CAMBRIDGE HEALTHTECH MEDIA GROUP
Bio·IT World
Indispensable Technologies Driving Discovery, Development, and Clinical Trials

Welcome to your digital edition of *Bio•IT World* magazine

We hope the magazine is meeting your expectations. We encourage you to share this issue with your team to stimulate further conversations. Share this issue.

We are genuinely excited about the prospects for this field—and the magazine—in the coming year. We've cemented our position as the flagship publication of Cambridge Healthtech Institute (CHI), and will continue to play a prominent role in hosting and reporting on the best CHI conference throughout the year.

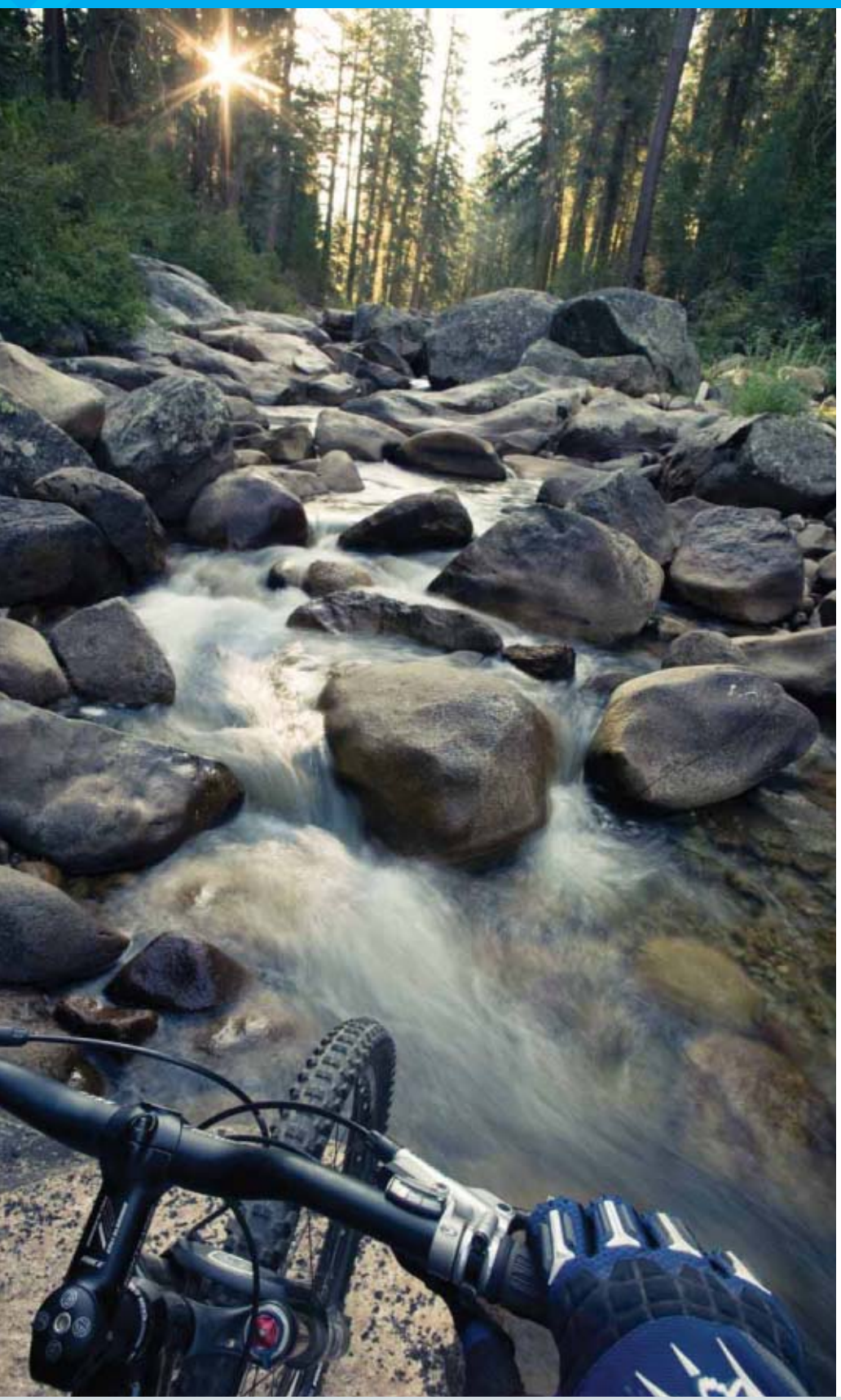
We pride ourselves on publishing critical insights and analysis of innovations across the drug discovery pipeline — from molecular modeling of popular drug targets to biomarkers that discriminate cancer responders, and from data handling for next-generation sequencing to new strategies for increasing the speed and efficiency of clinical trials. We will work hard to surpass those stories in the coming year, continuing our pursuit of the most critical tools and strategies that epitomize the world of "predictive biology."

We hope to continue to engage you with our editorial content, and within our network, and as always, we welcome any and all comments or suggestions—editor@healthtech.com

NAVIGATION KEY

- Previous/Next Page: mouse click (or keyboard) on left/right arrows to turn page
- Contents: takes you to the Table of Contents page
- To Zoom: In the browser version, simply click on the page to zoom in, and again to zoom out; in the enhanced PDF, use the zoom in/zoom out buttons in the top or bottom tool bar.
- Return to Front Cover: In the browser version, just click the double arrow button in the Nav bar; in the PDF, click on the Front Cover in the Tool bar.
- Search: Click on the search button to perform a full keyword or phrase search. You can search a single issue, or your complete archive.

Go beyond the everyday. EVERY DAY.



Symyx Notebook

The Freedom to Experiment.

Whether it's on the trail or in the lab, you want the freedom to take new approaches, routes, and paths to your goals. That's why there's Symyx Notebook. It's the only electronic laboratory notebook that can be deployed across the enterprise in multiple scientific disciplines. With Symyx Notebook, research teams share a single application to document, work, collaborate, and speed the experimentation workflow.

Symyx Notebook streamlines the capture of all experimental information and intellectual ideas. Everyday tasks such as data capture and note taking are optimized and automated. All of which gives you the time and freedom you need to experiment—and get back to doing science.

To learn more, visit www.symyx.com/notebook6

Symyx®

Cambridge Healthtech Media Group

www.bio-itworld.com

Bio·IT World

Indispensable Technologies Driving Discovery, Development, and Clinical Trials

SEPT. | OCT. 2009 • VOL. 8, NO. 5

QUAKE STRIKES

In Depth: Next-Gen Sequencing Informatics 20–39

Stanford's Stephen Quake, Norma Neff and Dmitry Pushkarev produce the first single-molecule human genome

Need to assess the novelty of substances?
SciFinder is the answer.

It includes CAS REGISTRY,[SM] the most comprehensive substance information available, integrated with relevant journal articles and patents.

Give your research team the highest quality and most timely scientific information resource.

Make SciFinder an essential part of your research process.

For more information about SciFinder, visit www.cas.org or e-mail help@cas.org.

an essential

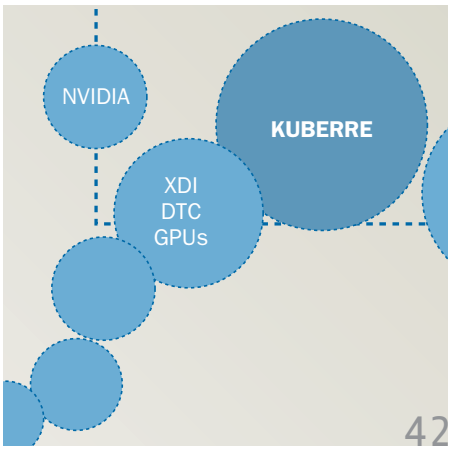
SciFinder®—Part of the process.™

CAS | SciFinder®

www.cas.org

CAS is a division of the American Chemical Society

Contents [09-10·09]



Special Report

Next-Generation Sequencing Informatics

Up Front

Clinical Trials

Computational Biology

Computational Development

IT/Workflow

In Every Issue

Cover photo by Linda Cicero

First Base

The Boys of Summer

KEVIN DAVIES

CHI's semi-annual Exploring Next-Generation Sequencing conference has been one of the highlights of a busy conference calendar, and the meeting in September in Providence—combined with a track on next-gen data analysis—won't be any exception. With that in mind, this issue of *Bio•IT World* contains a 16-page special report on next-generation sequencing informatics.

It's been an eventful summer of news in next-gen sequencing, highlighted by a steady stream of published human genome sequences. Over the past few months, we've seen not one but two Korean genomes published. The Genome Center at Washington University, St Louis, published the second leukemia genome in the *New England Journal of Medicine*, a feat unto itself but revealing some useful biological insights into the mutational basis of the disease.

In May 2007, Life Technologies' Kevin McKernan graced the cover of *Bio•IT World* beside one of the SOLiD prototype instruments. Two years later, his team published its first human genome analysis, that of a Yoruban HapMap sample. While it would be nice to see a detailed comparison between this analysis and Illumina's data on the same genome (published at the end of 2008), McKernan's group is pushing on with SOLiD 4.0 and a bake-off isn't on the cards.

Also this summer, Stephen Quake and two of his Stanford colleagues published details of his genome sequence—the first using single-molecule sequencing. Quake is the co-founder of Helicos Biosciences, and achieving its first human genome is a huge shot in the arm for the company, even if some commentators have grizzled at the misleading cost and number-of-author comparisons presented in the paper.

The financial markets perked up as well. Pacific Biosciences pulled in $68 million, while Complete Genomics found $45 million, which it says will help it sequence 10,000 human genomes next year. Those are still huge promises to keep, especially when the first goals fell short. 10,000 genomes is about 20 times more capacity than a leading genome center.

Cover boy, Kevin McKernan

The increasing pace of published human genomes is a useful indicator of the progress in next-generation sequencing, but it is by no means the only one. I'm grateful to Oxford Nanopore's chief informatician, Clive Brown, for drawing my attention to the stats maintained by the Wellcome Trust Sanger Institute (WTSI). This summer, WTSI surpassed 10 Terabases (Tb) of cumulative genome sequence data, and is generating an astonishing 400 Gigabases of sequence per week—that's the equivalent of about seven human genomes, and double what it was last winter. The NCBI's Sequence Read Archive is growing at 1 Tb a month.

Fall Forward

This Fall also promises to be a busy one for *Bio•IT World*.

- In October, we host our first conference in Europe, in conjunction with BioTechnica 2009 in Hannover, Germany. (See pages 26-27 and at www.bio-itworldexpoeurope.com.) We'll have highlights in the next issue of *Bio•IT World*.
- We're launching a series of live web symposia on a variety of topics ranging from data management for next-generation sequencing to cloud computing; remote data capture to trends in translational medicine. These will run from September through December every 1-2 weeks.
- We're introducing an exciting new e-newsletter that complements the stories you typically find here in *Bio•IT World*. Debuting in September, Pharma Services News, written and edited by John Russell, will examine the bio-IT and drug development world from the view of the services provider. From 'omics and software-as-a-service to medicinal chemistry and clinical trials, there is virtually no part of the drug discovery pipeline that cannot be outsourced to a group that can do the work faster, cheaper, and better than pharma. We look forward to highlighting many of these organizations and their interactions with the pharma industry.
- October also marks the launch of our Best Practices Awards for 2010. Following the success of this year's competition, which attracted a record 72 entries, we have high hopes for next year. Beginning in October, academic and industry organizations can submit accounts of outstanding case studies and best practices in any facet of life sciences/drug discovery. We particularly encourage nominations from the vendor community, who can encourage their clients to take the time to submit. As always, winners will be judged by an expert panel, feted at a gala dinner at next year's Bio-IT World Expo in Boston, and honored with detailed coverage in a subsequent issue of *Bio•IT World*.

Company Index

Advertiser Index

This index is provided as an additional service. The publisher does not assume any liability for errors or omissions.

VOLUME 8, NO. 5

Editorial, Advertising, and Business Offices: 250 First Avenue, Suite 300, Needham, MA 02494; (781) 972-5400

BioIT World (ISSN 1538-5728) is published bi-monthly by Cambridge Bio Collaborative, 250 First Avenue, Suite 300, Needham, MA 02494. *Bio IT World* is free to qualified life science professionals. Periodicals postage paid at Boston, MA, and at additional post offices. The one-year subscription rate is $199 in the U.S., $240 in Canada, and $320 in all other countries (payable in U.S. funds on a U.S. bank only).

POSTMASTER: Send change of address to Bio-IT World, P.O. Box 3414, Northbrook, IL 60065. Canadian Publications Agreement Number 41318023. CANADIAN POSTMASTER: Please return undeliverables to PBIMS, Station A, PO Box 54, Windsor, ON N9A 6J5 or email custserviceil@IMEX.PB.com.

Subscriptions: Address inquires to Bio-IT World, P.O. Box 3414, Northbrook, IL 60065 (888) 835-7302 or e-mail biw@omeda.com.

CAMBRIDGE HEALTHTECH MEDIA GROUP

Bio·IT World®

Indispensable Technologies Driving Discovery, Development, and Clinical Trials

EDITOR-IN-CHIEF
Kevin Davies (781) 972-1341
kevin_davies@bio-itworld.com

MANAGING EDITOR
Allison Proffitt (617) 233-8280
aproffitt@healthtech.com

ART DIRECTOR
Mark Gabrenya (781) 972-1349
mark_gabrenya@bio-itworld.com

VP BUSINESS DEVELOPMENT
Angela Parsons (781) 972-5467
aparsons@healthtech.com

VP SALES – WESTERN US, CANADA, EUROPE, PACIFIC RIM
Alan El Faye (213) 300-3886
alan_elfaye@bio-itworld.com

REGIONAL SALES MANAGER – NEW ENGLAND, NORTH EASTERN US, SOUTH EASTERN US, MIDWEST, INDIA
Kay O. Christopher (860) 693-2991
kchristopher@healthtech.com

SENIOR DIRECTOR OF MARKETING & OPERATIONS, PUBLICATIONS
Joan A. Chambers (781) 972-5446
jchambers@healthtech.com

PROJECT/MARKETING MANAGER
Lynn Cloonan (781) 972-1352
lcloonan@healthtech.com

ADVERTISING OPERATIONS COORDINATOR
Stephanie Cline (781) 972-5465
scline@healthtech.com

PRODUCTION MANAGER
Tom Norton (781) 972-5440
tnorton@healthtech.com

Contributing Editors
Michael Goldman, Karen Hopkin, Deborah Janssen, John Russell, Salvatore Salamone, Deborah Borfitz Ann Neuer, Tracy Smith Schmidt

Advisory Board
Jeffrey Augen, Mark Boguski, Steve Dickman, Kenneth Getz, Jim Golden, Andrew Hopkins, Caroline Kovac, Mark Murcko, John Reynders, Bernard P. Wess Jr.

Cambridge Healthtech Institute
PRESIDENT
Phillips Kuhl

Contact Information
editor@healthtech.com
250 First Avenue, Suite 300
Needham, MA 02494

CHI

twitter
Follow Bio-IT World on Twitter
http://twitter.com/bioitworld

Pathway Genomics: the New Kid

Another contender in the consumer genomics game.

BY KEVIN DAVIES

Twenty months after the debut of consumer genomics pioneers 23andMe and deCODEme, San Diego-based Pathway Genomics has launched its own comprehensive direct-to-consumer (DTC) genotyping service. Company founder and CEO James Plante says Pathway offers several key advantages over its competitors, including the depth, speed, and security of the genotyping analysis on offer.

Plante, a serial entrepreneur, became proactive about his own health when his father was diagnosed with polycystic kidney disease. He later died from organ transplant complications. "That was a wakeup call for me," said Plante. "I wanted to find a way for consumers to have access to genetic information at a low cost."

In 2008, Plante felt the time was right, the technology mature enough, to launch. He had little trouble raising capital, tapping investors in his previous start-up, including The Founders Fund, Edelson Technology Partners, and Western Technology Investment. "It's a great time to build a new company, particularly in San Diego, where we have depth of expertise in genetics, Illumina down the street, and pretty much everything we need within a few miles," said Plante.

Plante isn't daunted by the head start 23andMe, Navigenics and deCODE have enjoyed. "It takes a while usually for the major issues to get sorted out, before any new company starts to gather any meaningful new market share," said Plante. "We don't really believe that the folks that started 18 months ago have any significant market share yet," Plante continued. "We believe the timing is absolutely right for this."

Pathway's chief scientific officer is geneticist David Becker, formerly with TorreyPines Therapeutics. "I saw this as a great opportunity to take on one of the biggest challenges in this time of genomics—trying to translate this information into something that's useful... in a way that consumers or doctors can act on," said Becker.

Pathway's custom-designed, 10,000-square-feet CLIA-certified lab currently houses genotyping platforms from Illumina, Affymetrix, and Sequenom. Pathway vows to be technology agnostic, adopting whichever platforms provide the most value at the lowest cost.

"We can quickly adjust the technology we're using to be competitive, whether it's to have a broader offering, a better price, or whatever we feel is the important issue to provide a better product to the customer," said Becker, who stresses the secure environment offered by the lab. "We do all the processing here; we do all the genotyping here; we do all the data analysis here," said Becker.

> "We don't really believe that the folks that started 18 months ago have any significant market share yet."
>
> **James Plante**, Pathway Genomics

DNA samples from early adopters will be assayed on one disease-focused custom chip, which will supply information on complex diseases, carrier status for rare diseases, and pharmacogenetics markers. In addition, there will be "probably the most extensive SNP-based ancestry test that's available." Becker says there are plans to offer tests on the Affymetrix and Sequenom platforms as well.

Becker oversees an editorial team to review criteria from the latest peer-reviewed genome association studies. That team includes Victoria Magnuson, who trained with Francis Collins and John Todd and is an expert in type 2 diabetes genetics. A white paper will be published describing Pathway's criteria. "We've tried to be pretty conservative as to what is acceptable, validated research versus preliminary research markers," he said.

Retail Therapy

Pathway's Health Test retails for $249, while the Ancestry Kit is $199. Ordered together, the price is $348, making it the least expensive full-genome consumer genomics test currently on the market. "The conditions we have on the list are things people are most concerned about," said Becker, but with particular emphasis on rare disorders, carrier status, and pharmacogenetics.

Similar to Navigenics, Pathway features a full-time team of genetic counselors on staff, headed by Linda Wasserman, who formerly directed the UCSD clinical molecular genetics facility. However, there will be an extra fee for the counseling service because Pathway is pricing the genotyping chip so inexpensively. But that's an optional extra. "We think we can trust people with the genetic information. It's important to give them the full picture," says Becker.

Becker continued: "Obviously we'd love to just give everyone their full genome sequence, and tell them about every little detail. At this point, that's not really feasible. But we think people will be interested to know everything they possibly can about their overall health and that could motivate them to take action..." Plante also notes that Pathway Genomics will be a partner, like the other consumer genomics companies, for Illumina's new whole-genome sequencing service.

Becker said that reviewing his own data (as a beta tester) had motivated him to improve his own lifestyle in some respects, but he is trying to keep the genetics in perspective. "I have a 100% chance of dying," said Becker. "I'm not really going to fret over a few percent increase in risk over one thing or another!"

A REVEALing Study of Consumer Genomics Response

Alzheimer's risk does not seem to affect depression, anxiety.

BY KEVIN DAVIES

During a panel discussion at last year's Bio-IT World Expo, the Editor-in-Chief of the *New England Journal of Medicine*, Jeffrey Drazen, an early skeptic of the predictive power of personal genomics, outlined what steps he needed to see from the genetics community. "I'm from Missouri, and you have to show me," Drazen said. "You've got to do the study that shows that making a difference in [genetic] knowledge will make a difference in how people behave."

Drazen added, "We're not there yet... I wish you good luck, and send me your papers when you show that it works!" Sitting in the audience, Boston University neurologist Robert Green gladly seized the opening and informed Drazen he was preparing to submit just such a manuscript.

That paper, presenting the findings of the REVEAL (Risk Evaluation and Education for Alzheimer's Disease) study, was published in the Journal in July. It represents a milestone in judging the public's attitude to—and ability to cope with—the sometimes adverse results of personal genetic testing.

Green's group set out to examine attitudes of people with a family history of Alzheimer's disease to learning their all-important APOE genotype. The apolipoprotein E (APOE) gene on chromosome 19 is a well-known predictor of Alzheimer's risk. Individuals who inherit one copy of the ε4 allele have a 2-3 fold relative risk of the disease, whereas ε4 homozygotes have around a 15-fold greater risk.

The study was actually performed between 2000 and 2003. Green and his colleagues enrolled 162 adults who had one or both parents diagnosed with AD. All received counseling information before the trial began. 111 were told their APOE genotype, whereas 51 remained as controls in the nondisclosure group. Of the 111 individuals tested, 53 were heterozygous or homozygous for the ε4 allele.

Green and colleagues found few if any differences between the two groups with regard to the individuals' levels of anxiety, depression or distress, even up to one year after the study. "Subjects who learned they were ε4 positive... showed no more anxiety, depression or test-related distress than those who did not learn their genotype," the authors write. (There was a slight short-term but transient increase in anxiety in the ε4 group.) The individuals that showed the most dramatic, or clinically meaningful, changes in psychological profile were spread evenly between the control group and the disclosure group (regardless of ε4 carrier status).

Despite the lengthy genesis of the study, there are inevitably shortcomings. As Green et al. note, "If APOE genotyping had been provided without genetic counseling or to subjects who had no family history of Alzheimer's disease, the results might have been different. In addition, the exclusion of subjects with low neurocognitive scores and high depression scores may have influenced the results."

While advocating more expansive follow-up studies, nevertheless, Green's team draws satisfaction from the REVEAL findings that disclosing genotyping information to individuals who test negative is beneficial, and causes only transient, modest distress to those who end up testing positive. "These data support the psychological safety of disclosing data regarding genetic-counseling protocols" to Alzheimer's family members, "despite the frightening nature of the disease and the fact that the disclosure has no clear medical benefit." •

*Green, R.C. *et al.* "Disclosure of APOE genotype for risk of Alzheimer's disease." *New England Journal of Medicine* 361, 245-254; July 16 2009.

