genomic gaps. Library colonies are screened by PCR and the amplified products are sequenced using transposon specific primers and assembled using TIGR Assembler. Both methods gave promising results, which will be presented with respect to their effectiveness in genome gap closure.

Supported by NIAID grant RO1AI43567 to SRG.

P-043
Identification of Antibacterial Targets Using Large Scale Genome Comparisons

H.P. Fischer, GeneData AG, Postfach 254, CH-4016 Basel, Switzerland

The importance of completely sequenced genomes for the discovery of new drug targets against infectious diseases has been recognized. We have developed novel algorithms for large scale comparison of complete genomes in order to identify potential new antibacterial targets with defined selectivity and spectrum.

Starting with an all against all gene comparison of genomes, the algorithms cluster genes automatically into functionally equivalent gene families based on the connectivity of a similarity matrix, thereby avoiding the use of any absolute cut-off values for statistical significance.

This allows us to automatically identify weak sequence similarities in phylogenetically only distantly related species, the correct classification of multidomain proteins, and the separation of large gene families into biologically meaningful subfamilies.

Applying these algorithms a database has been generated that currently includes 22 complete genomes (20 bacterial, S. cerevisiae and C. elegans) representing more than 60'000 proteins and 2'528 protein families. Furthermore, we have developed algorithms that allow the quantitative evaluation of gene families as pharmaceutical drug targets.

P-044
The Use Of Homology-Modelled Malaria Proteins For Ligand Discovery

F. Joubert, A.W.H. Neitz & A.I. Louw Department of Biochemistry, University of Pretoria, Pretoria, South Africa.

A world-wide increase in malaria parasite drug resistance has emphasised the need for the design of new anti-malarial drugs, being either novel compounds or modifications to existing drugs. The Brookhaven Protein Data Bank contains structures of only 6 malaria enzymes, forcing researchers to employ other methods of structure-based drug design such as the exploration of the malaria genome. In this study, triosephosphate isomerase (TPI) and dihyfrofolate reductase (DHFR) were employed as targets for homology modelling and ligand discovery.

TPI-encoding mRNA was PCR amplified, cloned into a pET15b expression vector and expressed in *E. coli*. The recombinant protein was purified by means of a N-terminal His-tag, and shown to be enzymatically active. The three-dimensional structure of the malaria enzyme was initially predicted by homology-modelling techniques, and later verified against the newly available X-ray structure of *P. falciparum* TPI. The active site of malaria TPI was described using the DOCK package, and subsequently computer screened for novel inhibitors. Potential ligands were selected according to binding energy, and subsequently characterised in terms of inhibitory activity of the recombinant enzyme.

The high AT-content of malaria genes has hampered the expression of recombinant malaria proteins in bacterial systems. We have attempted the expression of a synthetic DHFR gene in a variety of systems, leading to either no products or insoluble products. It was recently shown by Sirawarapom *et al.* that correct folding and solubility of DHFR could be induced by expression of the complete DHFR-TS fusion protein. We have amplified the cDNA encoding DHFR-TS from indigenous malaria parasites, and subcloned it into the pGEM-T EASY™ system. Transfer of the DHFR-TS gene to the pET17b expression system is currently underway, with the ultimate goal of *in vitro* screening against computer-predicted inhibitors. The active site region of malaria DFHR was modelled by the MODELLER 4 package utilizing the structure of DHFR from other species as template. The active site was characterized and a site map prepared. Screening against the NCI-3D database for compounds binding to the active site utilizing the DOCK package is currently underway in order to discover new lead inhibitors.

P-045
Novel Fluorescent Reagents and Solid Phase Sequencing Chemistry for Genetic Analyses

Jingyue Ju, Hong Yan, Michael Doctolero, Michael Zaro, Don Sleeter, Stephen Lincoln, Eric Lachenmeier, Richard Cathcart. Incyte Pharmaceuticals, Inc. 3174 Porter Drive, Palo Alto, CA

The study of many genetic targets simultaneously drives the development of multiplex fluorescent tags. However, due to the limitations of the spectral region and therefore the availability of the detectors, the number of available fluorescent dyes that have distinguishable emission spectra are limited to ten or so. We have developed a novel fluorescent labeling approach, *Controlled Combinatorial Fluorescent Labeling* (CCFL)[1] that uses a limited number of fluorescent molecules to create a maximum number of fluorescent tags with unique fluorescence signatures. The approach is based on the fluorescence energy transfer principle and that the energy transfer efficiency is dependent on the separation distance between the donor and acceptor. Thus using two different fluorescent molecules, one as a donor, the other as an acceptor that is covalently bonded through a rigid linker, at least five fluorescent tags can be generated with unique fluorescence signatures. Such a library of five fluorescent tags can be easily detected with an instrument having two spectral channels. With the addition of one or more fluorescent molecules, an even larger library of fluorescent tags can be constructed. This library of fluorescent labels offers a valuable tool for multiple component genetic analyses, as they can be excited and detected with a simple optical system. Synthesis and characterization of these controlled combinatorial fluorescent labels as well as their application for DNA sequencing and multiplex PCR will be presented. Another advancement is a new sequencing chemistry[2] that produces much cleaner sequencing data on both slab gel and capillary array sequencers, eliminating the disadvantages of current dye primer and dye terminator chemistries. The procedure involves coupling fluorescent energy transfer (ET) primers that produce high fluorescent signals and solid phase capturable terminators such as biotinylated dideoxynucleotides. Following the solid phase purification, only the pure dideoxynucleotide terminated extension products are loaded on the sequencing gel. Such a procedure also offers an ideal platform for accurate DNA sequencing using matrix-assisted-laser-desorption-ionization time-of-fight Mass Spectrometry.

P-046
Rice Genome Sequencing Project : Sequencing of Rice Genomic DNA

Hiroyuki Kanamori, Kimiko Yamamoto, Jianyu Song, Noriko Kobayashi, Hui Sun, Zhong, Ari Kikuta, Kayo Machita, Yuko Nakama, Yumi Nakamichi, Michiko Ikeda, Kozue Kamiya, Satomi Hosokawa, Kazuko Yukawa, Harumi Yamagata, Michie Shibata, Sachie Onose, Mari Nakamura, Takashi Matsumoto, Yoshiaki Nagamura, Takuji Sasaki, Rice Genome Research Program, National Institute of Agrobiological Resources / Institute of the Society for Techno-innovation of Agriculture, Forestry and Fisheries, Tsukuba, Ibaraki 305-0854, Japan

We started the second phase of the Rice Genome Research Program (RGP) in 1998 with the aim of sequencing the entire genome. With a genome size of 430 Mb, whole-genome sequencing of rice is, an achievable goal as compared with other cereals. As the core of genome sequencing, a PAC (P1-derived artificial chromosome) genomic library derived from Oryza sativa ssp. japonica cultivar Nipponbare was established. Japan is in charge of sequencing chromosomes 1 and 6 as part of International Rice Genome Sequencing Project (IRGSP). For our initial sequencing efforts we are concentrating on gene-rich regions in these chromosomes. A shotgun sequencing strategy to ensure high-quality sequence information was developed. However, in order to accelerate the sequencing process, we are adopting a strategy such that sequencing of random subclones generated from a PAC anchored on the physical map is to be done mainly with ABI3700 sequencer. The sequence data is then assembled into contigs, and the finishing phase to close gaps between contigs and to resolve low quality regions is to be accomplished with ABI377XL. The detailed strategy will be described and sequence characteristics of some rice PACs will be shown.

P-047
Application of New Strategies to Increase Efficiency of BAC Closure in Arabidopsis thaliana Chrômosome II

Hean L. Koo, Kelly S. Moffat, Lowell Umayam and Terrance P. Shea. The Institute for Genomic Research, Rockville, MD.

The early completion of sequencing Arabidopsis thaliana chromosome II (19MB) by The Institute for Genomic Research (TIGR) involved a scale-up at all levels of production. Many innovations in closure were implemented to keep pace with increased random sequence production.

Prior to the scale-up, the average time a BAC clone remained in the closure phase was approximately one month. That average was reduced to eight days per BAC with a peak rate of up to 4 Mb total finished sequence per month. Principle among these innovations was the introduction of several key scripts to Haldol, the daily automated BAC-tracking program. At the 5X coverage stage (fully completed BACs are sequenced to 8X coverage), these scripts are automatically triggered to assess the TIGR Assembler-generated contigs. One script identifies clones for re-sequencing to close gaps resulting from short-read sequences, while another script identifies failed sequences for re-sequencing that may help close physical (unlinked) gaps. This automated identification of strategic sequences at the random stage reduced the number of gaps by 44% in BACs tested. The details of these methods and their application to the closure of chromosome II will be presented.

P-048
CHEMICALLY MODIFIED PRIMERS FOR ADVANCED DNA SEQUENCING

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

Sequencing off large insert clones (BAC/PAC, sub-Megabase size) and directly off genomic DNA from bacteria (Megabases) and eukaryotes (Gigabases) present a new challenge in technology development. The problems associated with the use of standard oligonucleotides as primers in genomic cycle sequencing protocols include insufficient specificity of primer annealing, non-specific amplification, low sensitivity and premature truncation at secondary structures in template DNA.

To overcome these problems we have developed a new method to generate and screen combinatorial libraries of chemically modified primers. The method is based on our proprietary monomers containing MOX or SUC reactive moieties that are incorporated into oligonucleotide precursors during modified coupling cycle (1). After oligo synthesis, a library of chemically modified primers is prepared by treating precursors with a variety of compounds. We assessed the effects of chemical modifications on primer annealing and extension by DNA polymerase and electrophoretic mobility of sequencing ladders for individual oligonucleotides and their small libraries. The screening of primer libraries in advanced sequencing protocols will be presented. Our results demonstrate advantages of using chemically modified oligonucleotides to solve very difficult problems of DNA sequencing.

Polouchine, N (1999) US Patent No. 5902879

P-049
A Fully Integrated cDNA Sequencing, Clone Management, and Microarray Fabrication Process

Eric Lachenmeier, Lyle Arnold, Tod Bedilion, JoeBen Bevirt, Noah Brinton, Teresa Carabeo, Michael Doctolero, Scott Eastman, Georgina Hodgson, Richard Johnston, Brad Krueger, Brett Krueger, Rachel Nuttall, Mark Reynolds, Eric Rollins, Bruce Wang, Drew Watson, Incyte Pharmaceuticals, Palo Alto, CA.

Polynucleotide microarrays for mRNA expression and genetic studies require a high level of quality and performance as the technology is adopted for high throughput applications. The integration of informatics and automated production systems is necessary in order to achieve these goals. Our system draws from over 170k highly curated human clones selected from the LifeSEQ Gold database, and prepares them for high-density microarray printing. This system combines automated clone handling and re-racking, PCR and sequencing reaction preparation, and capillary sequencing, with data tracking and QC measurements. Adaptive processing is used to account for typical process yields, ensuring complete GEM microarray coverage of selected genes. The hardware continuously updates the sample tracking database and an automatic labeling system eliminates errors and creates a log for each microarray manufactured.

P-050
Using the Invader™ Squared FRET Assay on a Panel of Familial Genomic Samples to Detect Single Base Changes at Multiple Loci

Scott M. Law, Brian Aizeinstein, Edward L. Beaty, Karl W. Nichols, Bruce Neri, and Monika de Arruda, Third Wave Technologies, 502 S. Rosa Rd., Madison, WI

The Invader™ Squared Assay employing FRET detection is a sensitive, low cost, and easy to use method that is capable of discriminating single base changes, insertions, and deletions directly on genomic DNA without the use of PCR. The assay consists of two target specific oligonucleotides(termed the Invader and the signal probe) hybridizing to the target, forming an overlapping complex only if the base of interest at the site of the SNP is present. This complex is a substrate for the structure specific Cleavase® enzyme, and the product of the cleavage event induces another cleavage event on a FRET oligonucleotide, freeing the signal fluorophore from the nearby quenching dye. Both reactions occur near the melting point of the oligonucleotides involved, so that multiple hybridization and cleavage events occur for each target molecule present, amplifying the signal but not the target. The fluorescence signal is detected directly on the reaction plates without the need for any further sample manipulation. This demonstration shows that the Invader Assay is a versatile, simple, and low cost method that is suitable for automated high throughput screening of SNPs.

P-051
A New and Rapid Method for Proteomic Profiling Using SELDI ProteinChip™ Arrays Demonstrated on Bacterial Lysates and Cultured Human Glial Cells.

James F. LeBlanc[1], Alfred M. Spormann[2], and Howard B. Gutstein[3], 1. Ciphergen Biosystems, Palo Alto, CA, [2]Stanford University, Palo Alto, CA, [3]M.D. Anderson Cancer Center, Houston, TX

We will present the generation and analyses of proteomic profiles for markers of changes induced by aromatic hydrocarbon substrates in bacteria used in bioremediation and human cultured glial cells treated with anesthetic drugs. Whole cell extracts and membrane preparations were incubated on a variety of Ciphergen's ProteinChip arrays and subjected to SELDI (Surface-Enhanced Laser Desorption /Ionization) analysis. After binding and washing, the mass and the amount (measured as peak height) for each of the bound proteins was determined by the Ciphergen mass reader to generate a protein spectrum. Comparison maps showing the changes in the protein profiles between the treated and untreated cells were generated using SELDI software. The analyses of the bacterial lysates showed a number of changes in the proteins produced due to the presence of specific organic substrates. Profiles of the bacterial membrane preparations revealed additional changes that were substrate specific. Proteins analyzed by SELDI ranged from less than 5 kDa to more than 100 kDa. We will show that difference marker candidates found on one ProteinChip surface were later confirmed on another chromatographic surface, illustrating the reproducibility of the technology. We will also show the production of a specific peptide marker candidate due to drug treatment of a human glial cell line that is reproduced on a second ProteinChip. The experiments themselves take only 1-2 hours to complete, from data collection to the analyses of the data with SELDI software that quickly generates difference maps to highlight specific changes in the protein spectra. We demonstrate that the ProteinChip platform is a versatile and reproducible method for finding changes in protein expression in a variety of different experimental systems.

P-052
Basecalling Software Capcall inside DNA Sequence Assembly Pipeline at JGI/LBNL

Yunian Lou, and Sam Pitluck, Human Genome Sequencing Department, Lawrence Berkeley National Laboratory/Joint Genome Institute

Capcall: a complete basecalling software package, originally developed to process the data generated by our capillary sequencer, handles the raw data from ABI-377 sequencers the same way as the ABI basecalling software. With compatible features including color-separation, automatic mobility shift adjustment and base calling, an additional striking feature of this powerful basecalling package is the incorporation of blind deconvolution, which has greatly enhanced the accuracy of basecalling. Capcall was also calibrated for the Phred scores for Big-Dye terminator chemistry by comparing the assembled contigs and the Capcall processed trace data. The raw data from ABI-377 machines were analyzed by Capcall then assembled by Phred/Phrap. The assembly results show that Capcall gives better results than ABI-377 basecalling. To take advantage of Capcall, we designed a new DNA Sequence Assembly pipeline at JGI/LBNL. The raw data from the ABI-377 sequencers were processed by both Capcall and the ABI basecalling software. Then, for each trace, the one with the most Phred Quality Score

(q) >20 bases was sent on to the assembly program. Based on 10 months of data collection, for more than 300000 traces processed, Capcall has averaged 90 more bases with q>20 for nearly half of the total traces when compared to ABI Basecalling Software.

This work was supported by the Director, Office of Energy Research, Office of Biological and Environmental Research, Human Genome Program, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

P-053
Automated Method for Purification of Dye Terminator Sequencing Products

Morten F. Lukacs, Arne H. Deggerdal, .Marie Bosnes, Dynal A.S, BioScience R&D, Oslo, Norway

The quality of sequencing results is determined by many factors but the purity of the template and the purification of the sequencing products have been shown to be especially critical factors. Dye terminator chemistry is rapidly becoming the preferred chemistry for high-throughput automated DNA sequencing. Excess unincorporated dyes and other contaminants such as salt, proteins and template DNA can interfere with the separation and detection of sequencing products, especially for capillary gel electrophoresis. They therefore require removal before loading on a sequencing machine. The traditional methods used for purification are ethanol precipitation or column purification, however these methods are difficult to automate.We have developed a method for rapid isolation of sequencing products and removal of excess dye terminators and other reagents from sequencing reactions. The method is based on biomagnetic separation technology using biotinylated sequencing primers and super-paramagnetic streptavidin beads. Sequencing reactions are performed in the presence of biotinylated primers and the biotinylated sequencing products are subsequently bound to streptavidin beads. The contaminants are washed away with 70% ethanol in a simple and fast one-step procedure. This method also has the advantage of eliminating the need for an additional purification of PCR reactions prior to cycle sequencing.For high-throughput labs, the protocol has been automated on different robotic workstations. Processing time for 96 reactions is less than one hour. Protocols have been optimised for use with the different commercially available dye terminator cycle sequencing chemistries.

P-054
A Large-Scale Strategy For Multiplex Sequencing

Joakim Lundeberg, Anna Blomstergren, Deidre O 'Meara, Anders Holmberg, Mathias Uhlén, Dept of Biotechnology, KTH-Royal Institute of Technology, S-100 44 Stockholm

Nucleic acid hybridization is an essential component in many of the standard molecular biology techniques. In a recent study, we investigated whether nucleic acid capture could be improved by taking advantage of "stacking hybridization". This refers to the stabilizing effect that exists between oligonucleotides when they hybridize in a contigous tandem fashion. Here we describe a solid-phase approach for purification of cycle sequencing products suitable for large-scale shot-gun sequencing projects based on this principle. Vector specific and re-usable capture beads have been designed for pUC18, pBluescript, pGEM plasmids that allow for selective purification of corresponding cycle sequencing products suitable for multiplex sequencing. The basis of the multiplex strategy is to perform cycle sequencing of multiple targets in a single reaction and then in an iterative and automated manner purify the individual sequencing reactions. A robotic magnetic workstation has been developed facilitating an automated sequence product clean-up with a throughput of 384 reactions (obtained from 96 quatroplex reaction) within two hours. The lack of salts and other impurities (template DNA) in the purified samples makes the approach suitable for high-capacity capillary instruments. Examples are taken from a whole-genome bacterial shot-gun project as well as from an expressed sequence tag project.

P-055
Developments in High Throughput BAC Ends Sequencing

Joel A. Malek, Bola A. Akinretoye, Sofiya Y. Shatsman, Jennifer E. Militscher, Sarah J. Goonasekeram, Maureen T. Levins, Stephany A. McGann, Keita D. Geer, Mark E. Hance, Getahun K. Tsegaye, Shaying Zhao, Tamara V. Feldblyum and William C. Nierman. The Institute for Genomic Research, Rockville, MD.

Libraries constructed in bacterial artificial chromosome (BAC) vectors have become the sequencing substrate of choice for medium to large scale genome sequencing projects. End sequences of BAC clones from a randomly constructed library may be used to select minimally overlapping clones for sequence contig extension. Recent completion of BAC ends sequencing projects, including Human, and completion of *Arabidopsis* Chromosome 2 have proved the ease, reliability and utility of the STC method. This strategy, however, relies on large numbers of end sequences from deep coverage BAC libraries. An efficient, 96 well procedure for BAC end sequencing has been developed at TIGR (Kelley *et al.*, 1999). The procedure is based on a modified Qiagen Alkaline Lysis DNA preparation. While the method has yielded excellent results it has also been improved continually through the Human, *Arabidopsis*, and *Trypanosome* projects at TIGR. Efforts have been aimed at increasing throughput to allow for rapid development of a BAC ends sequence resource for any gigabase size genome. We present developments including improvements in process speed, sequence quality, and costs.

P-056
Solutions For Sequencing Through Difficult Regions In BAC DNA Templates. Thermofidelase II For High-Throughput Sequencing

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

A new version of ThermoFidelase (TF-2) was developed for the advanced sequencing of difficult regions and long genomic templates. Enzymatic and DNA melting assays demonstrate that TF-2 rapidly unlinks DNA strands and accelerates primer annealing and extension. Using a novel BAC sequencing protocol with TF-2 we reduced the consumption of BAC DNA and eliminated a need for the addition of deaza-dGTP or dGTP for a number of G-rich islands. We have optimized cycle sequencing protocol for 384-well plate format. Total volume of sequencing reaction is reduced to 5 µl. This protocol can be applied for high-throughput sequencing of BACs, PACs and other templates. It was successfully used for closing gaps and sequencing through repeated regions in PACs from the ongoing Human Genome Project. New protocol has been extensively tested with BAC clone containing SEP15 human selenoprotein gene. We will present statistical results on the quality of individual reads and contig assembly. Feasibility of ensuring high quality of sequencing using BAC DNA templates will be discussed.

This work is supported in part by NIH grant 2R44GM55485-02.

P-057
High Throughput Sequencing with the MegaBACE 1000 System Today

J. Anthony Mamone, John R. Nelson, Bernard F. McArdle, John O. Schneider, John Bashkin, Bill Nielsen, Helen T. Franklin, Mark Lewis and Carl W. Fuller, Amersham Pharmacia Biotech. Sunnyvale, California and Piscataway, NJ

Several important improvements and developments have been made in the MegaBACE^{TM} 1000 DNA Sequencing System since its introduction. All of these improvements are aimed at increasing the amount and value of sequence data produced per run in the production scale sequencing environment.

The sensitivity of DYEnamic ET energy transfer terminators allows detection over a broader range of template amounts. Likewise, the robust performance of Thermo Sequenase^{TM} II DNA polymerase improves success rates with short cycling times. Exploiting the superior resolving power of linear polyacrylamide (LPA) separation matrix, we have developed simple protocols that allow PHRED 20 reads of over 1000 bases. Improvements in the basecalling software and retraining of PHRED for MegaBACE data further enhance the value of the system. Data on new developments such as compression-free dye primer sequencing with dITP and Thermo Sequenase II, and protocols for direct sequencing of BAC templates will be presented.

