



◀ Inside **Bio·IT** World

The Quarterly eBook of Bio-IT World's Most Trending Articles

Evolutions in

Next Gen Sequencing

▶ Part I: What's Next
for NGS



Cambridge Healthtech
Media Group

www.Bio-ITWorld.com

Evolutions in Next Gen Sequencing

- 3 Letter from the Editorial Director
- 4 Six Years After Acquisition, Roche Quietly Shuttters 454
- 5 Out of Many, One: Genohub's One-Stop-Shop for NGS Services
- 7 Uncertainties in Assembly: Communicating and Managing the Truth About Our Data
- 9 Bring Genome Analysis Tools to the Masses with GenePattern
- 11 Bina Launches Bina On-Demand, Exome Analysis
- 14 Arpeggi Adds Genome in a Bottle Consortium Data to GCAT
- 16 Genomics and the Healthcare Revolution
- 17 Too Much to Ignore: Anne Wojcicki's Plan for Health Care and Big Data

EDITORIAL DIRECTOR
Allison Proffitt (617) 233-8280
aproffitt@healthtech.com

PUBLISHER
Lisa Scimemi (781) 972-5446
lscimemi@healthtech.com

MANAGER, BUSINESS DEVELOPMENT
Jay Mulhern (781) 972-1359
jmulhern@healthtech.com

MANAGER, BUSINESS DEVELOPMENT
Katelin Fitzgerald (781) 972-5458
kfitzgerald@healthtech.com

MARKETING ASSOCIATE
Lisa Hecht (781) 972-1351
lhecht@healthtech.com

Contributing Editors

Deborah Janssen
John Russell
Ann Neuer

Cambridge Healthtech Institute

PRESIDENT
Phillips Kuhl

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

About Bio-IT World

Part of the Cambridge Healthtech Institute Media Group, Bio-IT World provides outstanding coverage of cutting-edge trends and technologies that impact the management and analysis of life sciences data, including next-generation sequencing, drug discovery, predictive and systems biology, informatics tools, clinical trials, and personalized medicine. Through a variety of sources including, Bio-ITWorld.com, Weekly Update Newsletter and the Bio-IT World News Bulletins, Bio-IT World is a leading source of news and opinion on technology and strategic innovation in the life sciences, including drug discovery and development.

Advertiser Index

Advertiser	Page #
Bio-IT World Product Directory	5
Bio-ITWorld.com/Product-Directory	
Subscribe to Bio-IT World	8
Bio-ITWorld.com	
Bio-IT World Free Download	10
NGS Survey Results	

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

Subscriptions: Address inquires to *Bio-IT World*, 250 First Avenue, Suite 300, Needham, MA 02494 888-999-6288 or e-mail kfinnell@healthtech.com

Reprints: Copyright © 2013 by *Bio-IT World*. All rights reserved. Reproduction of material printed in *Bio-IT World* is forbidden without written permission. For reprints and/or copyright permission, please contact Jay Mulhern, (781) 972-1359, jmulhern@healthtech.com.

Key

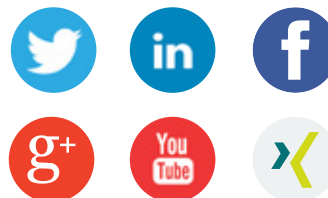
Look for the 

Click on this icon at the ends of articles to check out our RSS Feed!

Table of Contents 

Click this icon while reading, to return to this Table of Contents page at any time.

Connect with Us:





Allison Proffitt

What's Next for NGS

The next generation sequencing space has been moving quickly for a while now, and shows no signs of slowing. While Illumina is still proving to be the dominant platform in the market, it is by no means the only active player. Though Roche has chosen to close 454, from other corners of the market rumors of instruments still swirl.

As the field develops though, challenging new questions are rising to the forefront of conversation. What will the sequencing market look like? What kinds of questions should we take on now in human genomics and other species? What will be the business plan for researchers moving forward? How will analysis be done and by whom? What role will consumer genetics play, and how will ethical questions be decided?

Individuals and groups are emerging to take on these questions, each contributing a piece to the puzzle that is the growing industry.