Briefs

COLLINS HEADS NIH

President Obama appointed Francis Collins, former director of the **National Human Genome Institute** and leader of the international Human Genome Project, to head the **National Institutes of Health.** Collins was sworn in on August 17. Collins said that he plans to focus on "five themes" for the NIH include large biology projects, translational research and medicine, health care reform, global health, and empowering the biomedical research community.

GRAPHIC STANDARDS

The **European Molecular Biology Laboratory**'s **European Bioinformatics Institute** and colleagues from 30 labs worldwide released a new set of standards for graphically representing biological information: the Systems Biology Graphical Notation (SBGN), published in *Nature Biotechnology*. The project was initiated in 2005 and the team consisted of biochemists, modelers, and computer scientists. SBGN is made up of three orthogonal languages, representing different vision of biological systems. Each language defines a comprehensive set of symbols with precise semantics, together with detailed syntactic rules how maps are to be interpreted.

DNA MODELS

DNA Software has been awarded three Fast Track SBIR grants from the **National Institutes of Health** to develop original *in silico* technologies to predict 3D structures of RNA-based molecules, improve diagnostics via modified nucleotides, and model the reaction rates of DNA and RNA experiments. The company successfully completed its milestones for Phase I of each project and recently began work on Phase II.

What's in a Name?

BY ERNIE BUSH

When asked if I wanted to write a regular column for *Bio•IT World*, it took maybe a microsecond for me to accept the offer. For a person that loves to pontificate, such an opportunity is like offering an overnight stay in a candy store to a chocoholic. Unfortunately, the second question was much more difficult: What should we name it? A few initial thoughts such as: *Eye on Drug Safety, Drug Safety Insights, Pre-Clinical & Drug Safety Watch, Pre-Clinical & Drug Safety Trends, Pre-Clinical & Drug Safety Breakthroughs,* etc. seemed a little too 'predictable' and more importantly, none really addressed the objectives I had in mind. In fact, they demonstrate a fundamental issue with how we name the whole space around early evaluation of compounds intended for human medicines.

Several years ago, many in the pre-clinical drug development community realized that their systems and management practices were heavily fragmented; too often delivering disconnected and difficult to assimilate results. One solution to this problem was to combine all of the preclinical development activities into one operational unit such that their management, objectives and deliverables would be better aligned and more integrated. Accordingly, they combined the traditional *Toxicology, Drug Metabolism, Pathology, Pharmacokinetics, Bioanalytical,* and *Safety Pharmacology* groups into one entity. Among the multitude of issues that resulted from trying to integrate such an array of people, philosophies and business practices was again the problem of what do you call this group. Some of the names adopted were: *Non-Clinical Drug Safety, Pre-Clinical Sciences,* and *Early Development*. Unfortunately these names also fall short of communicating the real intend and objects of the integrated departments.

So what are these new departments about? In fact, the central goal and objective of these new departments is to predict whether a candidate molecule will make a value added human medicine, i.e. are the properties of this new molecule adequate for it to be a good drug? In other words, is the new molecule drug-like? Or, my personal favorite, what is its "drugness"? So this drugness property is really our primary objective therefore it should be part of the name of these new integrated departments and, by extension, part of the name of this column.

Given this is an informatics oriented publication and given that informatics has traditionally been a very weak area of safety (drugness) evaluation, it would seem clear that IT-sounding words should be part of the column's name as well. In addition, during my 27 years in this industry I would say the number one frustration and shortcoming of our discipline has been access to our history. Since GLP was adopted in the late 1970s there are literally millions of toxicology, drug metabolism and pharmacokinetic reports locked away, collecting dust, and hard to access, in company archives. Pharma has invested billions of dollars in collecting this data and it is absolutely shocking to realize how poorly we can leverage this investment, especially in terms of being able to utilize it for generation of new knowledge in the form of SAR analysis, modeling, or data mining. While words like "data warehouse", "knowledge management" or "safety informatics" all have a familiar ring to them, they again seem to be either too narrow or too tech-oriented for this column.

Decisions & Doctrines

For me, the real key to value creation in this space is "easy access", i.e. once you define a system, infrastructure, and process for making the test results easy to access, then all the other opportunities such as data sharing, data mining, etc. become straight forward extensions to this access engine. As such, I would like to see easy access to our drugness history as part of, or else implied in the column's name.

One last element is critical to capture in this column. In the end, the hoped for result of generating, collecting, and analyzing all this drugness data is to make good decisions. What is the best molecule to move forward, what is the best design of the clinical development program, how should this project be prioritized against the rest of the portfolio? These are the types of fundamental decisions being made daily, based in large part on this drugness data, and of course the existing pharma R&D attrition rates suggest we are not very good at making these decisions.

Putting this all together suggests a name like: *Better Pharma R&D Decision-Making Through Easy Access to Our Drugness History*. Hmm… doesn't exactly roll off the tongue, does it? To simplify this decision and to avoid having to make up new words, my editor has suggested the name *The Bush Doctrine*. It does have a certain ring to it. But before finally deciding on what to call this new column, I would like to hear your thoughts and feedback.

Ernie Bush is Vice President and Scientific Director of Cambridge Health Associates, the leading organizer and facilitator of biopharmaceutical collaborations in the safety evaluation space. CHA plans collaborative projects, roundtable summits, virtual meetings & seminars, and the Drug Safety Executive Council (an online community of industry leaders in safety assessment). Ernie can be reached at ebush@chacorporate.com

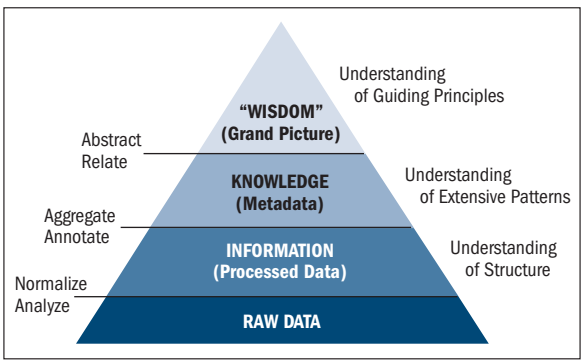
Making Sense of Data

HERMANN A.M. MUCKE

Data are the indispensable material for all analyses, but if they cannot be put into proper context and interpreted, their value is only in their potential. The famous statement by futurologist John Naisbitt, "We are *drowned* in *data* but *starved* for *knowledge*," might have been cited too often, but nevertheless it remains fundamentally true.

The structuring process that distills information from raw data and refines it until (ideally) a full understanding of the mechanisms underlying the observations is developed can be thought of as a pyramid. Layers of abstraction are added until the grand picture emerges; however, the process ultimately remains rooted in the raw data.

Because ferreting out connections between data and thereby extracting information that is relevant for a particular purpose is reminiscent of mining (which identifies and selectively removes ore from rock so that it can be processed for the desired metal), this approach is commonly referred to as data mining or knowledge discovery.



Distilling data into wisdom can be thought of as a pyramid.

Learning About the Unexpected

Data mining has been defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data" (*AI Magazine*. 1992;13:57–70). Indeed, exploratory analysis of large data sets without preconceived assumptions or quantitative models that look for emergent, fortuitous clustering (i.e., serendipity) is the "high realm" of data mining. In this case, the data are, to the extent possible, allowed to speak for themselves, and the analyst has to be prepared for surprises from emergent patterns.

Pattern discovery in large data sets requires a tremendous amount of methodological discipline. Most of the patterns that emerge during an exploratory analysis are meaningless, uninteresting in the context of the search goal, or even outright misleading because a correlation—even a statistically significant one—does not automatically indicate a causal connection between the factors. Only a thorough analysis of covariates and confounding variables can reveal if an actual connection exists.

Seeking Specific Signatures

The second approach to data mining is modeling according to a preexisting hypothesis, as opposed to "simple" information extraction, which aims to identify predefined specific classes of entities that contain explicit information of interest. In a broader sense, information extraction of any type is referred to as data mining as long as it includes a strong exploratory component.

In many cases, researchers already have concrete ideas about the structure of the data and the behavior of the system described by the data. This allows them to build a defined hypothesis that can be tested through a targeted analysis of the mined data, which searches for specific known (or specifically suspected) patterns. If the comparison shows significant misalignment between the expected and the observed patterns, new hypotheses can be derived from this disparity. In this case, data mining amounts essentially to an iterative process of targeted data extraction and processing.

In today's extremely competitive environment of academic and commercial life science applications, it is of essence to know as much of the prior art as possible before it can even be determined if a project can be undertaken, and how it should be structured. Published information will both guide and confound any planned efforts, and a better analysis of the prior art will allow the self-directed efforts to be targeted more accurately.

A new Insight Pharma Report examines the emerging role of the various flavors of data mining in translational drug development (*i.e.*, the formal stages of preclinical and clinical investigations) and pharmacovigilance (*i.e.*, the surveillance for potential side effects in the postmarketing stage). We present brief profiles of software and service providers that cater to the pharmaceutical and biotechnology industry, and point to the directions that the use of data mining in transitional drug development and postmarketing pharmacovigilance might take during the next decade. We conclude that the use of data mining in these fields of the life science industry is dynamically emerging yet has huge potential, which will be increasingly realized during the 2010s.

Further Reading:
Data Mining In Drug Development and Translational Medicine, by Hermann A.M. Mucke, PhD, was published by Insight Pharma Reports in July 2009. For more information, visit www.insightpharmareports.com/data_mining

Clinical Trials

DIA 2009 Trendspotting

Talking globalization, adverse events, and successful sites at the Drug Information Association annual meeting.

BY ANN NEUER

SAN DIEGO—At the DIA annual meeting in San Diego in June, clinical trial professionals met to discuss what's working, and what's not, in the clinical trials space.

Many conversations centered on the role of the investigative site in clinical trial operations. The site is the end user of clinical trial technology, so there is a need to better understand how it operates if its performance is to improve. Newly released versions of electronic data capture (EDC) tools boast simplicity as the best way to engage sites in collaborative efforts with sponsors and contract research organizations (CROs) seeking better performance. Steve Powell, Phase Forward, said, "Everyone is battling for the investigator space. The simpler you make it for them, the more beneficial it is for the sponsor running the trial."

Another trend focused on data management across the enterprise and the growing emphasis on standards. Terek Peterson of Octagon Research Solutions commented, "There is a stabilization of CDISC standards—including SDTM (Study Data Tabulation Model) and ADaM (Analysis Data Model). Some companies are somewhat resistant to adopting standards as it is disruptive to their process, but after years of talking about it, people are adopting standards."

Carmen Gonzalez, with Healthcare Communications Group, noted what was missing: "What wasn't present this year is the growing impact of social media. We know the patients are there, but we need some guideposts so we can do a good job using social media."

Global Solutions

Globalization was a central topic, the most visible example being presence staged by the Federation of Indian Chambers of Commerce and Industry (FICCI). From the Indian Pavilion in the Exhibit Hall to dedicated sessions to a special "India evening" featuring representatives from the Office of the Drugs Controller General of India (DCGI)—the Indian regulatory agency—the FDA, and industry, it was clear that India is positioning itself to be a major player. Dan McDonald, VP business strategy, Excel Life Sciences, a trial management organization focused on India, said the country is en route to being an accepted clinical trials locale. "Two or three years ago, there was a lot of misconception about India. But now, it's a destination that people consider."

Dan McDonald believes that India is poised to be a significant player as a clinical trials locale.

McDonald's observation is supported by the FICCI. According to company, the percentage of global clinical trials in India is estimated to be 8% in 2009, and forecast to jump to almost 14% by 2011, a 75% increase in two years.

At the India Evening, panelists discussed this growth as part of the country's vision of making India a global hub for pharmaceutical innovation. Much of this effort is slated to happen through government initiatives aimed at encouraging public-private partnerships. Murali Nair, a partner at Ernst & Young India, explained, "We will shortly be commencing work for the Indian government regarding what should be its focus and role in this effort."

One of the most telling developments that India is a growing force in the pharmaceutical sector is the recent opening of two FDA offices in that country—one in New Delhi, the other in Mumbai—eventually employing a dozen people. The offices provide technical experts and inspectors in regulated product areas such as drugs, devices, and food. Although their responsibilities extend well beyond the realm of clinical trials, the offices are expected to be a major plus in terms of working with Indian government authorities to ensure the quality of clinical research from India coming to FDA.

According to David Lepay, FDA's senior advisor for clinical science, the new offices are already proving valuable. "They have provided a communication channel to establish interaction, and over the past few months, we have had a very productive interaction with the DCGI office." Lepay said the Indian regulators have asked the FDA for help in setting up systems of clinical trials oversight and inspection that are compatible with international standards.

ASTER Watch

As always, patient safety and adverse event reporting was another hot topic. Michael Ibara, head of pharmacovigi-

(CONTINUED ON PAGE 16)

Clinical Trials

Site Gripes and Solutions for EDC

Of all the e-clinical technologies, electronic data capture causes the most site frustration.

BY DEB BORFITZ

Even the most progressive, tech-savvy investigative sites are aggravated by many of the e-clinical systems being embraced by study sponsors. Virtually no type of clinical trial technology escapes criticism, but EDC systems seem to be the biggest offenders, according to several sites that spoke to *Bio•IT World* about their experience. However, the same sites also made suggestions for easing the relationship.

Today's EDC systems typically have the horsepower to process large volumes of data simultaneously, but still manage to create more rather than less work and expense than they did in 1995, says Kelly Walker, director of operations at Research Across America (RAA), an independent site network based in Dallas.

The biggest problem is with web-based EDC systems that get placed on servers and networks "not nearly robust enough to handle the workload," says Walker. In some cases, web pages won't "refresh" correctly because the browser is trying to cache those pages to reduce server load and perceived lag. Although there are cache-control settings to rectify the problem, site personnel aren't instructed how to make the necessary adjustments.

Another ongoing issue is the variability with which EDC systems support different versions of foundational software such as Internet Explorer and Adobe Reader, says Walker. "It means study coordinators working on two different studies [often] need two different computers on their desk." Site staff can also inadvertently wreak havoc whenever they respond to automated promptings to update any of that underlying software.

That sponsors now have the ability to instantaneously reformat pages or add new questions to electronic case report forms (CRFs) has also created problems for RAA, says Walker. Sites are expected to gather the new information retroactively on patients who probably don't remember if that headache they had several months earlier was mild, moderate, or severe. Other times there is rampant indecision about what should count as an adverse event, leaving sites to alternatively delete and re-insert data. The high turnover rate among study monitors is a contributing factor, he adds.

Perhaps the biggest technical oddity at RAA is that two of the three PCs outfitting Walker's office won't display CRF pages correctly. Date field boxes are missing and no manner of technical support can figure out why. "Our people have had to learn how to enter data without the boxes....where they know [from experience] the answers should go."

Pedia Queries

At Pedia Research, the chief complaint about EDC systems is that they often generate unnecessary queries, says Richard Litov, director of the dedicated, three-site network based in Owensboro, Ken. To avoid answering umpteen automatically generated queries, site personnel wait until they have every last piece of patient visit information (including lab results) before entering any data at all. Often contributing to the query problem is that the software fails to help users at the sites visualize what information is missing or incomplete without the time-consuming process of opening each individual screen.

CRF pages written in-house by large companies using the Oracle database are the bane of Diane Palmer, site administrator and clinical research coordinator at Coastal Connecticut Research (CCR), a dedicated site in New London. With one recent study, the site had to enter data on 46 separate pages for every neurological exam done by the investigator. Each exam took only ten minutes, but data entry took a whopping 90 minutes. "We almost quit recruiting for the study because we were losing money."

Part of the problem is that headers are frequently not pre-populated with data from their previous visit, says Palmer, meaning someone at the site has to constantly re-key the visit date in the prescribed format—i.e. 03 May 2009 versus May 3, 2009 or, if the date is unknown, "not done" versus "ND" or "00." To make matters worse, each CRF page often contains only one or two data fields that have to be filled in and saved before moving on to the next set of questions.

Palm Pilot-type based electronic patient diaries also aren't particularly popular with the interviewed sites. One recent e-diary study at Pedia Research was a "complete disaster," says Litov. Patients ended up making entries on paper and the site struggled to re-key the information into the unwilling system. In this case, the central problem was a rushed timeline. "The poor vendor didn't have enough time to get the system ready to go, error-free." Patient misuse of the technology is part of the problem, says Walker, but the bigger issue is that e-dairies don't mimic clinical practice. If study participants fail to take a pill at the exact time specified the data they enter will be rejected. Questions can also be oddly or unclearly posed, confusing patients. Diaries appear at times to erroneously record entries about symptoms, disqualifying subjects who seem perfectly suited to a study.

The vendors, of course, blame the site or the user. But Walker says such predica-

(CONTINUED ON PAGE 16)

Another ongoing issue is the variability with which EDC systems support different versions of foundational software such as Internet Explorer and Adobe Reader.

Phase Forward Acquisitions Broaden Clinical Trial Solutions

The company's new solution includes ePRO and voice response services.

BY ANN NEUER

In a buying frenzy between April and July this year, Phase Forward made three acquisitions and launched InForm GTM—the latest version of its flagship electronic data capture (EDC) solution. It started with the $14-million purchase of Waban, a provider of platform solutions for the automation of clinical data analysis. In July, Phase Forward agreed to purchase Covance's Interactive Voice and Web Response Services (IVRS/IWRS) business for $10 million, and also spent $11 million to acquire Maaguzi LLC, a provider of Web-based, electronic patient reported outcomes (ePRO).

Martin Young, VP corporate development and marketing, says this intense activity supports the company's strategic vision of providing clients with a suite of solutions as the EDC market matures and moves beyond point solutions. "Our explicit strategy over the past three years has been to build on InForm's strengths, our point solution that has driven the initial automation of clinical research. We are now at the point that EDC will be used in approximately 55-60% of global trial starts this year, so we need to automate and integrate key areas such as randomization of patients, patient diaries, and provide a clinical data repository," says Young.

The growing use of ePRO is a good example of how more challenging studies impact the kinds of tools best suited for integration into a complete system. Maaguzi's solution is Web-based, requires no hardware, and can be used on any device with a browser, ranging from a desktop computer, laptop, even a smart phone. "There is a trend toward large late-phase studies with huge numbers of subjects, sometimes tens of thousands," says Young. "Often there is an ePRO component in those studies, so providing those subjects with PDAs can present huge logistical challenges and is cost-prohibitive. Some 25-30% of the cost of studies can be the cost of hardware. With the Maaguzi solution, there is no hardware cost."

Tied into the recent acquisitions is the June release of InForm GTM (Global Trial Management), an upgraded version that aims to raise the bar on this well-established EDC solution. Paul Boyd, director of product design, was intimately involved in developing InForm GTM. "When we watched our customers work, we saw how complex their jobs are, in terms of working with multiple sponsors on multiple studies with multiple EDC products. We saw that technologies are built into people's lives in a way they weren't ten years ago," Boyd says.

That initiative stimulated changes that are part of InForm GTM. The interface includes updated icons, streamlined navigation, and the ability to configure how many subjects per page the end user can see at once. He explains, "This allows information of interest to jump out at the study coordinator, such as whether there are unresolved queries. This is an easy and more efficient process than having to switch among pages in order to find this same information."

Sponsors Run to goBalto.com

Site provides feedback on drug development partners.

BY DEB BORFITZ

A newly launched website—goBalto.com—is providing the drug development industry with unprecedented intelligence about their would-be partners around the world. Initial response to the site, in beta testing since April, suggests life science companies have been hungering for the resource. Membership has been growing at the rate of hundreds per week and now numbers around 5,000, says Jae Chung, goBalto's founder and CEO. The company went from conception to incorporation in a week, with funding by angel investors.

The pharmaceutical industry has lagged in rating service providers, says Chung, who describes goBalto as a cross between consumer review sites Yelp and Expedia and the business networking site LinkedIn. In addition to reviewing past external partners, the goBalto user community can identify new ones—by name, region, and specialty—as well as ask questions, post requests for proposals, hold discussions, and collaborate.

"We can shave close to two months off the process of finding and evaluating vendors," says Chung. For service providers, notably CROs, goBalto is also a convenient way to "stand out from the crowd." The online directory of service providers includes close to 10,000 companies spanning more than 500 service categories, including all 2,000 or so CROs operating worldwide.

The website's name is derived from the Siberian Husky sled dog, Balto, who led his team on the final leg of the 1925 serum run to Nome to deliver diphtheria antitoxin.

Although not its original intention, Chung says goBalto is also being used by service providers to rate pharmaceutical companies on the quality of their outsourcing teams. The bulk of inquiries, up to 50 a day, are being generated by sponsors. Beta testing of goBalto is expected to conclude early next year.

BARNETT EDUCATIONAL SERVICES

The Path to Drug Development, Regulations, and Industry Trends

NEW! Medical Device Development: Regulation and Law

This practical reference provides the most comprehensive and updated analysis of US medical device and diagnostics development and approval requirements anywhere.

PAREXEL's Bio/Pharmaceutical R&D Statistical Sourcebook 2009/2010

Considered 'a must-have resource' for the drug development industry. This statistical sourcebook is filled with pharma/biotech R&D trends, data, market intelligence, articles, and graphs providing fresh insights into the developments reshaping pharma R&D and the drug development industry. Also available in electronic format.

The U.S. Drug Approval Trends and Yearbook 2008/2009

A one-of-a-kind compendium filled with hundreds of key trends and metrics used by professionals to gain industry insights, new benchmarks, and analysis to help plan their own R&D projects, improve company performance, and assess various drug approval options and strategies.

New Drug Development: A Regulatory Overview 2008

Addresses the most cutting-edge developments redefining how new drugs are developed and regulated today. The book is considered the "go-to" resource for regulatory, clinical, project management, training, and other drug development disciplines navigating the FDA's drug development approval processes.

Expediting Drug and Biologics Development: A Strategic Approach

Using a unique "reverse-engineering" approach, dozens of leading experts with extensive experience in all disciplines of drug and biologic development show how careful planning and a sharp focus on the end-goals can be used to expedite the most complex product development program.

Biologics Development: A Regulatory Overview

The only text book on biologics development and regulation in the post-CBER/CDER Consolidation Era! Offering an expansive examination of the FDA's regulation of biologic products, from preclinical testing to post-marketing regulatory requirements, and from user fees to electronic submissions.