P-058
The MegaBACE 1000 Methods Development Project: Analysis of the Factors Affecting Success and Read Length

Bernard F. McArdle, J. Anthony Mamone, Barbara Grossmann, David Spodick, Robert Feldman, Bill Nielsen, Maria Saluta and John Bashkin, Amersham Pharmacia Biotech. Sunnyvale, California and Piscataway, NJ

Since the introduction of the first 96 capillary DNA sequencing instrument by Molecular Dynamics, genome community expectations of what constitutes high throughput sequencing have markedly increased. Time-consuming processes of traditional slab gel systems such as washing plates, pouring gels, manual loading, and lane tracking are becoming obsolete. While it is clear that capillary sequencing systems offer great benefits in terms of automation and ease of use, we find that they are sufficiently different from slab gel separation methodologies to mandate reexamination of the protocols used to generate and use DNA sequencing templates.

We are currently pursuing an aggressive program to investigate the factors that affect success and read length using the MegaBACE^{TM} 1000 system. The plan is to characterize each step of the DNA sequencing process to determine the tolerances of the entire system. All of the steps along the process from stock cultures to finished data are being examined. Some of the major factors being looked into include cell culture growth conditions and template preparation methods. More specifically, we are examining components of the generic template preparation such as the identity and quantity of salts and proteins, and elucidating the most appropriate methods for determining these. The mass, form, and concentration range of DNA templates produced, the parameters of electrokinetic injection, and electrophoresis run voltage and duration are also being examined. This information is being compiled and distributed to MegaBACE users through training, support, and printed and electronic literature. The goal is to understand the complete range of possibilities for capillary sequencing.

P-059
T2: An Automated, Sub-microliter Thermal Cycler for High Throughput DNA Sequencing

J. Nakane, C. Hansen, A. Safarpour, N. Siu, T. Willis[1], and **A. Marziali** Department of Physics and Astronomy, University of British Columbia, Vancouver, Canada. [1]Stanford DNA Sequencing and Technology Center, Palo Alto, CA

We have demonstrated operation of an automated thermal cycler / reaction set-up instrument capable of 500 nl volume reactions and 576 samples / run. The instrument is based on automated assembly of cycle sequencing reactions inside disposable pipet tips that are subsequently sealed and thermally cycled 96 x 6 at a time in six temperature controlled air-flow cycling blocks. Assembly of the reactions and tip handling is performed by a custom-made small-volume 96 channel pipettor mounted on an XYZ gantry. Sample evaporation during cycling is eliminated through usage of a sacrificial water barrier at one end of the sample holder, and a heat-fused seal at the other end. Sample recovery is effected by automated slicing of the heat-sealed end of the pipet tips. One advantage of this technology as applied to dye-primer sequencing reactions is the ability to cycle four separate reactions, separated by air gaps, in a single pipet tip quadrupling the throughput of the instrument for this chemistry. Demonstrations of 300 nl and 500 nl multiplexed dye primer reactions using this technology in individual channels have been previously presented. In this presentation, performance of the fully automated instrument with 576 sample capacity will be discussed.

P-060
Current Status of Rice EST mapping

Takashi Matsumoto, Jianzhong Wu, Shinichi Yamamoto, Tomoko Maehara, Takanori Shimokawa, Chizuko Harada, Nozomi Ono, Yuka Takazaki, Fumiko Fujii, Jyunshi Yazaki, Kazuhiro Koike, Ayahiko Shomura, Tsuyu Ando, Izumi Kono, Katsumi Sakata, Takuji SasakiRice Genome Research Program (RGP), National Institute of Agrobiological Resources / Institute of the Society for Techno-innovation of Agriculture, Forestry and Fisheries, Tsukuba, Ibaraki, Japan

Rice is one of the major crops not only in Japan but also in Asia, Africa and Latin America. It is also recognized as the model plant of monocots. RGP has been constructing a rice expressed sequence tag (EST) map since 1997. This project aims to establish a comprehensive mapping of cDNA clones as PCR markers using EST sequences. The rationale of this mapping strategy is a YAC physical map with 71% genome coverage and vast rice EST collection, including 9400 3'-ESTs, derived from large-scale cDNA sequencing. These ESTs were efficiently identified in the rice YAC clones by PCR screening. So far, a total of 6285 ESTs have been subjected to PCR screening, 90% of which showed a single band by amplification. We have placed 4129 ESTs onto our physical map, including 770 ESTs on chromosomes 1 and 6. These mapped ESTs, together with

234 STSs designed by RFLP markers, are used for screening a PAC library for genomic sequencing. Furthermore, 397 ESTs could screen YACs that are not yet anchored on the YAC physical map. Genetic mapping of these position-unknown ESTs added 102 new DNA markers on our RFLP map and resulted in new YAC contigs of 33.7Mb mostly allocated in gap regions.

P-061
Utilization of the ABI 3700 DNA Analyzer for the Sequencing of Multiple Genomes

John E. Gill, Jennifer V. Colvin, Nicole T. Borkowski, Luke J. Tallon, Kristi J. Berry, Tamara V. Feldblyum. The Institute for Genomic Research, Rockville, MD

Whole genome sequencing has quickly been established as a viable method to aid in defining an organism's genetic makeup. The importance of elucidating the DNA sequence for a variety of organisms has been crucial in the flourishing areas of comparative and functional genomics. Utilization of whole genome sequence data continues to find new applications at an unprecedented rate. Furthermore, sequence data is becoming increasingly important in research approaches used in medicine, agriculture and biology.

The introduction of PE Biosystem's ABI Prism 3700 DNA Analyzer allows sequence data to be acquired at a considerably higher rate while reducing user interaction. TIGR is currently engaged in incorporating this capillary based sequencer into its sequencing process, and initial results are proving to be promising. This presentation will focus on the current use of this instrument on both eukaryotic and prokaryotic genomes. In addition, factors that appear to be critical in attaining sequencing data, and the potential effects the 3700 will have in high-throughput sequencing applications will be presented.

P-062
Genome Sequencing of Industrial Microorganisms: The Corynebacterium glutamicum ATCC 13032 Genome Project

Bettina Moeckel[1], Anke Weissenborn1, Walter Pfefferle[1], Jörn Kalinowski[2], Brigitte Bathe, Alfred Pühler[2], [1] Degussa-Huels AG, FA-FE-BT, Halle/Künsebeck, Germany, [2] Institut für Genetik,Universität Bielefeld, Bielefeld, Germany

The Gram-positive soil microorganism Corynebacterium glutamicum is well known as an industrial producer of economically important amino acids mainly used as feed additives. About 350.000 t/a of L-lysine are produced by fermentation with Corynebacterium glutamicum. A variety of experimental methods are at hand to further increase the amino acid yield in the fermentation process. Among them, screening procedures for antimetabolite-resistent microorganismes have been applied, but also cloning and expression of designated genes (1). The genome sequencing project will supply us with new information about metabolic pathways. Thus, rational construction of production strains seems attainable within the next future.

In the beginning of the sequencing project a physical map of the 3MBp circular chromosome of Corynebacterium glutamicum ATCC 13032 (2) and a cosmid library had to be constructed. The genome sequencing project of C. glutamicum will be the beginning of a new era in the development of industrial production strains.

(1) Eggeling, L. (1994) Amino Acids 6:261-272.
(2) Bathe, B., Kalinowski, J., Pühler, A. (1996) Mol. Gen. Genet. 252:255-265.

P-063
The Effects Of Oxidative DNA Damage On A Novel Repeat-Specific DNA-Binding Activity In Deinococcus Radiodurans

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

The Eubacterium Deinococcus radiodurans can withstand and repair large numbers of double-stranded breaks in its DNA, in contrast to most bacteria. Following damage, the genome is reconstituted in the correct order in less than thirty hours. While the majority of the repair appears to be recA-dependent recombinational repair, the earliest steps
(<6 hours) are recA-independent. Sequence analysis of the genomic complement (one chromosome and two megaplasmids) has revealed several families of DNA repeats. One of these repeats, poridge, is found 58 times across the genome. The repeats are spaced at intervals of, on average, 70 kilobasepairs. We have identified a DNA-binding activity that is specific for a thirty base-pair inverted repeat within the poridge repeat. This binding activity was ssayed during the recovery period following gamma-irradiation or oxidative DNA damage. Binding activity increases during the first four hours of recovery from damage, coinciding with the recA-independent phase of repair. Results were similar when either ionizing radiation or hydrogen peroxide were used to induce damage, indicating that this is a general response to DNA damage. These results suggest that the poridge repeat may play a role in aligning the chromosomal fragments correctly following damage.

P-064
High-throughput, Multiplexed SNP Genotyping by Mass Spectrometry

John Butler, Thomas Shaler, Stephanie Royer and **Joseph Monforte**, GeneTrace Systems Inc., Alameda, CA

It is essential that high-throughput, cost-efficient methods become available for analyzing single nucleotide polymorphisms, if they are to become a broadly applied method for genetic typing. We will present data from SNP analysis using methods developed to take advantage of the advanced capabilities of time-of-flight mass spectrometry. Mass spectrometry offers significant advantages both from the perspective of high throughput, 1 sample per second, and high levels of multiplex detection, e.g. 10 SNPs per sample. The mass spectrometric process is optimal when focused on gene-targeted association studies where SNP markers are developed for 10s to hundred of specific genes and analyzed for thousands of samples.

Supported by National Institute of Health grant R01 HG01700 to JM

P-065
Complete sequence analysis of 1,500 human cDNA clones harboring long and nearly full-length inserts

Nobuo Nomura,Takahiro Nagase, Ken-ichi Ishikawa, Reiko Kikuno, Makoto Hirosawa, Mikita Suyama, Nobuyuki Miyajima, Ayako Tanaka, Hirokazu Kotani, and Osamu Ohara. Kazusa DNA Res. Inst., Kisarazu, Chiba 292-0812 Japan.

One of the objectives of the Kazusa human cDNA project is to accumulate and examine entire sequence information of unidentified human cDNA clones harboring long and nearly full-length inserts. Among brain and KG-1 clones pre-selected either by in vitro translation method or northern analysis, we have so far determined the entire sequences of about 1,500 clones with average size of 5.0kb. Total nucleotide numbers sequenced have exceeded 7 Mb. Among them, 1,117 clones were well characterized and the sequence data were deposited in the databases under the gene name of prefix "KIAA". By searching Unigene database, the Kazusa clones were shown to comprise more than 50% of the publicly data-deposited clones with the insert size ranging from 4kb to 8kb. Approximately half of the cDNA clones were speculated to retain a complete ORF. Based on the results of homology search and motif survey, the functions of 636 clones were presumed and were allocated to several groups such as cell signaling/ communication (268 clones, 42.1%), cell structure/motility (141 clones, 22.2%), nucleic acid management (143 clones, 22.5%), protein management (44 clones, 6.9%), metabolism (30 clones, 4.7%), and cell division (10 clones, 1.6%). Expression and mapping analyses were also carried out.

P-066
Introducing Automated Dna Sequencing Into The American High School Biology Laboratory

Wesley D. Bonds[1], Sr. Mary Jane Paolella[2], [1]Department of Genetics, Yale University, New Haven, CT, [2]Sacred Heart Academy, Hamden, CT

We are engaged in a program to introduce automated cycle sequencing into the high school biology laboratory. To date, our sequencing efforts with students have emphasized the pedagogical benefits of manual sequencing. Initially, however, manual sequencing typically succeeds about half the time with high school students. With the need for back-up sequence and the desire for longer, more accurate reads for bioinformatics lessons, we begin sequencing lessons with parallel experiments in manual and automated modes. The students sequence double-stranded PCR products manually from one end and collect automated Big-Dye Terminator sequence from the other. Simplicity of operation, reliability, and safety features make the Perkin-Elmer 310 single-capillary instrument a natural choice for us. Gel-pouring technique and the concerns related to acrylamide exposure contribute to capillary sequencing as an obvious choice. Furthermore, student average read lengths jump from about 200 to 500 bps. Although we use the ABI Sequencing Standards to easily recover readable sequence, integrating the Big Dye Terminator Reaction protocols into our high school laboratory has proven to be a challenge, the genetic analyzer is an important tool for us in this landmark year of genomic sequencing.

P-067
Introduction of the ACAPELLA-5K Automated Fluid Sample Handling System

Deirdre R. Meldrum[1], **William H. Pence**[2], Harold T. Evensen[1], Stephen E. Moody[2], David L. Cunningham[2], Neal A. Friedman[1], Ethan B. Arutunian[1], and Mohan Saini[1].
[1]Genomation Laboratory, Department of Electrical Engineering, University of Washington, Seattle, WA; [2]Orca Photonic Systems, Inc., Redmond, WA.

The next generation ACAPELLA system, ACAPELLA-5K, will be introduced and initial experimental results presented. This system increases the throughput of the previous ACAPELLA-1K system from 1,000 to 5,000 capillaries in 8 hours. Microliter to submicroliter reactions such as restriction enzyme digests, PCRs, and sequencing reactions are prepared and performed inside glass capillaries using one capillary per reaction.

The throughput, reliability, and versatility of the ACAPELLA-5K system are improved over the -1K system through changes in the system hardware design and architecture. Hardware and software performance experiments will be presented for the ACAPELLA-5K system along with initial biological results. In the past year, key components of the new system have been tested extensively on the ACAPELLA-1K. These tests have included PCR screening of human inserts in BACs and sequencing reactions. These automated experimental results will also be presented.

This research is supported by NIH NHGRI grant number 1 R01 HG01497.

P-068
Novel Approaches to Closing Repeated and Difficult Regions of Microbial Genomes

Haiying Qin, Hoda Khouri, John Gill, Jessica Vamathevan, Terry Utterback, John Heidelberg, Tamara Feldblyum, Robert Fleischmann, Owen White, The Institute for Genomic Research, Rockville, MD

In a microbial shotgun sequencing project, the initial assembly process can have problems ordering repeats, and the random sequencing may leave gaps due to regions that are difficult to sequence under standard conditions (e.g.: GC- or AT-rich areas, homopolymers, secondary structures and unclonable regions). Sequencing through these repeats and difficult regions often delays the completion of genome projects. This poster focuses on novel strategies that were developed to solve these problems.

After the assembly, we detected repeats with a conbination of similarity searches and group the repeats into small, medium, large, inverted or tandem repeats according to size and type. Different strategies including completely walking the spanning clones, PCR amplification, cloning, or using transposons on a cluster of repeats, are then applied.

Methods used to deduce sequence from difficult regions include the addition of enhancers, optimization of PCR and sequencing reactions, Dye primer sequencing chemistry, cloning into low-copy number vectors, the use of the CLONTECH GenomeWalker kit, multiplex, and combinatorial PCR.

P-073
The Sequence Of Human Chromosome 22

Bruce A. Roe, University of Oklahoma, The Chromosome 22 Sequencing Consortium

Human chromosome 22 is being sequenced by a distributed consortium of four laboratories: The Sanger Centre; The Genome Sequencing Centre, St Louis; The Department of Molecular Biology, Keio University School of Medicine and the Department of Chemistry and Biochemistry, University of Oklahoma. The strategy we have used involves genomic sequencing of a minimal tile pathchosen from mapped bacterial clones (PACs, BACs and cosmids).

The consortium map covers 22q in 5 contigs from centromere to telomere, comprising of 495 clones. Work is continuing to close the remaining gaps, which are covered by YAC clones that are being sequenced, and estimating the gap sizes by fiber-FISH.

At the end of May 1999, over 80% of this sequence has been finished and has undergone preliminary analysis, while the remainder has begun shotgun sequencing. In addition, over 2 Mb of mouse genomic sequence syntenic to human chromosome 22 also has been mapped and is being sequenced. To date, over 70% of this region has been finished and has undergone preliminary analysis while the remainder is in the shotgun sequencing phase. Our projections suggest that all of the remaining human and mouse clones currently in shotgun sequence will be completed and analysed by the end of this summer.

P-074
Mapping, Sequencing and Analysis of *Legionella pneumophila* Genome

James J. Russo, Gil Segal, Sergey Kalachikov, Xiaoyan Qu, Anthi Georghiou, Minchen Chien, Hye Y. Park, Baohui Zhao, Jing Chen, Stuart G. Fischer, Pieter J. de Jong, Peisen Zhang, Eftihia Cayanis and Howard A. Shuman, Genome Center and Department of Microbiology, Columbia University, New York, NY

Legionella pneumophila, the causative agent of Legionnaires' disease, has a genome size of ~4 Mb and an intracellular existence in eukaryotic cells. We have initiated a project to sequence this bacterium by a combined 3x whole genome/5x clone-based shotgun approach. A high redundancy map was constructed by hybridizing a BAC library with random probes derived from whole genome and known Legionella gene sequences. BLAST searches and more advanced comparative phylogenetic strategies are used to suggest roles for newly identified ORFs. Based on initial sequencing, gene density and number of orthologs are expected to be similar to those found in other bacteria. As genes expected to play roles in pathogenesis or the organism's lifestyle are identified, an allelic exchange strategy is used to delete them in Legionella. A complete set of presumed virulence genes (*Agrobacterium vir* homologs) was removed and had no effect on intracellular growth or host cell killing, although they did reduce conjugative DNA uptake.

Supported by NIH grant AI23549 to HAS and Columbia Genome Center funds.

P-075
Directed Minimal Sequencing For Total Genome Analysis With Minimized Sequencing Efforts

Patrik Scholler[1], H. Voss[1], G. Casari[1], T. Schlüter[1], M. Arenz[1], B. Drescher[1], D. Schütte[1], J. Kämper[2], R. Kahmann[2], C. Basse[2], M. Feldbrügge[2], G. Steinberg[2], I. Häuser-Hahn[3], V. Vollenbroich[3], E. Koopmann[3], G. Seidel[3], K. Sievert[3], B. Jaitner[4], R. Ebbert[4], V. Li[4], M. Vaupel[4], P. Schreier[4] [1]LION bioscience AG, eidelberg; [2]Uni München; [3]Bayer AG ZF-BTB and [4]AG PF-MWF-Biotechnologie, Leverkusen

The demand for high throughput analysis of large genomes is currently met by massive upscaling of the shotgun sequencing technology. In order to minimize sequencing efforts and to maximize functional genome analyses, a new genome sequencing strategy, called DIRECTED MINIMAL SEQUENCING (DMS), has been developed. It is based on physical template mapping prior to sequencing in conjunction with proprietary clone and data handling technology. The combination of DMS with automated functional genome annotation by means of the bioinformatics platform bioSCOUT™ guarantees an unparalleled speed and accuracy in the generation of genomic information. We have demonstrated the efficiency of this approach in the 20 Mb genome sequencing project of the phytopathogenic basidiomycete *Ustilago maydis* and discuss further options for functional analyses and the applicability for gigabase genomes.

P-076
Dye Terminator Removal and Sequencing Reaction Clean up - Utilization of Superparamagnetic Particles

Frank Schubert, Wolfgang Zimmermann and Rolf Wambutt. AGOWA GmbH, Berlin Germany

High throughput DNA-sequencing requires cost-effective, high-speed technologies for template preparation, removal of dye-terminators and separation/detection of the reaction products. Several efforts have been made to increase the level of automation throughout the whole process.

The utilization of superparamagnetic particles in biological separations avoids filtration and centrifugation steps. Isolation of target molecules is achieved by alternate application of strong magnetic fields.

We have developed superparamagnetic particles with DNA specific coatings which adsorb DNA fragments from about 25 up to 2000 base pairs. These particles allow a fast and efficient isolation and purification of cycle sequencing products. More than 90% of the impurities (dye terminators, salts, the enzyme) are removed during the process. No particular chemical primer modification has to be done. The necessary buffers do not contain substances which may interfere with electrophoresis conditions, such as high salt concentration, detergents or polymers. Purified fragments are eluted in distilled water.

The JOBI-Disk™ (Jenoptik-Bioinstruments GmbH), a robot station which process 96 samples in PCR wells simultaneously has been utilized for clean up of the sequencing mixtures. It is equipped with buffer-reservoirs, a magnetic particle collector, a heating and a additional device, which moves PCR plates mechanically.

P-069
The Genomic Species Concept

Thomas Quinn, Evolutionary Theory Discussion Group, New Berlin, WI

There are many species models; however, classifying the wide range of living forms has proved difficult. The proposed genomic species concept attempts to provide a coherent taxonomic strategy for the classification of all living entities. Most field biologists, museum taxonomists, and professional bacteriologists still classify organisms using various morphological species concepts where phenotypic differences are used as the main criteria for recognizing distinct species; yet, it is the genetic information inside an organism that controls the expression of those traits. Since every species is a genetic system, the genomic species concept asserts that each species should be identifiable as relatively stable informational unit. Although the various genes between individuals can vary greatly, organisms within a specific species should possess nearly identical genomic organization at the chromosomal and loci levels. The continuity of any species through time and space is completely dependent on maintaining the same overall molecular structure of the genetic information from generation to generation. The members of single genomic species will have the same kinds of loci at corresponding locations. The recent bioinformation revolution makes classification via the genomic species concept technically feasible; however, it is still prohibitively expensive. This economic barrier is likely to dissolve in the near future as sequencing technologies improve. Until the entire genome of any organism can be easily accessed, a variety of methods currently used, such as karyotyping, overall base composition, CG content, DNA: DNA hybridization, codon bias, and comparing specific short DNA sequences (e.g. the 16S rRNA gene) will continue to be essential for analyzing genomic organization. The genomic species concept should allow one to classify viruses, bacteria, parthenogenic animals, sibling species, ring species, and species that arose via hybridization within the same framework. Using the genomic species concept could also lead to a better understanding of the evolutionary relationships between extant species since genomic reorganization is one driving forces behind the evolution of new populations, subspecies, and species.

P-070
DNA Binding Metrics for Microarray Applications

Mark A. Reynolds, Rick Johnston, David Chain, Mike Ruvolo, Steven Daniel, Paul Lee, and Lyle J. Arnold, Jr, Incyte Microarray Systems, 6519 Dumbarton Circle, Fremont, CA

cDNA microarrays enable massively parallel gene expression experiments. A fundamental precept of this technology is the ability to bind PCR amplicons at high density in a reproducible manner. Essential parameters include the functional density and uniformity of the solid substrate (in this case aminopropylsilanated glass microscope slides), arraying conditions (buffer, surface energy, etc.), length and sequence composition, and post-processing protocols. To optimize these parameters, we have developed a DNA binding model with a set of genes comprising a defined range of sequence composition and length. These clones were labeled with a fluorescent dye (Cy3) in order to monitor DNA binding through the various stages of post-processing and hybridization. Hybridization signals were monitored using probes labeled with a second dye (Cy5). Our data indicates a direct correlation from surface energy and surface uniformity to DNA binding and signal response. These studies are essential to generate cDNA microarrays of sufficient quality for the population of gene expression databases.