C. Titus Brown, an assistant professor in the department of Molecular Biology and Genetics at Michigan State University, in an interview with Bio-IT World challenges researchers to be open about uncertainty. "We have been told, through a succession of papers in high profile journals and with all of the various genome browsers, that here is THE genome of mouse, here is THE assembly of zebrafish," he said. "As a result, the unwary biologist (which is many of them) will unwittingly trust the assembly we have."

Arpeggi (now Gene by Gene) is hosting GCAT, a freely available cloud-based platform for evaluating the accuracy of next-generation sequencing (NGS) analysis pipelines that provides performance reports which users can share and discuss with the community.

GenePattern, from The Broad Institute, is providing researchers a platform for integrating any of the thousands of genomic analysis programs available on the web into a seamless pipeline. And Bina is offering its solution to questions of data volume and security with an on-site, on demand analysis appliance.

As the market's needs and preferences become clearer, the offerings will certainly be refined. In the meantime, it's an exciting time for research—full of new ideas and plans.

In order to keep abreast of this changing landscape, we're presenting some of the most interesting stories from the last quarter as Inside Bio-IT World. It's our hope that this will serve as an update on some of the questions and answers being posed in the industry: a glimpse of where we've been and where we're headed on a fast-moving ride.

Allison Proffitt
Editorial Director

“As the field develops though, challenging new questions are rising to the forefront of conversation.”



454
SEQUENCING

Six Years After Acquisition, Roche Quietly Shuttters 454

BY BIO-IT WORLD STAFF | OCTOBER 16, 2013

This month, Roche began the process of closing its wholly-owned subsidiary 454 Life Sciences, a once-dominant player in next-generation sequencing, and laying off the company's 130 employees. Manufacturing of 454 sequencers will continue through 2015, and the sequencers will continue to be serviced through mid-2016; the layoffs will be phased over this period. This announcement follows a series of downsizing measures from Roche in the area of genetic sequencing over the past year.

At the time that Roche acquired 454 Life Sciences in 2007, the company seemed poised to lead a reinvigorated gene sequencing market into the age of the \$1,000 genome. In a 2005 paper in *Nature*, 454 had described the successful use of a new technique of "sequencing by synthesis," synthesizing DNA fragments, separating them to be sequenced in parallel, and then virtually reconstructing the combined genome. The technique's efficacy was demonstrated by sequencing the nearly-complete genome of the bacterium *Mycoplasma genitalium*, at a fraction of the cost and time possible with the Sanger sequencing method that had been standard for a quarter

of a century. 454's method was significant not only because of its initial success, but because it bypassed limitations to pushing Sanger sequencing forward, promising refinements in the future that would continue to drive the price and time constraints of sequencing downward. The same year, 454 released the first commercially available next-generation sequencer, the GS20, and began its relationship with Roche with a \$60 million exclusive licensing deal. At the time, this string of achievements led 454's founder, Jonathan Rothberg, to tell Bio-IT World, "it's been a 25-year race, we're commercial, and we won."

In recent years, however, 454's technology has been eclipsed by other next-generation sequencers like Illumina's MiSeq and the Ion Proton by Ion Torrent, another Rothberg-founded company. Roche, meanwhile, has run into a series of walls in its aggressive attempts to regain the edge in genetic sequencing. A hostile takeover of Illumina was thwarted in April of 2012, and this April, Roche announced that it would close its Applied Science unit, firing 170 workers (including 60 working at 454 Life Sciences), and narrow the focus of this division strictly to next-generation sequencing. Two long-term R&D projects that might have pushed Roche toward "third-generation" sequencing, using semiconductors and nanopores respectively, were discontinued as part of the reorganization.

The closing of 454 marks Roche's latest pivot toward a more limited vision of its role in the sequencing market, at least for now. The company's largest remaining foothold in the area may be its exclusive licensing deal with Pacific Biosciences for forthcoming in vitro diagnostics products, using the SMRT sequencing system that powers PacBio's RS II Sequencer.

Out of Many, One: Genohub's One-Stop-Shop for NGS Services

BY MATT LUCHETTE | SEPTEMBER 25, 2013

The launch of Genohub, a company that helps link genomics researchers with sequencing providers, this past August marks a significant step forward in the Next Generation Sequencing industry.

Ten years after the end of the Human Genome Project, genetic researchers now have hundreds of companies, core facilities, and experts to choose from for sequencing their genetic samples. Each company offers unique services, but researchers have few resources to compare them. Genohub, a startup based in Austin, Texas, hopes to change that.