- Phase 4 commitments
- Emerging pharma/biotech R&D trends
- R&D performance and productivityt
- Phases of clinical development
- Medical device & combination product regulations
- The Biological License Application (BLA)
- Priority drug review destination
- CDER and the NDA review process
- Pediatric studies initiative
- Post-approval changes to marketed drugs
- Medical device compliance

Purchase your copy(ies) today for your colleagues, research staff members, and professionals involved in the management and conduct of clinical research.

To order, call 1-800-856-2556 or visit the "Publications" section on www.barnettinternational.com

www.barnettinternational.com

Barnett International: A division of Cambridge Healthtech Institute, www.barnettinternational.com
250 First Avenue • Suite 300 • Needham, MA 02494 USA • Phone: (800) 856-2556

CHI

EDC

(CONTINUED FROM PAGE 12)

ments happen often enough that the technology must be called into question.

Site Solutions

But sites also offered some solid remedies for improving the situation. Sponsors building electronic case report forms (eCRFs) in-house might want to start pre-populating data fields with information from subjects' prior visits, suggests Palmer. Better yet, site personnel could enter study visit data on a single page.

The Holy Grail of eCRFs would be for the industry as a whole to come up with standards for data entry, she continues. For sites like CCR doing multiple studies, it's difficult to remember company-specific requirements for inputting information. There are four alternatives for the date format alone: 5/26/2009, 26May2009, 26May09, or 5/26/09. Failure to follow data-entry rules results in automatically generated, time-consuming queries, which steals time from patient recruitment activities.

Palmer declares Phase Forward's Inform EDC system "one of the best" and worthy of mimicking, including its online training module that includes a printed certificate of completion.

EDC systems as a rule fail to provide the kind of information that could make life easier for sites, such as a study's enrollment status, says Litov. Instead, they deliver the intelligence six weeks late via hardcopy newsletters. "Sponsors and CROs can see that information [electronically]. Why not us too?" Similarly, adds Litov, central lab vendors continue to transmit lab values via fax. In an ideal world, sites would receive lab reports electronically, including graphics comparing current and previous lab values of a study subject so investigators don't have to sift through paper charts to detect important changes. Doing so "would lead to better oversight and evaluation of safety."

Litov says it would serve sponsors' interests to involve a few sites in the technology acquisition process and then conduct post-study surveys to learn how well a chosen EDC system performs in the real world. "We've seen a sponsor switch to a different EDC system for each of three studies conducted in the same program."

From consistency and accuracy, it would be advantageous for sponsors to provide a standard, study-specific template for source document forms rather than rely on each site to generate their own forms from scratch, especially since these forms mirror the CRFs that are already produced, says Litov. He estimates that Pedia Research donates between 5-20 hours of "secretarial work" per study to create the requisite forms. "Multiply that by 50 or more sites for a multi-center trial and that is a lot of wasted time diverting sites from starting up a study."

Kelly Walker believes it is time EDC systems were incorporating user-friendly handwriting or voice recognition features. Knowing it could be a long wait, he has taken matters into his own hands by developing a single-data-entry system called DB Pharma, with a programmer familiar with Microsoft tools.

DB Pharma, run off a wireless tablet, utilizes handwriting recognition technology and immediately transmits data to a web host. But it only gets used on a few studies a year, when RAA is responsible for either all of the data capture or all of the data analysis. The system's only downside, Walker says, is that it doesn't accommodate doctors' notoriously sloppy script. •

DIA 2009

(CONTINUED FROM PAGE 11)

lance information management at Pfizer, chaired a session on a highly original pilot study focused on a simple way for busy clinicians to report adverse drug events (ADEs) to MedWatch, the FDA's safety information reporting program. The study, named ASTER—ADE Spontaneous Triggered Event Reporting—has just wrapped up its three-month pilot, which was conducted earlier this year at two Partners HealthCare hospitals in Boston: Brigham and Women's and Massachusetts General Hospital (MGH).

"This was a proof of concept study involving real doctors and real patients. Our intent was to create a new business model using digitized data from electronic health records so doctors could quickly and easily report adverse events." Ibara says.

Ibara, through Pfizer, collaborated with other institutions with a strong interest in improving patient safety. His colleagues included Jeffrey Linder, representing Partners HealthCare, and principal investigator on the study; Landen Bain of CDISC; and Lise Stevens of FDA. CRIX International, a not-for-profit organization dedicated to building a common electronic infrastructure for the clinical research industry, provided the technology that enabled the information to be forwarded to FDA.

Linder explained that 26 doctors with affiliations at either Brigham and Women's or MGH were selected to participate in the study. Doctors were chosen based on two main criteria. "These are doctors who are very busy and who discontinue a lot of drugs in patients due to adverse events," Linder said. Despite their workloads, these doctors were keen to participate provided the reporting was straightforward.

The ASTER study signaled doctors electronically when they entered information into the electronic medical record about a patient discontinuing a medication due to an adverse event. At that point, a pre-populated form would pop up requesting the outcome of the event (death, hospitalization, etc.), and the earliest date of the adverse event. Once complete, the doctor would hit "OK", and the data would be sent off. Packaged with it would be additional information taken from the medical record, such as the adverse reaction, other medications taken by the patient, demographic data, and laboratory results.

Training was minimal; after a few attempts it took the doctors less than a minute to fill out the form. According to Linder, "We did a survey at the end, and overwhelmingly, they said, 'This is great. It was fast, and didn't interrupt my workflow at all.'" •

Computational Biology

Celera's Workflow Informatics

Automated workflows from InforSense are now a part of Celera's DNA.

BY JOHN RUSSELL

While widespread adoption of automated workflow tools has been sluggish in life sciences, Celera has jumped onboard, building on its several-years-old relationship with workflow platform vender InforSense. Nowadays, Celera's data handling comes in the form of genome-wide association studies (GWAS) and functional genomics work to identify and validate biomarkers.

As explained by John Sninsky, Celera's VP, Discovery Research, the turn to automated workflows for informatics analysis stemmed from its business plan evolution. Famously founded to sell access to its trove of genomic data, Celera transitioned its business model "to have solely a diagnostic and pharmacogenomic focus." Two years ago, Celera purchased a CLIA-approved clinical reference laboratory—the Berkeley Heart Lab—enabling it to offer services and generate *in vitro* diagnostic products.

In changing direction, Celera sought to exploit unmet medical needs. It initially settled on six areas ranging from autoimmune disease, neurological disease, liver disease, cancers, and cardiovascular disease. A discovery team for each area was established and work progressed. "Over the years we have found important associations but in some cases the therapeutics area hasn't developed as rapidly as we would like," says Sninsky. "For example, we were hoping that new drugs would have come on board for Alzheimer's disease so that our risk markers would have had value in early evaluation of treatment." Alas, those therapeutics did not materialize and the Alzheimer's program has been suspended.

Indeed, Celera has pragmatically pared back in areas where the payoff seemed more distant, focusing instead on three areas: cardiovascular disease, cancer, and liver disease.

Two major challenges prompted the adoption of automated workflows. One was the sheer industrial scale that Celera has traditionally managed, whether it be sequencing, SNP discovery, or mRNA profiling. "That industrial approach generates large amounts of data and complex data that you need to filter, sort, analyze, and interpret," Sninsky says.

Second, as the analysis tools were pushed out into the disease teams, "there began to be idiosyncratic modifications and decisions made about how one analysis would be done and what kind of filters would be used. We started ending up with a non-standardized analysis, very similar but different for different disease indications."

> Celera has pragmatically pared back in areas where the payoff seemed more distant, focusing instead on three areas: cardiovascular disease, cancer, and liver disease. "We went from eight different teams down to three teams."
>
> **John Sninsky**, Celera

Sninsky happened to know InforSense CSO Jonathan Sheldon from when they were colleagues at Roche. Before long, the two companies were collaborating. The idea is to be able to rapidly build, archive, and re-use workflows, which would bring efficiency, better control over the processes Celera scientists used.

David Ross, Celera's director of computational biology, says, "The standardization and uniformity led to remarkable improvements in how we as an organization dealt with the data." He cites one workflow developed for expression analysis over a couple of afternoons that would have taken weeks for a developer. "It's also reduced the amount of time that the patent group needed to address the question of how a particular analysis was done," he says. "They don't need to ask, 'Was this done? How was that done?' It was done in a standardized way."

Sninsky estimates efficiency has jumped five-fold, and notes that those productivity improvements are getting the attention of other divisions at Celera, including development and the clinical reference laboratory. "Although we served as the entrée for the Celera organization to the InforSense tools, my expectation is they are going to be embraced by a larger number of people," says Sninsky.

Slow Maturation

Despite their power, informatics workflow automation tools have taken time to catch on. InforSense and SciTegic (Pipeline Pilot) were both founded in 1999 to bring the technology to life science research. Teranode was formed in 2002 with a similar strategy. But scaling up business has proven difficult. Accelrys purchased SciTegic in 2004 and IDBS has just acquired InforSense.

Among the many possible reasons is that the platforms, though powerful, can be tricky to use. The pace of change in experimental and analysis technology has made companies reluctant to invest in automating workflow tools, suspecting there will always be too much manual work required. The ability to easily integrate a sufficient diversity of third-party

analytical tools is also important. Even conveying clearly what the platforms do can be challenging. (Both InforSense and Accelrys/SciTegic have increasingly positioned themselves as business intelligence/analytics platforms suitable for many industries.)

So what is an informatics workflow? Ross describes the elements of a workflow this way: "You grab data, most of the times from a database, which is both internal data and public data that we've put in; we manipulate that data—by that I mean pivot it or transform it or whatever we need to do to get it into the form we need to submit it for an analysis procedure... then they are displayed and they can be manipulated."

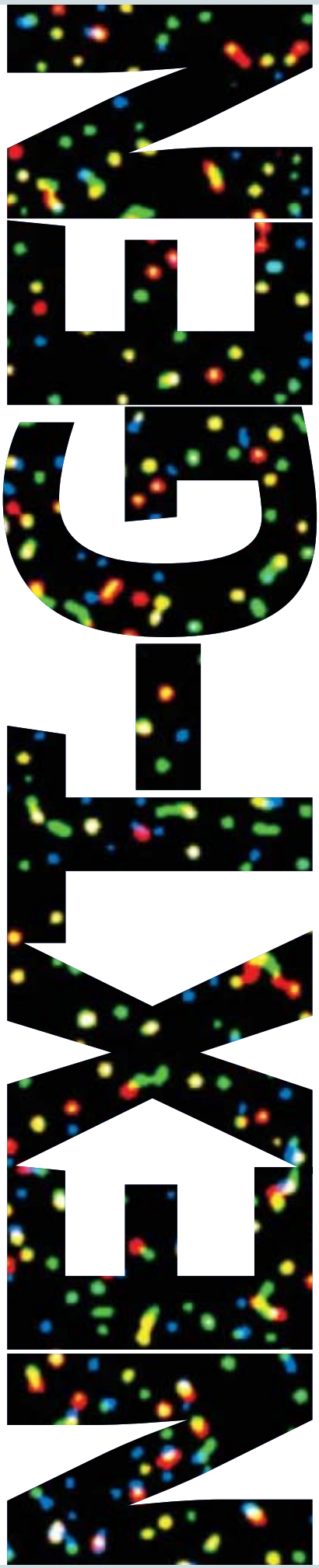
The InforSense platform "allows us to put a number of different analytical engines in the middle of that very easily." It's an organic system, he says. Manipulations can be done in SAS or R or Matlab. "We can easily grab those new analytical procedures and try them ourselves. We're particularly interested in the new semantic languages and databases."

Sheldon agrees. "In an area like genetics," he says, "there really isn't a set of five standard workflows that do genome-wide associations. You need a very open platform. You need the ability to rapidly integrate a whole variety of different algorithms from different data sources. Workflow has been the mechanism by which [Celera] could rapidly prototype different approaches to analyze the data."

Working together early has benefited both companies. InforSense was still developing what would become its Translational Research Solution (TRS), aimed at biomarker discovery activity. Celera was able to influence its direction, for example suggesting early on the inclusion of a SAS node to meet its needs.

Sheldon credits much of the input from Sninsky and Ross over several years to fine tune GenSense, such that InforSense can cope with genomics data types and include analytical methods appropriate for GWAS. The TRS includes modules for various 'omic analyses, as well as ClinicalSense for cohort identification and patient stratification, which Sheldon says is "typically one of the first steps that you carry out in a translational research study to identify biomarkers." VisualSense is the module for report generation, although Spotfire is also supported.

InforSense has also worked extensively with medical institutes such as the Mayo Clinic and Dana Farber, encapsulating that translational research insight into the product, says Sheldon. Sheldon is seeing signs that the surge in GWAS and translational medicine approaches will boost demand for informatics workflow tools. InforSense is working closely with two major pharma and several others at earlier stages. Meanwhile, Celera has at least one test approved (KIF6) and is busy looking for more. •



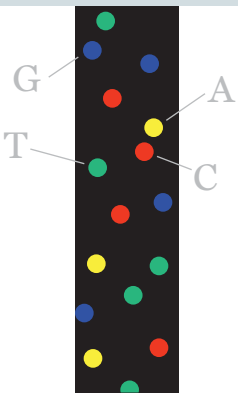
Special Report

A decade ago, researchers involved in the Human Genome Project held a party to mark the sequencing of one billion bases. This summer, the Wellcome Trust Sanger Institute surpassed 10 Terabases of mapped DNA sequence, with a current weekly output of about 400 Gigabases—equivalent to about 7 human genomes—and that's just one genome center. Close to 1000 2nd-generation sequencing instruments (454, Illumina, Applied Biosystems, Helicos) are in use around the world. Illumina dominates for now, accounting for 75% of the sequence in the NCBI Sequence Read Archive. The stream of complete human genome papers now includes two Korean genomes, the first human genome on the Life Technologies SOLiD platform, the second cancer genome, and Stephen Quake's sequence on the HeliScope.

This explosion of data puts added stress and pressure on the beleaguered informatics and IT teams. In this special 16-page report, coinciding with CHI's Exploring Next-Generation Sequencing conference (September) and Bio-IT World Expo Europe (October), Kevin Davies interviewed some of the key leaders and evangelists in 2nd- and 3rd-generation sequencing informatics. On the user side, **Bruce Martin** at Complete Genomics [page 21] has the trivial task of building the IT infrastructure to support 10,000 human genomes in 2010. Also interviewed are the software men at two 3rd-generation companies: **Clive Brown** (Oxford Nanopore) [p 28] and **Kevin Corcoran** and **Scott Helgesen** (Pacific Biosciences) [p 34]. Their past experiences at Solexa/Illumina, Applied Biosystems and 454 Life Sciences, respectively, should prove instructive.

On the government/academia side, **Jim Ostell** and **Martin Shumway** discuss the NCBI's Sequence Read Archive [p 36], the official repository of short-read sequence data. **David Dooling** describes how The Genome Center at Washington University, St Louis [p 30], is coping with continued expansion demands. Stanford's **Steve Quake** [p 25], co-founder of Helicos, discusses the impact of his own genome, produced by single-molecule sequencing. Davies also gets the perspective of a pair of software vendors which offer solutions for handling next-gen data—**Jan Lomholdt** (CLC bio) [p 32] and **Ron Ranauro** and **Richard Resnick** (GenomeQuest) [p 23].

What this report clearly illustrates is that the term "next-generation" has become obsolete. As one platform maker has noted, we really are talking "now generation."



Complete Compute: An Interview with Bruce Martin

Software engineer Bruce Martin's chief claim to fame came early in his career. Hired straight out of school in the late '80s, he was a member of James Gosling's team at Sun that developed Java. He moved onto other things as Java scaled up—mobile and email communications, banking and compliance—but nothing remotely close to life sciences.

Martin's latest challenge—building the IT and informatics infrastructure at Complete Genomics (see, "Will the Gene Microscope Change the World," Bio•IT World, May 2009) to build a human genome sequencing service capable of delivering 10,000 genomes in 2010—is a doozy! If he's having sleepless nights, he doesn't show it. "It's about crafting the right team with the right mix of skills and knowledge, and trusting them." Recruited by CEO Cliff Reid, Martin didn't hesitate. "For about a decade, I'd wanted to find something with a stronger footprint in the sciences, but also where I could contribute. So it seemed like a perfect marriage." Martin brought the expertise in software development and high-scale computing, and built a team of bioinformatics experts, experts in genomics, assembly, and large-scale scientific computing.

Bio-IT World *Editor-in-Chief* ***Kevin Davies*** *visited Martin at Complete Genomics (CGI) headquarters in Mountain View, California.*

***Bio•IT World*: Ok, where is the data center?**

MARTIN: For a variety of strategic reasons, we don't want to build out a data center here. A data center is capitalized over 25-50 years. Some pieces of it, like generators, have a 40-50 year depreciation path. That's just not typically something a small company wants to finance. The alternative is an outsourced data center. At this point, we have over 1000 compute cores and 1 petabyte (PB) data storage for R&D and pilot projects, with plans to scale up significantly for production. It is all located offsite at a co-location facility we lease in Santa Clara.

Think of the data center as a large warehouse with security, disaster recovery, back-up power, multiple power grids, etc. You go through these airlocks and security stations, biometrics, the whole nine yards. It's the size of a football field lined with chain-linked cages, trays of fiber optics and lots of computers. We have a chunk of the room, with our own security process to get into the CGI cage…

Once you use a megawatt of power (enough to supply a small town of 1000 people), you have access to a market where you get your own room! You get to decide a lot more about how you want it wired and organized. This is typically called the wholesale market. These facilities are used by large-scale compute consumers like a Yahoo or eBay, or companies who divide up the space and resell it on a retail market… We have a 10-Gigabit WAN connection to our data center, which is relatively fast (1000 times a cable connection). We push all of our data offsite to this compute environment.

What happens as you move from R&D into commercial sequence production?

For productization, our focus is scale-out and that ends up being a fairly complicated transition. As our sequencers get faster, we deploy more of them, our efficiencies get better, and we need more and more compute systems. We have preferred technologies and a software architecture than can distribute the workload into a large cluster of compute and storage. This enables us to scale as our business scales—we rack it out as we need it. From a facility standpoint, we're going to deploy into a wholesale space, where we have more flexibility and lower costs. We can pre-allocate a lot of space, build out the infrastructure as we need it, and have a much, much, much faster network connection to the data center.

One of the nice things in an area like Silicon Valley is that there's a lot of dark fiber in the ground—bundles of fiber optics that are not in use. If you're willing

(CONTINUED ON PAGE 22)

Bruce Martin

(CONTINUED FROM PAGE 21)

to bear the expertise and cost of running the network gear, you can cut your cost dramatically by using this infrastructure. This is what people do when they need to move a lot of data. We'll have very reliable, redundantly connected connectivity at hundreds of gigabits at a fraction of the equivalent Internet connection. As we grow, we will be able to scale into hundreds, or multiple hundreds of gigabits per second [gbps] in 2010. That's a function of how much data the instruments generate. Our cost, scale, and reliability analysis indicate that we are better off putting most of our compute offsite. These instruments throw off a lot of data!

Have you already selected specific storage and network partners?

In some cases yes, in others we're still in the evaluation phase. One of the things you try to do when you design an analysis pipeline is maintain the ability to switch vendors and evolve the technology platform to take advantage of new offerings. There will be technology evolution and multiple generations of hardware. The goal is to optimize around cost, efficiency, and the operational aspects such as quality control and product features rather than being tied to a single platform or vendor. You'd like to roll out an RFP every couple of quarters for any major technology purchase, and have the ability to pick the current best of breed.

Most our effort has been around determining the features, functions and architecture of our ideal software platform, rather than the final technology vendors. In some cases we have our initial preferred vendors, and we know from whom we want to buy our first few tranches of equipment. We're getting close on storage, and close on networking with computing and other elements to follow.

Which vendors have you selected?

We selected Isilon for our R&D storage platform. We have close to 1 PB of Isilon and we run all of our sequencing operations on the platform. They're under consideration for our production platform as well, but we haven't made our final selection. We're still having the bake-off.

In our selection process there are multiple steps. It starts with an initial phase where your IT architecture team runs around with a butterfly net trying to capture information about technology and vendors with a very open mind… what could solve our problem? Rapidly you reduce possible solutions to a handful of vendors where you do a deep evaluation and get into the specifics of the platform. What particular architecture would work? What are the power and space and other operational characteristics? How well does it fit into your reliability model and your scalability model? We're pretty far down the road with three vendors on storage, and it's unlikely it wouldn't be one of those three.

How do you ship the data off the machines to do the assembly?

Our pipeline is a subject of active R&D. A picture of what we're doing now: Life starts as an image—we go through a few steps of image processing, where we pull out intensity data from the images. We extract all of the different channels, light frequencies, and then we go through a base-calling process.

At this point, we make a probabilistic evaluation of what the most likely base is. We score it with a vector of information that will also tell us other less likely but possible calls. (Those initial steps are all done in real time—the scale and speed are so high we don't write that to disc. We stream it through a cluster of computers in a redundant and reliable way). That information gets poured into a pipeline that does things with a more traditional HPC cluster model. We save those calls, with the score information and the probability vectors to a large disk farm.

Then we go through a set of fairly traditional steps to filter and map the data, though we have some unique algorithms due to our read structure, and novel software enhancements. Our filtering process is relatively standard—we're looking at things like reads with insufficient quality to be useful downstream, systemic error modes, things we know about our production processes, etc. This filtering helps downstream analysis.

We then enter into the assembly pipeline, which is optimized for human resequencing. Because our business model is focused on human resequencing, we take advantage of that knowledge to optimize and reduce computing costs. We align the reads to a reference genome—similar to other high-speed short-read realigners (mappers). Mapping is used solely to pool and organize the reads by likely positions on the genome where they're probably going to be contributing information. We then assemble these partitioned reads using a combination of a local *de novo* process to generate sequence hypotheses, a Bayesian model to evaluate the quality of the hypotheses, and an optimization loop to find improvements in the calls.

We're trying to find the best-fit hypothesis for a given set of reads. Once we get a set of very fit hypotheses in every region of the genome, we then proceed to variation calling. We'll align the potential variations to the reference and make a call on the variation. The output of that is a genotype file.