P-071
Automated, Multiplex Detection of Genetically Modified Organisms (GMOs) on your Grocery Shelves

Richard B. Rhodes, Daniel D. Kephart and Brent A. Spoth
Promega Corporation, Madison, WI

A diverse array of biotechnology methods have been developed to improve crop quality and increase desireable agronomic traits. The ability to monitor the presence of introduced transgenes has important applications in both reasearch and commercial applications. However, public concern about the safety of GMOs has led the European Community to institute the Novel Food Ordinance, creating a critical need for a high-throughput and cost effective method for detection of trace amounts of GMO contamination in food stuffs.

We used Promega's READIT™ technology to generate a multiplex screen for simultaneous detection of the cauliflower 35S promoter and A. tumefaciens NOS terminator regions, two genetic elements commonly used for transgene expression. We demonstrate the absence of these sequences in GMO negative control samples, but detect the elements in GMO positive cotnrols and in a variety of commercially available food stuffs derived from soy beans and corn. The sensitivity of this detection system is demonstrated by our ability to detect limited quantities of GMO contamination in a non-GMO background using READIT™. Quantitative, non-gel based analysis enables automated calling of samples. The READIT™ approach to GMO detection is sensitive, modular, and scalable to provide a cost-effective analysis tool for GMO detection in a variety of sample materials.

P-072
Quantitation of the Inefficiencies of Closing Gaps by Walking

Jared C. Roach, Institute for Quantitative Systems Biology, University of Washington, Seattle, WA

The maximum efficiency of map or sequence walking is obtained when characterized clones are linked end to end with the minimum possible overlap. This overlap is determined by the minimum length necessary to establish connectivity from the underlying data such as, sequence, restriction sites, hybridization, STC linkage, or STS content.

Two inefficiencies are introduced when gaps are closed by walking. The first is caused when spanning clones are linked by overlaps larger than the minimum. This inefficiency is due to clone library randomness and imperfections in the underlying data. The second inefficiency, potentially much greater than the first, occurs when the distal end of the final clone in a walk overlaps the far end of the gap by more than the minimum length. This occurs once per gap. The location of this inefficiency may be shifted towards the middle of the gap if walking is simultaneous from both ends.

These inefficiencies are modeled with theory and simulations. The inefficiencies depend largely on the number of gaps, and on the distribution of gap lengths and clone lengths. Relevance to the human genome project is discussed.

The system works with cycle sequencing products of plasmid preparations as well as PCR-fragments, it takes about 20 minutes for 96 parallel isolations. Cycle sequencing products are adsorbed on the magnetic beads under defined buffer conditions, collected magnetically, washed, dried and eluted with distilled water. The eluates may be directly placed at the ABI 3700 DNA – sequencer. Analysis on the ABI 377 automated sequencer requires evaporation to dryness and solution in loading buffer.

P-077
An Efficient Method for BAC, PAC and P1 DNA Purification

Eiichi Sengoku, David I. Wheeler and JoAnne H. Kerschner
Sigma-Aldrich Corporation, St. Louis, MO.

BAC, PAC and P1 clones are difficult to purify compared with conventional plasmids due to their low copy number and large size (100-300kb). Current protocols for isolating large DNA constructs are often time-consuming, inefficient, and may require further cleanup for downstream applications. We addressed these issues by designing a new device for single large volume BAC purification.

The new system incorporates the traditional alkaline lysis procedure with bind, wash and elute steps under vacuum. The time from bacterial pellet to DNA pellet is greatly reduced due to the addition of column purification under vacuum (patent filed). Actual time to pass 50 ml cleared lysate through a conventional column can take upwards of 1 hour whereas with our system this procedure takes less than 10 seconds.

Furthermore, we examined BAC DNA shearing using our system versus a solutions-only method (standard alkaline lysis) with favorable results. We used this device/method to test downstream applications such as automated dna sequencing and restriction digestions, with excellent results.

P-078
Identification of Thousands of Single Nucleotide Polymorphisms (SNPs) in the Human Genome

Nila Shah, Cindy Chen, Sangeetha Kondapalli, Vivian Reyes, Chunmei Liu, Michael Savage, Michael Janis, Maria DeGuzman, Richard Watts, Anthony Berno, Naiping Shen, Jyoti Baid, Jim Snyder, Claire Marjoribanks, Howard Lee, Daryl J.Thomas, Robert Lipshutz, Nila Patil, and Janet A. Warrington, Affymetrix, Inc., Santa Clara, CA

We are screening thousands of human genes across 40 individuals to identify frequently occurring SNPs using GeneChip® arrays in an automated high throughput laboratory with a custom laboratory management system that integrates every component of the process including subject information, sample processing, and SNP discovery outcome. Currently, the use of automated sample preparation and array handling enables us to screen 1.8 MB of sequence per day. We will present a summary of all projects completed to date as well as detailed information from one of the initial projects in which we are screening genes expressed at high levels in lymphoblast cell lines. In this project, we are amplifying ~ 700 genes across 40 unrelated males and females of Caucasian, African-American and Asian origin using RT-PCR. Samples are assayed using high-density GeneChip® variation detection arrays, designed to screen 30 kb of sense and antisense sequence simultaneously. To date we have identified 1006 candidate SNPs at a density of 562 bps in the lymphoblast project. Confirmation of the SNPs is in progress.

Supported by NIH R01 MH60007-01 (CWR) R01 HG01919-01 (Stanford), R01 HG01953-01 (Whitehead Inst.)

P-079
Ion-Pair Reversed-Phase HPLC (IP-RP-HPLC) Approach for Cloning of PCR Products and Colony Screening

Carla M. Shaw-Bruha and Kimberly A. Lamb. Transgenomic Inc., Omaha, NE

Polymerase chain reaction (PCR) products are frequently used for cloning. However, isolation of a single DNA molecule for ligation into a plasmid vector is time consuming. PCR products are routinely separated on and purified from agarose or polyacrylamide gels. The limited resolution capabilities of gel separation and the technical difficulties of removing single size fragments following gel analysis have been demonstrated to result in inadvertent cloning of multiple products. Colony screening for recombinant DNA clones is time consuming and laborious as well as costly. Although PCR-based colony screening strategies may reduce screening time, improved fragment separation and purification approaches that allow for sub-cloning of only the single fragment of interest offer the biggest time-saving potential. In this regard, ion-pair reversed-phase HPLC (IP-RP-HPLC) analysis of nucleic acids offers significant improvements in nucleic acid fragment resolution.

We demonstrate that PCR products analyzed and isolated using the IP-RP-HPLC approach (WAVE™ DNA Fragment Analysis System) can be cloned directly into a linearized plasmid vector, without additional processing prior to performing the ligation reaction. We demonstrate further that extension of IP-RP-HPLC analysis to the colony screening process helped to reduce the overall colony selection and screening process to 1.5 - 2 days.

P-080
Specific DNA Measurement Using a New, Enzyme-Based System, the READIT™ System

John W. Shultz, Donna Leippe, Ken Lewis and Michelle Mandrekar Promega Corporation, Madison, WI

The revolution in molecular biology has resulted in a need to measure very low levels of DNA and to rapidly determine the genotype of a DNA sample. In this presentation, a new system for the measurement of both genomic double-stranded DNA and specific DNA species will be described. The system is based on two incubations using various enzymes. In the first incubation, an amount of ATP is made proportional to the amount of DNA in the sample. In the second incubation, the amount of ATP produced in the first incubation is measured using a Luciferase/Luciferin based assay system.

The use of the system for the measurement of as little as 20pg of genomic DNA will be presented. In addition, methods will be described that allow single nucleotide polymorphisms to be determined rapidly in a manner that clearly allows simple determination of the genotype of an unknown sample.

P-081
Further Upgrades to the ABD 377 Sequencer for High Throughput DNA Sequencing and Fragment Analysis

Chris Silk, Eric Lachenmeier, Kevin Smith, David King, Georgina Hodgson, Jingyue Ju, Hong Yan, Don Sleeter, Akbar Khan, Scott Saywell, Clark Tibbets, Rick Cathcart Incyte Pharmaceuticals, Inc, 3174 Porter Dr., Palo Alto CA 94304

Incyte's DNA sequencing capabilities have doubled annually over the past four years. This trend has been greatly facilitated by our internal efforts to build upon preexisting technologies. The ABD 377 Sequencer has provided us the foundation to focus our efforts on improving data quality and throughput over this period. Initially, we focused on increasing the data sampling rate which has enabled us the spatial resolution to run 96 lanes on a single gel. We have since implemented custom hardware and software that allow for complete control of all relevant parameters during data collection. This flexibility has enabled: Dynamic placement of any number of virtual filters, more accurate calibration, custom spectral deconvolution, on the fly scan rate control, and enhanced signal acquisition processing. These improvements have led to significant increases in signal to noise, greater resolution over longer reads, and the ability to multiplex five or more dyes during SSCP mutation detection.

P-082
Covalent Attachment of Sequence-optimized PCR Products for DNA Microarrays.

Mike Chen, John ten Bosch, Ken Beckman, Sepehr Saljoughi, Chris Seidel, Nico Tuason, **Ralph Sinibaldi**, and Bob Saul, Operon Technologies, Inc., Alamedia, CA

We have optimized a method for the covalent attachment of nucleic acids to glass slides. Experiments comparing our attachment chemistry with other published methods, including poly-L-lysine, aldehyde, and amino alkylsilanes indicate a very stable attachment and low rate of wash-off during hybridization and washes. Our chemistry of attachment works well with both oligo-sized and larger DNA products. In addition, we have examined the suitability of various DNA targets to measure changes in gene expression. Array technologies which depend on the spotting of DNA samples to measure changes in gene expression may be hindered by sequences with a high degree of homology. Using S. cerevisiae as a model we have created a set of sequence-optimized gene segments (OPTs) for spotting onto arrays. These segments were chosen to be free of homology with other yeast genes. We have printed arrays containing full length ORFs (500-2000 bp), OPTs (400 bp), and oligos (35-45 nt) and used them to examine changes in gene expression during diauxic shift in yeast. OPTs provide a high degree of sensitivity and reproducibility when compared with oligos and ORFs.

P-083
The Filarial Genome Project: *Brugia malayi* Contains An *a*-Proteobacteria Endosymbiont

Barton E. Slatko[1], Mehul Ganatra[1], Jennifer Ware[1], Jeremy Foster[1], David Guiliano[2] and the Filarial Genome Project. [1]New England Biolabs, Inc., Beverly, MA 01915 USA; [2]ICAPB, University of Edinburgh, Edinburgh. Scotland, EH9 3JT.

Since 1992, the World Health Organization has sponsored a Filarial Genome Project. Filarial nematodes, responsible for lymphatic elephantiasis and cutaneous filariasis, infect over 100 million people worldwide causing widespread morbidity in endemic populations. This international collaboration has implemented a program of gene discovery, genome mapping, and post genomic analysis. cDNA and a large insert genomic DNA library of *Brugia malayi* serve as the core of the gene discovery and genome mapping initiative. Over 17,000 *Brugia* EST sequences (and 7000 *Onchocerca*) ESTs have been deposited in dbEST, derived from 8 independent life-cycle stage cDNA libraries. Redundancy analysis identifies 6000 "clusters", about 35% of the expected number of "genes" in the organisms. The cDNA libraries have been gridded as high density filter arrays and are being screened with non-radioactive probes for subtractive library sequencing. The *Brugia* BAC library also contains sequences of an obligate intracellular *Wolbachia* endosymbiont. BAC-end sequencing and mapping indicates sequence similarity to the *Rickettsia prowazekii* genome, although the physical maps differ significantly in organization. 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.

P-084
Changing Gel Based SNP Assays to Non-Gel Based, Digital Read Out Assays

Rick Smith, John Shultz and Ken Lewis, Promega Corporation, Madison, WI

A large number of genetic polymorphisms have been analyzed using methods, such as ARMS and RFLP analysis, that rely on gel fractionation for identification of the specific alleles in a sample. However, such methods take a substantial amount of time to complete the analysis and can give ambiguous results. In this presentation, strategies will be presented for the conversion of such tests to the READIT™ System - a non gel based system that provides a numeric readout for the amount of specific DNA sequences present in a sample.

In the first example, a system will be described for the conversion of an RFLP determination of the genotype of a particular SNP.

A second example will describe how an ARMS system can be converted to a READIT™ System test. This example will also show how restriction endonuclease sites can be incorporated into the primers used in such a system for specific detection of the products of primer extension.

P-085
A Novel Chemical Affinity System For Cleanup Of Cycle Sequencing Reactions

Amy L. Springer, Wendy Ankener, Jing-Ping Chen, Anna S. Gall, Karin A. Hughes, Robert J. Kaiser, Guisheng Li, Deborah D. Lucas, Jean P. Wiley, Mark L. Stolowitz, Prolinx, Inc. Bothell, WA, USA

The Prolinx® Chemical Affinity System is based on a very specific interaction between two families of small molecules, the simplest representatives of which are phenylboronic acid (PBA) and salicylhydroxamic acid (SHA). Counterparts in these two small-molecule families react specifically to form a complex under a variety of conditions, the only by-product of which is an equivalent of water. The Prolinx® affinity system has been used to develop a novel technique for cleanup of cycle sequencing reactions that yields highly purified sequenced product suitable for capillary electrophoretic analysis. The process removes template, dye terminators, dNTPs, and the purified products are eluted under low ionic strength conditions. It is a simple rapid procedure that is easily automated. Sequencing extension products are prepared using a primer modified with a PBA derivative. The extension products are then captured on an SHA-modified solid support, where they can be washed. Finally the extensions products are released at low ionic strength, in water. The data for this application have been generated using SHA-modified magnetic particles but the system may be developed using other solid supports as well. Prolinx has successfully applied this method to the purification of cycle sequencing reactions from a variety of templates, including M13, pUC18 and PCR products.

P-086
Sequencing Human DNA at the Stanford Human Genome Center

Nancy E. Stone, **J. Shang**, J. Schmutz, C. Caoile, K. Elliot, L. Fischer, D. Fotopulos, K. Litton, J. Logan, J. Lopez, C. Medina, E. Prakash, L. Ramirez, T. Shin, O. Tan, G. Tong, M. Tsai, S. Vartanian, D. Vitale, J. Yang, D. R. Cox and R. M. Myers, The Stanford Human Genome Center and Department of Genetics, Stanford University School of Medicine,
Palo Alto, CA

The Stanford Human Genome Center has determined the complete nucleotide sequence of 7.5 Mb of chromosome 4 at an accuracy of fewer than 1 error in 135,000 base pairs. We have also followed a semi-random BAC identification strategy, where we screen the RPCI-11 library by hybridization with widely spaced STSs from the Radiation Hybrid map. This generates non-overlapping BAC clones for sequencing reagents. Our sequencing strategy uses bi-directional shotgun sequencing from plasmid templates at 7-fold sequence redundancy, followed by a automated finishing strategy that includes direct BAC sequencing, PCR sequencing and transposon hopping.

Thus far, we have identified BAC clones for more than 400 unique regions on the chromosome, corresponding to more than 90 Mb. We also have 2.5 Mb of contiguous BAC clones on the p arm (STS4-631 - STS4-1167) in the process of finishing as well as 30 Mb proximal to this region nearly contiguous in BAC clones in the process of shotgun sequencing. To date, we have finished 7.5 Mb of sequence at an error rate of 1 in 135,000 bp with no gaps.

P-087
Sequential Double-Priming; A Highly Efficient Method to Generate 5'-Biased cDNA Libraries

Ruoying Tan, Christine Kim, Richard Goold & Karl Guegler
Incyte Pharmaceuticals, Inc., Palo Alto, CA

The low representation of 5'-ends in cDNA-libraries is one of the limiting factors preventing large scale EST-sequencing programs from efficiently assembling full-length genes. The assembly of approximately 4 million ESTs from the Incyte and public domain databases results in ~10,000 full-length genes. The vast majority of partially assembled genes are complete at the 3'-end but lack the starting methionine. Since most of the published protocols to make full-length cDNA libraries require large amounts of mRNA and are hard to reproduce we evaluated a number of novel approaches.

Here we report the results of a cDNA library construction method improving the discovery rate of 5'-ends of genes in high-throughput EST-sequencing. Reverse transcription of mRNA is accomplished through a two-step process, priming first with oligodT and then with a random oligomer. Only the cDNA's generated in the second step are used for cloning of the library. The analysis of randomly sequenced clones from such libraries showed that 1) aeveral folds increase in the discovery rate of novel sequences not found in any database, and 2) a significant increase in representation of 5'-ends of known genes.

P-088
Analysis Of The Effects Of Different DNA Sequencing Methods On Accuracy And Quality And Expansion Of A Web-Based Sequencing Resource: Results Of The ABRF DNA Sequencing Group 1999 Study

Theodore Thannhauser[1], Pamela S. Adams[2], Mary K. Dolejsi[3], George Grills[4], Susan Hardin[5], and Margaret Robertson[6]. [1]Cornell University, Ithaca, NY; [2]Trudeau Institute, Saranac Lake, NY; [3]Fred Hutchinson Cancer Research Center, Seattle, WA; [4]Albert Einstein College of Medicine, Bronx, NY; [5]University of Houston, Houston, TX; [6]University of Utah, UT, Association of Biomolecular Resource Facilities.

The effects of different instrumentation and methods on the quality of DNA sequencing were studied using three different sequencing templates. Results of sequencing a common standard template and two difficult GC-rich templates were submitted to the study by a wide variety of sequencing laboratories. The difficult templates were prepared and distributed to participating laboratories. Data and protocols were submitted by HTTP or FTP. Results were judged by both accuracy and quality. The effects of different types of instrumentation and chemistries were examined, including different configurations of the same machine type, different enzymes and dyes, and various reagent dilutions and cleanup methods. Both common and new technologies were analyzed. Results are continually posted on the web (http://www.abrf.org) for quality control, trouble shooting and decision making.

P-089
Analysis of cDNA clones from oligo-cap libraries and the improved oligo-capping method for cloning full length cDNA clones

Ai Wakamatsu[1], Toshio Ota[1], Kaoru Saito[1], Yuri Kawai[1], Shizuko Ishii[1], Jun-ichi Yamamoto[1], Tetsuo Nishikawa[1], Tomoyasu Sugiyama[1], Koji Hayashi[1], Yutaka Suzuki[2], Sumio Sugano[2], Yasuhiko Masuho[1], Takao Isogai[1], [1]Helix Research Institute, Kisarazu, Chiba, Japan, [2]Institute of Medical Science, University of Tokyo, Tokyo, Japan

Full-length cDNA clones are important tools in order to elucidate the functions of the genes experimentally. Full length enriched cDNA libraries were constructed from human cell lines and tissues by oligo-capping method (1), and then the 5'-end sequences of about 55,000 clones were analyzed in combination with ATGpr (2). The fullness ratios of 5'-end sequences including the intact ORFs were estimated by comparisons with sequences encoding known proteins. The fullness ratios were as follows: 61% (no selection), 71% (ATGpr: over 0.3), 77% (ATGpr: over 0.5), 82% (ATGpr: over 0.7) (sensitivity: 79%).

The ratio of full length cDNA clones obtained by oligo-capping method was extremely increased in combination with ATGpr. But it was still difficult to obtain full length cDNA clones shown low ATGpr values effectively.

It was difficult to construct cDNA libraries with highly full-length rate from small and low quality samples by the oligo-capping method previously reported by Suzuki et al. In this study, we have purified the tobacco acid pyrophosphatase, and improved the oligo-capping method. As the result, we succeeded to constructing cDNA libraries with extremely highly full-length rate from samples of about 1/100 or less than those used in the previous method. We constructed cDNA libraries of more than 10 kinds of human tissues and cell lines by this improved method. Full length rates in the 5'-ends of cDNA clonesof these libraries were distributed between 89% and 99%.

(1) Y. Suzuki et al. (1997) Gene 200:149-156; (2) A.Salamov et al. Bioinformatics (1998) 14: 384-390.

P-090
Complete Sequencing of the Ribosomal RNAs in Small Genomes

Jessica J. Vamathevan, Hoda M. Khouri, Haiying Qin, Herve Tettelin, Steve Gill, and John F. Heidelberg, The Institute for Genomic Research, Rockville, MD

In many bacteria, there are multiple ribosomal operons, each consisting of a 16S rRNA gene, 16S-23S intergenic spacer region (variable between operons), 23S rRNA gene and a 5S rRNA gene. The large size (approximately 6-kb) and a high degree of sequence homology between ribosomal operons, result in this region being problematic for correct assembly and closure. Further complications include the high amount of secondary structure, which can make PCR across difficult, especially when operons are located in tandem. Also, the promoter can be fully or partially toxic to *E. coli*, and the DNA sequence flanking the 16S rRNA genes is not represented in the clone library used for the random sequencing phase.

The techniques used to locate rRNA operon sequences differ between genomes, depending on the presence or absence of the promoter sequence, following the random phase. When the promoter is represented and the entire operon sequence is determined, a PCR product across the operon, from the unique flanking regions, is generated to confirm the assembly and intergenic region. If the promoter is not represented, many strategies are attempted to locate the 16S flanking sequence. These include screening a large insert, low-copy vector library, genomic walking, the CLONTECH GenomeWalker kit, combinatorial or multiplex PCR.

P-091
Trouble-Shooting In Genome Sequencing Projects, And The Construction Of Temperature-Sensitive Mutant Alleles For Functional Analyses Of The *Saccharomyces Cerevisiae* Genome

Guido Volckaert, Jan Van der Schueren and Johan Robben
Laboratory of Gene Technology, Katholieke Universiteit Leuven, Kardinaal Mercierlaan 92, B-3001 Leuven, Belgium.

Final assembly of individual contigs into a gapless chromosome or genome sequence is often, if not usually, hindered by the presence of regions that escape from routine sequencing. We have been focussing in recent years onto solving such difficult-to-sequence regions by optimization of PCR strategies to overcome problems such as polymerase stops, palindromic structures, perfect and imperfect repeat regions of various sizes, etc. Examples and solutions towards a gapless *Arabidopsis thaliana* chromosome IV sequence will be presented.