"There's a disconnect between sequencing providers and [the] researchers who need their services," said a molecular biologist and founding member of the company, who I'll call Aaron. (He asked not to be named in this article because of other professional commitments.)

To find reputable sequencing companies, many researchers rely on their colleagues' word of mouth, Aaron explained. And when these scientists are lucky enough to have a genomics core at their own institution, there are often "queues for their own core facilities."

The role Genohub hopes to have in the NGS market is akin to companies like Kayak in the airline industry. A consultant looking to catch the Sunday night red-eye from San Francisco to New York has specific guidelines to meet when she purchases her airline ticket. But with tens of thousands of flights per day worldwide, all of them served by well over a hundred airlines, she could end up wasting valuable time tabbing through

airline websites until she found an adequate flight. Kayak is a one-stop-shop that lets her filter out departure and arrival cities to find the best ticket for her needs.

For researchers looking for an NGS provider, the search can be just as onerous. Scientists need to find reliable companies that provide the specific

“The number one market is going to be small and medium sized labs”

Martin Gollery, Tahoe Informatics

sequencing service they're interested in at the right price. On Genohub, clients can refine those parameters to quickly find the provider they need.

Pouya Razavi, the company's CEO and co-founder with Aaron, noticed that of the NGS company directories available, none of them allows clients to compare services and prices between providers. He and Aaron first met at SUNY Buffalo in 2000 where Razavi was studying computer engineering and Aaron was studying molecular biology. Razavi went on to receive his master's

from the University of Illinois while Aaron stayed at SUNY to finish his Ph.D. The friends relocated to Austin soon after.

During frequent lunch conversations, Razavi heard from Aaron about many of the challenges in the still-nascent NGS industry. "We specifically became interested in the sequencing-as-a-service trend," Razavi said, "and how a professional online marketplace like Genohub could make a huge impact." The two felt that the industry could benefit from a service that would bridge the gap between the sequencing providers and their clients, creating a comprehensive database of NGS companies in the process.

"After several conversations with actual service providers and researchers," Razavi says, "we decided to move forward and start Genohub."

"Other companies," Aaron explained, function like a phone book of providers, "but you can't compare services. They get a lot of offers, but not a lot follow through." Genohub hopes that by making service criteria transparent and tunable, the company can create better matches between NGS companies and potential clients, with less back-and-forth between the parties before a deal is made.

And like Kayak, which profits partially from airline referrals, Genohub's profits come from successful transactions. Researchers can browse and buy services for free; companies can list services on Genohub for free, and they're charged a small service fee for completed projects. The specific fee is still being refined, says Razavi.

Looking for products & services in NGS?

Visit the **Bio-IT World** Product Directory

www.Bio-ITWorld.com/Product-Directory





»» Read a related article on new NGS & Bioinformatics Services

From the Beach to the Skies

The mere fact that the biotech industry might need a company like Genohub marks an important step in the commoditization of genome sequencing. In the decades after Kitty Hawk, airplane travel went from an aristocratic adventure to a common commute. Similarly, improved techniques in the years since the Human Genome Project have taken genome sequencing from a multi-million dollar, multi-center project, to a couple-thousand-dollar procedure, depending on the technique.

While Genohub is still in its early stages, and far from an industry staple, one could imagine it having an impact like Kayak, creating an open NGS marketplace, streamlining the connection between providers and clients, and improving quality through competition between companies.

But who is using Genohub now? And what's the feedback been like so far?

Genohub's initial development began around May of 2012. This past March, "We attended the ABRF trade show and conference in Palm Springs [...] when we started opening up our service on a limited basis to service providers," Razavi said. "Since then we've also directly approached some of them. However an increasing number are now hearing about and contacting us on their own.

"Initial feedback was great, and it's gathering momentum," Razavi explained, adding that providers "like that they focus on next gen sequencing." In particular, he said, providers enjoyed Genohub's interface for listing services and prices in a structured way, freeing up wasted time quoting prices for customers on routine services. Since their initial opening, Genohub now hosts 30-40 providers.