What's interesting about the assembly approach is that we're computationally very efficient. We take advantage of the reference genome to organize reads, but we're not bound to it for the variations we call. We're able to detect fairly large variations, much larger than the 1-2-base indels typically found by current mapping algorithms. Increasing our ability to detect larger and rare variants is a major R&D project for Complete Genomics. Ultimately we expect that our sequencing and software technologies will be able to find very long and complex variations.

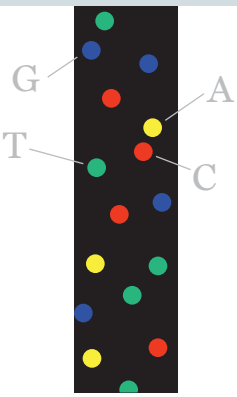
What is the minimum number of reads for you to be confident of a sequence?

As is typical for analysis algorithms, there is a trade-off between read length, error rate, coverage depth, bias and other factors. In our case, we typically require between 15X and 20X coverage per haplotype to get a very high quality genome. Sequence characteristics also can affect required coverage, and in many cases we do well with far less coverage.

What happens after assembly?

Post-assembly, we have a genotype file. We do a few additional steps to make it

(CONTINUED ON PAGE 39)



Taking Next-Generation Sequencing Data to the Cloud

BY KEVIN DAVIES

"IT here's a lot of interest in the cloud. In a sense, GenomeQuest has built the first commercial application-specific cloud for biocomputing." So said Ron Ranauro, GenomeQuest president and CEO, marking the July launch of GenomeQuest 6.0Beta, a sequence data management solution that provides a web-accessible, cloud computing environment for researchers to "align and mine" next-generation sequencing (NGS) data.

From the Internet, users can access the GenomeQuest cloud, namely "a 500-CPU compute farm for processing that's purpose-built for processing volumes of sequence data."

GenomeQuest is best known for mining sequence data for intellectual property. The strategy netted more than 100 customers, including most of the big pharmas, and several agricultural science customers. The strategy, Ranauro says, was always to create an enterprise sequence data management platform. But when would the market be ready? The arrival of the NGS era sparked what Ranauro calls the "catalytic event for causing the enterprise and academic markets to rethink the way they're managing sequence data."

GenomeQuest VP software Richard Resnick and CEO Ron Ranauro

EASY BUTTON

GenomeQuest 6.0Beta addresses the needs of three key constituencies—the researcher, the bioinformatician, and the IT manager. "The early visionary market for next-gen sequencing wants to do everything, but the mainstream market wants 'the easy button,'" says Ranauro. "They also want some flexibility to tune parameters. They're not interested in managing data but want common workflows." GenomeQuest delivers production workflows for gene expression and variant (SNP) discovery. "Any researcher can self register and upload a file, or use the sample file and start getting results very quickly." (See, "You want results...").

Bioinformaticians, on the other hand, typically access data models and algorithms through an application programming interface (API). "We've put a tremendous investment in exposing the [API] at multiple levels. Since it's a web application, there's a URL API used to script and access any data or workflow or database in the system. There's a scripted command line API which most bioinformatics developers will prefer."

As for the IT manager, scalability is critical. "The volumes of these next-gen machines just continue to escalate," says Ranauro. "A system that won't scale is going to be a difficult investment to justify."

Ranauro half-jokingly says GenomeQuest is becoming a web company. "Researchers can come to the site, self register, use a sample data set or upload their own, run workflows and mine the results." The available sample data include metagenomic pathogen data (454), variant detection workflows, and gene expression data (Illumina, Life Technologies).

GenomeQuest 6.0 fits into the next-gen workflow from the generation of the raw data, picking up up the raw FASTA files, post image processing, and uploading that file. A multigigabyte file might take several hours. For bigger files, the sneakernet will suffice. (Ranauro says GenomeQuest is open to leveraging data-transfer services from companies such as Aspera.)

From a simple web application, the end user selects a reference genome and level of annotation. Are novel variants found in dbSNP? Are they found in coding regions? The result file is a sequence database of the assembly, which can be mined according to those properties. "Being able to mine and filter the results is the secret sauce of the scalable engine," says Ranauro. "Now the biologists can do this work without needing to be a programmer, through a very simple web application. That's the contribution we're making—allowing a broader, mainstream audience to participate fully in next-generation sequencing."

Biologists can select and create custom views of the appropriate reference sequence or subsets thereof. "It's providing data management, but data isn't really moving around or up and down from the server to the PC. All the manipulation is happening in the cloud but the user is able to manipulate [it]." The web archi-

(CONTINUED ON PAGE 24)

You want results or not?!

After a stint with Eric Lander's group at the Whitehead Institute in the mid '90s, Richard Resnick tried his hand at industry, starting Mosaik (eventually acquired by NetGenics and LION), and a spell in the entertainment industry, before joining GenomeQuest. "GenomeQuest has been an IP software company, but when I was consulting for the founder in '02–'03, they had something I'd have died to have used as a bioinformatics director in '96: this incredible powerful platform to do massive sequence data management and comparison."

NGS creates a huge data management and analysis problem, but GenomeQuest addresses the next step: how do I make my new transcriptome the reference sequence? Or take 100 Illumina runs and identify variation across them? "Nobody knows how to do that. It's such a nascent market."

While Resnick feels GenomeQuest could prove useful to the genome centers. "They're doing a lot of janitorial work they don't need to be doing," he says, "but they're not our immediate commercial customers." Instead, he is targeting—successfully—core labs with 3, 5, 10 machines running regularly, supporting university groups or medical schools.

Resnick sympathizes with the researchers presented with a stack of FAST-Q reads. Without a bioinformatics friend or group, they're lost. "The core lab managers are struck with this problem—managing the sequence data and how do I give these results to our users? They're not the Broad or Wash U—they don't have that level of resources—but that makes them really hot customers for us."

Resnick just signed the University of Florida, the first 454 customer, which he is installing as an "on the premises" platform behind their firewall. Another strong source of interest is agricultural biotech. Interest from pharma and biotech is picking up slowly, with Resnick highlighting areas such as toxicology, oncology, metagenomics, and infectious diseases.

GenomeQuest's four workflows cover variant detection, RNASeq, rapid mapping/annotation, and assembly. There is demand for ChipSEQ and microRNA characterization, which will likely be added. The reference databases are continually updated. "Having the reference databases available, easily minable, and filterable is *sine qua non* for having accurate results," Resnick insists.

He selects human chromosome 7, and within seconds, he is filtering the list of high quality SNPs found from a sequencing run that affect a protein translation and are totally novel. The answer is 19, and the list—and the filtering workflow—are saved for future use. Some clients still use sneakernets to upload their data. "Trucks and planes are sometimes faster than the Internet," says Resnick. "Our transport speeds are increasing to line speed, but we'll ship you a disk if you want one."

GenomeQuest supports all the major platforms, including Helicos. Interested users can create a basic account and do sequence comparisons "in a massive way... in perpetuity," says Resnick. "You can take anything you've uploaded, and compare it to any database you want." The only part for which GenomeQuest charges is the high-end NGS workflows. "It's less than the cost of the reagents," says Resnick. He asks rhetorically: "You've invested thousands of dollars in a run. Do you want to get the results or not?"

GenomeQuest

(CONTINUED FROM PAGE 23)

tecture enables everything to be shared, including workflow, result databases, and selected views on results. "Those can be used as hypothesis drivers for the next set of experiments," he says.

UPSIDE AND ROADMAPS

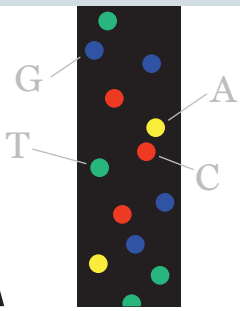
While Ranauro has his sights set on mainstream users, he sees upside elsewhere. "In the fullness of time, a genome center is going to want to get onto the cloud, because they have to lower their costs, just the same as anybody else, to get to the $1000 genome. It might be that GenomeQuest's platform provides a smoother path onto the cloud than taking all the in-house infrastructure and trying to recreate it on Amazon... We see ourselves providing the on-ramp to the cloud."

Ranauro adds: "We're actively looking at scaling options that might include Amazon. Hosting this on Amazon is a very real possibility, but it's not currently on our roadmap."

De novo assembly functionality is on the roadmap, however, for the second half of 2009. "We'll provide the computational and alignment engine but we'll rely on the industry for the assembly. There are important assemblers, such as 454 Newbler, today. For short reads, later this year—there we'll rely probably on Velvet or Abyss."

Ranauro also sees a rich environment for next-gen software companies such as CLC bio and DNASTAR to add value. "Those tools have a very rich feature set. There's always going to be a researcher that can benefit."

Ranauro says his product is alone in that it can process the data and then mine it using an easy-to-use web-based platform. "There's a reason why the IT industry went from client-server to web-based. It provides centralized management, local control, more of a tractable knowledge engineering environment for an enterprise. We don't see our customers wanting to move data up and down between PCs and servers or across networks. They really want to have it stored centrally but be able to manipulate it easily. We're really the only company offering that." •



A Single Man: Stephen Quake Q&A

Working on a single HeliScope instrument, Norma Neff, a research associate in the lab of Stanford professor Stephen Quake, generated the first single-molecule human genome sequence in just four runs in a month while Ph.D. student Dmitry Pushkarev handled the informatics. According to Quake, a.k.a. "patient zero" and co-founder of Helicos Biosciences, his group's success—published in Nature Biotechnology *in August—is proof of the growing democratization of genomics. "Literally three people did this work," says Helicos president Steve Lombardi. "That's a real harbinger of what we see the direction of this market going. It'll be very interesting to see what Francis Collins, in his officially appointed role [as NIH Director], does with that!"* ***Kevin Davies*** *asked Quake about his landmark personal genome publication.*

***Bio•IT World*: You talk a lot in the paper about the democratization of gene sequencing.**

Quake: There's a table in the supplement which indicates the effort that's been needed to sequence human genomes up until now. Our work is important at this time in that this is the first case [in which] you haven't needed a genome center to sequence a human genome. What we've shown is that you can do it with a pretty modest set of resources—a single professor's lab, one person doing the sequencing, one instrument, lower cost. Those are all order-of-magnitude improvements over what's been published recently.

That being said, the DNA sequencing industry is certainly competitive. Everything is moving fast, very much in flux. All the manufacturers are improving their platform by a factor of two per year. I'm just saying, at this point in time, Helicos is the best platform, and they're going to be in a dogfight to try to keep that title—which is good for the scientific community.

There's very little mention of Helicos in the paper. Did you deliberately keep this a separate effort?

Yeah, it's complicated. One of the reasons is the conflict-of-interest rules of my institutions... Stanford and the Howard Hughes Medical Institute. They have almost orthogonal, non-overlapping conflict-of-interest rules, very constraining. One upshot of that is I'm not allowed to collaborate with a company... It's much more strict on the Hughes side; it's like one of the Ten Commandments.

You didn't buy the HeliScope, you collaborated with Stanford faculty?

Exactly. The machine was purchased by the stem cell institute at Stanford. The purchasing process was very transparent... The reason they bought it was not to sequence my genome, but to sequence cancer, tumor stem cell genomes. That's what's up next. Mine was just to practice, to show that we could do it and to get the informatics into place.

The supplementary information put the price of your genome at $48,000. Can you elaborate?

Those were just the reagent costs. The amortized machine cost is about another $10-20,000.

Why didn't you name yourself in the paper as Patient Zero?

Well, you know, we wanted to retain some semblance of dignity for the scientific literature! It's really irrelevant for the purposes of the paper.

Did your grad student write the variant calling algorithm out of necessity?

Helicos wrote a mapper, but not a base caller. We used their mapper, which is tuned to the error profile of the instrument. All the mapping softwares have been written with particular instrument performance in mind. For example, MAQ and ELAND are written basically for the Illumina platform, where the dominant error is substitution. For Helicos, the dominant error is deletion, and that has consequences for how you do the algorithm. We used the Helicos mapper [IndexDP], but then, all the base callers are tied to the mapping software. So ELAND and MAQ will call the bases, but it's all linked into how they do the mapping. So we ended up writing our own base caller.

The genome coverage was 90%. Would you get higher coverage with more reads?

There are very repetitive parts of the genome that don't map well. Most people aren't mapping to the whole thing. The Chinese one was also 91-92%, something like that.

The paper notes the frequency of deletion errors...

Yeah, that's the primary source of error—deletions due to these 'dark bases.' One of the reasons this is an interesting result for the academic literature is: Is it possible to sequence the human genome with reads that are a little shorter and different dominant error mode than you have on other platforms? We show that it's definitely possible.

You've done some preliminary analysis of genetic conditions. Did you use Trait-o-matic?

That's right. George [Church] was very kind, and ran it through Trait-o-matic. That's where we got a preliminary annotation... We're preparing another paper on the annotation. In fact, my medical

(CONTINUED ON PAGE 39)

FEATURED SPEAKERS

Phil Butcher, Head of Systems, Sanger Centre

Antonello Covacci, M.D., Global Head, Systems Biology, Novartis Vaccines and Diagnostics

Chris Dagdigian, Founding Partner and Director of Technology, BioTeam, Inc.

Panos Deloukas, Ph.D., Senior Investigator, Human & Medical Genetics, The Wellcome Trust Sanger Institute

Jakob DeVlieg, Ph.D., Global Head, Molecular Design & Informatics, Schering-Plough

Carol Goble, Ph.D., Professor of Computer Science, University of Manchester; Principal Investigator, myGrid Project

Jan Korbel, Ph.D., Group Leader, Gene Expression, European Molecular Biology Laboratory

Jerry Lanfear, Ph.D., Head, Data Support and Management, Research CoEs, Pfizer

Hermann Lederer, Ph.D., Head, High Performance Application Group, Garching Computing Centre of the Max Planck Society; Member, DEISA Coordination Team

Hans Lehrach, Ph.D., Founder, Alacris Pharmaceuticals; Director, Molecular Genetics, Max Planck Institute; Professor in Biochemistry, Free University of Berlin

Carsten Möller, Ph.D., Lab Head, Therapeutic Research Group, Women's Healthcare, Bayer Schering Pharma AG

Sándor Szalma, Ph.D., Senior Research Fellow, R&D Informatics, Centocor R&D, Inc.

Matthew Trunnell, Manager, Research Computing, Broad Institute

Jörg Kurt Wegner, Ph.D., Scientist, Integrative Chem-/Bio-Informatics, Tibotec (J&J, Belgium); Blogger, Mining Drug Space; Project Administrator, Open Source Development

Hans Winkler, Ph.D., Senior Director, Global Head, Oncology Biomarkers, Ortho-Biotech Oncology R&D, Inc.

For a complete list of speakers and topics, and to register, visit www.bio-itworldexpoeurope.com

FEATURED TOPICS

High Performance Computing Trends

Scaling Up for the Data Deluge

Grid & Cloud Computing

Mapping DNA through Next Generation Sequencing

Comprehending DNA Regulation

Applying Genotype to Phenotype Correlations

IT & Business Process Information

Open and Collaborative Platforms

Data Modeling, Storage, Management and Analysis Tools

Leveraging Informatics to Drive Innovation & Productivity in Drug Discovery

Leveraging Databases and Data Mining to Improve the Drug Discovery Process

Enabling Translational Research and Biomarker Discovery through Informatics

Integrating Chemogenomics and Systems Biology Data

Knowledge Management for Pathway Analyses

Ontologies and Semantic Web

Pre-Conference Workshop*

Co-organized by CHI and BioTeam

Monday, 5 October • 14:00-17:30

Sequencing Data Storage

Matthew Trunnell, Manager, Research Computing, Broad Institute

Chris Dagdigian, Founding Partner and Director of Technology, BioTeam, Inc.

Guy Coates, Ph.D., Group Leader, Informatics System Group, The Wellcome Trust Sanger Institute

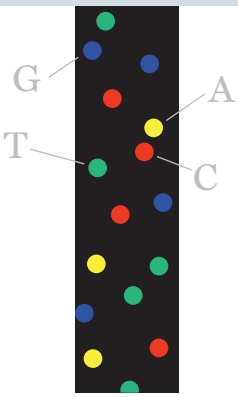
Carter George, Vice President of Products, Ocarina Networks

Additional Speakers to be Announced

BioTeam has been on the frontlines of next-generation sequencing integration, having helped several organizations with the unique next-gen IT, storage, and data management challenges. This workshop will present real-world customer experiences straight from the trenches. You'll get practical information about the storage support needed by research organizations in the next-gen world.

**Separate registration is required*

Please mention keycode U02 when registering!
+1-781-972-5400 or toll-free in the U.S. +1-888-999-6288. Register online at Bio-ITWorldExpoEurope.com



What Can Brown Do For Oxford Nanopore?

BY KEVIN DAVIES

Whether Clive Brown, vice president of development and informatics for Oxford Nanopore Technologies (ONT), is indeed "the most honest guy in all of next-gen sequencing," as one blogger recently dubbed him, is perhaps debatable. But as someone who has already stamped his mark on the sequencing world, his views certainly count for something.

Five years ago, Brown was the director of computational biology and IT for Solexa, helping to spearhead the British company's successful entry into the next-generation sequencing market, which spurred a $650-million acquisition by Illumina. After a spell at the Wellcome Trust Sanger Institute, Brown joined fellow Solexa alum, vice president research John Milton in moving to Oxford to commercialize nanopore sequencing. An intriguing subplot to the business of next-gen is whether Milton and Brown can catch lightning in a bottle again.

Oxford Nanopore is based around the pioneering nanopore research of Oxford University chemistry professor Hagan Bayley. CEO Gordon Sanghera remains close lipped about the firm's platform specs, but an elegant paper in *Nature Nanotechnology* earlier this year showed that ONT's nanopores can neatly discriminate between the four bases of DNA, based on the degree of current inhibition across the lipid bilayer (see, "Breathtaking Biology," *Bio•IT World*, Mar 2009). ONT received further validation by inking an $18-million marketing deal with Illumina.

Under the watchful eye of Oxford Nanopore's communications director, Zoe McDougall, Brown has to be more circumspect than is his true nature. "Things are on track—without telling you what the track is," he says helpfully. What he will say is that many of the key risks in ONT's technology have been addressed, and his team has built the entire informatics subsystem of the instrument.

Among the next priorities are to couple an exonuclease enzyme to the nanopore so that it can successively snip off bases from the end of a DNA strand, which will then tumble into the pore and the sequence read. ONT was recently granted a patent for its stable bilayer design. "This is a core element of our nanopore sensing system, not just for DNA sequencing," says McDougall. Milton calls these bilayers "the workhorse of our nanopore chemistry. We use the bilayer chip to focus on single nanopores and we also operate multiple-channel versions for higher throughput experiments."

WHAT THE TRACK?

Before ONT can produce an instrument, Brown has to become essentially a small genome center to test the product for months in house. Brown hired his former Sanger Institute colleague Roger Pettett to build up that infrastructure, as well as the software that goes on the instrument. It is "revolutionary new stuff, but we're reluctant to talk about that at the moment," says Brown, though he would say, "It does break the conventional instrument software paradigm."

Brown says the data throughput on ONT's sequencer will be high, "many tens of Megabytes/second." Not as high as some high-tech military applications, but "significantly higher than traditional lab equipment."

"Even before we were running chemistry," Brown says, "we made software that simulated data streams at launch spec rates. We designed interconnects and wiring, computer boards and live software that would process that data. We did it all in parallel. So when it came to plug-

Oxford Union: (L to R) John Milton, VP research; CEO Gordon Sanghera; and Clive Brown, VP development and informatics

ging a chip in, it all more or less worked." But Brown knows from experience that the system must be "very, very flexible to change."

Another priority is move the data processing close to the point of data generation. "We have already put a huge effort into not outputting raw data, but outputting optimized processed data instead." Brown has considered running some of the algorithms on GPUs, but worries about the power consumption demands. "The other option is to use FPGAs. They're good accelerators, very low power requirements, but a bugger to program and so not very agile." Brown says FPGAs might be used at the end of product development, but not before. "So far we haven't had any problems in terms of compute speed when dealing with our data, either at the instrument level or centralized datasets."

Brown says the data processing simulations have been instructive. "It's quite early, and we're not scared," he jokes. Meanwhile, Brown is quietly checking out potential software partners, which he hopes will deal with the quality scored DNA sequence output. In addition to genome centers and large-scale laboratories, ONT is also targeting the bench-top. "In order to have a bench-top sequencer," says Brown, "we have to provide pretty easy to use bioinformatics solutions. Otherwise, it's just not going to happen."

"One of the problems with all these existing sequencers is, even if you automate the sample prep and make the sequencer easy to use, you still end up with a file with a billion short reads in it. This is still beyond the capability of most non-bioinformatically trained postdocs to do anything sensible with." ONT aims to generate even more sequence with longer individual reads.

BENCH-SIDE MANNER

Brown's goal is to provide, for want of a better term, a "turnkey" bioinformatics solution sitting alongside the sequencer. Brown has met with several potential partners, including one unnamed company that demonstrated that its "software can deal with a whole human genome-type workflow in a day or 6 hours on a typical workstation." Brown says that looks quite promising.

He also plans to find a partner to liaise with user IT groups and "help us to smooth the early adoption of lots of our systems. I'm more worried about the bench-top side than the high-end side." Once ONT is fully launched, Illumina will have a large say in that part of the workflow.

Besides targeting the genome centers and the bench-top sequencer sitting next to a lab researcher, Brown thinks that service organizations such as Complete Genomics might prove another fertile market—in other words, "very large sequencing centers that are not traditional genome centers," focusing on medical sequencing applications. "I think Complete Genomics is a perfect customer for us. In fact I think our machine's better suited for what they want to do than theirs is!"

As Brown talks, it sounds as if ONT stands for "on track." Surely there are problems somewhere? "I don't want to oversell things, and remember we are still very stealthy as a company," he responds. As at Solexa, "things just aren't linear in a company like this. You have days when things work beautifully, and long dry periods when things aren't working. Half of it is just keeping your nerve."

Certainly Brown has assembled a strong team to build the IT/informatics infrastructure. Nava Whiteford, another Sanger Institute recruit, is adapting existing algorithms and developing a novel file format called Fast5 for scored sequences. Physicist Stuart Reid is driving data quality measurements and some of the basic science feeding into the platform. Lukasz Szajkowski joined from Illumina to manage the writing of the instrument software, which Brown calls "one of the most risky areas, but it's all on track thanks to him." Molecular modeler Mick Knaggs has implemented much of his software on GPU-enabled systems.