As one of our contributions to the functional analysis of the *Saccharomyces cerevisiae* genome, we are developing a general strategy for the construction of mutant strains with temperature-sensitive (*ts* and/or *cs*) phenotypes in essential ORFs. Efficient and generally applicable protocols to create such alleles have been developed and optimized by combining error-prone PCR technology and classical yeast genetics. As a first model, we used the essential gene *YNL006w* (involved in transport of permeases from the Golgi to the plasma membrane). Mutation rates of 1 per 100 bp and 1 per 400 bp were obtained reproducibly. Mutants alleles were introduced on a centromeric plasmid. From 3000 clones of a haploid *YNL006w*-disruptant strain, eight clones were confirmed to be *ts* due to mutations in *YNL006w*. This strategy is amenable to systematic large-scale generation of *ts* alleles of (essential) yeast genes.

Supported by the European Union projects BIO4-CT96-0338, BIO4-CT97-2294 and BIO4-CT98-0274.

P-092
DNA Microarray Probe Analysis and Quality Control

Bruce Wang, Rachel Nuttall, Gabor Bartha, Michael Doctolero, Teresa Carabeo, Eric Rollins, Chris Silk, Rick Johnston, Thomas Theriault, Georgina Hodgson, Eric Lachenmeier, Incyte Pharmaceuticals, Corporate Technology Development, Palo Alto, CA

As part of our continual effort to improve the data quality of our GEM microarrays, we have developed a gel-based electrophoresis method to examine fluorescent probe composition. Through a systematic examination of probe composition and hybridization signal, several critical parameters have been identified for the probe synthesis process, which affect data quality. To implement this predictive assay into our GEM production operation as a quality control measure, an in-house gel analysis software program was produced that automatically captures and analyzes the critical probe parameters. In anticipation of increasing throughput requirements for the probe QC assay, we have capitalized on Incyte's early efforts to increase DNA sequencing throughput. These efforts led to the previously described software and hardware modifications ("Metamorph") of the ABI 377 sequencing platform. Additional hardware modifications have now been made to the Metamorph platform in order to perform the probe QC assay. Probe analysis can now occur in a streamlined fashion at a throughput in excess of 1000 probes/day. Finally, a complete high-throughput automated sample preparation system was designed that will process tissues or cell samples into hybridization-ready probe.

P-093
Semi-Automated Solid-Phase Sequencing Reaction Purification

Haiyang Wang[1], Rohini Dhulipala[1], Peter Hewitt[1], Anders Holmberg[2], Scott Duthie[1], and Johan Wahlberg, [1]Amersham Pharmacia Biotech Inc., Piscataway, NJ, USA, [2]Kungl Tekniska Högskolan, Stockholm, Sweden

We have investigated a solid-phase purification procedure to separate sequencing fragments generated using either dye-labeled terminators or primers from unincorporated material. The method was developed using manual routines and then adapted to 96-well format on a robotic platform. Solid-phase purification involved the hybridization of a deoxyoligonucleotide covalently attached to magnetic particles with a complementary sequence just 3' from the sequencing primer. Following high affinity capture by magnetic beads, the sequencing fragments were separated from the remaining material using a magnet. The purified fragments were eluted from the beads and loaded onto either MegaBACE® 1000 or ABI PRISM® 377 DNA sequencing instruments for analysis. We present data showing that up to four different sequencing reactions may be multiplexed in one reaction tube allowing the possibility of sequencing two different templates simultaneously from each end. Additionally, the beads do not show signs of cross-contamination between captured reactions and have been reused up to 15 times without loss of sequence quality. In a 96-well plate format, this protocol will allow the simultaneous sequencing of 384 templates with reagent usage normally used for only 96 templates with average read lengths of >600 bases at 97% accuracy using pBluescript™ as the cloning vector.

P-094
ABI PRISM 3700 DNA Analyzer: A Fully Automated and High-Throughput System for GeneScan Applications

Yiwen Wang, Mimi Roque-Biewer, Ariana Wheaton, Katherine Rogers, Melissa Rivera, Anthea Dokidis, Allen Swei, Nathan Caffo and Penny Dong, PE Biosystems, 850 Lincoln Centre Drive, Foster City, CA

The 3700 DNA Analyzer is a fully-automated and high throughput capillary electrophoresis system for fragment analyses and DNA sequencing. It combines advanced robotics for flexible sample handling with high-sensitivity sheath flow detection and 2-D CCD imaging. With 96 samples analyzed per cycle, up to 15,000 genotypes can be generated per day with unattended operation.

We have developed the GeneScan application, GeneScan NT and Genotyper NT software for the 3700 DNA analyzer. High-performance fragment analyses have been routinely achieved using uncoated capillaries filled with replaceable polymer. Within a run the sizing precision is less than 0.15 nt standard deviation and 0.3 nt standard deviation between runs and instruments. The investigation and optimization of different aspects of experimental parameters will be presented. Furthermore, the adaptation of linkage mapping set reagents with the 3700 will be demonstrated.

P-095
Optimized Polyacrylamide Gel Matrix for High Throughput DNA Sequencing

David I. Wheeler, Chandra Krishnan and JoAnne H. Kerschner
Sigma-Aldrich Corporation, St. Louis, MO

Increasing throughput while maintaining data quality is important to any DNA sequencing project. The purpose of this project was to determine factors that influence sequencing read lengths, data quality and gel run times.

A series of matrix formulations were designed to study the relationship between gel percentages and run parameters such as voltage, run times and well to read length. Control sequencing reactions were run using the ABI Prism™ 377, 36cm and 48cm plates. We found that base resolution is related to the sieving characteristics of the gel and that further manipulations of the above factors can significantly increase throughput. The results of this project helped us develop a group of new gel formulations, AutoPAGE and AutoPAGE™ Plus, as well as optimized run parameters that can reproducibly resolve high quality bases with increased speed and greater gel stability.

P-096
Sequencing and Analysis of Full Length cDNAs in the Course of the German Genome Project

Stefan Wiemann, Wilhelm Ansorge, Helmut Blöcker, Helmut Blum, Andreas Düsterhöft, Karl Köhrer, Werner Mewes, Brigitte Obermaier, Annemarie Poustka, Rolf Wambutt, [1]Molecular Genome Analysis, German Cancer Research Center, Im Neuenheimer Feld 506, D-69120 Heidelberg,Germany, and the German cDNA Sequencing consortium

A consortium of eight sequencing laboratories and Germany's leading bioinformatics institute has formed in the frame of the German Genome Project. We aim at the sequence analysis of 3,000 to 4,000 complete novel cDNAs, comprising eight megabases of finished sequence. Sequencing started in September 1997 and a progress report of the consortium will be presented. The libraries generated in the course of the grant „Generation of full length cDNAs in the course of the German Genome Project„ are the primary source for sequencing. EST sequences of 12,000 independent clones are generated to identify novel genes. The EST sequences are analyzed for the likelihood of the clones to be full length (e.g. by the presence of CpG clusters) in order to obtain a minimal set of full length clones for efficient complete sequence analysis. Clones identified to be full length are sequenced and further analyzed by members of the consortium. The sequences are analyzed for possible function *in silico*. Functional analysis projects have started using the clones analyzed by the consortium as resource. All clones and data generated in the project are made publicly available via the Resource Centre of the German Genome Project (RZPD).

P-097
Development of a cDNA microarray manufacturing platform utilizing a thermal ink jet deposition system

Debra Wiest, Mike Caren, Doug Amorese, Jay Bass, Mike Bittner, Laurakay Bruhn, Herb Cattell, YiDong Chen, Larry DaQuino, William Fisher, Diane Ilsley, Paul Meltzer, Kyle Schleifer, Richard Tella, Jeff Trent, Peter Webb, Mark Westall, Hewlett Packard Company, Palo Alto, California and National Human Genome Research Institute, Bethesda, MD

We have developed a system for automated, high-volume manufacture of cDNA microarrays. The system consists of an on-the-fly deposition platform utilizing a thermal jet delivery device, an automated mechanism for loading the thermal jet device, and an inspection process for drop delivery verification. The integrated platform efficiently, reliably and accurately deposits large numbers of unique cDNAs into tightly packed arrays (>3,500 features per $1cm^2$). Arrays produced on the thermal jet system perform equivalent in hybridization assays to arrays generated using conventional pen technology, but offer advantages in feature quality, control of spot size, and large scale manufacturability. A description of the cDNA array-manufacturing platform as well as experiments assessing detection limits in two-color hybridization assays on thermal jet cDNA arrays will be shown.

P-098
Iterative Optimization of Oligonucleotide Arrays that Measure Gene Expression in *Saccharomyces cerevissiae*

Paul K. Wolber[1], Karen W. Shannon[1], Robert H. Kincaid[1], Andrew S. Atwell[1], Julja Burchard[2], Alan P. Blanchard[2] and Stephen H. Friend[2] [1]Hewlett-Packard Co., Palo Alto, CA and [2]Rosetta Inpharmatics, Kirkland, WA

Microarrays of surface-linked oligonucleotide probes are an emerging technology for exploring the genome-wide relationships among gene sequences and functions. Array probe sequence optimization must balance probe specificity and sensitivity. We have developed and applied a probe-picking algorithm that specifies the design of arrays fabricated by a novel, parallel printing process and an iterative refinement procedure that optimizes probe sets for desired genes. Arrays designed by these methods were used to measure the expression levels of ten yeast genes chosen to explore a range of mRNA levels and gene family sizes. The results show that multiple sensitive probes can be identified from multiple targets, in parallel, in 3-4 design iterations. Probe specificity can be simultaneously assessed *in silico* (by homology search algorithms) and experimentally (by use of appropriate control probes). The results demonstrate the power of fast turn-around, custom array fabrication to enable new approaches to oligonucleotide probe design and optimization.

P-099
Contamination of Human Genomic Libraries used for Large Scale Sequencing by E. coli IS186 Insertion Elements.

Andrew W. Womack, Mark O. Mundt, & Norman A. Doggett DOE Joint Genome, Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM

The E. coli insertion element IS186 is a 1343 bp transposable element which is present at three to four copies in the E. coli K12 genome. The transposon is flanked by a 23 bp inverted repeat and has been shown to insert preferentially into GC-rich targets. We have discovered 8 BAC and P1 clones from several widely used human genomic libraries which have a single copy of this insertion element included in the finished Genbank submission. These clones were sequenced at six different sequencing centers and in no case was the insertion element annotated as being derived from E. coli. The average G+C content of a 100 bp window on either side of the insertion site in all clones is very high at 75.8% and appears to be within CpG islands. In two of the 8 cases the insertion site is flanked by G+C-rich SVA repeat elements (a retroviral LTR class of repeat). Earlier studies of IS186 insertions into plasmids have shown that target duplications of 8 to 12 bp occur at the insertion site. By analyzing the flanking regions of the insertion sequence, we found an 11 bp duplication in one clone and 10 bp duplications in the remaining clones. We estimate the frequency of this insertion in finished clones to be approximately 1 in 1000 but the actual frequency could be much higher if this element has been removed from some finished sequences prior to submission.

Supported by the US DOE.

P-100
HPLC Purification of Differentially Expressed Gene Fragments

Lily Y. Wong, Victor Belonogoff, Victoria L. Boyd, Nathan M. Hunkapiller, Peter M. Casey, Sueh-Ning Liew Katherine D. Lazaruk, Susanne Baumhueter, Celera Genomics, Foster City, California, U.S.A.

GeneTag™ expression profiling is an amplified fragment length polymorphism (AFLP)-based technique that generates a large number of gene fragments, each of which is distributed into a unique bin by combining a multiplex PCR approach with size separation by capillary electrophoresis. In the GeneTag™ process cDNA is restricted with two different restriction enzymes and synthetic oligonucleotide "adapters" are ligated onto the sticky ends to serve as primer binding sites. The complexity of the PCR reaction is reduced by systematically lengthening the 3' end of the primer pairs by one or more nucleotides, thus decreasing the number of gene fragments in any one PCR reaction. Incorporation of a fluorescently labeled primer during the last PCR step allows the visualization of all gene fragments when separated by size using a 310 Genetic Analyzer. The difference in the expression level of a gene is determined by comparing peak heights. Reversed phase ion-pairing HPLC is used to isolate selected dsDNA fragments that represent differentially expressed genes. The purified fragments are sequenced directly, and the gene is identified using blast searches against various sequence databases.

P-101
Reaction Additive Improves Sequencing Through Difficult Templates

Lisha Xu, Alice C. Young and Robert W. Blakesley, Life Technologies, Inc., Rockville, MD, USA

The role of reaction additives in sequencing has been examined using a set of problematic DNA templates with both original dye chemistry and BigDye chemistry. The sequence characteristics of those problematic templates include GC-rich, AT-rich, and various direct or inverted repeats. Due to the complex secondary

structures created by these motifs the sequencing of those templates becomes a challenging task. In
the absence of reaction additives the sequence of these templates was of poor quality or in some cases totally uninterpretable. With the addition of reaction additives, a significant improvement in sequence quality was observed in most cases. Numerous additives were tested and shown to be effective to various degrees. Occasionally additive/template combinations produced poorer quality sequence. While the mechanisms of action of these additives are still not fully understood, the benefit of the additives on difficult DNA templates is clear, especially with GC-rich templates.

P-102
Simpler Plasmid Isolation Using Solid Phase Cell Lysis

Alice C. Young and Robert W. Blakesley, Life Technologies, Inc., Rockville, MD.

Alkaline Lysis is the most widely used method of producing lysates for plasmid purification. The reagents required for this procedure are simple and inexpensive; however, careful handling is required to obtain plasmid DNA free of *E.coli* genomic DNA. Excessive numbers of cells can overload the system causing inefficient lysis and reduced yields. The numerous mixing steps make this method variable and difficult to adapt to high throughput formats.

Here we describe a solid phase alternative to alkaline lysis, requiring no mixing or harsh chemicals. Alcohol precipitation completes the plasmid isolation procedure. The resulting plasmid DNA is low in RNA and genomic *E.coli* DNA, and it acts as an excellent substrate in automated fluorescent DNA sequencing and restriction endonuclease digestion. This solid phase method is adaptable to single tube, 96-well and 384-well formats and is compatible with automation. In addition, the chances of cross-contamination are minimized since there are no mixing or vacuum filtration steps. The simple handling and few reagents make this solid phase lysis method for plasmid isolation very rapid and robust.

P-103
Determination of the complete genomic DNA sequence of Thermoplasma volcanium GSS1

Hideaki Koike[1], Tsuyoshi Kawashima[1], Yoshihiro Yamamoto[2], Hironori Aramaki[3], Takeshi Kawamoto[4], Kazuko Nakamasu[4], Mitsuhide Noshiro[4], Tatsuo Nunoshiba[5], Koji Watanabe[6], Masaaki Yamasaki[6], Yoshie Ohya[1], Naoki Amano[1,7], Joeg M. Suckow[1], Masaru Tateno[1], Kozo Makino[8], and Masashi Suzuki[1] [1]AIST-NIBHT CREST Centre of Structural Biology [2]Department of Genetics, Hyogo College of Medicine [3]Department of Molecular Life Science, Daiichi College of Pharmaceutical Sciences [4]Department of Biochemistry, School of Dentistry, Hiroshima University [5]Biology Institute, Graduate School of Science, Tohoku University [6]Bioscience Research Laboratory, Fujiya Co. [7]Doctoral Program in Medical Science, Tsukuba University [8]Research Institute for Microbial Diseases, Osaka University

The complete genomic DNA sequence of Thermoplasma volcanium GSS1 of 1.6 M bases has been determined. Libraries constructed by using the BAC, l, and cosmid cloning systems were used for the determination of the sequence, together with products of LA-PCR, that bridged several gaps. The mean redundancy of the sequence determination was 8.7. On the basis of the determined sequence, genes, pseudo-genes, and operon structures have been identified by analyzing transcription and translation signals, using an informatical method. The transcription signals are the nucleotide sequences that are recognized by transcription factors, TBP and TFB, while the translational signals are the Shine-Delgano sequences that
are recognized by 16S ribosomal RNA. The identified genes are being analyzed in relation with some biological characteristics of this organism described in what follows. The optimum growth temperature of Thermoplasma volcanium GSS1 is 60°C and its optimum growth pH is 2.0. Because of its lack of a rigid cell wall, it has been discussed that this organism might be close to the organism that later evolved to the nuclei of eukaryotic cells through the endo-symbiosis with proto-mitochondrion (Searcy et al., 1981). All the archaebacteria whose genomic DNA sequences have been determined so far are adapted to single types of environments, either anaerobic or aerobic. In contrast, Thermoplasma volcanium GSS1 is aero/anaero-facultative and survives in both types of environments. Thus, the analysis of the genomic DNA sequence of Thermoplasma volcanium GSS1 will facilitate understanding of the transcription network in archaebacteria, since a systematic transcription regulatory mechanism is expected to this organism, in order to correspond to the alternative environmental changes.

P-104
A High-Throughput Robotic System for the Extraction of Nucleic Acids

Christian C. Oste[1], Luigi Jonk[2], Keith Osiewicz[3] and Dan Roark[4]. [1] BioScope International; [2].Packard Instruments, NVV, [3]Packard Instruments, Inc.; [4] CCS Packard

The last few years have seen an enormous growth in the number of sequencing projects, in a variety of fields, such as plants, animals, viruses, not to forget, obviously, the Human Genome Project.

Such progress has been made possible by remarkable advances in sequencing chemistries, as well as by the recent availability of high-throughput, capillary electrophoresis-based, sequencing platforms.
What has been lagging behind somewhat is the development of high-throughput methods for the extraction of the nucleic acids to be sequenced.

Here, we present a modular robotic system, based on a linear track, which has been married to a non-silica based chemistry for the sample preparation. The system, which has been designed initially to work in the 96-well format, provides a throughput of at least 10 plates-equivalent per hour.

The two-steps chemistry, which relies on protein scavenging as the first step, is very flexible and adaptable, and allows to process a wide variety of samples, such as plasmids, BACs, genomic DNA from whole blood or other biological fluids, and RNA. Data about the purity and integrity of the processed nucleic acids will be presented.

P-105
State of the Art High Throughput Sequencing on MegaBACE 1000

Michael A. Zaro, James Chapman, Lori Ferguson, Adam Friedman, Tom Goralski, Emma Katzman, Akbar Khan, Gregg Jones, Jingyue Ju, Khoi Nguyen, Careyna Peralta, Chris Silk, Victor Tong, Hong Yan, and Richard Cathcart, Incyte Pharmaceuticals, Inc., 3174 Porter Dr., Palo Alto, CA

The race to sequence the human genome has inspired the development of the next generation of DNA sequencers. At Incyte, we have met the demands of that race by optimizing the MegaBACE 1000 to reproducibly generate the longest read lengths in the industry while maintaining excellent data quality and base call accuracy. Thoroughly integrated into our high throughput operation, MegaBACE 1000 routinely sequences short and long EST, G/C rich, and genomic templates across a variety of organisms. This robust platform has proved its ability to provide extremely consistent data while requiring minimal maintenance and has enabled us to greatly increase our sequencing capacity. Data will be presented demonstrating the platform's talents in a high throughput sequencing operation, highlighting its long reads, excellent sequence quality, rapid turn around, consistency between runs, and instrument reliability.

P-106
Plasmodium falciparum: large-scale sequencing of chromosomes 10 and 11

Cummings, L.C., Gardner, M.J., Carucci, D.J., Shallom, S. J., Smith, H.O., Fujii, C., Mason, T.,Bowman, C., Craig Venter,J.. Cle, Fraser, C.F. and Hoffman, S.L., The Institute for Genomic Research, Rockville, Maryland , Malaria Program, Naval Medical Research Institute, Rockville, Maryland, Celera Genomics, MD

As part of an international consortium to sequence the entire genome of the malaria parasite Plasmodium falciparum, The Institute for Genomic Research (TIGR) is sequencing chromosomes 10 and 11. This initiative is funded by the National Institute of Allergy and Infectious Disease, National Institutes of Health. Unlike many genome sequencing programs carried out at TIGR, the Plasmodium falciparum genome is being sequenced using a chromosome-based approached. Malarial chromosomes are fractionated by pulse-field gel electrophoresis, mechanically sheared and DNA fragments cloned into plasmids.

The shotgun sequencing phase for chromosome 11 has been completed and approximately 76,000 sequence reads are available to the malaria research community through our web site (http://www.tigr.org/tdb/parasites/index.html). Closure of sequence and physical gaps on chromosome 11 is in progress, aides by an NheI optical restriction map created by Dr. David Schwartz (University of Wisconsin, Madison) and 49 chromosome-specific microsatellite markers that allow sequence contigs to be ordered. At present, 85% of the chromosome length is represented in assembly groups larger than 80 kb and work is underway to generate additional microsatellite markers for chromosomes 10 and 11 (Xin-zhuan Su and Thomas Wellems, National Institute of Allergy and Infectious Diseases, NIH).

The shotgun sequencing of chromosome 10 is in progress and approximately 23,000 sequences (~ 4x genome representation) are available on our website; the shotgun phase will be completed by October with 60,000 sequences available at which point, chromosome 10 closure will begin.

When complete chromosomal sequences are available, they will be annotated using a variety of software tools developed for eukaryotic sequence analysis. Fully annotated sequences will be accessible through the TIGR website upon publication.

Other research centers participating in the malaria genome sequencing are The Sanger Center, U.K. (http://www.sanger.ac.uk/) and Stanford University (http://sequence-www.stanford.edu/).

P-107
Diversity in Fungal Proteomes: A Comparative Analysis of Public Fungal Sequences

Qiandong Zeng, Marco Kessler, Guillaume Cottarel, Anthony Caruso, Andrew DePristo, and Skip Shimer. Genome Therapeutics Corporation, Waltham, MA

We compared fungal proteomes from the complete sequence of the *Saccharomyces cerevisiae* genome, the nearly complete genomic sequences of *Candida albicans* and *Schizosaccharomyces pombe*, and the relatively large collection of cDNA sequences from *Aspergillus nidulans*, *Neurospora crassa* and other fungi in the public domain. Our analysis show that a substantial portion of the genes in *Candida albicans*, *Schizosaccharomyces pombe* and other fungi examined are not shared with *Saccharomyces cerevisiae*. We also found human gene homologues in several fungi, but not in *Saccharomyces cerevisiae*. These observations suggest that extensive divergence exist among in the fungal kingdom, and, in addition to *Saccharomyces cerevisiae*, other fungi should also be examined in future studies for anti-fungal drug development.