By making it easier for clients to find reputable companies, Genohub could also expand NGS

services to new labs. "The number one market is going to be small and medium sized labs," said Martin Gollery, a bioinformatics scientist with 15 years experience. His bioinformatics consulting company, Tahoe Informatics, provides researchers and companies with "data analysis and higher-level planning."

Gollery thinks Genohub will be particularly useful for labs with specialized needs. "There're enough labs that do sequencing [now] that couldn't do it before," he said. Without companies they have tried and trust, though, these labs risk spending thousands of dollars on shoddy results if they choose an unreliable company.

Instead, says Gollery, "these labs may say 'I don't know who to go to. I don't want to buy a system. I'll just go to Genohub and pick what I want to do.'" Having Genohub as a middleman, mediating conflicts if problems with a company's quality or turn-around time arise, could provide the peace of mind these smaller labs need to take the plunge.

"It may not change the work that's being done," Gollery said, "just make it more efficient," adding that it could also help drive prices down and increase transparency in the services specific companies provide.

A One-Stop-Shop

Razavi noted that in the near future, he's hoping to add a variety of bioinformatics analysis services to Genohub. The company has "heard from a lot of full-service providers and they would like to have bioinformatics" on the website, said Razavi.

Gollery was a bit more skeptical. "Bioinformatics will wind up on Genohub," he said, but "it doesn't work as well." NGS services are typically standardized and priced based on the amount of sequencing done, but bioinformatics is "less of a cookbook," said Gollery. The services are typically

customized for specific projects, and the pricing can be just as variable, making it less amenable to a site like Genohub.

"Genohub will be a very useful site for those people who want to get sequencing done," said Gollery.

“The balance we’ve done is to allow providers to list bioinformatics services and not enforce a single pricing structure.”

Pouya Razavi, CEO, Tahoe Informatics

While Razavi agrees that pricing bioinformatics services would be less straightforward, the company plans to make pricing for these services optional, instead allowing customers to "compare general factors," such as the company's technical expertise, turnaround time, and reputation, without forcing them to fix their prices on the site.

"The balance we've done is to allow providers to list bioinformatics services and not enforce a single pricing structure," he explained. For researchers to know the exact price of a service, they would request a quote from a company that fits their criteria. Razavi says these new bioinformatics features are currently in development.

What's Next?

In addition to adding bioinformatics services to the website, Genohub has a number of long term goals they're working on, including providing data storage services for clients. Furthermore, with an eye on the growing importance of personal genomics in medicine, Genohub hopes to create an interface for linking physicians to sequencing companies that are approved to handle patient samples.

"The exciting part of the NGS market," said Razavi, is that "the variety of NGS applications is growing." With Genohub, he hopes to expand access to NGS services, and bring NGS to researchers "who aren't just in genomics," like behavioral scientists or pharmaceutical developers.

"There's a lot of exciting questions to address," Razavi remarked, and with Genohub, he hopes to give researchers the means to answer them. 📧

Uncertainties in Assembly: Communicating and Managing the Truth About Our Data

BY ALLISON PROFFITT | AUGUST 26, 2013

Published in late July in *GigaScience*, the Assemblathon 2 paper has been in the works since June 2011. Three genomes and about 17 GB of compressed data later (doubled if you include the contig files that were extracted from the scaffolds), the findings can be summarized in five words: “We’re doing it all wrong.”

At least according to C. Titus Brown. Brown is an assistant professor in the departments of Molecular Biology and Genetics; and Computer Science and Engineering at Michigan State University. Brown was an early reviewer of the paper, and took to his blog, *Living in an Ivory Basement*, soon after to hash out his thoughts.

“It took about a week after I first wrote the review and submitted it to sort of understand the implications,” he told *Bio-IT World*. “Never before had I seen a straight up systematic comparison of two different assemblies of the same genome from the same data.”

The Assemblathon had 43—16 fish assemblies for a Lake Malawi cichlid (dataset provided by Broad); 12 assemblies for a Red tailed boa constrictor (dataset provided by Illumina); and 15 assemblies for a Budgerigar or common pet parakeet (dataset provided by Erich Jarvis at Duke University and the BGI and Pacific Biosciences).



C. Titus Brown, Assistant Professor, Michigan State University

“There are many genome assembly programs out there, but it is not always clear as to which is the best,” the Assemblathon organizers explained in the published rationale for the project. “Part of the problem is that it is not easy to define what ‘best’ is and an assembler that might work well in one situation (e.g. assembling a high-repeat-content genome) might not fare as well in other situations,” they continued.