I don't suppose the Sanger Institute is too happy about some of their top people being poached. "Yeah, we did have a chat about my recruitment methods," says Brown honestly. •

More Online

For a full version of this interview, go to www.bio-itworld.com/NGS-brown.html

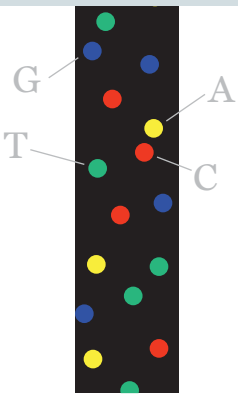
Quake Rumblings

Clive Brown was not alone in feeling the paper by Stanford's Stephen Quake in *Nature Biotechnology* and colleagues was "a little bizarre." While the first report of a single-molecule human genome was "perfectly worthy of a good *Nature* paper," with a "respectable throughput" of about 2 Gigabases/day, Brown was mystified by the "ridiculous back-door marketing" in the paper.

"Their own cost seems to be exactly what Illumina is citing for their service sequencing now [$48,000]," although the paper referenced outdated 2008 figures of $250,000 and up. Says Brown: "They're using the number of names on the Solexa paper [*Nature* 2008] as evidence of how many people are required to run an instrument! Well, that Solexa paper was the culmination of 8 years work encompassing the entire development of the platform, so it had everybody's name on it. The CEO's name was on there, and he didn't run any instruments." He adds that the original Helicos paper (*Science* 2008) had more than 20 authors for the sequencing of a tiny virus.

"They're setting themselves up for a tagline: 'Look, we only need three people to run a Helicos machine. And Illumina needs all these people and it's much more expensive.' If they'd just stuck to the high ground, i.e. they've got a working system that does single-molecule genomes, they'd be a lot better off...."

"I think Helicos deserves some kudos. They've stuck with it, they've had a rough time as a company, and they've made it work about as good as it can work with single-molecule fluorescence, with the cameras they have. People have taken the technology outside and they've used it successfully. And that's not trivial.... They should stick to the high ground – you can quote me on that."



David Dooling: Gangbusters at The Genome Center

David Dooling joined The Genome Center at Washington University at St Louis in 2001 from Exxon Mobil, where he'd been developing chemical reaction models. He started as a programmer, writing a lot of software, with no life science familiarity, and picked things up as he went along. He now oversees about half of the informatics group, including Laboratory Information Management Systems (LIMS); the Analysis Developers group, which creates an automated pipeline for the bioinformaticians; and the IT group—infrastructure, network computing, and storage. He's also the author of the "PolITiGenomics" blog. ***Kevin Davies*** *spoke to Dooling the same week as his group published the second cancer genome paper, in the* New England Journal of Medicine, *an important study that identified recurrent mutations in genes not previously associated with acute myeloid leukemia (AML).*

***Bio•IT World*: How has life changed at The Genome Center in the time you've been there?**

DOOLING: Well, things were good for a while, we had 10 Terabytes of disk, everything was great! Now we have about 3 Petabytes. When I started it was [AB] 3700s. Then we replaced that fleet with 3730s, Megabytes a day. Then 454 came along, Illumina, SOLiD. It's been gangbusters ever since.

What's the current platform setup at The Genome Center?

Right now, we have 454s and Illuminas. We don't have any SOLiDs any more... We'd purchased one, and [were using] a couple of others. We carried both platforms forward, but there's a significant expense with each of them, manual labor costs, library preparation, emulsion PCR, DNA input requirements, etc. In cancer research, you just don't have 3-5 micrograms of DNA. The Illumina has much lower DNA requirements, which we've driven even lower. Carrying on the informatics, lab pipelines, analysis pipelines, we carried both platforms forward but made a decision to concentrate on Illumina... Wrestling with two at a time is troublesome.

Wouldn't the SOLiD set up two-color space be advantageous for cancer genomics?

That's true. With the color-space correction, the reads are more accurate. I think the accuracy is marginal. The coverage values you need to be confident you're sampling both alleles is sufficient that the marginally higher error rate you see with Illumina is washed out in the consensus.

You just published a second cancer genome. Where does that fit in with your other projects?

We aim for about 300 genomes in the next eight months. It's the Washington University Cancer Genome Initiative—150 tumor-normal pairs. 300 genomes, 150 patients. About 1/3 will be AML, 1/3 lung cancer, and 1/3 breast cancer, with a few others probably. That's completely separate from the 1000 Genomes Project. We'll also be doing some glioblastomas and ovarians as part of The Cancer Genome Atlas (TCGA). In addition to just tumor-normal pairs, we have a breast cancer quartet where we have the tumor, normal, and a biopsy from a brain metastasis to see the difference between the primary tumor and the metastasis.

Can you describe the new data center?

We took possession in May 2008. We're now completing a second phase of construction. The building is over 16,000 square feet. The data center is about 3600 sq ft. About one fourth was outfitted with cooling and power... Less than a year later, we're getting the rest of that equipment installed so we can fully utilize the data center. At full capacity, it'll consume about 4 MegaWatts of power. It'll have capacity for around 100-110 racks of high-density computing equipment. Average 15 kiloWatts per rack, which is high. Current fully loaded blades are around there.

Are you working with any specific vendors?

For storage, we're using a software solution called PolyServe, developed by a company that was purchased by HP. We like it, compared to something like Isilon (see, "Isilon's Data Storage Odyssey," *Bio•IT World*, May 2009), because it's hardware agnostic. We can buy whatever servers, SAN switches, discs we want. If we decide to go away from it, we can still use those discs. It's a proprietary file system, so we'd have to move all the data off, but we'd have to do that anyway. It's a parallel file system on the back end that any number of heads can address. It has failover capability... We've had pretty good success with it.

On the hardware side, we've been purchasing HP storage, which has been the cheapest. We're using HP and Dell servers. Blades, pretty much all Dell. It's not like we throw stuff away! Over time, we go with whatever works best.

Did you consider commercial LIMS?

I manage about a dozen people in the LIMS group. The LIMS has been developed over a decade... We have evaluated [commercial systems] on several occasions, but not recently. Actually, we've talked to the folks at WIkiLIMS and Geospiza, but they're not really designed for our scale. We're topping tens of millions of transactions per month. We have tables in our database with billions and billions of lines.

You're an open source advocate. How does that relate to your role at TGC?
Why open source? It's just better software. Our entire system runs on Linux, Perl, PHP, Apache. We use Oracle but also MySQL and PostgreSQL. We have several thousand cores in our computer cluster and 250 desktop workstations that all run GNU/Linux, maintained by 1.5 system administrators. You're talking about thousands of systems that can be maintained by 1.5 FTEs. You can't get that with a Windows solution or a Mac solution. Granted these guys are highly skilled, but if there's a problem, they can dig into it. At the scale we operate, we're always breaking things. Whatever people bring in here, it breaks. We need to have the capability to tweak and to have the source code there and the communities that develop around free software. When we have problems, Google is our friend. 99 times out of 100, you'll find someone who had that problem. With the proprietary solutions, there's not a lot out there. They may not care about you.

Do you use any commercial software?
We've spoken with CLC bio, we were one of the first people to partner with the Synamatix search tool. We've worked with Novocraft. There's also Real Time Genomics, formerly SLIM Search (see, "The Quest to Make Sequence Sense," *Bio•IT World*, Nov '06). We've had that for a couple of years and are talking to them about their next-generation alignment and analysis tools. We look at them, but it's a tough nut to crack for those folks given the pace at which this field is changing.

Do you have a need for cloud computing?
Yes and no. We're interested in making our tools more useful to as many people as possible, releasing them open source. Part of that is making them useful in HPC environments, whether clouds, or Open Science Grid or BOINC (the engine behind SETI@Home). The one we're most aggressively pursuing is Open Science Grid (OSG), a federation of grids that provide end-users computing resources through a granting process. It's not like Amazon, where they charge by the hour...

The other side is that the utilization of our infrastructure goes through ebbs and flows. It'll be much more efficient to have a system that could overflow onto OSG in times of stress, rather than have things pile up or build a much larger infrastructure just to support the heaviest utilization periods. We're also talking to Sanger. In March, we had a Genome Informatics Alliance meeting. Amazon was there, Google, OSG, Microsoft. One of the action items was to work with those folks. Sanger took the lead with Amazon.

A specially constructed, fortress-like building on the campus of Washington University School of Medicine in St. Louis houses a computing cluster of more than 3,000 cores with 3-plus petabytes of disk storage.

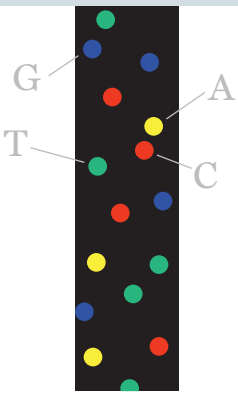
What are the bottlenecks you anticipate in the next 12 months?
I'd be lucky if I could pick the bottlenecks for the next 8 hours! Essentially, to get to where we are right now, we've created a very well balanced system. There isn't one aspect of the pipeline I'm concerned about—I'm concerned about them all in equal measure. Initially, you're getting the data and so you buy a lot of disc space. Then you buy more compute nodes, but you can't get the data to the compute nodes, so you upgrade your network. Now you're not efficiently using your CPUs, so you rewrite the algorithm in C, make the computation more efficient. Then you find the disc I/O is bad, so you need a more distributed system for higher throughput. We're getting sustained 10-15 Gigabits per second out of our disc system now. It's crazy! A year ago, you don't do that. So each time you dial one up, you have to dial the others up. Now they're all at 11. It's just a matter of keeping that stuff in balance, enhancing your monitoring'/ troubleshooting techniques.

For the current generation of sequencing technologies, we're on a good path. Everything scales really nicely. For PacBio etc. it's going to be 1-2 years for a production instrument to really gain a foothold. I'm very interested to work with any of these 3rd generation sequencers at very early stages and figure out what the problems are. They're going to have to deliver data in a very different way. You're never going to have the equivalent of images—it's just not possible at that scale. It's likely that's going to be much more information than you need, but you won't know what you need. What sort of systems will be in place? By the end of this year, you'll have dozens of whole-genome sequences. Where are the tools to do whole-genome vs. whole-genome comparison? Linking that up with phenotypic information? That's the other huge challenge.

Could you ever outsource sequencing to someone like Complete Genomics?
Sure, why not? By the time they hit that $5000 mark, other vendors will be hitting that mark. SOLiD said $30,000 for their genome. We're looking somewhere around what they're charging now per genome in the not too distant future ($20K range). That's a fully loaded cost—including instrument depreciation. •

More Online

For a full version of this interview, go to www.bio-itworld.com/NGS-dooling.html



CLC bio Satisfies Next-Gen Bioinformatics Cravings

BY KEVIN DAVIES

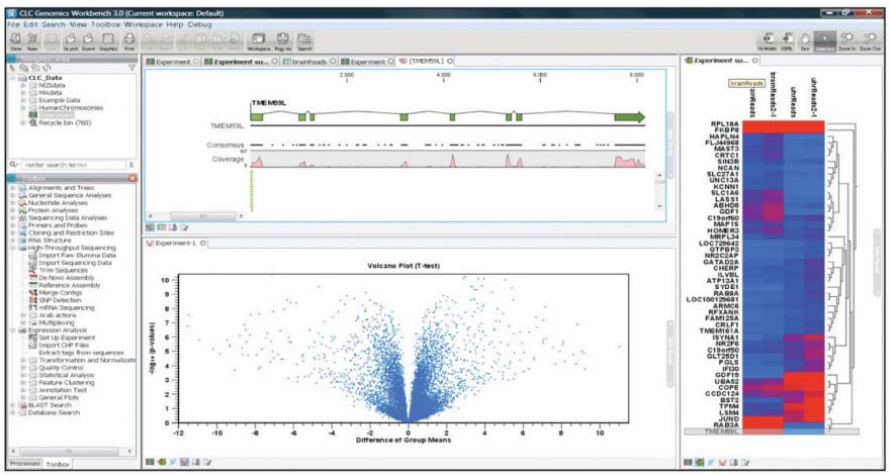
From humble beginnings in 2005, Danish software company CLC bio has emerged as one of the leading software providers for the exploding genomics and next-gen sequencing market. The Danes say they aim to be "among the most innovative bioinformatics companies in the 21st century."

After earning his Ph.D. from the University of Florida, Bjarne Knudsen saw a business opportunity to improve the quality, efficiency, and user-friendliness of life science software. He returned to Denmark, recruited his brother Thomas (now CLC bio's CEO) to the cause, hired a few local software developers, and CLC bio was born.

The origin of the name is a closely guarded secret. "'Cake Loving Company' is a good guess, but it's incorrect," says CLC bio North America CEO Jan Lomholdt.

CLC bio initially took a "Microsoft approach" to bioinformatics. "Everyone has to have it on their desktop, download it from the Internet. It was global thinking from the beginning," says Lomholdt. The resulting Workbench suite, going to head to head with the likes of DNASTAR, Vector NTI, and Gene Codes, proved competitive. It was platform independent and "people could just right-click and get [GenBank] directly into their program," says Lomholdt. By 2006, CLC bio had serviced 100,000 free downloads.

> **"We come from Denmark near LegoLand! We do bricks—build your own special plug-ins and increase the value of your program."**
>
> **Jan Lomholdt**



The Genomics Workbench 3.0 offers interactive and zoom-able analyses.

Adding more functionality, CLC bio attracted its first industry customers. With next-gen sequencing platforms emerging, Lomholdt and colleagues worked with the major vendors—Roche/454, Illumina, Applied Biosystems (AB), and Helicos—to ensure it could handle all types of sequence data.

Released in 2008 at the Bio-IT World Expo, the Genomics Workbench operates as a desktop application for next-gen sequencing analyses. Early this year, CLC bio added a three-tier server architecture with CLC Genomics Server, providing a server structure for CPU-intensive jobs, a database component, and the ability to use Genomics Workbench as a thick client and develop and execute customized plug-ins centrally.

Lomholdt says that the major next-gen platforms are "very good in hardware technology [but] they are not good in doing software—and they know it. Customers demand downstream analysis. When we said, 'Hey, we can do this,' they said, 'Finally, there is one we can point to when the customer gets mad!'"

Lomholdt admits that CLC bio was a little late in supporting AB's color space, but does now support SOLiD data. Moreover, the vendors recognize that many customers use multiple platforms. "We can be the one that merges their different datasets," says Lomholdt. "We can handle long or short reads and merge them to get a higher quality result."

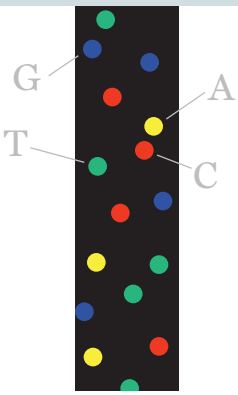
VENTER VALIDATION

A coup was attracting the J. Craig Venter Institute (JCVI). "He [Venter] found out what we were doing, and said this is exactly what I need," recalls Lomholdt. "We come from Denmark near LegoLand! We do bricks—build your own special plug-ins and increase the value of your program."

Granger Sutton, JCVI's senior director of informatics, said JCVI would implement the full enterprise platform as it integrated workflows "across different technologies and geographical sites."

JCVI incorporated CLC bio's accelerated versions of HMM algorithms into its pipeline for metagenomic annotations. According to JCVI director of bioinformatics software, Saul Kravitz, the single instruction, multiple data (SIMD)-accelerated tools increase analysis capacity and "take advantage of our annotation pipeline without any further hardware investments."

(CONTINUED ON PAGE 39)

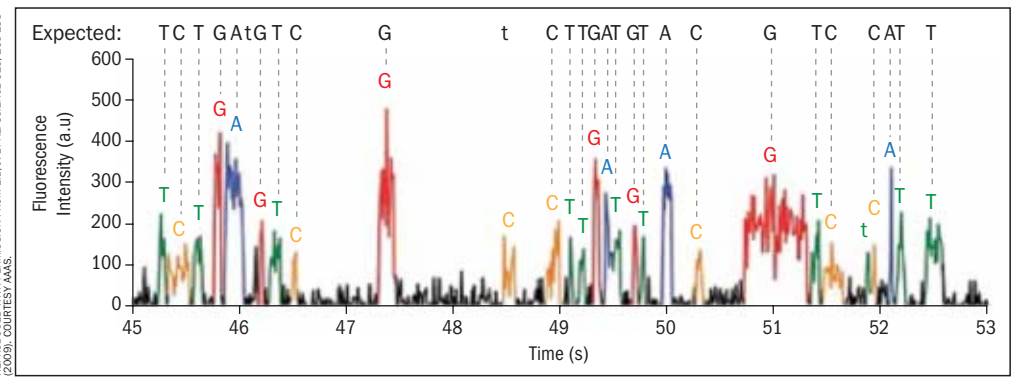


SMRT Software Braces for the Pacific Biosciences Tsunami

BY KEVIN DAVIES

Earlier this year, Pacific Biosciences founder and CTO Stephen Turner ran an animation of a real-time single-molecule sequence trace as a crawl at the foot of his slides for the duration of his talk, demonstrating not merely the impressive length of DNA reads the company could generate, but also its slightly hypnotic quality. "I do hope that some of you will watch the rest of the talk," Turner said.

The cute animation was devised by a member of Scott Helgesen's software group at PacBio, which is not surprising. A decade ago, Helgesen and Brad Carvey composed the opening dragonfly CGI sequence for *Men in Black*, before Helgesen traded New Mexico for New England and a job with 454 Life Sciences.



REPRODUCED WITH PERMISSION FROM EID, J. *ET AL. SCIENCE* 323, 133-138 (2009). COURTESY AAAS.

The PacBio software group has to interpret single-molecule, real-time sequencing traces (lower case bases are miscalls).

Helgesen is now part of the software team, headed by VP Kevin Corcoran, that PacBio is depending on to handle the data from PacBio's single-molecule sequencer next year. The third-generation sequencing system eavesdrops on a grid of DNA polymerase enzymes, tethered to the bottom of nanoscopic wells, as they synthesize DNA strands in real time. As each fluorescently tagged base is snared by the polymerase prior to being incorporated in the new DNA strand, its signal is detected (see "PacBio Sparks Florida Fireworks," *Bio•IT World*, Mar 2008). The method is dubbed single molecule, real time (SMRT).

The vital job of capturing that information and producing the informatics pipeline that converts those signals into pure sequence falls to Corcoran, who together with Helgesen, has ties with almost every competitor in the market. Corcoran formerly ran the sequencing business for Applied Biosystems (AB), and was involved in AB's due diligence of its own next-gen sequencing acquisition, Agencourt Personal Genomics. He was previously the CEO of Lynx, which merged with Solexa in 2004, two years before Illumina acquired the new entity. Helgesen, director of software engineering, primary analysis and simulation, spearheaded software development at 454 for several years, until leaving in 2006, just as the first Genome Sequencer was released.

REAL TIME ANALYSIS

PacBio has a fairly large instrumentation software group that writes hard-core firmware and builds real-time operating systems. Once acquired, the data are passed onto Helgesen's group, which handles the primary analysis—image processing, signal processing, base calling and quality value assignments. From there, the sequence data are subjected to secondary analysis, including consensus calls and assembly.

"The biggest challenge is real-time processing of the data," says Helgesen. "The data rates—the amount of data that comes flooding from the sensors—are really high compared to 454, much higher because we're looking at real-time events." Unlike at 454, where Helgesen used a CCD camera to integrate photons over time and under his control, "Now, we're not in charge of the events happening—the molecules just do their thing and we have to watch them."

One of Helgesen's passions at 454 was the use of FPGA (field programmable gate array) technology. He's circumspect on whether he sees a niche for the hybrid computing solution, but Corcoran says: "We'll either go down the FPGA route or some of these other alternatives. Graphics GPUs are becoming very affordable and more easily programmable."

For the prototype research instrument, which measures 3,000 DNA polymerase enzymes running in parallel, the software team has to capture the data in real time but doesn't need to process in real time. When the commercial instrument launches in 2010, however, "the spec for the production system for shipping is capture in real time and process in real time, to keep the throughput going," Helgesen says.

Helgesen says the processing throughput trade-off is the length of each DNA fragment and the numbers of parallel fragments going on simultaneously. The scheme is scalable, which is neat. "The big issue is the data are coming in so fast, you don't have time to store it to disc. You cannot capture the original raw signal data, so you have to process that—first-level

data reduction in real time. Even when I interviewed here and heard the number, I was like, 'Man...!'"

Handling that problem is a concerted IT strategy, involving computers with internal blades, data reduction strategies, algorithm optimization, and more.

READS AND ERRORS

Late last year, PacBio published examples of its first single-molecule sequencing results in a paper in *Science*. The single-read errors were on the order of 15-20%, but those data were generated almost 12 months ago. "The interesting thing about single-molecule sequencing is that errors are random verses systematic," says Corcoran. "In Sanger reads, the errors start to get worse the farther you go out. We don't see that phenomenon."

Another benefit of PacBio's approach is molecular consensus. By circularizing the DNA template into a so-called SMRTBell, the polymerase could figuratively "take a couple of laps around the circular molecule, [so] you get phenomenal consensus accuracy of one particular molecule." Turner reported individual reads of several thousand bases earlier this year.

Kevin Corcoran brings a wealth of experience from Lynx and Applied Biosystems to PacBio's software development team.

Corcoran says a priority is to drive up the raw accuracy rates as high as possible. The consensus sequencing mode would by definition reduce throughput, but provide additional fidelity when searching for rare mutations. "In Scott's pipeline, he has huge amounts of raw movies," says Corcoran. "He has to identify where all the pulses are, assign a base to that pulse in real time, and then, if it's a molecular consensus run, assign a consensus value to that particular read."

As for the sequence traces themselves, Corcoran says customers will have the option of saving them, "but they'll have to be saved off onto some system they provide. We'll stream them off the instrument in real time as we're processing." More likely, they will go down a level and save the base calls and associated confidence values.