We thank Stanford University and The Sanger Center for making the *Saccharomyces cerevisiae*, *Candida albic*ans and *Schizosaccharomyces pombe* genomic sequences available to the public. We thank the University of Oklahoma for releasing the fungal cDNA sequences.

P-108
The Umpire System — Automated Processing of DNA Sequences

Don Sleeter, John Wilkinson, Eryk Vershen, Tim Baxter, Tai Nguyen, Sid Cowles, Richard Linton, Rick Cathcart, Incyte Pharmaceuticals, Inc. Palo Alto, CA

DNA sequencing efforts are often bottlenecked by data processing and analysis tasks. The Umpire system is Incyte's enterprise-wide chromatogram analysis system. It improves efficiency of DNA sequencing by speeding up the processing of data and providing more accurate and reliable process information, at reduced cost.

Umpire automates the analysis and processing of large volumes of DNA sequence data that otherwise would need to be monitored and analyzed by trained experts. It is a scalable client/server system that processes chromatograms from any DNA sequencing platform. Results are recorded in relational databases and distributed across the Internet using small data files called scorecards.

The server-side Umpire processing daemon performs basecalling, quality scoring and uses a flexible rule-based QC system to categorize the pass/fail status of chromatogram data uploaded from any source.

A Java-based Umpire client application provides a user interface for human review of analysis results over the Internet, including chromatograms and process-oriented statistics. The client/server system provides a wealth of information to technicians and management as early as possible, reducing errors and workload.

P-109
OGI™ - Java Based Software for Gel Analysis

Hong Guo, Mark S. Welsh, Martin D. Leach. Bioinformatics, CuraGen Corporation, New Haven, CT.

Large scale sequencing projects require high quality gel analysis without compromising on speed. To meet such needs, CuraGen has developed OGI™ (Open Genome Initiative), a web-based client-server application in Java for high-throughput gel analysis. This client-server design allows an operator, using any web browser, to control processing on many OGI™ servers, each of which takes output from several sequencers. Currently, OGI™ supports sequencing on the ABI 377 and MegaBACE™ 1000 machines. Within a web browser, the Java applet communicates with the server using RMI (Remote Method Invocation). A multi-threaded Java application on the server schedules CPU-intensive image processing steps. Sequence traces are analyzed using CuraGen's versatile DOLPHIN™ trace processor, and then base-called using PHRED (Ewing *et al.*, 1998). OGI™ has been designed as an open and extensible framework, which will accept new processing steps and whole new data-flows with ease. The ability of OGI™ to coordinate data processing and analysis using the internet makes it ideal for sequencing facilities. OGI™'s Java and ANSI-C executables will be made available through our web site: www.curagen.com.

This research was supported by a grant from HIH.

P-110
The Mouse Genome Sequence Database: A Resource for Integrating Sequence and Biology.

Carol J. Bult, Judith A. Blake, Janan T. Eppig and the Mouse Genome Informatics Group. The Jackson Laboratory, Bar Harbor, ME.

The laboratory mouse has been and will continue to be one of the primary model systems for studying the molecular basis for human biological processes and for investigating mammalian gene function and phenotype. The preeminence of mouse as a model system for human biology stems from three major factors: 1) the evolutionary relatedness between mice and humans, 2) the physiological and biochemical similarities among mammals generally, and 3) the ease and cost-effectiveness with which mice can be manipulated genetically.

As the efforts to sequence the human and mouse genomes accelerate, we will soon be in an era of direct sequence comparison between the complete genomes of these two species. The Mouse Genome Sequence (MGS) database is being developed as informatics resource to connect biologically significant features identified in mouse genomic sequence to existing biological knowledge (e.g., phenotype, expression data, nomenclature, etc.) about the laboratory mouse. MGS is part of the informatics infrastructure that is necessary to exploit the explanatory power sequence-level comparative genomics can provide to the research community.

Supported by Department of Energy grant DE-FG02-96ER62327 and NIH grant HG01559 to CJB.

P-111
Mouse Genome Informatics: Expanding Phenotypic Resources

J.A.Blake, J.E.Richardson, J.T.Eppig, and the Mouse Genome Informatics Group. http://www.informatics.jax.org, The Jackson Laboratory Bar Harbor, ME

The Mouse Genome Database (MGD) is a comprehensive database of mouse genetic and biological information (see *Merriam* poster). MGD has always supported the description of gene function and of mutant phenotypes; now rapid expansion of the number of genes, the complexity of their relationships, and the desire to ask biological rather than sequence driven questions is shaping the expansion of phenotypic representations. Controlled vocabularies to describe gene function, biological process and cellular location of gene products are being incorporated as part of a collaboration with yeast and *Drosophila* genome databases.

Additional controlled vocabularies for describing mutant phenotypes and other phenotypic terms are being developed. Collaboration with domain experts on gene family classification and nomenclature has also significantly improved gene representations Similarly, close coordination of the relationship between gene objects in MGD and Swiss- Prot results in the linking of MGD gene family information with Swiss-Prot protein family information. Multiple classification systems and their various presentations in MGD are presented and the advantages and disadvantages discussed.

Supported by NIH Grant HG00330

P-112
Auto Primer: Generating Primers Automatically Using Stored SYBASE Data

Jeffrey R. Buchoff, Terry Shea and Terry Utterback, The Institute for Genomic Research, Rockville, MD, Anne Deslattes-Mays, Celera Genomics, Rockville, MD

As we begin working with larger genomes, the time needed to design each primer and check for uniqueness manually has slowed the closure process significantly. We are developing Auto Primer to speed up the process of designing unique primers to fill physical gaps, sequence gaps, and walk off the ends of existing oligos. Auto Primer is designed in Java to run on a variety of platforms, including UNIX, LINUX, NT, and Macintosh.

Auto Primer will speed up the process of gap closure by designing multiple primers for each sequencing and physical gap stored in our project databasest. Auto Primer will also be able to design new primers on existing walks without reassembling the project. By using the constraints offered by Primer3, an existing primer designing program, we can increase the accuracy of the primers developed. Once the primers are designed, Auto Primer will check each primer against its assembly or the entire project, depending on what type of gap we're working with, for uniqueness. These unique primers will be used to generate new sequences, which will help close the remaining gaps in our projects to get a complete assembly.

P-113
Integration of ABI PRISM® 3700 DNA Analyzer Generated Sequence Samples with phred, phrap, cross_match, and consed Analysis Tools in a Relational Database Model

Greg A. Harrington, Jacob E. Holland, PE Informatics, a Division of PE Biosystems, San Jose, CA

With the very large data sets generated by the ABI PRISM® 3700 DNA Analyzer, new techniques are required to successfully manage this flow of data and to provide easy, reliable access for all consumers of this information. A relational database model is one possible solution for an increasingly connected and dynamic laboratory environment. We have demonstrated that this model is feasible and offers significant benefits.

This model is valuable in that it allows for data centralization, improved data security and data access. Mechanisms are readily available that allow for real-time access to core data and easy generation of summary information that is difficult to tabulate with a traditional file-based approach. For example, a quality assurance summary could include data reported by instrument type, a specific instrument, run time , run conditions, etc.

Analysis tools such as phred, cross_match, phrap and consed have been integrated (without modification) into the relational database model to demonstrate that useful and common tools can be included without a degradation of performance or other metric.

P-114
The Genomic, Comparative and Functional Analysis of Hyperthermophilic Cren- and Eury-archaea

Yutaka Kawarabayasi[1,2], Yumi Hino[1], Hiroshi Horikawa[1], Koji Jin-no[1], and Hisasi Kikuchi[1], [1]Biotechnology Center, N.I.T.E., Shibuya-ku, Tokyo, Japan; [2]N.I.B.H., Tsukuba, Ibaraki, Japan

In our institute the entire genomic sequences of an anaerobic hyperthermophilic euryarchaeon, *Pyrococcus horikoshii* OT3[a]) (1,738,505bp), and an strictly aerobic hyperthermophilic crenarchaeon, *Aeropyrum pernix* K1[b]) (1,669,695bp), were determined. The entire nucleotide sequence and additional information of these two microbials are released on our internet homepage (URL:http://www.mild.bio.go.jp). Gene functions of 25% of total ORFs are estimated by similarity analysis. From the results of comparative analysis, it is indicated that these two strains are more close than these positions shown in the *evolutional tree*. From the comparative analysis among three *Pyrococcus*, it is indicated that 66% of total ORFs are common, but approximate 20% of total ORFs in *P. horikoshii* OT3 are absent in both of two other species. On the other hands the functional analysis of these two strains prediction are performed by collaboration with many laboratories. Already some kinds of proteins in *P. horikoshii* OT3 are expressed in *E. coli*, analyzed of these activities and dissolved of these 3D structures. These results will be discussed in this meeting. Now we are testing which strains in *Crenarchaeota* is fit for determination of the entire genomic sequence. In this meeting probably we will discuss the comparative analysis among two species in *Crenarchaeota*.

a) DNA Res. Vol.5, No.2, 55-76 & 147-155 (1998)
b) DNA Res. Vol.6, No.2, 88-101 & 148-155 (1999)

P-115
Annotation Scheme of the Rice Genome Research Program

Isamu Ohta, Hideki Nagasaki, Atsuko Idonuma, Yoshiyuki Mukai, Masatoshi Masukawa, Manami Negishi, Baltazar A. Antonio, Katsumi Sakata and Takuji Sasaki.Rice Genome Research Program (RGP), National Institute of Agrobiological Resources / Institute of the Society for Techno-innovation of Agriculture, Forestry and Fisheries, Tsukuba, Ibaraki, Japan.

With the start of genome sequencing, the Rice Genome Research Program (RGP) has entered a new stage and is expected to play a pivotal role in the International Rice Genome Sequencing Project. Initially RGP concentrates on completing the sequence of chromosomes 1 and 6. The sequence data are analyzed for relevant genetic information and immediately submitted to public databases. We also release the sequence data to the public domain through our database INE (INtegrated rice genome Explorer, http://www.dna.affrc.go.jp:82/) that can be accessed on our website. Our annotation scheme makes full use of various sequence analysis programs and integration of results to predict the genes in the genomic sequence. The annotated sequence we release in our database includes the position of the predicted genes as well as the details of the output of each of the programs used for analysis. In addition, the annotation of genomic sequence is displayed in our database INE together with the corresponding genetic and physical mapping information. Therefore this can provide relevant information that can be utilized by researchers interested in a particular gene or a particular region of the chromosome.

P-116
Sequence Database Management using Virtual Network File Servers

Roger A. Sayle and J. Scott Dixon, Metaphorics LLC, Santa Fe, New Mexico.

Previous attempts to curate bioinformatics data in relational or object-oriented database management systems are plagued by the problem of backward compatibility. Virtually all-existing bioinformatics tools work on flat files. Many applications, such as BLAST, SRS or GCG, even require their own binary file format. This creates the administrative problem of keeping the duplicated information synchronized. The need to subset and superset databases, for ESTs, non-redundancy or by organism, exaggerates the problem for already large data sets. The apparent alternative is modification of the application source code to use the DBMS schema, negating the potential benefits of having an application specific file format or index.

One potential solution is the Virtual Network File Server. A VNFS server emulates the behaviour of a remote file system, using a protocol such as NFS or Microsoft SMB. Like a WWW server, a VNFS server receives requests to read or write files from a client application and generates the appropriate responses and acknowledgements. Unlike CORBA, no modifications are required to the legacy application software or even to the client machine. The server is free to choose the internal representation of the data, rapidly generating flat-file formats, such as FASTA or BLAST indices, on the fly. This allows a VNFS server to act as a gateway to a relational database or a CORBA ORB.

P-117
High-throughput DNA Sequencing with the ABI Prism® 3700 DNA Analyzer

Trisha A. Moore[1], Jeffrey Bates[2], Mary M. Blanchard[1], [1]Monsanto Company, St. Louis, MO 63167, [2]Cereon Genomics, Cambridge, MA 02139

DNA Sequencing is an integral component of Monsanto's genomic endeavor. A high-throughput approach is being utilized to sequence DNA for product discovery in the areas of agriculture, health, and nutrition. The 3700 DNA Analyzer presents an exceptional opportunity to increase capacity while decreasing the hands-on-time and subsequent inconsistencies of slab gels without sacrificing sequence quality.

Since installing the instrument, numerous parameters have been optimized to achieve the maximum read lengths possible with plasmid DNA while retaining exceptional quality. These include instrument controls such as run temperature, voltage, and sample injection time. In addition to instrument optimization, modifications were made in sample preparation. Such modifications include sample precipitation and resuspension.

In accordance with its high-throughput objective Monsanto has recently implemented sequencing in 384 well plates in conjunction with the 3700 usage, thus enabling the maximization of available resources.

P-118
Accelerating Discovery Through Life Science Domain Integration

David George, Ph.D., Director, Customer and Strategic Consulting, NetGenics, Inc., Cleveland, OH

There is a growing awareness that life sciences research is fundamentally dependent on an enterprise's ability to manage, interrelate, and interpret the rapidly accumulating body of research information. Interrelating disparate information constitutes a large-scale problem in the integration of heterogeneous databases and software tools that were not originally designed to communicate across the bounds of their domains. Typically, databases and software tools are distributed across a variety of servers deployed on multiple computer platforms. NetGenics has developed the SYNERGY™ framework, a CORBA-Java architecture designed to address information integration problems in life sciences research. We will examine the SYNERGY framework in the context of integrating gene expression, biosequence and metabolic pathway data and statistical analysis tools. The ability to combine gene expression and biosequence data provides a basis for identifying unknown gene expression targets via homology modeling. Combining pathway data with measurements on induced gene expression patterns can provide insights into induction mechanisms. Conversely, modeling gene expression patterns can provide insight into novel regulatory pathways.

P-119
DNA Microarray Tools for Genomic Sequence Data Mining

Mark Schena, TeleChem International, Inc., 524 E. Weddell Drive, Suite 3, Sunnyvale, California 94089-2115, USA.Ronald W. Davis, Department of Biochemistry, Beckman Center, Stanford University Medical Center, Stanford, California 94305-5307, USA.Todd Martinsky, TeleChem International, Inc., 524 E. Weddell Drive, Suite 3, Sunnyvale, CA

The rapidly expanding human sequence databases, coupled with the already available genomic sequences from many prokaryotic and eukaryotic organisms, provide an unprecedented opportunity for comprehensive analysis of complete genetic blueprints. Microarray technology is increasingly becoming the platform of choice for hybridization-based functional genomics. Microarrays of oligonucleotides and cDNAs allow massive, parallel analysis of gene expression patterns and single nucleotide polymorphisms (SNPs) on a genomic scale. Continued innovations in microarray manufacturing technology, clean-room and demonstration facilities, surface and hybridization chemistries, fluorescence detection, and data mining software provide a complete "tool kit" for microarray researchers. DNA sequencing and microarray experimentation will revolutionize our understanding of human health and agriculture by providing deeper insights into disease onset and progression, drug action and toxicity, and the mechanisms of crop plant development and herbicide action. Microarrays might also find use in diagnostics applications by providing cost-effective information to health care providers.

P-120
Software Methods For Large Scale Sequence Reconstruction

James D. Candlin, Paracel, Inc., Pasadena, CAGennady Denisov, Paracel, Inc., Pasadena, CAEric Gaidos, Paracel, Inc., Pasadena, CAXiaoqiu Huang, Paracel, Inc., Pasadena, CALicen Xu, Paracel, Inc., Pasadena, CATim Hunkapiller, Paracel, Inc., Pasadena, CA

The constant introduction of new technologies in sequencing, the ever-increasing volumes of data, and a strong need to pursue gene discovery and characterization on partial data sets pose substantial challenges for software that base-calls, assembles and clusters reads from automated DNA sequencers.We are developing a series of components to address these needs, including chromatogram analysis, screening of unwanted sequences, identification of repeats, accurate and rapid calculation of pairwise overlaps, clustering and layout, and generation of contigs. These components are built for maximum sensitivity while being scalable to large problems. We have also put in place a flexible framework to connect them together in specific applications, and we will illustrate an EST clustering application built in this way.It is important for a complex process applied to a large scale problem that it be parameterized and validated well. We have designed a series of models and metrics for assessment, both for the ensemble as well as the individual components. We will present the results and compare the performance with previously available tools.

P-121
GATEWAY Cloning Technology: A High-Throughput Gene Transfer Technology for Functional Analysis and Protein Expression

Dr. Michael Brasch, Senior Scientist. Dept of Gene Expression and Protein Analysis, Life Technologies Inc., Rockville, MD

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

P-122
Navigation, Visualization, And Query Of Genomes: The Genome Channel And Beyond

R. Mural, M. Parang, M. Shah, D. Hyatt, M. Land, J. Snoddy, E. Uberbacher, and the Genome Annotation Consortium. Computational Biosciences Section, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN

The Genome Channel Browser is a Java based viewer capable of representing a wide variety of genomic-sequence annotation and links to a large number of related information and data resources. It relies on a number of underlying data resources, analysis tools, and data-retrieval agents to provide an up-to-date view of genomic sequences as well as computational and experimental annotation. The current version of the Genome Channel Browser (v2.0) provides a diverse set of functional features. New features in this version of the Genome Channel include: additional features such as tRNA and BAC ends, additional organisms including microbes, genetic and radiation hybrid maps, extended and detailed listing of features and generation of summary reports, text-based searches and query of underlying data, BLAST searches against individual or combined assembled sequences and products, and pattern searches against genomes that return genome location and context of related sequences (Visual Genome Search Server). We are also interested in forging new collaborations to add value to the genome sequence and annotation framework.

Supported by the Office of Biological and Environmental Research, US DoE, contract DE-Ac05-84OR21400 with Lockheed Martin Energy Systems, Inc.

P-123
Generation of Transcript Profiles by a High Throughput Pyrosequencing Strategy

Charlotta Malmqvist[1], Maria Sievertzon[1], Anders Holmberg[1], Anders Alderborn[2], Magnus Larsson[1], Mathias Uhlén[1] and Joakim Lundeberg[1], [1]Department of Biotechnology, KTH, Royal Institute of Technology, Teknikringen 34, 100 44 Stockholm, Sweden [2]Pyrosequencing AB, Uppsala, Sweden

Transcript profiling has become one of the key components to elucidate gene function. One possibility is to employ microarray technology. However, one of the limitations with this is that it requires micrograms of mRNA. We have developed a complementary strategy suitable for small tissue samples. The strategy relies on a new high throughput liquid based DNA-sequencing principle, pyrosequencing, and 3'tagged cDNA libraries. The pyrosequencing principle is based on an iterative sequencing by synthesis strategy in which single specific nucleotides are added to an extension substrate in presence of a DNA polymerase. Successful incorporation is detected through an enzymatic cascade that produces a quantitative light signal, measured by a luminometer. Approximately 20-30 bases of each cDNA are sequenced, which is sufficient for identifying a gene by database comparison. The obtained transcript profiles from compared libraries are visualized by virtual chip technology. For the analysis of only a few cells, a cDNA amplification step, keeping the relative transcript levels, is used in the generation of the libraries. As little as 1 ng total RNA (approximately 1000 cells) may be used as starting material for the amplification strategy. The present objective is to study the development of atherosclerosis. Two 3'tagged cDNA-libraries have been investigated derived from macrophages treated with oxidized LDL representing foam cells, which are key components in atherosclerotic plaques.

Index to Electronic Posters

Sunday, September 19
1:00 – 3:00pm

Littlejohn, Tim	E-01	Network Computing For Bioinformatics
Tatusova, Tatiana	E-02	From Genome To Protein Sequence To 3D Structure: Protein Neighbors In Entrez Genomes
Saeed, Alexander I.	E-03	Tools For Display And Analysis Of High Density Array Data
Piétu, Geneviève	E-04	The Genexpress Image Knowledge Base : A Prototype Resource For Characterization Of Gene Transcripts Involved In Muscle Or Brain Functions
Fernandes, Elma R.	E-05	Gene-Based Drug Discovery In The Genomics Era
Burgess, C.E.	E-06	To Be Or Not To Be - Novel Human Orthologs Which Replace Accepted Human Orthologs
Bandaru, Raj	E-07	Integrating Multivariate Methods And Evolutionary Approaches For Predicting The Function Of Novel Genes
Uberbacher, E.	E-08	An Infrastructure For Genome Analysis And Annotation
Sasinowski, Maciek	E-09	A Software Tool To Construct And Edit Physical Maps From Fingerprinting And Multi-Level Hybridization Data
Prayaga, Sudhirdas K.	E-10	Needle In A Hay Stack: Identifying Functional Genes Within Genomic Sequence By Novel Sequence Analysis And Validation Strategies
Pruitt, Kim D.	E-11	Integrated Information About Genes
Slaven, Bradley E.	E-12	Annotation Master: A Parallel Computing System To Facilitate High-Throughput Genome Annotation

Monday, September 20
1:00 – 3:00pm

Sharov, Vasily A.	E-01	Image Processing Software For Microarray Expression Analysis
Upton, Jonathan F.	E-02	The TIGR Gene Indices: Estimating The Transcript Content Of Genomes
Holt, Ingeborg	E-03	Incorporation Of RH Mapping Data And Orthologous Pairs In The Gene Indices
Davis, Brian P.	E-04	The Bioknowledge™ Library - A Collection Of Curated Model Organism Proteome Databases With Applications For Comparative And Functional Genomics
Merriam, Jennifer J.	E-05	The Mouse Genome Database - Mgd
Liang, Feng	E-06	Building Reliable Consensus Sequences For Gene Indexing: A Comparison Of Sequence Assembly Programs

Leach, Martin D.	E-07	Rapid Identification, Classification, And Organization Of Gene-Based SNP's
Ermolaeva, Olga	E-08	EGAD: A Nonredundant Database Of Genes And Proteins
Hardin, Susan H.	E-09	Octamer-Primed Sequencing Technology: Development Of Primer Identification Software
Han, Cliff S.	E-10	Selection Of Unique Sequences From The Unigene Database For Hybridization Purpose
Cuticchia, A. Jamie	E-11	The Continuation Of The GDB Human Genome Database And Its Evolution With The Human Genome Project
Vovis, Gerald F.	E-12	Haplotyping Genes Important To The Pharmaceutical Process And Drug Response
Gordon, David	E-13	Consed And Autofinish In The Phred/Phrap/Consed System

Electronic Poster Abstracts

Electronic Poster Session I

Sunday, September 19
1:00 – 3:00pm
Champagne Room

E-01
Network computing for bioinformatics

Tim Littlejohn, eBioinformatics Inc (eBioinformatics.com), Pleasanton, CA., USA / Sydney, NSW, Australia

The genome science community has lead the sciences in the development and adoption of WWW based resources. Web-based access to bioinformatics tools (databases, software, computing) has been widely embraced by the genomics research and education community due to the ubiquity of browsers and their ease of use as well as the relative simplicity of publishing resources using this technology. However traditional approaches to delivery and use of bioinformatics resources through the web all fail to address many of the needs of genome analysis systems including: 1) high throughput automated analyses; 2) off-line remote processing; 3) virtual desktops for remote storage of data; 4) common user interface with a comprehensive array of tools and databases; 5) automatic searching and researching of growing data sources; 6) secure, collaborative data sharing environment; 7) alternative methods for accessing bioinformatics resources (based on perspectives held by researchers eg. "tool", "data type", "problem" or "package" views); 8) consistent architecture suitable for novice and expert users alike and; 9) coordinated, centralized scalable computing resources. The eBioinformatics *BioNavigator* system (bionavigator.com) addresses many of these deficiencies and offers an optimal approach to the broad-based access to genome bioinformatics tools, providing a virtual bioinformatics facility accessed on a time-share basis, founded around the application service provider (ASP) software delivery model.