Brown said he wasn’t surprised that the assemblies were different. “I thought, ‘Well I would have expected that,’” Brown said. “But what I realized that it meant was that whenever somebody produced the assembly of something, they were failing to take into account that if you’d used a different technique, you might have gotten a subtly different answer. Until this paper came out, I hadn’t put that into words or really realized that.”

Of course we can’t judge which assemblies were “right” because we don’t know. “There may not be one answer,” Brown said. “The question isn’t, ‘Is the result correct?’ It’s, ‘Is the result useful?’”

For example, Brown was most struck by the analysis of core gene repertoires, “genes that every eukaryotic genome should have.” Each assembly showed about 95% of the catalog, a measure of how complete the assembly was.



"But the thing that struck me was that the 5% that was missing was different between the different genomes," Brown said. "I would not have expected that. You could put them all together and get a 100% catalog, but they were each missing a slightly different subset of the core genes. I couldn't really think of a really good, straightforward computational reason why that would happen."

This presents a twofold problem, Brown believes. On the one hand, it's not surprising to a bioinformatician that assemblies are different. Different assemblies address different goals—and that's fine as long as you know that's the case.

Yet as Brown points out in his blog, "we have been told, through a succession of papers in high profile journals and with all of the various genome browsers, that here is THE genome of mouse, here is THE assembly of zebrafish. As a result, the unwary biologist (which is many of them) will unwittingly trust the assembly we have."

In Brown's own experience working on the sea urchin genome, 8% of the raw reads never made it into the published assembly. "We said, 'Here's our final assembly. This is what we analyzed. It's as good as it's going to get for this paper. We're done.'"

That's commonly the case, he says, but it's not commonly discussed. "It doesn't make it into the paper; it doesn't make it into the reader's head. It doesn't make it into the discussions between the old guard who largely decide a lot of the funding and initiative efforts," Brown says.

Leaving out that data has real repercussions. "It short changes anyone who is actually trying to improve the genome," by presenting an inaccurate view of what is completed.

"Biology computation is enough of a black box for enough biologists that they simply don't realize how uncertain we really are," Brown said.

He recounted a conversation over lunch with a colleague trying to choose the "right" option from three slightly different test results. "Why aren't you content with just having three results?" Brown asked. "Because then we won't know which result is the correct one," was the response.

“The goal of assembly projects is not just to generate the assemblies but to generate the population of researchers that can make use of them”

"Why," Brown countered, "when you only had one result, did you think that was the correct one?"

It's a paradigm that computer scientists are comfortable with, he said, but biologists are very wary of.

Once the uncertainty inherent in genome assembly is fully open, what can be done to address it? How does one work with 12 different boia constrictor assemblies from the same dataset?

"There are no good ways to combine those assemblies and no good ways to compare those assemblies. We're really facing a lack of tools," Brown said. "It's much more exciting to write your own nifty tool that does something subtly different from what everyone else has done, than to actually do a good job of comparing and evaluating tools."

Brown joked that any conversation with a scientist would inevitably return to funding, and proved his point by calling for funding to develop the tools needed to compare and combine the results we have. "It all goes hand in hand, right? We can't get the funding to develop the tools until we have the cultural outlook that says building the tools is useful."

It's not just tools that are necessary, but the skills to tackle these problems as well.

"The goal of assembly projects is not just to generate the assemblies but to generate the population of researchers that can make use of them. Those skills are also what you learn when you have people involved in these software development projects."

Projects like Assemblathon will drive the maturity of the field, Brown believes. "That's going to be a lasting legacy of this paper. Who cares what assembly we were using in 2013, but we started to have a much more mature conversation about how we should compare genomes." 📡

Bio-IT World

- ✓ DAILY NEWS
- ✓ FREE INDUSTRY WHITEPAPERS, PODCASTS AND WEBINARS
- ✓ MARKET SURVEY RESULTS
- ✓ COMPLIMENTARY NEWSLETTERS
- ✓ ONLINE PRODUCT DIRECTORY

Subscribe Today!
www.Bio-ITWorld.com

Bridging the Gap: Bring Genome Analysis Tools to the Masses with GenePattern

BY MATT LUCHETTE | AUGUST 22, 2013

This past month, researchers from the Broad Institute released a software update to GenePattern, the Institute's open-source genome sequence analysis software (and winner of a 2005 Bio-IT World Best Practices Award), which will allow programmers to upload their own analysis tools to an open database.