I asked Helgesen how PacBio compared to the nostalgic days at 454? "It's definitely more challenging." At PacBio, Helgesen has a team that is "really strong on simulation, figuring out everything beforehand." "The best thing about Scott," adds Corcoran, "is that I was explaining what I was looking for, [and] he instantaneously knew what all my problems were!" •

Been There...

Kevin Corcoran was a software engineer at Applied Biosystems who became head of Genetic Analysis software group. In 1992, AB spun out Lynx (see, "Just Bead It," *Bio•IT World*, Feb 2004) to develop antisense therapeutics, but redirected efforts to develop short read sequencing based on technology developed by Nobel laureate Sydney Brenner. "It was the first massively parallel sequencing in a big way—we were doing 2 million events," says Corcoran. The MPSS (massively parallel sequencing system) produced 24-base reads, used mainly for transcriptome profiling. The technology had its challenges, but "as a service, it worked very well."

However, the technology was way ahead of its time. "You were talking to people and trying to explain the benefits of digital expression. Today, everybody gets it! The technology was ten years too early."

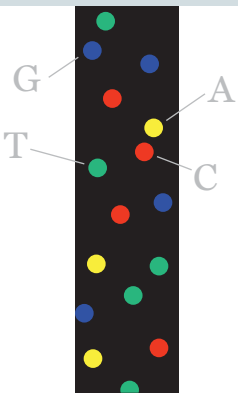
In 2003, Lynx and Solexa jointly bought the assets of a Swiss company called Manteia. With Lynx running out of money and Solexa in need of engineering expertise, they entered into a transatlantic reverse merger. "It made perfect sense," says Corcoran. "Since we both jointly owned the Manteia technology we both had guns pointed at each others' heads. They had cash; we were a public company." The newly public Solexa was then swallowed up by Illumina.

Corcoran opted to return to AB and run the sequencing business for a couple of years. One of his duties, along with Andy Watson, was to identify prospects for AB's next-gen platform. "AB had a big program, looked at a wide range of technologies. We settled on Agencourt Personal Genomics. We did due diligence on a lot of technologies."

After that, Corcoran took ten months off and "recuperated." But with several ex-colleagues reveling at PacBio, he inevitably got the call. Still, he admits to being "very curious about our friends at Oxford Nanopore," having gotten to know Clive Brown and John Milton during the Solexa merger.

Helgesen's interview at PacBio was far different than his job interview at 454, where all anybody wanted to know was how he came to create the special effects for the first two minutes of Men in Black! Joining 454, his first taste of biotechnology, Helgesen had no idea if building next-gen sequencing software was possible. "Now, after going through that experience, I'm used to that situation. I'm not as nervous about it. I'm a software engineer."

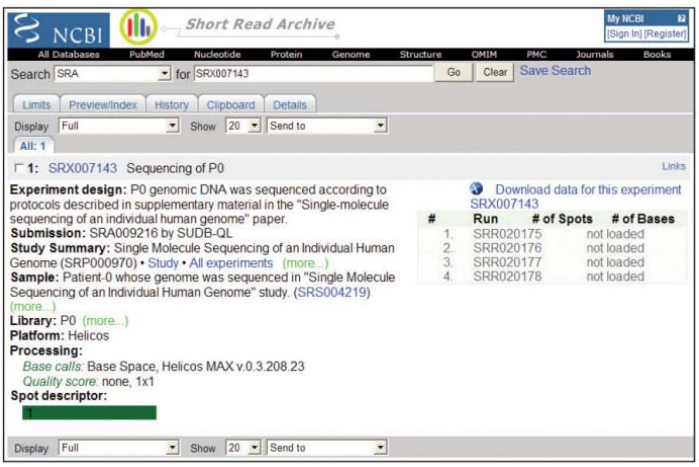
Before joining PacBio, he did talk to 454 founder Jonathan Rothberg about his latest venture, Ion Torrent Systems, but Rothberg couldn't seal the deal. "No way I want to move back to the East Coast," said Helgesen honestly. Now he gets to enjoy the California climate, and more importantly, as Corcoran points out, join more than "200 people who understand where you're going. Everybody has this idea of their responsibility."



NCBI's Sequence Read Archive: A Core Enabling Infrastructure

Two years ago, in Spring 2007, Cold Spring Harbor submitted the first DNA sequence data—Jim Watson's 454 sequence reads—to the Short Read Archive at the National Center for Biotechnology Information (NCBI). Since then, the SRA has become a critical component of the genomics community infrastructure, providing two-way access to enormous datasets, integrating with European and Japanese repositories, and storing sequence information on nearly 1,000 different organisms. As current and future technologies push the read-lengths closer to Sanger territory, the SRA has even undergone a name change: it now stands for "Sequence Read Archive."

***Kevin Davies** spoke to NCBI's Jim Ostell (chief, Information Engineering Branch) and staff scientist Martin Shumway about the past, present and future prospects of the SRA.*



The SRA allows for storage and rapid retrieval of reads with quality scores, metadata, secondary analyses, and intensity data.

***Bio•IT World*: How did the SRA come about?**

OSTELL: The SRA came out of the Trace Archive activity, which was originally started for the mouse genome [shotgun sequencing] project. People thought they'd need a place to combine reads from a number of different centers and have one collection that you could try with different assemblers. That very quickly grew in demand, expanded, and became international when the Sanger Center Trace Archive joined. It was recently moved to EBI [European Bioinformatics Institute], and DDBJ [Data Bank of Japan] created a resource as well. We started exchanging data. With the advent of short-read technology, it was clear that the database design of the trace repository would not scale to short reads, so NCBI architected the design basically to accommodate the bulk expected from short reads but also to learn from some of the history of the design of the trace archive.

For example, in the trace archive, a lot of the metadata about each experiment is associated with each trace. That could be an issue when there's a correction of the information, and you have to update thousands of traces. In the SRA, [we are] careful at what level of the hierarchy information was placed, so the least amount of information is touched during updates...

The SRA has a column-oriented design—that has several advantages. We can use the same architecture for people to submit data with and without certain elements, but it also means that we can take one of those columns and store it a different way. We're not expecting people to want the intensities very often, so we can store those on a low-cost tape farm, which is slower, and store the reads and the base calls on a disc drive. There's a meta-database which is orchestrating those different columns, so from the point of view of a user, it's all one thing.

Does "short read" mean shorter than traditional Sanger traces?

OSTELL: That's a good point. SRA started out meaning Short Read Archive, but now it means Sequence Read Archive. The reason is exactly the point you raised. It's a superior architecture to the trace repository. We'd expect next-gen technology to generate longer and longer reads but still producing massive amounts of data. Even as the reads get longer, we'll still store them in SRA.

What can you tell us about the IT Infrastructure?

SHUMWAY: We use SQL Server—that's the meta-database, if you will. That's keeping track of where all the pieces are. The pieces themselves are disk files that use a directory system. We use Panasus...

OSTELL: That's like a NetApp, lots of discs distributed on a network and a virtual file system. It was selected by our systems group, on the basis of price and performance. It seems to perform and scale well... We're also installing a tape farm, a tape jukebox, like a really slow disc drive, for the intensities. You get a virtual view spanning this room full of disc drives connected to a tape jukebox.

How does the SRA relate to GenBank?

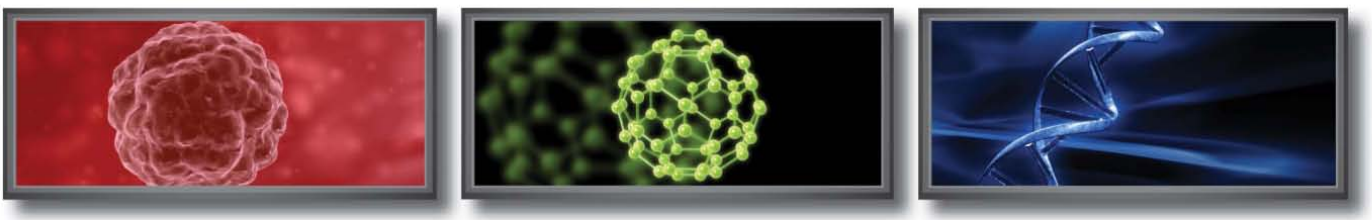
OSTELL: There are two relationships. One is a structural biological relationship—SRA has the reads and it will also have eventually the alignment of the reads to the reference genome. It doesn't have the assembled sequence—that goes to GenBank. Ultimately, structurally you'd like to have the assembly, the reads, and the alignment all connected together, even though they go to repositories with different repositories and different needs. The ID's will match them up.

(CONTINUED ON PAGE 38)

Cambridge Healthtech Institute's 17th International

Molecular Medicine Tri-Conference 2010

SHAPING FUTURE MEDICINE



DIAGNOSTICS CHANNEL
Molecular Diagnostics
Personalized Diagnostics
Cancer Molecular Markers

CHEMISTRY CHANNEL
Mastering Medicinal Chemistry

INFORMATICS CHANNEL
Adopting R&D Informatics Systems
Cancer Profiling and Pathways

BIOLOGICS CHANNEL
Stem Cells
RNA Interference
Cancer Biologics
Delivery of Biologics

CANCER CHANNEL
Cancer Molecular Markers
Cancer Profiling and Pathways
Cancer Biologics

EXECUTIVE CHANNEL
Translational Medicine

PLENARY KEYNOTES

When Drug Research is Personal
John F. Crowley, Founder, Novazyme Pharmaceuticals, Inc.

Chips, Clones and Living Beyond 100
Paul J.H. Schoemaker, Ph.D., M.B.A., Professor, Wharton School of Business

Corporate Sponsors

Applied Biosystems
Fate Therapeutics
gene
PreClinOmics adding value to research

Lead Sponsoring Publications

Bio-IT World
GEN Genetic Engineering & Biotechnology News
International DrugDiscovery
nature
Science AAAS
TheScientist

Co - Sponsors

The State of California

Conference Dates: February 3-5 | Exhibit Dates: February 3-4
Moscone North Convention Center | San Francisco, CA

Mention Key Code U02 when registering!

Tri-Conference.com

Organized by:
Cambridge Healthtech Institute • 250 First Avenue, Suite 300, Needham, MA 02494
Telephone: 781-972-5400 or Toll-Free in the U.S. 888-999-6288 • Fax: 781-972-5425

Ostell and Shumway

(CONTINUED FROM PAGE 36)

The political thing is that, since the TA and SRA began as NCBI projects and then got picked up by [Wellcome Trust] Sanger [Institute]... There's been consolidation on the European side. The trace archive moved from Sanger to EBI. In addition, DDBJ, the third partner in the GenBank partnership, said they'd have a short read archive. At our last meeting, we agreed to officially make the trace archive and SRA part of the GenBank-EMBL-DDBJ collaboration. We're beginning to integrate them together.

EBI has an SRA running—they've taken a copy of the software for our SRA and they are running it there after discussion of design and architecture. We're attempting to have a common codebase for that, which obviously simplifies a lot of things. Historically, that's not the case for GenBank—the formats are different and the databases are different.

How fast is data in the SRA growing?

SHUMWAY: We're currently at 8.5 Terabases (Tb) of biological sequence under management. We're growing by about 1 Tb/month. The major contributors are the 1000 Genomes Project, The Cancer Genome Atlas. We're also bringing up a controlled access instance of SRA within the dbGAP resource—to provide the same privacy protection to research participants as the NIH GWAS studies have. The human microbiome project and epigenetics projects are other major contributors. Right now, the growth [rate] is about linear at 1 Tb/month. We may see a growth curve later this year or next year, as the American stimulus funds translates into sequencing data.

The archive contains 85% human data, whereas the old trace archive is 15% human. There are still a lot of genomes in the SRA—about 800-900 single organisms. As for complete human genomes, we have the two Korean genomes, EBI received the Han Chinese genome. We've received three whole genomes from Illumina—a Yoruban trio. One of those [Yoruban] individuals, Hap Map sample NA18507 was also sequenced to deep coverage on the SOLiD platform. We also received a whole human genome sequencing dataset done on the Complete Genomics platform. Finally, we have been archiving the output of the 1000 Genomes Project, which consists of a number of deeply covered human genomes.

How are the various next-gen platforms represented in the SRA?

SHUMWAY: Illumina data currently occupies about 75% of the repository. The other platforms share 25%. We have good coverage from 454, a number of datasets from the SOLiD platform, and then we have one dataset each from Helicos—a yeast genome—and Complete Genomics.

How do users interact with SRA, such as accessing the Complete Genomics dataset?

SHUMWAY: There are those who download the complete dataset. It's a challenge, no doubt. We've adopted the Aspera technology as a replacement for FTP, and it's worked very well for submissions and otherwise. The data are huge, and most people seem interested in working with the data in the alignment form. If you're on Internet 2, it would take a few hours [to download a human genome dataset]. If you're not, it would probably take overnight.

Are you trying to drive the SRA agenda or adopt a more reactive posture to serve the community's needs?

OSTELL: Oh, we're very aware of that difference! We're trying to strike the appropriate balance between those two extremes. This is a public service—we're doing this for other people. But there are a number of constituencies. The mega projects have their deadlines and needs and ways to deliver the data. The small mid-scale projects, university sequencing centers, may not have an automated pipeline for providing data. We have people doing multiplexing lots of smaller samples, so their needs is more like a run-based archive, how do I give you a tenth of my run? Professor X is ready to give you his part but Professor Y isn't.

We work with the vendors who, of course, are anxious to have their platforms well represented. We try to make sure, for example, the archive is capable of representing AB color-space data, which obviously wouldn't be the highest priority for a different vendor like Illumina. [laughs] But I think the vendors appreciate our even handedness among them. At the same time, we're using our experience to judge what's likely to be important in the longer term. For example, a certain type of mega-user dominates right now, they download everything over Internet 2. But we recognize that as the data accumulate, other classes of user will have questions, e.g. 'I'd like all the short read sequences under this particular gene, I'm not interested in the entire genome.' That's a vertical slice through the archive across many submissions—that's the utility of working through an alignment. We think that's going to be needed and we're working towards providing that for the next wave of people interested in using the SRA.

Are you talking to future third-generation sequencing providers?

OSTELL: Oh we try and talk to them. Most of them know to talk to us, and we strongly encourage that as early as possible.

SHUMWAY: Each has their idiosyncracies, but in a very basic way they all do the same thing. We're gratified that our data modeling has held up as long as it has. For example, the PacBio platform actually looks like the old Sanger platform, in terms of data management—not technology.

So where does the SRA go from here?

OSTELL: We're very aware that sequencing will be the tool to be applied to any biological problem—it'll be cheap and fast and quantitative. We see this impacting many resources—gene expression is going from microarray to sequence based; genotyping is going from microarray to sequence-based. The flow of sequence data through NCBI, it's hard to think of that now as just being the sequence database, because it's part of GEO, dbGaP, all sorts of things. It's becoming an architecture where sequence flows in at various levels—the read, the alignment, the assembly, the genome... We see the SRA as a core enabling infrastructure spanning lots of projects and reaching across most of bioscience. •

More Online

For a full version of this interview, see www.bio-itworld.com/NGS-NCBI.html

Bruce Martin

(CONTINUED FROM PAGE 22)

easier to use. We're working on a variety of annotation steps, e.g. annotate variations by the dbSNP accession ID, functional annotation, and the like. Reports and annotation will make the data easier to use.

After annotation, there is a validation stage. Like most sequencing systems, we do QC at every step. We want to catch failures early. But there's a certain class of analysis you really can't do until you get the complete data set. Until you have all variations and aggregated the metrics from every stage, there's a certain class of issues you're really not going to detect. We do a computationally intensive final QC pass—the validation stage of the pipeline results. The goal is to look at all the metrics captured, the final result, anything else we know. The automated validation will allow us to increase our sequencing capacity to hundreds of genomes per day and maintain high quality. It is a lot cheaper and more efficient to throw computers at QC than to have people manually perform the task, and our scale allows us to make the software investments required to carry out this task in a fully automated manner.

What's the format on the hard drives?

Our results contain both read and called variations files. We haven't made a final selection on file formats and we are in discussions with customers, collaborators, the NCBI, and others to try to determine the best format. What should we support as our default? We may end up supporting more than one. The current format is a CGI-designed file format, but this is likely to change as we get more feedback. The short read data is encoded in a compressed binary file format designed to be small on the disk and easy to compute on. We designed it for use in our native pipeline. The variations file format is similarly one of our own design; and it has been used by some of our collaborators with good success. It's fairly straightforward: at a given position, we found this variation. Both file formats are designed to be easy to compute upon, because that's what we do every day at CGI.

There's an open question in the community about the right formats. There are as many choices as vendors and institutes. We certainly would like to see one standard emerge; I think it would be better for everybody, but the science and technology is a moving target. We may be able to contribute to solving that problem. •

More Online

For a full version of this interview, see www.bio-itworld.com/NGS-Martin.html

Stephen Quake

(CONTINUED FROM PAGE 25)

colleagues have gotten really interested in this. There's a small army doing a hand annotation for things that aren't covered in Trait-o-matic yet, like pharmacogenomics. That's going to be quite a lot of fun.

What other research uses do you foresee using the HeliScope?

We already have three more genomes in the can related to leukemia and cancer. We're neck deep trying to analyze those and understand what they mean.

After a tough 2008 for Helicos, this must be a very timely publication?

It's hard to say whether there will be any impact. It's kind of a David vs. Goliath battle. There are four commercial platforms out there right now, and three of them are billion-dollar companies. The fourth is Helicos, which is a scrappy little bunch—they're trying to hang on! I think they're fantastic, and I'm hoping they're going to end up at the top of the heap. •

More Online

For a full version of this interview, see: www.bio-itworld.com/news/08/10/09/stephen-quake-interview.html

CLC Bio

(CONTINUED FROM PAGE 32)

One reason for CLC bio's success is that it has consistently worked on algorithm optimization. Customers appreciate the improvements in RAM allocation allowing *de novo* assembly of a human genome on a single computer with 32 Gigabytes of RAM in 17 hours, as opposed to the massive RAM requirements by open source alternatives, says Lomholdt.

"We have an assembler that can challenge MAQ—higher coverage and speed, less RAM consumption," says Lomholdt. Last June, CLC bio unveiled CLC Genomics Machine, bundling next-gen sequencing software with the hardware. "IT is a very important component. The bioinformatician is important, the scientist is important. Now we have a solution for all three."

DANISH DELIGHT

Among other notable clients for CLC bio is Saudi Biosciences. The sequencing of the first Arab genome was outsourced to BGI Shenzen, which transmitted the raw sequence data to CLC bio for the assembly as well as some "special analysis" for the Arabs (Lomholdt declined to elaborate). Head of bioinformatics, Ruiqiang Li, said Genomics Workbench is "simply in a league of its own when it comes to flexibility."

Customers can obtain the software in many models—renting, leasing, owning, or site licenses. The Albert Einstein Epigenomics Center in New York is using the CLC portfolio for teaching, and the software used as an educational tool at Harvard and elsewhere. "One of our efforts is to help scientists get the Nobel Prize! Some say publications, I say Nobel Prize—why not?! To do that, you have to visualize the analysis."

The CLC Genomics Server won a Best of Show award at the 2009 Bio-IT World Expo. University of Pittsburgh's Michael Barmada called the CLC Genomics Server an ideal platform and said, "it's nice to see complex computational algorithms and routines presented in a user-friendly environment with a very elegant interface."

The company now has 50 staff, with bureaus in Singapore, Brazil, India, the U.K. "We think we're in a very sweet position between the vendors, the customers, the HPC, the analysis, and the algorithm development," says Lomholdt. •

Computational Development

IO Informatics' Working Solution

The software company has built a Personalized Medicine working group to tackle workflow bottlenecks.

BY ALLISON PROFFITT

"My view of the world is that if you have a tough question to answer, it takes a village of diverse expertise to really answer it," says Bruce McManus, head of the Prevention of Organ Failure Centre of Excellence (PROOF Centre) in Vancouver, Canada.

In this case, the village is the Informatics for Personalized Medicine working group, organized by IO Informatics.

"We're interested in one focus, which is how do informatics, analytics, and technologies like ours contribute to personalized medicine," says Robert Stanley, chair of the working group and IO Informatics' president and CEO (see, "Building a Google for Bioinformatics," *Bio•IT World* Jul 2007). The mandate of the working group is "developing practical applications of informatics technologies for personalized medicine." Stanley concedes that there is certainly a strong bias, "but we think it's a bias of general value."

IO Informatics started the working group in January, their second such group, as a combination of customers, prospects, and other colleagues. The 11-member group comprises "individuals representing the spectrum of translational medicine," said Stanley. There are representatives from early drug discovery, research hospitals, technology experts, systems biology experts, and three representatives from IO Informatics.

"To me the way that the group is set up is beneficial in that, as opposed to other working groups I've been a part of in the past, IO Informatics really listens to real world feedback. They're a small company and I don't see as many big egos in place." says Kathy Gibson, who after 13 years at Pfizer started her own consulting firm, Helio Consulting, and earlier this year joined the IO Informatics' advisory board. "The biggest part is not as much how they're organized as their operating philosophy,"

It's the philosophy that's at the core of the working group, says Bill Hayden, director, international business development at IO Informatics. "Knowledge integration is fine within a group, it's fine within a company, but if you really believe in that—which we do, obviously, and all the members do—integration of data throughout the whole spectrum should be better, should be the best, where you're pulling from bench to bedside and back again. When you integrate all that data, you really do get the full picture."

The topics up for discussion are brought by the working group members themselves. "Basically the working group offers an interface with clinician scientists who are creating large, complex, uneven datasets that need to be analyzed in a very rapid fire way, in a continuously refined way," explains McManus. "IO Informatics benefits from this partnership—an intellectual partnership—by seeing the various settings and all of the challenges we're faced with, and trying to extract information from the data."

The working group conversations have resulted in talks given at scientific conferences and could, Stanley hints, lead to publications in the future.

An area focus for the current group is bringing biomarkers to point of care and early toxicity detection. "If you're looking at different compounds, we can help the customer create profiles for different types of toxicity, and do that same kind of screening," explains Stanley. "Go out and get different data sources and pull data though and get automated alerts—you know, this experimental compound is starting to show some toxicity in those assays—and do that much earlier and cast a broader net for toxicity detection."