E-02
From genome to protein sequence to 3D structure: protein neighbors in Entrez Genomes

Tatiana Tatusova, Yanli Wang, Steven Bryant. National Center for Biotechnology Information National Library of Medicine National Institutes of Health, Bldg. 38A, 8600 Rockville Pike, Bethesda, MD 20894

Entrez information system is based on a set of integrated databases containing complete genomes, nucleotide sequences, protein sequences, taxonomy database and 3D structures.

Entrez Genomes displays data from small viral and organelle Genomes,complete and near-complete genomes from bacteria and lower eukaryotes. Flexible web based views, pre-computed relationships, and immediate access to analytical tools provide scientists with the new insights to be gained from completed genome sequences.

Using protein sequence similarity information from BLAST search, each gene from the 21 complete genomes in GenBank database was searched against "nr" database(all non-redundant GenBank CDS translations + PDB + SwissProt + PIR). Neighbor relationships to the proteins with known 3-dimensional structures were detected.

The detected homologs were classified into three major Phylogenetic groups, Eukaryota, Eubacteria and Archaea. In addition the neighbors to protein sequences from PDB database were selected and linked to Cn3D viewer and MMDB (The Molecular Modelling Database). Cn3D viewer allows to display simultaneously three-dimensional structures, sequences, and alignment. MMDB contains experimentally determined biopolymer 3D structures obtained from the Protein Data Bank (PDB), and provides the users with a pre-computed structure neighbors with VAST(the Vector Alignment Search Tool), the NxN record of "neighboring structures" that often identify distant homologs.

E-03
Tools for Display and Analysis of High Density Array Data

Alexander I. Saeed, John Quackenbush, The Institute for Genomic Research, Rockville, MD

Microarrays provide the opportunity to study gene expression patterns using thousands of genes in a single assay. The interpretation of such data requires advanced tools for data analysis and visualization. We have developed a package of Java tools to facilitate microarray analysis that allows the user to display graphical representations of hybridized slides, retrieve experiment information from a database, perform statistical tests on the data to identify differentially expressed genes, and view and export the results. Expression data are retrieved from either a database or a flat file and a representation of the hybridization is generated. Display elements representing the arrayed genes are colored to represent relative expression; options include false color, green/red overlay, and an easy-to-interpret ratio display. Users can click on array elements to retrieve information about the underlying gene from a database using JDBC. Fluorescence intensities can be normalized using a number of strategies. This software, freely available to academic researchers, represents a first-generation visualization tool for analysis of expression measurements. Future enhancements include multiple experiment views and dendograms illustrating relationships in gene expression patterns.

E-04
The Genexpress Image Knowledge Base : A Prototype Resource For Characterization Of Gene Transcripts Involved In Muscle Or Brain Functions

Geneviève Piétu, Régine Mariage-Samson, Eric Eveno, Charles Decraene, Nicole A. Fayein, Christiane Matingou, Fariza Tahi and Charles Auffray. Genexpress, CNRS ERS 1984, BP 8, 94801 Villejuif, France.

The Genexpress team has conducted an integrated approach for the analysis of human muscle and infant brain cDNA libraries by collecting sequence and mapping data. Partial sequences of cDNA clones corresponding to the same transcript have been clustered and registred in the Genexpress Index. We have exploited this catalogue representing some 15,000 distinct gene transcripts by characterizing their expression profiles. Using macro-array, we have collected the expression profiles of 910 gene transcripts expressed in skeletal muscle and about 6,000 gene transcripts expressed in infant brain. Differential expression of the gene transcripts was detected with complex cDNA probes derived from various tissues allowing us to identify novel gene transcripts with muscle-restricted or brain-restricted patterns of expression. A systematic effort was undertaken to further integrate these expression profiles with available cDNA sequence clustering and gene mapping informations from the Genexpress Index. These results are available on a dedicated Web site at http://idefix.upr420.vjf.cnrs.fr/EXPR/welcome.html to form the brain and the muscle modules of the Genexpress IMAGE Knowledge Base, a prototype integrated resource for the study of muscle and brain physiology and pathologies.

E-05

Gene-Based Drug Discovery In The Genomics Era

Elma R. Fernandes, Sudhirdas K. Prayaga, Catherine E. Burgess, Raj Bandaru and Richard A. Shimkets. CuraGen Corporation, New Haven, CT

CuraGen has generated a catalog of approximately 1.7 million expressed human gene sequences to accelerate the pace of drug discovery. We have strategically sequenced human samples to get adequate representation of rare tissues and transcripts and integrated this with public data to generate a comprehensive human tissue expression database. We have analysed this SeqCalling™ database using CuraTools™, our integrated suite of DNA and protein sequence analysis tools. This has yielded several novel homologs of secreted and membrane-associated genes for development as protein therapeutics, diagnostics and drug targets. The secreted proteins include novel homologs of growth factors and cytokines in clinical trials as well as proteins which have no homology to known proteins and were predicted as secreted proteins using sub-cellular localization analysis. Curagen's SeqCalling™ database also includes several novel homologs of conventional drug target classes like G protein-coupled receptors, ion channels, and nuclear hormone receptors. Here we describe the strategies we use to identify and validate these novel genes from our SeqCalling™ database.

E-06

To Be or Not To Be - Novel Human Orthologs Which Replace Accepted Human Orthologs

C.E. Burgess, R. Bandaru, E.R. Fernandes, R.A. Shimkets. CuraGen Corporation, New Haven, CT 06511

CuraGen has employed a novel normalization strategy to establish a human gene expression database consisting of more than 3.5 million EST sequences (public and private). Scanning this database using CuraTools, our sequence analysis package, we have identified hundreds of novel homologs/orthologs of known genes and frequently found that our novel protein was more similar to a species ortholog than the currently described human ortholog. One of many examples is the human neurotrimin gene (CG_NTRI) which has 94% amino acid (aa) similarity and 87% nucleotide (nt) identity to NTRI_RAT. It replaces OBCAM, the accepted human ortholog, which shows only 86%aa-similarity and 65%nt-identity to the rat gene. Similarly, CuraGen's D-dopachrome tautomerase (CG_DOPD) has 83% aa-similarity and 93% nt-identity to DOPD_MOUSE and replaces DOPD_HUMAN which has only 80%aa-similarity and 63%nt-identity to the mouse gene. Functional assays are currently underway to confirm the assignment of these proteins and others as the appropriate human orthologs.

With less than 12% of the human genome considered 'finished' sequence, it is likely that a number of accepted orthologs of mammalian genes will be relegated to 'family member' status as new and more orthologous gene sequences take their place. These new orthologs may ultimately account for the functional differences we see in the currently accepted mammalian orthologs.

E-07

Integrating Multivariate Methods and Evolutionary Approaches for Predicting the Function of Novel Genes

Raj Bandaru, Catherine E. Burgess, Sudhirdas K. Prayaga, Elma R. Fernandes, and Richard A. Shimkets. Curagen, Corporation, New Haven, CT

Predicting function when only a small fragment of the gene is available can often be difficult. Simple homology algorithms and motif / pattern matching may yield either misleading or inconclusive associations. Additionally, manual analysis of all available information to predict function can be tedious, especially with many gene fragments having little or no homology to known proteins. Here we describe an efficient approach which can rapidly cluster and annotate a large number of novel gene fragments with higher confidence. This analysis assigns gene fragments to functionally similar clusters of known proteins based on evolutionary and amino acid sequence characteristics. Assumptions are made based on evolutionary theory, such as correlations between proteins with similar functions and selection pressures, expression and codon usage, mutation bias, etc. These factors, combined with other protein characteristics and with the evolutionary distances to known proteins from each of the defined functional groups, are used to perform a multivariate clustering of the novel gene fragments. Using this method we show the clustering of many novel gene fragments from Curagen's SeqCalling EST sequence database with known protein families as well as predict putatively novel families with undefined function.

E-08

An Infrastructure for Genome Analysis and Annotation

E. Uberbacher, S. Petrov, M. Galloway, S. Martin, M. Land, M. Parang, F. Larimer, D. Schmoyer, M. Shah, I. Vokler, V. Olman, R. Mural, J. Snoddy and the Genome Annotation Consortium. http://compbio.ornl.gov/ Computational Biosciences Section, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6480

We have developed a large-scale infrastructure to comprehensively analyze the genome sequences from human, mouse, microbial and other organisms. This system consists of a number of online analysis services for genome centers and end users, and also several information resources containing comprehensive human genome annotation and annotation for many other organisms. A number of analysis servers, including GRAIL EXP, have been placed online for end users in a *Genome Computational Analysis Toolkit*. The *Annotation Pipeline* utilizes these tools to provide comprehensive analysis and annotation services to data producers and genome centers prior to GenBank submission. An in-house process uses the pipeline to perform first-pass, computational analysis of publicly available genome sequence data for multiple genomes. These analysis results are periodically updated and integrated into a genome sequence and feature framework, stored in a local warehouse. Users can access results through several browsing and query interfaces. This analysis process has led to the discovery of many credible gene and protein models that are not available from Genbank annotation. Robust linkages between these models and other experimental information (EST's, protein homologs, etc) are created.

Research sponsored by the Office of Science, US DOE under contract number DE-AC05-96OR22464 with Lockheed Martin Energy Research Corporation.

E-09
A Software Tool to Construct and Edit Physical Maps from Fingerprinting and Multi-Level Hybridization Data

Maciek Sasinowski and Heather Sasinowska, INCOGEN and Clemson University, Clemson, SC

Comprehensive physical maps containing information about relative positions of clones along a chromosome play a key role by serving as "road maps" in sequencing efforts. We have developed an interactive software tool that constructs robust physical maps of whole chromosomes from data obtained from restriction digestion experiments and hybridization experiments. The program accepts hybridization data in multiple formats, such as an unordered hit/no-hit hybridization matrix, and fingerprinting data containing band coordinates for each clone. The algorithm takes into account multi-level hybridization events and a Bayesian statistics approach to generate weights that are used to order the clones. The program produces robust contigs that can further be reduced to a minimum tile coverage of the chromosome. The data and results can be manipulated via a Java-based web browser interface.

E-10
Needle in a hay stack: Identifying functional genes within genomic sequence by novel sequence analysis and validation strategies

Sudhirdas K. Prayaga, Raj Bandaru, Catherine E. Burgess, Elma R. Fernandes and Richard A. Shimkets. CuraGen Corporation, New Haven, CT.

The human genome sequencing is progressing at an accelerated pace and promises to make available 3 billion base pairs of human genomic sequences by 2002. The next major task will be deciphering the genomic information and to identify the associated genes. Current algorithms can predict putative ORFs and exons in genomic sequences but do not always identify correctly spliced genes.

At CuraGen, we have developed a unique three-step process to identify and validate novel genes within genomic sequence. The first step involves identifying genomic clones with potential novel genes by a database blast and key word search. The second step is to predict the correct initial, internal and terminal exons and to align them into a complete putative novel protein. This involves aligning the predicted exons with public ESTs, CuraGen's database and known members of related protein family. The third step validates the predicted genes by various expression analysis methods, a key element in assigning biological value to the predicted gene sequences. Using our multi-tiered strategy, we have identified several novel genes from the genomic database. We present our results employing this strategy including novel glycoprotein hormones and a novel KGF.

E-11
Integrated information about genes

Kim D. Pruitt, Donna R. Maglott, Fasika Aklilu, Kenneth S. Katz, Hugues Sicotte, National Center for Biotechnology Information, NIH

Public sequence databases include sequences that are redundant, incomplete, or present dated information. This hinders use of the data as there is no view of the information that reflects our current state of knowledge. RefSeq and LocusLink address this problem by providing reference sequences as well as a single point-of-access for integrated information about genes.

Both RefSeq and LocusLink are built via a collaborative effort that curates: the correct association between a named gene and a sequence; official and alternative names; map position; and links to OMIM. LocusLink also integrates data from UniGene, dbSNP, PubMed, MMDB, and external web sites. Human mRNA and protein RefSeq records - accessed via BLAST, Entrez, FTP, and LocusLink - are available in two states, *provisional* and *reviewed. Provisional* RefSeq records provide the gene name-to-sequence association and information derived from LocusLink. *Reviewed* records are enhanced with additional annotation and/or sequence.

Together, RefSeq and LocusLink support annotation of the human genome and gene expression, mutation, and polymorphism studies by providing a curated, non-redundant view of genes and their products.

E-12
Annotation Master: A Parallel Computing System to Facilitate High-Throughput Genome Annotation

Bradley E. Slaven, Hanif G. Khalak, Bill T. Lee, Daniel Kosack, and Erin K. Hickey, The Institute for Genomic Research, Rockville, MD

Recent success in the dramatic acceleration of genomic sequence data processing has, in turn, illustrated the need for large scale annotation systems. Continuing the framework of distributed data processing implemented for homology searches, a new system has been developed to facilitate high-throughput annotation of multiple genomes.

We obtained supercomputer class performance for microbial genome annotation by networking a cluster of personal computers running the Linux Operating System (Red Hat Release) using Parallel Virtual Machine (PVM) software (Oak Ridge National Laboratory).

The AnnotationMaster was implemented on a pvm-cluster of linux workstations, which incorporated homology searching using both wu-blast and a praze implementation of the Smith-Waterman Algorithm followed by an automated preliminary functional annotation on a single workstation. Results from the application of AnnotationMaster on microbial genomes demonstrates the profound utility of parallel algorithmic implementations of bioinformatics tools. We implemented heuristic scoring thresholds and compared AnnotationMaster results to traditional curation procedures to provide useful insights into building a reliable system to assess gene function on a large scale.

Electronic Poster Session II

Monday, September 20
1:00 – 3:00pm
Champagne Room

E-01
Image Processing Software for Microarray Expression Analysis

Vasily A. Sharov and John Quackenbush. The Institute for Genomic Research, Rockville, MD

Microarrays provide the opportunity to study gene expression patterns on genomic scale. Thousands of genes are arrayed on a microscope slide and relative expression levels are determined by measuring fluorescence intensity of labeled mRNA hybridized to the arrays. A crucial first step in analyzing the data generated in such an assay is image processing to identify the array elements, estimate background, and calculate the fluorescence intensity for each gene. While there are a number of commercially available software packages for micorarray image processing, we have found that these are not completely satisfactory for our research needs. Available packages generally require significant user intervention to properly align the a sample grid for the array, are unable to process large images due to inefficient memory models and many require expensive third party engines such as MATLAB.

We have developed an image processing software tool *TIGR Spotfinder*, which meets the most of the technical requirements for rapid, automated, reproducible analysis of microarray images containing 10,000 or more spots per slide.

Features in *TIGR Spotfinder* include:
Automatic grid selection without any preliminary user drawing or grid alignment using an adaptive thresholding algorithm.
The ability to handle large image files (25MB or more in each of the Cy3 and Cy5 channels). Data output data to standard Excel workbook sheets. Gene expression ratios calculated using integrated intensity, mode, or median values with background correction. A simple-to-use graphical user interface.

This software, freely available to academic researchers, represents a first-generation image processing tool for analysis of expression measurements. Future refinements will include the ability to interact directly with databases and to interface with other data analysis software.

E-02
The TIGR Gene Indices: Estimating the Transcript Content of Genomes

Jonathan F. Upton, Ingeborg E. Holt, Feng Liang, Olga Ermolaeva, Geo Pertea, and John Quackenbush. The Institute for Genomic Research, Rockville, MD

The TIGR Gene Indices are a representation of transcribed sequences from worldwide collection of expressed sequence tag (EST) data. The Gene Index assembly process treats ESTs and coding sequences as elements of a shotgun sequencing project and uses them to assemble Tentative Consensus sequences (TCs). ESTs are downloaded daily from dbEST, cleaned to remove untrimmed vector, linker, ribosomal, mitochondrial, low quality, and poly(T) sequences. Cleaned ESTs and TCs from the previous build are compared pair-wise to identify overlaps. Sequences sharing a minimum of 95% identity over a 45 bp or longer region are grouped into a cluster. Within each cluster, THCs are replaced by their component ESTs and all sequences are assembled at high stringency. In the process, chimeric, low-quality, and non-overlapping sequences are flagged and stored as singletons. A second round of clustering and assembly using the first-round TC sequences further reduces redundancy. TCs are annotated to provide a provisional functional assignment and the resulting Gene Index is released through the TIGR web site (<http://www.tigr.org/tbd/tdb.html>). In addition to the Human Gene Index, TIGR maintains Gene Indices for a variety of important organisms, including mouse, rat, zebrafish, *Arabidopsis*, tomato, rice, and *Drosophila.*

E-03
Incorporation of RH Mapping Data and Orthologous Pairs in the Gene Indices

Ingeborg Holt, Feng Liang, Olga Ermolaeva, Geo Pertea, Jon Upton, and John Quackenbush, The Institute for Genomic Research, Rockville, MD

The TIGR Gene Indices identifies non-redundant transcripts within the publicly available EST data by assembling them into Tentative Consensus sequences (TCs). The Indices provide additional annotation including putative id, a list of top matches, and EST library information. To this we are adding mapping data and building orthologue-pair links. These are being developed first for the mammalian Gene Indices – human, (HGI), mouse (MGI), and rat (RGI) – but will later be incorporated into the other Indices.

The mammalian Gene Indices represent a unique resource for comparative analysis of genes and can provide insight into gene function, regulation and evolution. The TCs are generally longer than individual ESTs and consequently more likely to contain coding sequence. We are using the TCs to generate orthologue sets and these are being used to link TCs in HGI, MGI, and RGI.

Mapping places transcripts into a genomic context. Radiation Hybrid Mapping data from The 1998 Gene Map of the Human Genome (Deloukas *et al.*) has been incorporated into the TC Reports for HGI. Similar data are available for rat and should soon be available for mouse. When combined with the orthologue data, this should allow the construction of high-resolution synteny maps.

E-04
The BioKnowledge™ Library - A collection of curated model organism proteome databases with applications for comparative and functional genomics

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

The Yeast Proteome Database (YPD), used by thousands o academic and corporate researchers, has become the standard in-depth genome annotation. Knowledge is extracted fror research literature by Proteome's expert curators and integrated with yeast genomic and functional genomic datasets. YPD is used by functional genomics researchers who must quickly evaluate lists of genes from expression experiments, by drug discovery researchers who must select and validate targets, and by bioinformaticists who use the yeast genome to help assign functions to unknown genes from other organisms.

Proteome, Inc. has recently announced a similar Proteome Database, WormPD, for the nematode *C. elegans*. WormPD and YPD are tightly integrated. Software tools allow queries of protein function, sequence similarity, molecular interactions, and pathway information across species lines. Using the databases within the BioKnowledge Library, proteins implicated in human disease can be traced to homologs in model organisms, where experiments to test for drug effects can easily be designed and implemented.

E-05
The Mouse Genome Database - MGD

Jennifer J. Merriam, Janan T. Eppig, and the MGD group, The Jackson Laboratory, Bar Harbor, ME 04609 USA

The Mouse Genome Database (MGD) is a comprehensive database of mouse genetic and biological information. The Human Genome Initiative has emphasized the distinct importance of the mouse as a model system and has accentuated the necessity of a parallel organizational effort.

Established on the World Wide Web (WWW) in 1994, MGD provides researchers with a tool for quick and easy access to information on genetic loci with standardized nomenclature, extensive sequence links, summary phenotypic/functional information, mapping data, mammalian homologies, inbred strain and polymorphism information, and annual Mouse Chromosome Committee Reports. Affiliation with the Gene Expression Database (GXD) provides an expression data resource.

To date, information on more than 10,000 genes, 25,000 loci, and 54,000 references are presented in MGD. Data are updated continuously from the published literature, electronic submissions, and bulk data downloads. The database continues to evolve to meet the changing needs of the community it serves.

Supported by NIH Grant HG00330

E-06
Building Reliable Consensus Sequences For Gene Indexing: A Comparison Of Sequence Assembly Programs

Feng Liang, Ingeborg Holt, Norm Lee, Steven Salzberg and John Quackenbush, The Institute for Genomic Research Rockville, MD

The construction of gene indices is a very effective method of reducing the redundancy present in expressed sequence tag (EST) data. ESTs from dbEST are first cleaned to remove extraneous and low quality sequences and then compared with each other and known expressed sequences to generate clusters. For each cluster, the component sequences are then downloaded and assembled at high stringency to produce Tentative Consensus (TC) sequences. Individual TCs are then annotated the resulting gene index is released through the TIGR web <http://www.tigr.org/tdb/tdb.html>. The gene indices provide a unique resource for investigation of the properties of transcribed genes; the TCs can be used to annotate genomic sequences, place transcripts on Radiation Hybrid (RH) maps, and build orthologous gene groups.