"We want to let them release their own tools into the wild," said Dr. Michael Reich, the Director for Cancer Informatics at the Broad.

After being publicly released in 2004 by Jill Mesirov's lab at the Broad, GenePattern has found a niche in providing researchers a platform for integrating any of the thousands of genomic analysis programs available on the web into a seamless pipeline.

Reich, a researcher in Mesirov's lab, estimates that there are over 10,000 genome analysis tools available online from a number of developers, and many of the tools do not support the same file formats, making it difficult to use multiple programs to analyze genomic data—a small problem for a computer scientist, but a rate-limiting step for biologists who may not know their bytes from their Booleans. GenePattern makes it easier for the less tech-savvy researcher to combine these programs together, like Lego pieces, into a single workflow.

“GenePattern provides tools to help researchers without programming experience automate workflow creation online, without downloading any programs, and even run the analysis on the Broad's computer servers”

"The input to any step," said Reich, "makes a compatible output" for the next tool in the pipeline. And they seem to be on to something: the software supports nearly 50,000 total users and handles thousands of analyses per week.

The update, said Reich, provides improved integration between GenePattern and GParc, GenePattern's third-party software archive, making it easier for outside developers to "disseminate data easily" by contributing their own analysis programs to GParc. Reich hopes that combining GParc's growing repository of third-party applications with GenePattern's pipeline creation tools will further the program's mission of making genome analysis software easy to use for non-programmers.

Additionally, the update allows users to easily add modules from the GParc repository to their pipeline, and it can tell the user which repository a module came from after an analysis is completed.



In his role at the Broad, Reich plans to pursue software projects that are "designed to serve the requirements of the world-wide genomics community." GenePattern is one such way he hopes to serve that community by making powerful programs for genome analysis more accessible, bridging the gap between biology and bioinformatics.

Compared to the web's other open-source genome analysis programs, such as the UNIX-based Tuxedo Suite of applications from Johns Hopkins University, which require users to have programming experience to build the analysis workflow, GenePattern provides tools to help researchers without programming experience automate workflow creation online, without downloading any programs, and even run the analysis on the Broad's computer servers. The feature is particularly useful for users who may have not have the computer science experience to build their project from a command-line interface. The program, however, still allows users manipulate its modules in MATLAB, R, or Java if they wish.


“Making bioinformatics tools more accessible to a wider audience may be a difficult problem for programmers, but through GenePattern, Reich and his colleagues hope to bridge the divide and ‘bring powerful technology together.’”

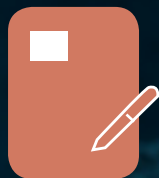
Aside from its module integration tools, one of the major benefits of GenePattern, Reich said, is recording all of a pipeline's analysis steps, allowing users to easily summarize their analysis methods. Mesirov calls this technique "accessible reproducible research." Reich stressed that this feature is especially important for allowing other researchers to reproduce and check a lab's work. "A paper is not research," he said. "It's an ad for research."

GenePattern is also designed to integrate with other web-based genome analyzers. Galaxy, for instance, "is complementary to GenePattern," says Reich. Both programs are a part of the GenomeSpace Initiative, a collaboration between

Mesirov's and Aviv Regev's labs at the Broad, which provides an environment for integrating multiple analysis programs from different developers into a single workflow.

Since its public release in 2004, GenePattern has been cited in over 20 papers that have expanded the program's initial applications to a number of others, from flow cytometry to microRNA expression analysis.

Making bioinformatics tools more accessible to a wider audience may be a difficult problem for programmers, but through GenePattern, Reich and his colleagues hope to bridge the divide and "bring powerful technology together." 



Free Download:
Bio-ITWorld's
NGS Survey Results

[Click Here to Download Survey](#)

Bina Launches Bina On-Demand, Exome Analysis

BY ALLISON PROFFITT | AUGUST 19, 2013

Bina Technologies has launched two new capabilities making their big data genomics platform accessible to more users. Whole Exome Analysis and Bina On-Demand expand the user base for the company's hardware appliance that quickly handles assembly and alignment of raw reads and variant calling in next generation sequencing.