In the working group, McManus finds the discussions helpful. "IO Informatics comes at this in a very scholarly way, and they have some terrific people within that organization that are sort of intellectual relatives of our computational scientists within our own teams," he says. "Having access to the network of people who are either on the working group or who are connected to the working group and having the free flow of information from a variety of media... has been very beneficial to our team here." McManus has found the conversations to be a sort of "arms length validation of some of our own computational conclusions."

Gibson agrees. "Most of us are coming at this having been in health care or biomedical research for a number of years, and know very well the opportunities and frustrations of having tons of data out there... and not being able to make effective use of it." She says the problem is one she saw firsthand at Pfizer. "Lots of people have talked about this over this over the years, and to be frank not much has hap-

> "My view of the world is that if you have a tough question to answer, it takes a village of diverse expertise to really answer it."
>
> **Bruce McManus**, Prevention of Organ Failure Centre of Excellence

An Absorbing Proposal

Absorption Systems offers custom ADME-tox screening on demand.

BY REBECCA PALMER

It is no secret that a perceivable shift has been seen in the pharmaceutical industry. "Big Pharma" is not so big anymore, shrinking due to smaller pipelines, strong political pressures against rising health care costs, and increasingly difficult economic times. Pharmaceutical companies around the world are decreasing their discovery efforts and outsourcing more of that work to contract research organizations (CROs). Key partners are now needed to help get their compounds to investigational new drug (IND) submission with the FDA in a manner that is cheaper and faster.

Enter Absorption Systems. The Exton, Penn.-based company was founded over 10 years ago with the goal to "de-risk" compounds going forward in the approval process, according to CEO Patrick Dentinger. The group's approach eventually came to focus on ADME (absorption, distribution, metabolism, and excretion), pharmacokinetics, and toxicology.

The company now offers custom work for their clients with an emphasis on human-derived models, including Caco-2 cell line-based systems. Custom testing for absorption through skin, intestine, and nasal barriers with complementary pharmacokinetic studies in various animal models set this group apart in the industry. Although great strides have been made in these ADME and pharmacokinetics areas, toxicology still remains a "black box" for those in the biopharmaceutical market.

Tox Focus

To expand the breadth of their offerings, particularly in the toxicology arena, Absorption Systems acquired Perry Scientific in July 2009. Perry Scientific, the oldest and largest CRO in southern California, was founded 10 years ago to carry out preclinical studies for the pharmaceutical industry.

The partnership with Absorption Systems was obvious, but the new team's outlook is unique. "We are a biology company—it's what we do," says Dentinger, "but tox assays blend nicely with our current offerings and they enable us to keep an eye on that IND filing with the FDA. I think it is unrealistic to envision a 'one-stop-shop' for CROs, but this acquisition allows us to expand and emphasize our tox offerings and provides us with a strong presence on the west coast of the U.S."

To head this emphasis on toxicology is newly hired director of toxicology and pharmacology, Sarath Kanekal. Kanekal, most recently of Supergen (Dublin, Calif.), has experience taking several drugs from discovery to new drug application (NDA) with the FDA. As a former consultant for Perry Scientific, Kanekal is quite familiar with Perry's offerings and goals for the future.

"For our clients, we offer overall guidance on their IND programs and then recommend and carry out specific experiments on a wide range of experimental models. The end result is a recommended safe starting dose for the first human clinical trials. We see that the testing program moves forward efficiently, and we do it in a cost-effective manner. In the end, the client is ready to take the candidate directly into Phase I clinical trials."

With a new west coast presence and expanded toxicology offerings, Absorption Systems expects to grow by 50% in employee head count at the San Diego site within the next six to nine months. The combined resources of the new team are well placed to now help "Big Pharma" focus on improving pipelines. •

---

pened to improve mining large data stores and having it be intuitive to people. I think this working group is the best opportunity to whittle away at the problem."

ASKed and Answered

From IO Informatics' perspective, the working group has provided access to user questions and data. "It's a two way street," says Stanley. "Our products should be talking about their problems... [group members] talk methods, goals, and problems. We talk about solutions."

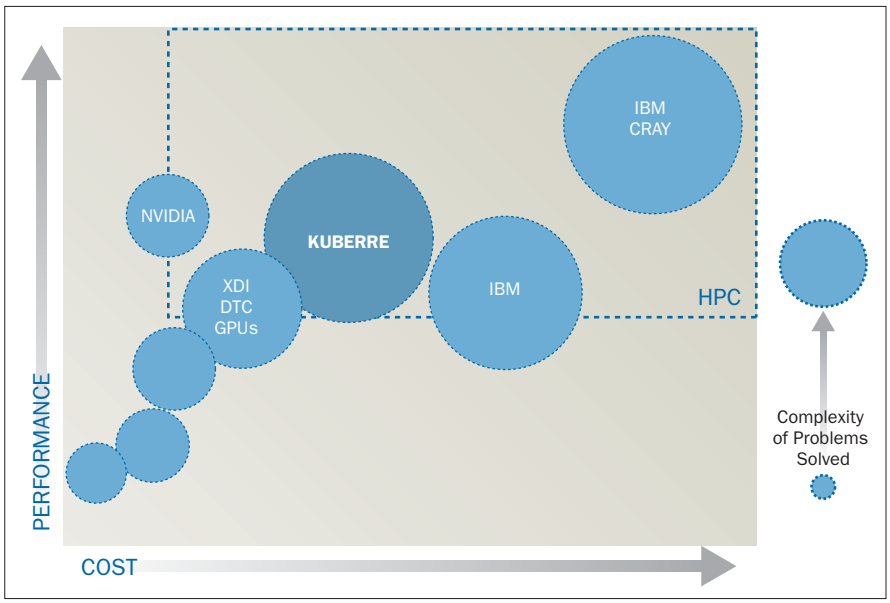
Those relationships were instrumental in shaping IO Informatics' new product, the Applied Semantic Knowledgebase, or ASK. "It's a different type of knowledgebase," says Stanley. "It's not the kind of knowledgebase that you normally think of that has all of the information in a therapeutic area, all the proteins, genes, and pathways. It's a knowledgebase containing the patterns that 'make a difference', the clusters of biomarkers that can be used for stratifying drugs according to their activity, or compounds according to their activity... It's a practical hypothesis base."

ASK is an enterprise product that works with other IO Informatics products to streamline toxicity profiling, target validation, patient stratification, and other tasks relying on a semantic database and arrays of SPARQL queries. It can automate screening and queries. Stanley said that the company had some ideas for a product like ASK, but time spent with the working group learning the users' challenges and problems refined the concept. "It's a learning relationship," he says, "we share a high level vision for predictive biology and systems biology to change the face of drug discovery and health care."

McManus echoes Stanley's analogy. "[Our] only goal eventually is to find out ways to segment patients to care for them better... IO Informatics and their interest in trying to solve these complex problems fits perfectly with the [needs of] the independent working group members." Besides, he continues, "It's a lot of fun. Very few people get a chance to think about the world as a system and the biology of health as a system and I'm getting the privilege of doing that, and this working group is one of the vehicles for that privilege. In the end knowing that it's very likely we're going to be able to help." •

Kuberre: Think Outside the Box

FPGAs move from financial services to life sciences.



BY KEVIN DAVIES

At about 3 cubic feet, the box sitting in a corner of Kumar Metlapalli's modest office in Andover, Mass., doesn't necessarily strike me as a "next-generation supercomputing platform." But the box, or HANSA, might be the biggest thing to hit high-performance computing in a long time.

The founder of Kuberre Systems, Metlapalli is an avid proponent of FPGAs (field programmable gate arrays), a chip that can be programmed to provide far greater specificity and efficiency than traditional CPUs, yet offer a more affordable solution than large grids with hybrid blades a super computers.

HANSA has a scalable architecture that can include from 4 to 64 FPGAs in a 9u cabinet ranging in price from $50,000 to $500,000. It combines a new hardware design and rich software stack for use in the HPC market with a memory that scales up to 256 GB. It delivers the equivalent of a 768-CPU server grid or a 1,536 core supercomputer, at 1/3 the cost, with 2% of the energy requirements, and 1% of the floor space.

While Kuberre has carved a niche in the financial services sector, cracking the life sciences market is both a top priority and a tough challenge. "We're getting there, but we need to build those relationships," Metlapalli admits.

Passage from India

An electrical engineer by training, Metlapalli moved to the United States from India in 1991, got his Masters degree in computer science from the University of South Carolina, specializing in image processing. He worked for XyVision before being recruited as a "quant" for Wall Street, predicting trends based on historical data.

The idea for HANSA, which means swan (think "Lufthansa") but also stands for Hardware Accelerator for Numerical Systems and Analysis, came in 2006, while Kuberre was providing a unified platform for financial markets. Kuberre's sister company in India had built an accelerator card for BLAST utilizing FPGAs. Metlapalli quickly targeted the benefits of FPGA technology to financial services, but recognized the danger of becoming a black box. Given the programming flexibility of FPGA platforms, Metlapalli opted to design a software stack on top of the FPGAs, so that users can write algorithms in their native languages (see, "Swan Structure"). "We can't be building singleton solutions," Metlapalli said. "The library must work across verticals and provide flexibility."

The HANSA architecture makes use of the scaLAPACK library, originally written for supercomputers or clusters, which breaks up matrices into submatrices, and lets them compute and combine.

One application Metlapalli is convinced will work on HANSA is GWAS (genome-wide association studies.) GWAS calculations are giant matrices with hundreds of thousands of values (SNPs). On HANSA, Metlapalli says "one doesn't have to take shortcuts. You have 256 GB memory to host the data you need, the compute power you need."

FPGAs have been around since the '70s, but seen little application in life sciences. One exception is Scott Helgesen (see p. 34), who featured them in the original software for the first 454 Life Sciences sequencer. The massive parallelism of FPGAs is finding particular use in military applications such as digital signal processing and Fourier transforms.

FPGAs are one increasingly popular flavor of hybrid computing, which complements CPUs with chips such as a GPU or FPGA. Metlapalli says he's flipped the role of the CPU, so that the CPU becomes the co-processor, and the majority of operations are performed on the FPGAs.

Unlike a CPU with millions of gates die cast, an FPGA has millions of gates controlled via software. "You're programming the chip to perform what you want to perform, in the most optimal fashion. I take FPGA, put a software code on top of it, and everything runs on the hardware." More logic in hardware translates to more acceleration. With FPGAs, "essentially you can transform one's personality based on the application you're solving." A binary search might take 1000 gates, but because the FPGA has 1 million gates, one

(CONTINUED ON PAGE 45)

Cambridge Healthtech Institute's Seventh International

Discovery on TARGET

Diverse Pathways • Multiple Targets • One Event

November 2-4, 2009
InterContinental Hotel, Boston, MA

BOSTON

Seventh Annual
RNAi for Screening Cellular Pathways and Targets

Third Annual
HDAC Inhibitors

Fourth Annual
Ion Channels as Therapeutic Targets

Fourth Annual
GPCR-Based Drug Discovery

Third Annual
RNAi for Developing Targeted Therapeutics

Third Annual
Kinase Inhibitors

Second Annual
Targeting Diabetes with Novel Therapeutics

SHORT COURSES:

Sunday, November 1

Strategies for Effective RNAi Screens

Targeting GPCRs and Ion Channels with Antibodies

Strategies for Optimizing RNAi Delivery

Combating Diabetes with Strategies for Enhanced Pancreatic Beta Cell Survival and Regeneration

Structure-Based Design of Ion Channels

Cardiovascular Safety in Drug Development - From Preclinical to Phase I

To customize your sponsorship or exhibit package, contact: Jon Stroup, Manager, Business Development • Tel: 781-972-5483 • Email: jstroup@healthtech.com

Make sure to mention keycode U02 when registering!

Image Courtesy of IRBM P.Angeletti
Image courtesy of QIAGEN

Cambridge Healthtech Institute • 250 First Avenue, Suite 300, Needham, MA 02494
Telephone: 781-972-5400 or Toll-Free in the U.S. 888-999-6288 • Fax: 781-972-5425

DiscoveryOnTarget.com

Out of the Gate

Mitrionics' hybrid computing speeds genomics apps.

BY RYAN DEBEASI

We use CPUs for general computing, and GPUs are for more than just graphics, but Mike Calise, executive VP of Mitrionics, is hoping that next-generation sequencing companies will add a third type of chip to their arsenal: the Field Programmable Gate Array, or FPGA.

An FPGA is a processor in which the logic gates can be rearranged to create a chip that's specialized for a particular application. Mitrionics sells FPGA-based servers along with software that makes them easier to program. Calise anticipates a $500 full genome sequence within the next five years, but in order to get there, sequencing companies will need to crunch masses more data for less money. This trend, he says, will drive adoption of his company's hybrid computing platform.

Calise defines a "hybrid computer" as "a heterogeneous mix of processors in a single system with working software... That could be a cell processor from IBM; that could be a GPU from Nvidia or others; that could be an FPGA; or a combination of those all in the future." Of crucial importance is the software. "A heterogeneous computer is simply a computer with multiple processor types, but a full hybrid computer is a heterogeneous computer with all the software worked out to get the maximum benefit for any algorithm that goes through it."

For example, to do more efficient calculations in programs such as BLAST the FPGA can be configured as a 2-bit processor, rather than a 64-bit processor. Each base can be represented in two bits, so an FPGA configured this way would process bases individually rather than in groups of 32, improving accuracy, and if running calculations in parallel, speed.

Jag Boleria, a senior analyst at the Linley Group, explains the importance of bit width with a multiplication problem. If you're multiplying two and two, he says, each of the twos can be represented in two bits: "10." In a 32-bit processor, however, those numbers will be represented as "10" preceded by 30 zeroes. That's a waste of 30 bits for each number. A two-bit processor, on the other hand, would only use two bits for each number. "Architecturally, that's much more efficient for the functionality than a general-purpose processor would be," says Boleria. According to a Mitrionics whitepaper, not all algorithms can be sped up by using the processors, and most FPGAs top out in the hundreds of megahertz, compared to 2 or 3 gigahertz in consumer-level CPUs.

Super Soft

Founded in 2001, Mitrionics differentiates itself by its software applications are written in Mitrion-C, not a chip design language, and are run in the company's Mitrion Virtual Processor software, which runs on top of the physical FPGA. This "Mitrion" platform differs from what companies such as Celoxica offer, says Boleria: Celoxica provides C-programmable FPGA hardware, but it can only be used for a specific set of financial calculations.

By contrast, the Mitrion platform is general-purpose. "It's a complete reconfigurable solution," says Calise. "Hardware stays constant and the MVP [is] what changes. It's all virtual changes so you can run a job on BLAST, you can run an SSearch, you can reprogram for HMMer, you can do all of this stuff flexibly with no hardware change and no being pigeonholed into a closed 'black box' system."

Mitrionics has about 30 customers, some of which bought the platform from SGI, which sold servers that included Mitrionics software. In April, SGI filed for bankruptcy and was purchased by Rackable for $25 million. "Mitrionics is working with a number of premier FPGA accelerated systems suppliers," says Calise. The company now sells Mitrion-based hybrid computing servers itself. "I'd say our sweet spot in deal size is between $20,000 and $200,000."

Currently, fewer than half of the company's customers are biotechnology companies. "The challenge we have," says Calise, "is that the people within [genome] centers are trying to figure out how to do more genomes; they're not necessarily the people who know how to take an accelerator and do something exciting to solve that problem with [it]." Potential customers in genomics centers will "wait until they have a fast-running application on an accelerated system. Then it's like, 'Great, whatever's under the hood [is] OK by me, but we won't tinker under the hood.' Once we see that, there'll be more tinkerers and it'll be a self-fulfilling growth. But it's not quite there yet."

To attract more customers in the sequencing space, Mitrionics has ported the BLAST-N application to the Mitrion-C language and made the program open source. Also in the pipeline are a Smith-Waterman program, Hidden Markov Models code, and generic DNA kernels for next-gen sequencing as well. Calise says that in tests using the Mitrion platform on SGI-supplied servers, BLAST-N ran 60 times faster than it would have on a non-accelerated server. Calise also notes that FPGAs allow users to boost performance without increasing power usage or size.

Calise says, "[FPGA] technology did not come to be standardized and cost effective until just about over a year ago." He expects units to become smaller and more energy-efficient, and predicts the number of logic gates in the processors will increase over time.

Mitrionics could face competition from pay-per-hour cloud computing services such as Amazon. While Calise sees the cloud as a competitor in the short term, he calls it a "phenomenal opportunity" in the long term.

"We actually don't care where... that FPGA hybrid computer, exists," he says. "Frankly, if it's in the cloud, that makes it more pervasive and more interesting, and potentially people could charge different subscription rates for different results, so all of that is exciting for us. Should we be talking to Amazon? Absolutely. Should we be talking to the large pharmas about the problems they'll see in pharmacogenomics? Absolutely. Do we care where that data center is? ... Nope. And if it's in the cloud and that generates sales, even better." •

Kuberre

(CONTINUED FROM PAGE 42)

can optimally dedicate a particular number of cores for the search, while performing secondary searches in parallel.

Get a Life

Metlapalli doesn't want to be constrained to a single industry. "To pick one vertical, we'd be doing a disfavor to the platform," he says. "I want pharma to know this solution exists."

One early prospect is an outsourcing vendor in India that works with most big pharmas. Kuberre is putting together a "business initiative document" under NDA. "They already have an idea of what they want to build," says Metlapalli, indicating molecular comparisons using tools such as JCHEM. "Think of drug discovery as a funnel—the narrower you make the funnel, the faster the process. That requires more sophisticated computation."

As for genome centers, Metlapalli says, "We strongly feel that the genome centers need a box like HANSA." Metlapalli says he's had encouraging discussions with Matthew Trunnell at the Broad Institute, but "the challenge has been allocating the research resources to look and see how the solutions will be built on this platform."

With HANSA providing the equivalent of 2.5 racks of nodes, 80 inches tall, at your desk, Metlapalli is convinced that HANSA's efficiencies can knock out clusters. The challenge is in "motivating these people and getting enough of their time to look at the box and build a solution on top."

Despite all the hype over cloud computing, Metlapalli says HANSA offers a cost-effective alternative by providing the compute processing at the point of collection. "Take it, collect it, process it... It has the computational power to bypass cloud computing."

"If you have a cluster or cloud, HANSA could be one of the nodes on that. If you need a departmental supercomputer, this is what you need." It provides the equivalent of 1500 processors or 700 blades.

For example, Metlapalli claims HANSA offers a 1000-fold improvement in the BLAST search algorithm. Based on work for a previous client on a single board with 4 FPGAs, Metlapalli saw an 80X performance boost. "We have 16 boards in HANSA. So it's 16x80, or 1280 or so." If you take out the latencies between the boards, maybe 1000X. But life science customers "really don't care" because BLAST makes up a small piece of their workflow. "They'd rather know how HANSA can solve their own workflow issues."

As a privately funded company, Kuberre runs "a very lean and mean operation," which is why Metlapalli is reluctant to build demo units. Instead, he challenges potentially interested researchers: "Give us a problem you're not able to solve. We'll do the legwork, build the prototype. Tell us that you're going to buy it! That's all we need. Just need an hour's worth of time, saying what the problem is, give us the sample data, this is how the algorithm should work. Boom! We'll do the rest." •

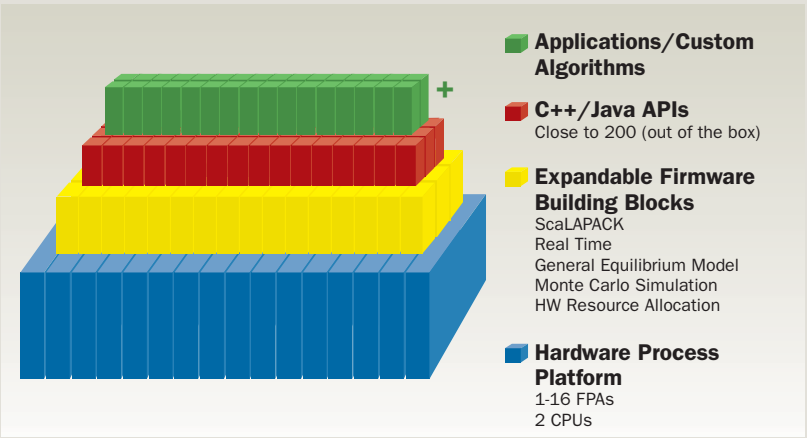
Swan Structure

The HANSA architecture consists of four layers. On the bottom is the physical hardware—16 boards, each containing four FPGAs. (Each FPGA has 12 processors talking to one memory bank, 12 talking to the other.) The next layer consists of expandable firmware building blocks (for example a binary search algorithm), so users do not have to deal with VHDL. Then comes a C/C++/JAVA API layer, so one of these APIs could be used in multiple building blocks underneath it to execute the programs. The icing on the cake is the user's own applications and custom algorithms.

"What we've done is provide level of flexibility they need to build their own algorithms in their own native languages," says Metlapalli. "No one has thought about building a supercomputer utilizing so many FPGAs together in a single box, or how to utilize with a software stack to solve problems."

Out of the 16 FPA boards in the box, five could be doing Monte Carlo simulations, six doing intense numerical algorithms. The other boards might capture streaming data. "That's what you can partition through the software. In one box, you're dividing the personalities of HANSA into sub personalities." One might be numerical algorithms, another might be pattern matching.

HANSA contains programming capability for C/C++, MatLab, R, and Java. "Imagine running 768 legacy C/C++ programs in parallel without having to make any changes to the legacy code, just do a recompile," says Metlapalli. Users might want a core library such as BLAST, Smith-Waterman, etc. "We don't want to build the entire conformation on the FPGA side. I want to provide a library so they can write their own algorithms." Kuberre provides the ScaLAPACK Library for use out of the box. "But if you want your own custom algorithms, we'll build those for you." K.D.