In order to provide the most reliable consensus sequences for individual EST clusters, we have evaluated three widely used sequence assembly programs – TIGR-Assembler, Phrap and CAP3 – for the assembly of EST clusters. Using rat ESTs as a model, more than 100,000 sequences were clustered, and sequences from individual clusters were assembled and analyzed. The results from the three programs were compared with respect to the number of consensus and singleton sequences produced, the fidelity of consensus sequences, and the performance of the programs. Analysis of the results suggests that while both TIGR-Assembler and CAP3 can construct high-fidelity consensus sequences from ESTs, CAP3 produces fewer consensus sequences and singletons, resulting in a gene index with a lower level of redundancy. This indicates that CAP3 is an extremely effective tool for constructing high-fidelity consensus sequences and highly nonredundant gene indices.

E-07
Rapid Identification, Classification, and Organization of Gene-Based SNP's

Martin D. Leach and Richard A. Shimkets, CuraGen Corporation New Haven, CT

Single nucleotide polymorphisms (SNP's) within the mRNA- and protein-coding regions of genes can be rapidly detected in a high-quality cDNA sequence database and validated quickly by many methods. The sequence variants, particularly those that alter the predicted protein sequence, have wide application to the fields of disease genetics and pharmacogenetics. CuraGen's scheme for the identification and classification of SNP's will be presented. In addition, tools for the visualization of SNP's and association with gene expression and proteomic data will be presented.

E-08
EGAD: A Nonredundant Database of Genes and Proteins

Olga Ermolaeva, Ingeborg E. Holt, Feng Liang, Geo Pertea, Jonathan F. Upton, Michelle L. Gwinn, Robert J. Dodson, Owen White, and John Quackenbush, The Institute for Genomic Research, Rockville, MD

As the body of sequence data in GenBank continues to grow, the identification of a high confidence, nonredundant set of coding sequences becomes a significant challenge. However, the creation of such a resource will be essential for the future classification of gene function and for the functional annotation of the significant body of genomic sequence that will be available in the near future. The Expressed Gene Anatomy Database (EGAD) was created at TIGR as a means of producing such a collection of representative DNA sequences and their encoded proteins [Adams *et al.* (1995) *Nature* 377 (Supp), 3-174].

EGAD is populated by extracting CDS features for full-length gene and mRNA sequences submitted to GenBank. Redundant entries are identified by sequence comparisons and one representative is chosen although links to alternative GenBank Accessions numbers are maintained. The annotation of these Expressed Transcript sequences (ET; alternatively, Human

Transcript – HT – for human sequences) is checked for consistency and the records are loaded into database.

We are currently in the process of restructuring the database and developing tools for rapid sequence loading and nonredundification so that EGAD continues to maintain an accurate picture of the coding potential of microbial and eukaryotic genomes.

E-09
Octamer-primed Sequencing Technology: Development of Primer Identification Software

Susan H. Hardin, Gangwu Mei, Anelia Kraltcheva, Dept. of Biology and Biochemistry, University of Houston, Houston, TX

Octamer-primed Sequencing Technology (OST) is a primer-directed sequencing strategy in which an individual octamer primer is selected from a pre-synthesized octamer primer library and used to sequence a DNA fragment. However, selection of candidate primers from such a library was time consuming and presented a bottleneck in the sequencing process. To accelerate the sequencing process and to obtain high quality sequencing data, a computer program, electronic OST or eOST, was developed to automatically identify candidate primers from an octamer primer library. eOST integrates the base-calling software PHRED to provide a quality assessment for target sequences and identifies potential primer binding sites located within a high quality target region. To increase the sequencing success rate, eOST includes a simple dynamic folding algorithm to automatically calculate the free energy and predict the secondary structure within the template in the vicinity of the octamer binding site. Several parameters were found to be important, including base quality threshold, the window size of the template sequence segment, and the threshold point of the DG value. Additionally, improved OST reaction conditions will be detailed. OST, coupled with the eOST software, can be used to sequence short DNA fragments or in the finishing assembly stage of the large-scale sequencing of genomic DNA.

Supported by NIH grant HG01151 to S.H.H.

E-10
Selection of Unique Sequences from the Unigene Database for Hybridization Purpose

Cliff S. Han, Linda J. Meincke, Judy G. Tesmer, Larry L. Deaven and Norman A. Doggett; Life Sciences Division and Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM

We are starting with the unigene sequences from NCBI and selecting the best unique sequences for hybridization purpose from each of them. As we all know, repetitive sequences presented in the probe or template are the biggest problem in hybridization based research. Most of the false positives are from the signal produced by the repetitive sequences. Repetitive sequences appear everywhere in the human genome, including the genes. At present microchip based hybridization are engaged in many gene expression studies. The repetitive sequence represented in the expressed sequence should be addressed. We use the program which were written to select overgo (overlapped oligonucleutide) from STSs' sequences to analysis the unigene data base from NCBI. About 146 thousand expressed sequences are analyzing. About 90% of the sequences were successful in choosing at least one 40mers for overgo design and hybridization.

Supported by the US DOE, OBER under contract W-7405-ENG-36.

E-11
The Continuation of the GDB Human Genome Database and Its Evolution with the Human Genome Project

A. Jamie Cuticchia, Christopher J. Porter, C. Conover Talbot, Jr., and Weimin Zhu. Genome Data Base, Hospital for Sick Children, Toronto, ON, and Johns Hopkins University School of Medicine, Baltimore, MD

With the recent termination of public U.S. funding for the GDB Human Genome Database, support was obtained through private groups and corporations for the continuation of the project. Staff located both in Canada and the U.S. continues to collect and curate human gene mapping and polymorphism information. Though the monolith that was once GDB and funded in excess of $7M annually has been set aside, a new "leaner and meaner" database focused on presenting curated mapping information continues. Relying on the Human Genome Organisation's (HUGO) Human Gene Mapping Committee as the advisory board, GDB continues in its role as the primary source of human gene mapping information. We remain committed to provide a freely accessible public view of the data. Moreover, work has begun to integrate and collaborate with other bioinformatics organizations such as the National Center for Biotechnology Information (NCBI) and the Genome Annotation Consortium (GAC) to integrate GDB data with sequence information.

E-12
Haplotyping Genes Important to the Pharmaceutical Process and Drug Response

Gerald F. Vovis, Joel Claiborne Stephens, Vincent Schulz, Krishnan Nandabalan and Gualberto Ruaño, Genaissance Pharmaceuticals, New Haven, CT

Human gene diversity is an important factor in how patients react to pharmaceuticals. We have targeted genes that are useful to the pharmaceutical process and are identifying the variant forms present within a diverse human population. We sequence genomic DNA to determine the polymorphisms that are present in the promoter, exons, exon/intron borders and the 3' UTR. Using a set of proprietary technologies, we determine the gene haplotypes, i.e. the organization of the polymorphisms as they appear in each chromosomal locus of the gene. The initial set of genes haplotyped had a greater number of polymorphisms than expected from previous estimates. All of the genes had a large number of different haplotypes. We used our DecoGen^{TM} Informatics Platform for phylogenetic analysis. Examples of closely related and divergent haplotypes were seen. The progenitors of these divergent haplotypes probably arose in antiquity. In addition, there were examples of clear differences in the major haplotype that was found in individuals with different geographical backgrounds. The phylogenetic analysis of gene variation, only possible with the haplotype approac reveals clustering as well as distant isoforms, thus allowing d reduction and inference on possible biological significanc resulting picture of gene variation can be used to identify which haplotype or group of related haplotypes, if any, harbors the variant causing a specific phenotype such as a positive response to a pharmaceutical product.

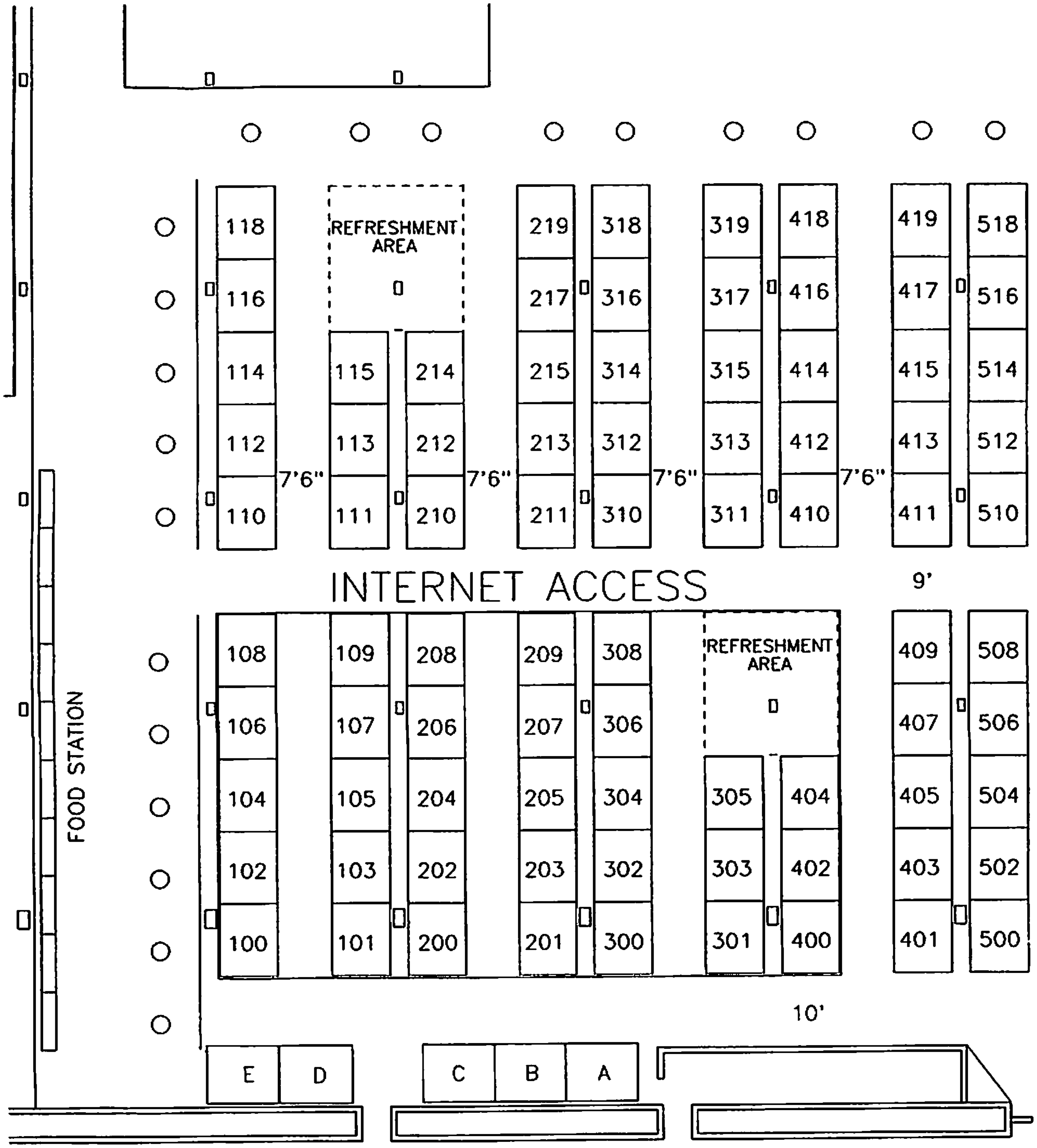
E-13
Consed and Autofinish in the Phred/Phrap/Consed system

David Gordon, University of Washington, Seattle, WA

Consed is an editor and viewer for phrap sequence assemblies. It is now in use in over 100 sites worldwide. Recent enhancements include integration with POLYPHRED for viewing and tagging of sequence variants, ability to automatically incorporate additional reads without reassembly, tearing and re-joining of contigs, manual moving of reads, primer and template picking for finishing reactions, and "autofinishing".

Consed's autofinish feature is able to automatically call most of the finishing reads necessary to complete a project, thus allowing finishers to concentrate their time on the most difficult problems. Autofinish identifies gaps, regions of low data quality, and regions spanned by single subclones, and determines the reads (custom primer or universal primer forward or reverse) necessary to fix these. User-specified costs for the different read types are used to estimate the most economical choice of reads necessary bring the consensus sequence to a desired target accuracy level.

Autofinish has been in use for over 8 months on an evolving basis in the University of Washington Genome Center and elsewhere, and has eliminated a substantial fraction of the finishing effort -- many BACs are now completely finished using Autofinish with no human editing or other decision-making required. Algorithmic details, success rate data, and new features will be presented.


118
REFRESHMENT AREA
219
318
319
418
419
518
116
217
316
317
416
417
516
114
115
214
215
314
315
414
415
514
112
113
212
213
312
313
412
413
512
7'6"
7'6"
7'6"
7'6"
110
111
210
211
310
311
410
411
510
INTERNET ACCESS
9'
108
109
208
209
308
REFRESHMENT AREA
409
508
106
107
206
207
306
407
506
FOOD STATION
104
105
204
205
304
305
404
405
504
102
103
202
203
302
303
402
403
502
100
101
200
201
300
301
400
401
500
10'
E
D
C
B
A

11th International Genome Sequencing and Analysis Conference

Addenda

Plenary Speaker Abstracts

Tuesday, September 21, 9:00 am
From Genome To Proteome: A New Look At Nature Complexity

Denis F. Hochstrasser*(1); Pierre-Alain Binz(1); Amos Bairoch(2); Ron D.Appel(2); Jean-Charles Sanchez(1). (1) Central Clinical Chemistry Laboratory, Geneva University Hospital, 1211 Geneva 4, (2) Swiss Institute of Bioinformatics, University Medical Centre, 1211, Geneva 14, Switzerland

How many final protein products expressed and controlled by a genome and its epigenetic network should be found in a living organism? What are the technological difficulties to display many if not all polypeptides at a given time in the life of an organism (a proteome)? Does the current technology allow to efficiently identify and partially characterise so many proteins? Last but not least, is it useful? Preliminary experiments demonstrate that the number of final post-translationally modified proteins could be 3 to 6 times greater than the number of genes for higher organisms. On the contrary to nucleic acids, proteins display extremely diverse and heterogeneous physical-chemical characteristics. They are often poorly soluble, in extremely variable concentration, often labile and cannot be amplified. Therefore no single technology, but a combination of them will allow the full characterisation of a proteome. Mass spectrometry, genomic information and bioinformatics are key elements of proteomic research. Proteomics alone has not allowed any major medical breakthrough yet, but in combination with genomic information should provide much deeper insight into functional genomics.

Ref: "Proteome Research: New Frontiers in Functional Genomics" M.R.Wilkins, K.L.Williams, R.D. Appel, D.F. Hochstrasser (Eds.) (1997) Springer ISBN 3-540-62753-7

Sunday, September 19, 11:15 am
Sequence Analysis of a Drosophila Centromere

Janice Wahlstrom, Hiep Le and **Gary H. Karpen**. The Salk Institute, La Jolla, CA

Heterochromatin is an important and mysterious region of the genome, but it is excluded from 'whole' genome sequencing efforts due to the presence of repeated DNAs. We need to develop approaches to map and sequence heterochromatin, since it contains centromeres and other elements responsible for nuclear organization and function.

We have investigated the structure and function of heterochromatin using a Drosophila minichromosome, Dp1187. We have localized the fully-functional centromere to a 420 kb region, and determined its molecular structure and composition with strategies that circumvent the problems normally associated with analyzing repeated DNA. A complete restriction map has been generated, and 25% of the centromere has been sequenced to date. Our results reveal striking features of the centromere: it is primarily composed of uniform satellite arrays and single, complete transposable elements. Whole genome analysis reveals that other regions of Drosophila heterochromatin are organized in a similar fashion. Surprisingly, the Dp1187 centromeric satellites and transposable elements are neither unique to centromeres nor present at all centromeres. This work constitutes the first detailed molecular structure of a functional centromere in a multicellular organism. The impact of these results on our understanding of heterochromatin structure, and on the determinants of centromere identity and function, will be discussed.

Monday, September 20, 7:00 pm
The Mouse Genome Project

Miriam H. Meisler, Dept. of Human Genetics, Univ. of Michigan, Ann Arbor, MI

Two major goals of the Mouse Genome initiative are sequencing of the mouse genome and discovery of gene function by phenotype-driven mutagenesis. The NIH program (www.nih.gov/science/models/mouse) proposes a "sequence first, map later" strategy with draft sequencing of BAC clones from a C57BL/6J library. Clones will be distributed to a network of sequencing centers from a central server at NCBI. Alignment of human and mouse genomic sequence will be used to identify conserved functional coding and noncoding sequences. Automated alignment of genomic sequences can now be obtained electronically at http://globin.cse.psu.edu/pipmaker. Random mutagenesis by the chemical mutagen ethylnitrosourea (ENU) generates point mutations and can produce both loss-of-function and gain-of-function alleles. With an *in vivo* mutation rate of 10-3/locus, screening several thousand mice can generate multiple alleles at any locus if an efficient screen to detect mutants is available. Novel screens for interesting biological functions may be incorporated into ongoing screening at mutagenesis centers in Europe (Harwell, Munich), Japan, Canada and the U.S.

Tuesday, September 21, 11:45 am
Genes and Behavior

Larry J. Young. Department of Psychiatry and Behavioral Sciences, Emory University, Atlanta GA

Understanding the complex relationship between genes, the brain, and behavior is one of the great challenges of the next century. Selection of appropriate model species is essential for the progress of this field. We have chosen two closely related rodent species as a model for understanding the molecular basis of social behaviors. Prairie voles are highly gregarious and monogamous. In contrast, montane voles are much less social, and are not monogamous. The neuropeptide vasopressin modulates affiliative behavior, the formation of long-lasting pair bonds between mates, and parental care in male prairie, but not montane voles. Prairie and montane voles have dramatically different distributions of vasopressin receptor expression in the brain, providing a potential explanation for the differences in social behavior. Changes in the vasopressin receptor locus, for example gene duplication and accumulation of a repetitive expansion in the 5' flanking region, are associated with the

differential expression of the receptor in these species. Mice transgenic for the prairie vole vasopressin receptor gene express the receptor in a neuroanatomical pattern similar to that of the prairie vole and respond to vasopressin by showing elevated levels of affiliative behavior. These results support the hypothesis that genetic changes which modulate the tissue-specific expression of behaviorally relevant genes may result in the evolution of social behavior, and could potentially influence individual variability in human behavior.
Supported by NIMH 56897

Concurrent Session Abstracts

Bioinformatics

Tuesday, September 21, 3:40 pm
Needle in a hay stack: Identifying functional genes within genomic sequence by novel sequence analysis and validation strategies

Sudhirdas K. Prayaga, Raj Bandaru, Catherine E. Burgess, Elma R. Fernandes and Richard A. Shimkets. CuraGen Corporation, New Haven, CT.

The human genome sequencing is progressing at an accelerated pace and promises to make available 3 billion base pairs of human genomic sequences by 2002. The next major task will be deciphering the genomic information and to identify the associated genes. Current algorithms can predict putative ORFs and exons in genomic sequences but do not always identify correctly spliced genes.

At CuraGen, we have developed a unique three-step process to identify and validate novel genes within genomic sequence. The first step involves identifying genomic clones with potential novel genes by a database blast and key word search. The second step is to predict the correct initial, internal and terminal exons and to align them into a complete putative novel protein. This involves aligning the predicted exons with public ESTs, CuraGen's database and known members of related protein family. The third step validates the predicted genes by various expression analysis methods, a key element in assigning biological value to the predicted gene sequences. Using our multi-tiered strategy, we have identified several novel genes from the genomic database. We present our results employing this strategy including novel glycoprotein hormones and a novel KGF.

Genomics

Sunday, September 19, 3:00 pm
Sequencing and Analysis of Full Length cDNAs in the Course of the German Genome Project

Stefan Wiemann, Wilhelm Ansorge, Helmut Blöcker, Helmut Blum, Andreas Düsterhöft, Karl Köhrer, Werner Mewes, Brigitte Obermaier, Annemarie Poustka, Rolf Wambutt, [1]Molecular Genome Analysis, German Cancer Research Center, Im Neuenheimer Feld 506, D-69120 Heidelberg,Germany, and the German cDNA Sequencing consortium

A consortium of eight sequencing laboratories and Germany's leading bioinformatics institute has formed in the frame of the German Genome Project. We aim at the sequence analysis of 3,000 to 4,000 complete novel cDNAs, comprising eight megabases of finished sequence. Sequencing started in September 1997 and a progress report of the consortium will be presented. The libraries generated in the course of the grant „Generation of full length cDNAs in the course of the German Genome Project,, are the primary source for sequencing. EST sequences of 12,000 independent clones are generated to identify novel genes. The EST sequences are analyzed for the likelihood of the clones to be full length (e.g. by the presence of CpG clusters) in order to obtain a minimal set of full length clones for efficient complete sequence analysis. Clones identified to be full length are sequenced and further analyzed by members of the consortium. The sequences are analyzed for possible function *in silico*. Functional analysis projects have started using the clones analyzed by the consortium as resource. All clones and data generated in the project are made publicly available via the Resource Centre of the German Genome Project (RZPD).

Sunday, September 19, 3:40 pm
Identification of Thousands of Single Nucleotide Polymorphisms (SNPs) in the Human Genome

Nila Shah, Cindy Chen, Sangeetha Kondapalli, Vivian Reyes, Chunmei Liu, Michael Savage, Michael Janis, Maria DeGuzman, Richard Watts, Anthony Berno, Naiping Shen, Jyoti Baid, Jim Snyder, Claire Marjoribanks, Howard Lee, Daryl J.Thomas, Robert Lipshutz, Nila Patil, and Janet A. Warrington, Affymetrix, Inc., Santa Clara, CA

We are screening thousands of human genes across 40 individuals to identify frequently occurring SNPs using GeneChip® arrays in an automated high throughput laboratory with a custom laboratory management system that integrates every component of the process including subject information, sample processing, and SNP discovery outcome. Currently, the use of automated sample preparation and array handling enables us to screen 1.8 MB of sequence per day. We will present a summary of all projects completed to date as well as detailed information from one of the initial projects in which we are screening genes expressed at high levels in lymphoblast cell lines. In this project, we are amplifying ~ 700 genes across 40 unrelated males and females of Caucasian, African-American and Asian origin using RT-PCR. Samples are assayed using high-density GeneChip® variation detection arrays, designed to screen 30 kb of sense and antisense sequence simultaneously. To date we have identified 1006 candidate SNPs at a density of 562 bps in the lymphoblast project. Confirmation of the SNPs is in progress.