Educational Opportunities

This section provides a variety of educational events in the life sciences industry that will help you with your business and professional needs. To list an educational event (print & online), contact **marketing_chmg@chimediagroup.com**. To preview a more in-depth listing of educational offerings, please visit "Events" on **www.bio-itworld.com**

Featured Events

CHI Events

For more information on these conferences and other CHI events planned, please visit **www.healthtech.com**

Bio-IT World Conference & Expo EUROPE
October 6–8, 2009 • Hannover Germany

PEGS Europe
October 6–8, 2009 • Hannover Germany

Drug Repositioning Summit
October 13–14, 2009 • Boston, MA

Mechanism of Action
October 15–16, 2009 • Boston, MA

Clinical Training Forum
October 26–27, 2009 • Boston, MA

Immunogenicity Summit
October 26–29, 2009 • Philadelphia, PA

GPCR-based drug discovery
November 2–3, 2009 • Boston, MA

Strategic Resource Management
November 2–3, 2009 • Philadelphia, PA

Discovery on Target
November 2–4, 2009 • Boston, MA

Kinase Inhibitors
November 3–4, 2009 • Boston, MA

Portfolio Management
November 3–4, 2009 • Philadelphia, PA

Post-Approval Drug Safety Strategies
November 4–6, 2009 • Philadelphia, PA

The Science of Biobanking
November 16–17, 2009 • Philadelphia, PA

Advances in Gene Expression Profiling
November 17–18, 2009 • Philadelphia, PA

PEPTalk 2010
January 11–15, 2010 • Coronado, CA

Barnett Educational Services

Visit **www.barnettinternational.com** for detailed information on Barnett live seminars, interactive web seminars, on-site training programs, customized eLearning development services, and publications.

Live Barnett Seminars

Clinical Trials for Pharmaceuticals: Design and Development
October 20–21, 2009 • San Diego, CA

Drug Development & FDA Regulations
October 20–21, 2009 • San Diego, CA

Fraud in Clinical Research
October 20–21, 2009 • San Diego, CA

Working with CROs
October 20–21, 2009 • San Diego, CA

Patient Registry Programs
October 21–22, 2009 • San Francisco, CA

Managing and Conducting Global Clinical Trials
October 26–27, 2009 • San Francisco, CA

Drug Approval Process
October 27–28, 2009 • Philadelphia, PA

Pharmacokinetics
October 29–30, 2009 • Philadelphia, PA

Data Management in the Electronic Data Capture Arena
October 29–30, 2009 • Philadelphia, PA

IRBs: The Changing Landscape & the Effect on Conduct of Clinical Research
November 10–11, 2009 • San Diego, CA

Medical Device Postmarketing Vigilance Reporting
November 12, 2009 • San Diego, CA

Adverse Events Managing and Reporting for Pharmaceuticals
November 19–20, 2009 • Philadelphia, PA

Conducting Clinical Trials in Developing Regions
December 1–2, 2009 • Boston, MA

Teambuilding for the Cross Functional Global Team
December 3–4, 2009 • Philadelphia, PA

GMP for Pharmaceuticals
December 8–9, 2009 • Philadelphia, PA

Patient Recruitment and Retention
December 10–11, 2009 • Philadelphia, PA

Mastering Cost Management for Global Clinical Trials
December 10–11, 2009 • San Diego, CA

Interactive Web Seminars

Examining the Impact of the eCTD on the Regulatory Submissions Process
October 7, 2009

The Pharmaceutical and Medical Device Industries Today
October 12, 2009

Operational Modeling and Simulation in Clinical Trials
October 13, 2009

Sponsors/CROs Preparing Clinical Research Sites for FDA Inspections
October 13, 2009

Study Feasibility: Eliminating Low and Late Enrollment
October 14, 2009

Cost Reduction Without Increasing Regulatory or Business Risk
October 19, 2009

21 CFR Part 11 Compliance for Electronic Records and Signatures
October 23, 2009

New Developments in FDA's Human Subject Protection (HSP)/Bioresearch Monitoring (BIMO) Initiative
November 5, 2009

Critical Decision Points in Design & Conduct of Patient Registries
December 3, 2009

Biomarkers of Drug Efficacy and Safety: Fundamentals and Applications in Preclinical and Clinical Development
December 4, 2009

Examine Approaches for Investigator Initiated Trials
December 9, 2009

White Papers, Webcasts, and Podcasts

Keep abreast of current industry trends and developments. Browse our extensive list of free Life Science white papers, webcasts, and podcasts on **www.bio-itworld.com**. Interested in learning about developing a multi-media solution to generate leads? Please email **marketing_chmg@chimediagroup.com**

White Papers

Chromatography Software Provides Easy Access to MS Detection

Sponsored by Waters

Waters® Empower™ 2 Software is an acquisition control & data processing portal for single/tandem quadrupole systems. Comparable data handling for MS & UV, detect broader range of compounds & co-eluting peaks.

Biomarkers: An Indispensible Addition to the Drug Development Toolkit

Examining the Potential of Biomarkers
Sponsored by Thomson Reuters

Biomarkers are becoming an essential part of clinical development. In this white paper, Thomson Reuters provides insight from experts in industry and academia, and explores the role of biomarkers as evaluative tools in improving clinical research and the challenges this presents. Discover the potential of biomarkers to improve decision making, accelerate drug development, and reduce development costs.

Scientific Data Lifecycle Management: Preparing for Storage in an Uncertain Future

Sponsored by BlueArc

Managing vast and overwhelming streams of gene sequencing data today requires ultra-high performance systems and processes. With continued rapid advancement and improvements in gene sequencing, expect tomorrow's instruments to output quantities of genomic information that will dwarf current levels. Help your organization maintain data control and prepare for the future of sequencing through this informative paper.

Podcasts

Web-Based Dashboard for Analyzing Chromatography Performance Data

Sponsored by Waters

Analyzing chromatography performance data can provide significant insight into the operational efficiency of analytical laboratories. However, consolidating information into a format that is easy to view and interpret can be time consuming. Waters® Empower 2 Business Intelligence Manager™ is a web-based dashboard comprised of pre-built analysis modules that provide intuitive, interactive, intelligence for laboratory managers and system administrators. Learn about key features and applications of this new option for the Empower™ 2 Chromatography Software.

A Renaissance in Clinical Trials: Connecting the Functional Dots from Source to Submission

Sponsored by MaxisIT

The Pharmaceutical and Life Sciences industry suffers from a fragmented software landscape that lacks a single, holistic solution that would provide the requisite end-to-end functionality in an integrated and automated fashion. This podcast addresses this problem and proposes a preferred solution that is web-based and integrated. The solution offers streamlined processes, reusable infrastructure and requires minimum user-intervention to produce timely deliverables ranging from protocol to the eCTD – anytime, anywhere, on-demand.

Webcasts

Adobe Web Conferencing Case Study: MedPoint Communications

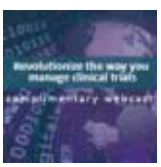
Sponsored by Adobe

With the potential for approved pharmaceuticals to earn upwards of a million dollars a day, there is tremendous pressure to infuse new products into the sales pipeline and recoup drug research and development costs.

To help pharmaceutical companies educate medical professionals about new drug treatments, MedPoint, a leading provider of specialized education services to biopharmaceutical and global healthcare industries, sought out a reliable web conferencing solution that would give participants easy access to meetings and support rich presentations filled with video, live images, and audio.

An Inside Look at Best Practices in Clinical Trials Design

Sponsored by Tessella



Adaptive designs are an exciting development that allows drug companies to better select doses, potentially bring novel therapies to the market faster, and provide real savings. While currently growing in popularity, Bayesian adaptive designs are still at the early adoption stage and it is believed this situation will change radically and quickly when the first adaptive trial blockbuster emerges. This will result in many more adaptive trials across the industry, and ultimately, significant patient benefit.

New Products

Microfluidic Perfusion for Live Cell Imaging

CellASIC has introduced the ONIX Microfluidic Perfusion System, delivering a better method to culture cells on the microscope stage and control real-time media perfusion during live cell timelapse experiments. The instrument's easy to use format and performance allows any user to run live cell imaging experiments with confidence. Cutting-edge microfluidics enable precise changeover between media solutions, parallel experiments for cost-efficient data collection and high quality CO_2/temperature control. Cell migration and chemotaxis studies are also made possible by a proprietary chemogradient. The ONIX adds on to existing microscopes to deliver a powerful method for environment control during live cell microscopy.

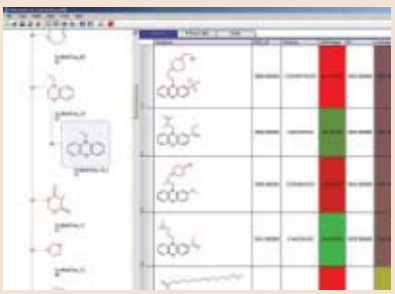
Company: CellASIC
Product: ONIX Microfluidic Perfusion System
Available: Now
For more information: www.cellasic.com

Merging & Sharing Data

Tripos has released a new software solution, Benchware Pantheon, that enables discovery scientists to rapidly and accurately merge data from multiple sources, and then analyze and visualize that data for faster and better decision making. Researchers can integrate and translate various spreadsheets, merge biological data with chemical structures, and then use a range of tools to sort, analyze, and visualize data in 2D and 3D. For instance, a researcher could visualize the various R-groups that have been synthesized on a particular scaffold and e-mail a colleague a set of interesting protein-ligand structures and comments on them.

Company: Tripos
Product: Pantheon
Available: Now
For more information: www.tripos.com

Building Compound Scaffold Trees



Altoris has introduced SARvision Plus v2.8, a desktop application to visualize, mine, and organize chemical data. The software automatically identifies chemotypes in a chemical library and organizes data in a compound scaffold. SARvision Plus comes with capabilities to filter data by scaffold type and by any other associated data such as HTS results, docking scores, or physicochemical data, and it identifies the chemotypes that satisfy user-selected criteria. Multiple tools including powerful graphics facilitate the task of identifying relevant scaffolds and compounds.

Company: Altoris
Product: SARvision Plus v2.8
Available: Now
For more information: www.chemapps.com

Next-Gen Cancer Biochip

febit has announced the launch of the first human cancer biochip for HybSelect, febit's highly automated technology for sequence capture, enabling targeted Next-Generation Sequencing. The new catalogue cancer biochip features 115 important genes which are reported to be associated with common types of cancer by the Wellcome Trust Sanger Institute. A 2Mb exon cancer biochip will be available this summer and a 30Mb biochip is planned already for release in 2010.

Company: febit
Product: HybSelect cancer biochip
Available: Now
For more information: www.febit.com

Precision Microarray

The NimbleGen MS 200 Microarray Scanner for high resolution (down to 2 micron) scanning with a 48-slide autoloader and advanced automation capabilities helps ensure high density arrays are analyzed with utmost precision and accuracy giving consistent and robust data. The NimbleGen MS 200 is a state-of-the art DNA Microarray scanner optimized to provide excellent performance when used with NimbleGen arrays. With a completely isolated and ozone-protected slide magazine in addition to high-quality PMT detectors, the NimbleGen MS 200 provides the enhanced performance required to extract valuable data from even your most demanding microarray experiments.

Company: Roche NimbleGen
Product: NimbleGen MS 200
Available: Now
For more information: www.nimblegen.com

The Russell Transcript

Certara's Translational Vision

JOHN RUSSELL

Modeling software and services pioneer Pharsight is about to launch PKS (Pharsight Knowledgebase Server) online, a hosted version of its enterprise database and collaboration platform. It is an early step in a journey to build a 'translational science company" around the nucleus of Pharsight and Tripos, both acquired by Vector Capital and placed under a new umbrella brand, Certara. The move to increase software as a service (SaaS) offerings is part of Certara's long-range strategy, says Dan Weiner, a Pharsight veteran who now serves as CTO for both companies. The move follows successful efforts by tool makers Phase Forward and SAS to do the same. PKS online is aimed largely at smaller pharmas and biotechs with tight budgets and fewer internal IT resources, but cash-squeezed larger companies may also find the offering attractive.

"[SaaS] is a growing trend and we are evaluating the prioritization of deploying other products in a hosted way. We also view PKS hosting as a component of that larger translational science solution and the mechanism for capturing and sharing clinical pharmacology information," says Weiner.

Pharsight, like Tripos, has for years produced drug discovery and development tools (see "A Virtual Pharmacopeia," *Bio•IT World*, Nov 2002) but encountered the same market-limiting forces bedeviling most informatics companies. The rise of translational approaches—which emphasize breaking down silos to provide synergistic collaboration throughout the R&D process—may create the opportunity to band together distinct tool makers into firms with broader offerings.

Tripos and Pharsight bring strengths in cheminformatics and statistical analysis for preclinical data and clinical activities that involve modeling and simulation. Weiner says there are still pieces missing, such as bioinformatics, safety assessment, and metabolic prediction. "Our goal is to grow organically into some of these spaces, make some acquisitions, but we also envision some significant partnerships to provide expertise."

Mark Hovde, SVP marketing for Certara, says, "Stage one of the branding is just to introduce a subordinated brand with Tripos and Pharsight still being the primary brands... Over three years, we're going to move so the Certara brand will be much more prominent. Pharsight and Tripos will probably never fully go away but eventually Certara will take on meaning as we introduce more products into the translational space."

The company says most of the "top 30" have purchased PKS and lists Wyeth, Roche, Sanofi Aventis, Schering-Plough, and Centocor among its clients. The FDA's Office of Clinical Pharmacology uses PKS to support modeling of QT safety data but is looking to expand that into more disease modeling.

Online Advantage

PKS online has essentially the same features as the internally deployed version, but as a hosted service. Pharsight has partnered with a server farm to deliver PKS. By stripping out the overhead costs, which for 5-to-10 users can be 4X the license fees, Pharsight has substantially cut the PKS cost of ownership.

"Some companies of don't have Oracle so there's that expense (license for Oracle-based PKS). Even if they have Oracle, there are costs for a database administrator, the installation cost, and validation costs which can be really significant depending on their own SOPs. It's all these costs associated with support and validation that we can largely, though not completely, eliminate," Weiner says.

There are other advantages as well. Deployment is faster, as it's a web-based service. It's portable, meaning users can work from wherever there is adequate internet service. Sharing access to the data with partners is also easier. PKS product manager Peter Schaefer notes that if you strip off the overhead, the cost per user is roughly the same for each version. For PKS online, he says break-even is probably around 5-to-10 users.

While the translational science umbrella presents an opportunity to knit diverse tools together, Weiner still expects a distributed data environment to exist: "No company, no IT group, wants to have one massive database. We want to keep the databases federated but be able to do queries that would extract only the relevant information from each database and compile that in such a way that they can use it for models if they are doing analysis or for reporting purposes."

Another product, D360, developed by Tripos, has many of those hallmarks. Weiner calls it "a dash-boarding product that has very powerful querying capabilities and we would view PKS online as being one of the sources that D360 would go out and query. Then you might want to match that information with what were the results of some early safety assessments on the same molecule that's coming from different sources."

The vision is to build a fuller translational science solution. Integrating the offerings from both companies is an early step. So is sprucing up the respective product lines. In June, Pharsight rebranded its desktop suite under the Phoenix name with the launch of Phoenix WinNonLin. A new population PK/PD offering, Phoenix NLME, is expected soon. And Hovde also promises something new in molecular modeling, possibly addressing the needs of the medicinal chemist.

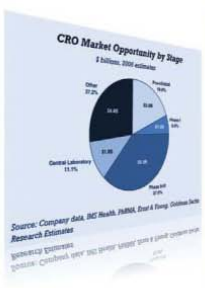
PAREXEL's Bio/Pharmaceutical R&D Statistical Sourcebook 2009/2010

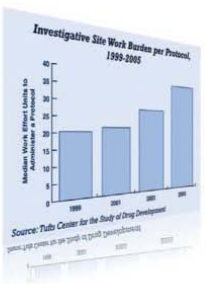
*"PAREXEL's Bio/Pharmaceutical R&D Statistical Sourcebook is **one of the most important** sources for accurate industry and regulatory data, and for some tables, it is the only source."*

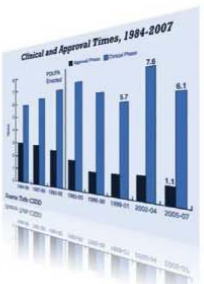
— Bert Spilker, President
Bert Spilker & Associates, LLC

PAREXEL's Bio/Pharmaceutical R&D Statistical Sourcebook 2009/2010 is the leading resource for statistics, trends, and proprietary market intelligence and analyses on the biopharmaceutical industry. Supported by thousands of graphs, illustrations, and analysis, the Sourcebook provides the latest intelligence on every aspect of biopharmaceutical development – from product discovery, to R&D performance and productivity, to time-to-market trends. With real-world analysis and key contributions from leading consultancies and experts, the Sourcebook includes:

- New proprietary analysis on US clinical trial starts, segmented by therapeutic category, as well as overall active clinical trials.
- Emerging data on worldwide and company-specific R&D pipelines and product launch trends.
- An all-new and comprehensive analysis of clinical research off-shoring revealing which pharma companies are now locating their new clinical trials overseas.
- New analysis on emerging trends in pharma and biotech licensing deals and other partnerships critical to industry's efforts.
- Drug approval statistics compiled from FDA, EMEA, and other regulatory agencies.
- New global R&D spending trends and other international R&D data from key markets.
- And much more!







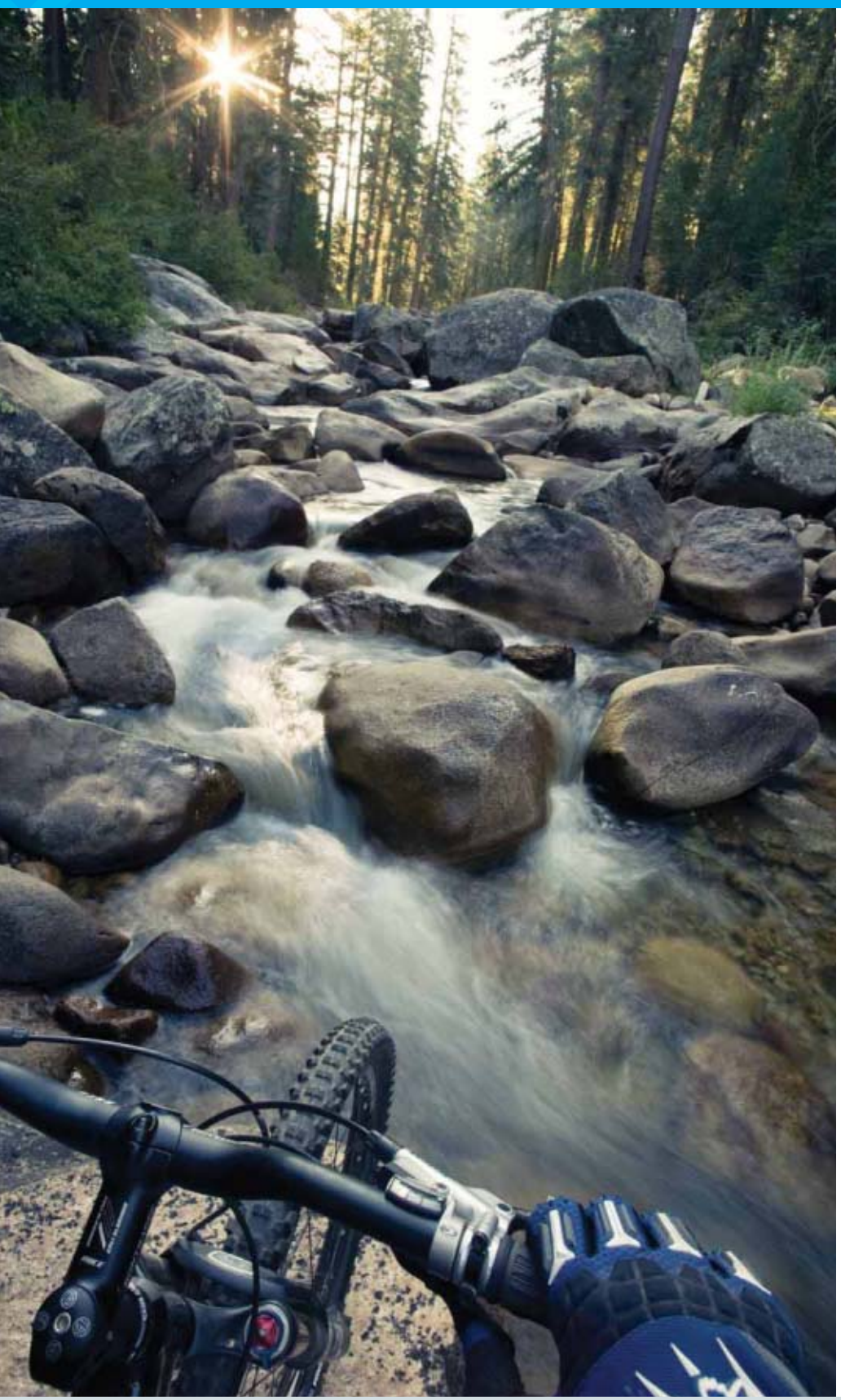
PAREXEL's Bio/Pharmaceutical R&D Statistical Sourcebook 2009/2010 is a 'must-have' resource for the drug development industry as it puts real-world data sets at your fingertips for presentations, reports, business development efforts, strategic meetings, and critical decision-making analyses.

The 2009/2010 edition will once again be offered in electronic format for individual users, small groups, business units, or for company-wide access.

To purchase *PAREXEL's Bio/Pharmaceutical R&D Statistical Sourcebook 2009/2010* **call:** 1-800-856-2556, **E-mail:** customer.service@barnettinternational.com **or visit the Publications section on** www.barnettinternational.com

BARNETT EDUCATIONAL SERVICES
250 First Ave., Suite 300, Needham, MA 02494 · 800-856-2556

Go beyond the everyday. EVERY DAY.



Symyx Notebook

The Freedom to Experiment.

Whether it's on the trail or in the lab, you want the freedom to take new approaches, routes, and paths to your goals. That's why there's Symyx Notebook. It's the only electronic laboratory notebook that can be deployed across the enterprise in multiple scientific disciplines. With Symyx Notebook, research teams share a single application to document, work, collaborate, and speed the experimentation workflow.

Symyx Notebook streamlines the capture of all experimental information and intellectual ideas. Everyday tasks such as data capture and note taking are optimized and automated. All of which gives you the time and freedom you need to experiment—and get back to doing science.

To learn more, visit www.symyx.com/notebook6

Symyx®