Supported by NIH R01 MH60007-01 (CWR) R01 HG01919-01 (Stanford), R01 HG01953-01 (Whitehead Inst.)

Sunday, September 19, 4:00 pm
Rice Genome Sequencing Project: Sequencing of Rice Genomic DNA

Hiroyuki Kanamori, Kimiko Yamamoto, Jianyu Song, Noriko Kobayashi, Hui Sun, Zhong, Ari Kikuta, Kayo Machita, Yuko Nakama, Yumi Nakamichi, Michiko Ikeda, Kozue Kamiya, Satomi Hosokawa, Kazuko Yukawa, Harumi Yamagata, Michie Shibata, Sachie Onose, Mari Nakamura, Takashi Matsumoto, Yoshiaki Nagamura, Takuji Sasaki, Rice Genome Research Program, National Institute of Agrobiological Resources / Institute of the Society for Techno-innovation of Agriculture, Forestry and Fisheries, Tsukuba, Ibaraki 305-0854, Japan

We started the second phase of the Rice Genome Research Program (RGP) in 1998 with the aim of sequencing the entire genome. With a genome size of 430 Mb, whole-genome sequencing of rice is, an achievable goal as compared with other cereals. As the core of genome sequencing, a PAC (P1-derived artificial chromosome) genomic library derived from Oryza sativa ssp. japonica cultivar Nipponbare was established. Japan is in charge of sequencing chromosomes 1 and 6 as part of International Rice Genome Sequencing Project (IRGSP). For our initial sequencing efforts we are concentrating on gene-rich regions in these chromosomes. A shotgun sequencing strategy to ensure high-quality sequence information was developed. However, in order to accelerate the sequencing process, we are adopting a strategy such that sequencing of random subclones generated from a PAC anchored on the physical map is to be done mainly with ABI3700 sequencer. The sequence data is then assembled into contigs, and the finishing phase to close gaps between contigs and to resolve low quality regions is to be accomplished with ABI377XL. The detailed strategy will be described and sequence characteristics of some rice PACs will be shown.

Technology

Sunday, September 19, 3:40 pm
Directed Minimal Sequencing For Total Genome Analysis With Minimized Sequencing Efforts

Patrik Scholler[1], H. Voss[1], G. Casari[1], T. Schlüter[1], M. Arenz[1], B. Drescher[1], D. Schütte[1], J. Kämper[2], R. Kahmann[2], C. Basse[2], M. Feldbrügge[2], G. Steinberg[2], I. Häuser-Hahn[3], V. Vollenbroich[3], E. Koopmann[3], G. Seidel[3], K. Sievert[3], B. Jaitner[4], R. Ebbert[4], V. Li[4], M. Vaupel[4], P. Schreier[4] [1]LION bioscience AG, eidelberg; [2]Uni München; [3]Bayer AG ZF-BTB and [4]AG PF-MWF-Biotechnologie, Leverkusen

The demand for high throughput analysis of large genomes is currently met by massive upscaling of the shotgun sequencing technology. In order to minimize sequencing efforts and to maximize functional genome analyses, a new genome sequencing strategy, called DIRECTED MINIMAL SEQUENCING (DMS), has been developed. It is based on physical template mapping prior to sequencing in conjunction with proprietary clone and data handling technology. The combination of DMS with automated functional genome annotation by means of the bioinformatics platform bioSCOUT™ guarantees an unparalleled speed and accuracy in the generation of genomic information. We have demonstrated the efficiency of this approach in the 20 Mb genome sequencing project of the phytopathogenic basidiomycete *Ustilago maydis* and discuss further options for functional analyses and the applicability for gigabase genomes.

Poster Session Abstracts

P-046
From Stop to Start: Scanning Whole Genomes for Transcripts

H.G. Khalak, and H.O. Smith, The Institute for Genomic Research, Rockville, MD

Pre-mRNA transcripts in bacteria are delineated in genomic DNA as units of expression bounded by sites where transcription is initiated and terminated. These sites are accompanied by signals in the DNA defining intrinsic structural features or areas where protein interaction occurs. Discrimination of such signals for transcriptional termination has been addressed, for both intrinsic features (stem-loops) as well as rho-dependent sites. With respect to initiation (promotion) sites, though signal patterns have been identified, the variation in both sequence motif and inter-pattern spacing yields quite low sensitivity when scanning whole genomes. Any further delineation of promoter regions into classes is subsequently difficult, particularly when using multiple sequence alignment.

To address this problem, likely termination sites were predicted and subsequently used to indentify upstream candidate regions for prediction of promoter sequences. A number of techniques were evaluated to predict promoter regions, including motif searches, neural network classifiers, and dynamic programming. The sequential approach compared favorably to parallel scanning, as demonstrated for genomes of H. influenzae and E. coli.

P-075
PAC physical mapping for rice genome sequencing in RGP

Satoshi Katagiri, Tomoya Baba, Shoko Saji, Masao Hamada, Marina Nakashima, Masako Okamoto, Yoshino Chiden, Mika Hayashi, Ryoichi Tanaka, Kazuhiro Koike, Jianzhong Wu, Takashi Matsumoto Takuji Sasaki, Rice Genome Research Program (RGP), National Institute of Agrobiological Resources/Institute of the Society for Techno-innovation of Agriculture, Forestry and Fisheries, Tsukuba, Ibaraki, Japan

As the core of rice genome sequencing, RGP has constructed a PAC (P1-derived artificial chromosome) genomic library from Oryza sativa ssp. japonica cultivar, Nipponbare, in order to establish a sequence-ready physical map. This PAC genomic library consists of about 71,000 clones with average insert size of 112 kb and 16-fold genome coverage. To construct PAC contigs, we screen our PAC library by PCR using STS primers of mapped DNA markers as well as ESTs. At present we concentrate our efforts on ordering of PACs on rice chromosomes 1, 5, 6 and 10. So far, a total of 245 DNA markers and 565 ESTs mapped on these chromosomes have been used for screening. As a result, we have selected 3,790 PACs aligned in 257 contigs with a total physical length of about 35 Mb, corresponding to approximately 30% of these chromosomes. We have also started the complete genomic sequencing of some of these ordered PACs. Initial results of physical mapping with PACs as well as genomic sequencing have shown the uneven distribution of genes in the rice genome. Thus in order to identify the most number of genes at the outset of our sequencing efforts, we will focus our PAC physical mapping strategy on gene-rich regions of the genome.

P-078
The BioMerge Relational Database. Managing Distributed Sequence Data and Annotation from Public, Proprietary and Third Party Sources.

Jeffery R. Mathis, T.Preston Boyd, Gary A. Thompson, Russell B. Kepler, Padraic I. Hannon, and Emily G. Dickinson.
PE Informatics, Santa Fe, NM.

One of the central problems confronting researchers is management of sequence data from disparate sources. Results from analysis packages must be associated back to the sequence and made available to other researchers.

BioMerge is a management system designed to further metabolic and drug discovery efforts by solving this problem. Public (EMBL, SwissProt, TrEMBL-SP, ProSite, PFAM), private and 3rd party sequence data can all be stored and retrieved. Support for chromosome mapping, multiple sequence alignments and protein families are incorporated into the schema. Sequences and analysis results can be stored and automatically propagated to remote BioMerge sites with our Publish/Subscribe mechanism.

The system implements a multi-level access control mechanism to enforce viewing and editing permissions, and has an auditing capability to track annotation history. The DiscoveryTools web application provides a query mechanism and analysis results management system. A java-based API removes platform dependencies and simplifies read/write integration with analysis packages from in-house or commercial vendors, such as the Wisconsin Package and HMMER.

P-096
Gossypium L. Retrotransposons Families Identified by Reverse Transcriptase Domain Analysis

Essam A. Zaki, Ph.D., Nucleic Acids Program Genetic Engineering & Biotechnology Research Institute (GEBRI) Mubarak City for Scientific Research, Alexandria, Egypt

Retroviruses consist of populations of different but closely related genomes refereed to as quasispecies. A high mutation rate couples with extremely rapid replication cycle allows these sequences to be highly interconnected in a rapid equilibrium. It is not known if other retroelements can show a similar population structure. I have initiated a study to study the tempo and pattern of *copia* and *gypsy*-like retrotransposons evolution in allopolyploid and diploid species of cotton (*Gossypium* L.). Nucleotide sequence data generated from 53 accessions representing the spectrum of *Gossypium* diversity have shown that reverse transcriptase fragments of *copia* and *gypsy*-like retrotransposons are not unique sequences but represented by population of different but closely related sequences. This highly variable population is not in a rapid equilibrium and could not be considered as quasispecies. The analysis of *copia* and *gypsy*-like retrotransposons sequence variability within and between *Gossypium* species shows that mutations frequently occur in important regulatory elements and that defective element are often produced.

P-118
Increased Sequence Quality Through the Use of an Improved Acrylamide Gel Solution

Bruce C. Reinemann, Arnold R Kana, Hongmei Lee, Tay Ho, Ying Wang, Lynn A. Doucette-Stamm. Genome Therapeutics Corporation, Waltham, MA

A common method of DNA sequence separation is through the use of slab gel electrophoresis. The overall sequence resolution is dependent upon the properties of the gel matrix. Modifications to the gel matrix can lead to an overall improvement in the quality of the data. Several new commercially available gel solutions claim increased resolution as well as decreased run times.

At the GTC Sequencing Center at Genome Therapeutics Corp. gel solutions were evaluated for their ability to increase quality (as measured by quality scores based on the program PHRED) and their ability to reduce gel run times. Sigma AutoPAGE™ PLUS 4.5% acrylamide was found to significantly increase Phred 20 and 30 base calls by >5%. Readlengths also increased allowing run times to be significantly reduced with equivalent results. This resulted in improved efficiency in the production sequencing facility.

In summary this enhanced gel solution allows the GTC Sequencing Center to improve data quality while simultaneously improving efficiency. Data will be presented demonstrating the overall quality improvements as a result of the implementation of novel gel matrices in the production laboratory.

EXHIBIT DESCRIPTIONS

Beckman Coulter, Inc. will present an automated solution for sequencing that was developed by examining sequencing as a process, rather then as a series of individual method steps. Our integrated approach provides true time savings, resource efficiency and upgrade pathways without sacrificing an intuitive, easy-to-use operator interface.

DNX Transgenic Sciences is a contract research organization that specializes in genetic modification of mice and rats using gene addition, gene targeting, homologous recombination and microinjection techniques. DNX offers over 10 years of extensive experience for gene expression analysis and phenotypic characterization of the mice and rats it produces for their clients. DNX has produced more than 7,500 transgenic mice and rats from over 1,300 gene constructs.

LION Bioscience AG provides integrated genomics and BIO-IT solutions for accelerated and improved Life Science research. LION provides a new platform of IT solutions and develops and implements under the i-biology program enterprise wide, customized biological data and information management systems for multinational Life Science companies.

Spotfire, Inc. provides software solutions that empower scientists and engineers—and their enterprises—to make decisions in eTime that get products to market first. Our Web-based offerings are used by life- and material-sciences companies worldwide for such activities as high-throughput screening, genomics, lead optimization, combinatorial chemistry, formulation development, and bioinformatics.

EraGen Biosciences is dedicated to producing software and databases for functional genomics. MasterCatalog™ is a unique, browsable sequence database organized by evolutionary families, providing user-friendly analysis of biological function for interesting as well as completely novel genetic sequences. MasterCatalog™ will aid both specialist "gene finders" and the broad non-specialist spectrum of life science researchers alike.

Genetic MicroSystems Inc adeleo@geneticmicro.com High performance systems that enable scientists to make and use DNA microarrays. The GMS 417 Arrayer uses Pin and Ring™ spotting technology for superior consistency, flexibility. The GMS 418 Array Reader employs Flying Objective™ scanning technology for improved speed and sensitivity. System includes software, accessories, consumables, instrument services, applications support.

InforMax is a leader in the development of bioinformatics software for accelerated discovery, enabling over 10,000 researchers worldwide to achieve greater insight into genetic function through biological data mining and integrated sequence analysis. Presently, InforMax offers Vector NTITM Suite desktop sequence analysis and Software Solution for BioMedicineTM enterprise bioinformatics systems.

JENOPTIK Bioinstruments GmbH is a sales, marketing, service and management corporation offering laboratory automation solutions. One highlight is the JOBI-Disk™ automated pipettor with 384 tips for very fast, reliable, easy handling of microplates with up to 1536 wells and deep well blocks. In its version with special plate transfer unit and a sideways transport mechanism, the JOBI-Disk™ is particularly well suited to magnetic bead handling.

MWG Biotech is a genomic services company. We manufacture HPSF oligos which are always purified and salt-free. MWG's DNA sequencing lab offers both custom and genomic projects. MWG manufactures a family of thermal cyclers...the newest one being a linear multi-block system. MWG also offers an automated workstation for cycle sequencing, PCR, and comb loading.

Molecular Simulations Inc., a subsidiary of Pharmacopeia, Inc., is a leading provider of molecular modeling, simulation, and informatics software products that provide a broad range of solutions for accelerating the drug discovery process. Our featured products for protein bioinformatics include Gene Explorer for target discovery and target evaluation and GeneAtlas™/AtlasBase™ for high throughput protein analysis and functional assignment of genomic data.

NEN Life Science Products provides labeling and detection products and technologies for life science research and drug discovery. NEN features MICROMAX, the first commercial glass slide microarray system for differential gene expression analysis, as well as an extensive line of fluorescent and hapten-labeled nucleotides. The Kodak Digital Science Image Station 440CF will also be highlighted.

Nalge Nunc International NUNC™ Brand Product Description Nalge Nunc International, under NUNC™ Brand products, produces disposable plastic lab products used primarily in research. In addition however, many NUNC Brand products are used in diagnostic kits and other clinical applications. The products are segmented into four major categories, Cell Culture, Molecular Biology, Immunology and Cryopreservation.

Nature America, publishers of the world's leading sources of news in basic, applied nad medical research. Including Nature, the world's most frequently cited weekly scientific journal; Nature Biotechnology, the premier monthly covering applied industrial research; Nature Genetics, Nature Structural Biology, Lab Animal, Nature Medicine; Nature Neuroscience, and the new Nature Cell Biology.

Neomorphic develops advanced bioinformatics applications for the analysis and annotation of biological databases. Neomorphic's approach encompasses both computational biology techniques for mining genomic databases and interactive user interfaces that enable scientists to explore DNA and protein sequence databases and to share their discoveries with others. For more information about Neomorphic, please call 510-704-1030, email info@neomorphic.com or visit the company's web sits at http://www.neomorphic.com/. Neomorphic, making science out of data.

Operon Technologies has become a leading supplier of synthetic oligonucleotides and genes to customers worldwide. Operon provides high ***quality custom synthetic oligonucleotides***, specialty oligonucleotides, custom synthetic genes, stock primers and special contract manufacturing services as well as microarray products to leading research, development and production groups. All products come with an unconditional guarantee and extensive customer service support to assure customer satisfaction.

PE Biosystems is the life sciences division of the PE Corporation, integrating the products and services of four business units: PerSeptive Biosystems, PE Informatics, Applied Biosystems and Tropix, into one comprehensive organization. PE Biosystems supplies instrument systems, reagents and related services to the life science industry and research community.

Sigma-Aldrich Research and recently acquired partner Sigma-Genosys have developed and optimized a broad range of products for genomic applications. Featured are products for increasing automated sequencing throughput, nucleic acid purification, PCR, RT-PCR, expression profiling, as well as, custom services for oligonucleotides and specialized solutions for molecular biology.

Transgenomic manufactures and markets analytical solutions for the analysis of nucleic acids (DNA/RNA), amino acids, proteins, and carbohydrates. Our mission is to provide the researcher with enhanced tools that produce rapid, automated turnkey solutions. **Transgenomic's** core platform technology, the **WAVE™ Nucleic Acid Fragment Analysis System,** offers researchers a sensitive (UV and FL) and quick, cost effective workstation for mutation screening and scoring. The **WAVE™ System** allows rapid separation and analysis of sequence variations without sequencing. In addition, these variants can be collected for further analysis. **WAVE™ System** applications include **SNP, Mini- and micro-satellite (STR), VNTR,** and **Differential Display** analysis.

"celebration of the weed"

1 Rabbit Opsins

Predicting "transcription factors" is not very informative

Are the single exon genes the bacterial genes.

X.LIN

Chromosome IV

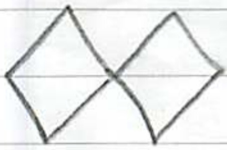
angiogenomics

sequencing strategies

BAC-end

FPC contigs

MPI-MP end proc.



- 2 gaps in ~~heter~~ochromatin
- no seq. of rDNA
- many markers
- 81 tRNA
- 3751 prot

Distribution of repeats (using blastn)

- clustered around centromere
- chromomere 43 Kb tract of 22 repeats each of which are ~1.3Kb tandem repeats
- 60 Kb cluster of PAC1 repeats

- similar to repeats on chr II
- many tandem duplications { prot kinases / prot kinases } 3 major clusters

Landmarks for centromere

- hyb. whole bac to chrIV
- also hyb AAL1 repeat -

one probe lights up heterochrom on chrIV

lower levels of gene expression in heterochromatin

- plot EST distribution on chromosome
- also reduced # of predicted genes
- increase # of hypotheticals
- A few genes in there though

Centromeres

- pAL1 + Athilla repeats
- most genes are hypothetical

Chr IV

- 64% no EST match
- 75% of ESTs match 6% of genes

Sequence analysis

Annotation = gene model

Crude f(x) catalog.

F(x) categories

-50% = sig sim to something

- -most of these = defense, signal, metab, txn

Comp to other organisms

·divide this up by f(x) categories

analysis only looked at hits not highest hits

19% = chloro p predicted transit peptides

some bias of them ... more of them hit Syn. sp than other genes (w/o signals)

Interesting Genes

-BRCA2 homolog?

-Werner's

-human PTPA

-RNAse L inhibitor

Protein Folds

-prot. kinases #1

-NATPASES

-

-#4,5,6 = v. low in C.elegans → RING finger
→ cyt. P450

S. Somerville

Arrays
T-DNA KO's
plant specific genes

} useful in beginning to ID functions

D. Preuss

-centromeres - detail in lower euks about sequence → f(x)
- detail in higher euks about structure → f(x)

-centromere f(x)
-tetrad analysis v. useful in yeast

Do repeats provide centromere f(x)
OR
Do repeats accumulate in centromeres

comparing unique regions

sc/v

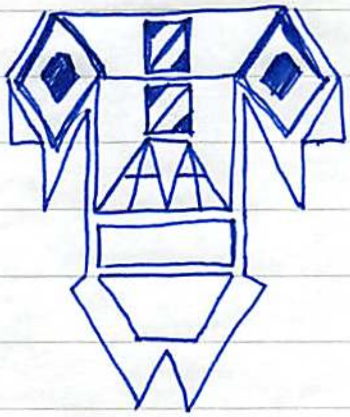
G. Rubin

Intro to Drosophila

⊙⊙ Hox clusters typical in that there are 4x copies in vertebrates than in Drosophila

BDGP = Berkeley Drosophila Genome Project

BAC based physical map



① STS data

↓

② screened bacs ③ sequenced bac ends

↘ ↙

map of genome

M

M.Adams

Goal - complete + high quality sequence
Emphasis on euchromatin

	HFlu	Drosophila	x
size	1.8mb	140mb	77
reads	26000	3x10^6	116
month	4	4	1.7
staff	24	40	10
time for assembler	1day	1day	

- want at least 70% to be
in pairs

- if you

Screen out certan regions
↓
Overlapper
↓
Assembler
-unitigr (assemble unique contigs)
-identify how unique

1200 interconnected alphas

Gene Myers
2:1 ratio of short - long good

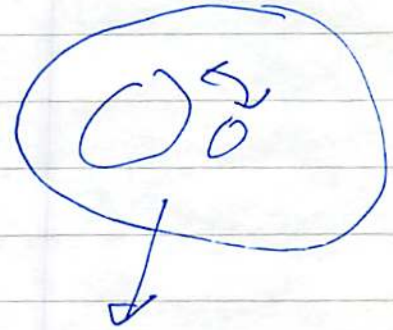
Mark Adams
Annotation
Distinguish authoratation vs. tentative

Data - many more homologs of human disease genes than in C.eleg + Yeast

Gene Ontology
classify in variety of ways

Insertional Mutants

NEED FOR GENETICS



JH - V. cholerae

- small chromosome important

Origins

- derived from main chromosome
- captured megaplasmid + gen addition
- captured chromosome + gene loss

-

S.Brenner

compare annotation

↓

minimum error rate (based on cases where there are disagreement among groups)

How does he know structural similarity is not convergence

How distinguish convergence + homology

~1000 major superfamilies

Criteria

- widespread
- heavily studied
- easy to characterize
- human disease

Blast scores exaggerated significance

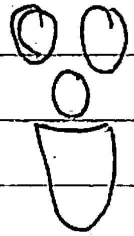
e value v. useful in iding homology

SOBrien

TProlla

Critical targets of aging

- some suggest replication
- he suggests post-mitotic more important
-



Muscle tissue

① many stress genes induced

② heat shock { HSP70 fams
HSP40 fams

repair [
- GADD45
- ERCC1
- MLH1 - sugg. also involved in O_2 damage
- AP Endo
- 8-oxo-dGTPase -
- type II DNA topo

metabolic stress [
- sarcomeric creatine kinase
- ATP synthases + subunits
- cyc C oxidase

③ decreased expression
proteosome/ubiquitin

many Δ's prevented by caloric restriction

Caloric Restriction

HSPs suppressed
Detox suppressed
DNA repair suppresses (RAD50, XPG, Pol β)

FBI

RFLP
PCR-DQα
~1/10 cases in which DNA testing
exonerated people convicted

STR's -
13 loci

IDENTITY

-

CODIS - combined DNA index system

202 324 543

Larry Young

- vasopressin + receptor
together play key role

- two copies of VR in prairie voles

these two have sim. behavior

- prairie voles has many repeats in promot
- so do pine voles
- but not as long