It's been a big year for Bina. The company announced their offerings in February at AGBT. Since then the team has completed a series B round of financing worth \$8 million led by Sierra Ventures, Columbus Nova Technology Partners (CNTP), and AME Cloud Ventures.

Narges Bani Asadi



Headcount has also grown. Narges Bani Asadi, Bina's founder and CEO, welcomed Sharon Barr (formerly of Couchbase and Yahoo) as Vice President of Engineering; Take Ogawa—of Complete Genomics, RainDance Technologies, Roche and Invitrogen—as Director of Field Sales; and Thomas Kanar (formerly with Authonet Mobile) as CFO.

The capabilities announced today mark the next step in Bina's evolution.

Bina's first product, the Bina Genomic Analysis Platform, dealt exclusively with whole genome analysis. "We solved the hardest problem first," Asadi says. The hardware piece sits right next to the sequencer, collecting all the data seamlessly. The company's algorithms can assemble, align, and handle variant calling

on a whole genome sample in three hours, Asadi said.

The new pipeline is specifically for whole exome analysis and can handle analysis in a half an hour.

"In theory you could think, we could run a whole genome, so it's very easy also to run a whole exome," Asadi explained. "But there are detail differences. For the exome... The depth of coverage is different, the capture is different, which introduces some noise usually, so you have to be careful where you align the read and where you call the variants."

Asadi says the exome analysis platform is the first of many new workflows that can be expected in the coming months. "We want to make sure we address all the different demands that a user of next generation sequencing data has."

The result, the company hopes, is an expanded audience for Bina, which is where the Bina On-Demand business model comes in.

Genomics Vending Machine

The Bina Box can be compared to a cable box. The use of the box—as well as the channels and services it delivers—is part of your a fee.

“For smaller researchers that have projects that might not keep a Bina Box fully busy, [Bina On-Demand] puts the technologies in their hands in an affordable way.”

Mark Sutherland,
Senior Vice President,
Bina Technologies

Mark Sutherland



"Our standard product offering is a subscription-based agreement," Mark Sutherland, Bina's Senior Vice President of Business Development explained. "The entire package—the hardware platform called the Bina Box, along with the operating system, the applications, the training and technical support—is all delivered as an integrated package, delivered to the customer's site for a monthly fee. The use of that box is unlimited during that box, and they can process as many samples as they wish of any type with any frequency."

But just as all users are not doing whole genome sequencing, all users are not working at a volume to keep a dedicated Bina Box engaged. "For

“For smaller researchers that have projects that might not keep a Bina Box fully busy, [Bina On-Demand] puts the technologies in their hands in an affordable way.”

Mark Sutherland, Senior Vice President, Bina Technologies

active genome centers and cancer researchers that have a lot of data flowing through the lab, that works really well because it gives them the lowest effective price per sample as long as the box is kept fairly occupied.”

“For smaller researchers that have projects that might not keep a Bina Box fully busy, [Bina On-Demand] puts the technologies in their hands in an affordable way.”

Instead of the cable box, Sutherland suggests a vending machine. The Bina Box is still delivered to the customer’s site, but there’s no monthly fee, no subscription. “If you walk by it and you don’t use it, you don’t pay. Whereas if you want a Diet Coke, you put in a dollar and you get the product.”

The Bina on Demand offering allows users to simply share. “It allows us to make the product available to individual researchers—scientists, labs, departments—across an academic setting where any one of these investigators by themselves may not have a workload sufficient to justify a Bina Box for their sole use, but by being a part of a

user community for Bina on Demand, they can send their projects there and get very quick turnaround, very high speed, very high accuracy, ease of use, robustness that a dedicated Bina solution would provide, but pay only for what they use.”

Bina records the highest level information about usage—users and job titles—and submits that to the host organization with the monthly bill. The host manages the user level billing.

Privacy Software

The shared user community, of course, has its challenges. Asadi says that before the Bina on Demand could launch, the company developed “very innovative software” to address the privacy concerns inherent in a shared workspace. “Each of the different scientists has their data completely secure and in isolation from the other scientists,” she said.

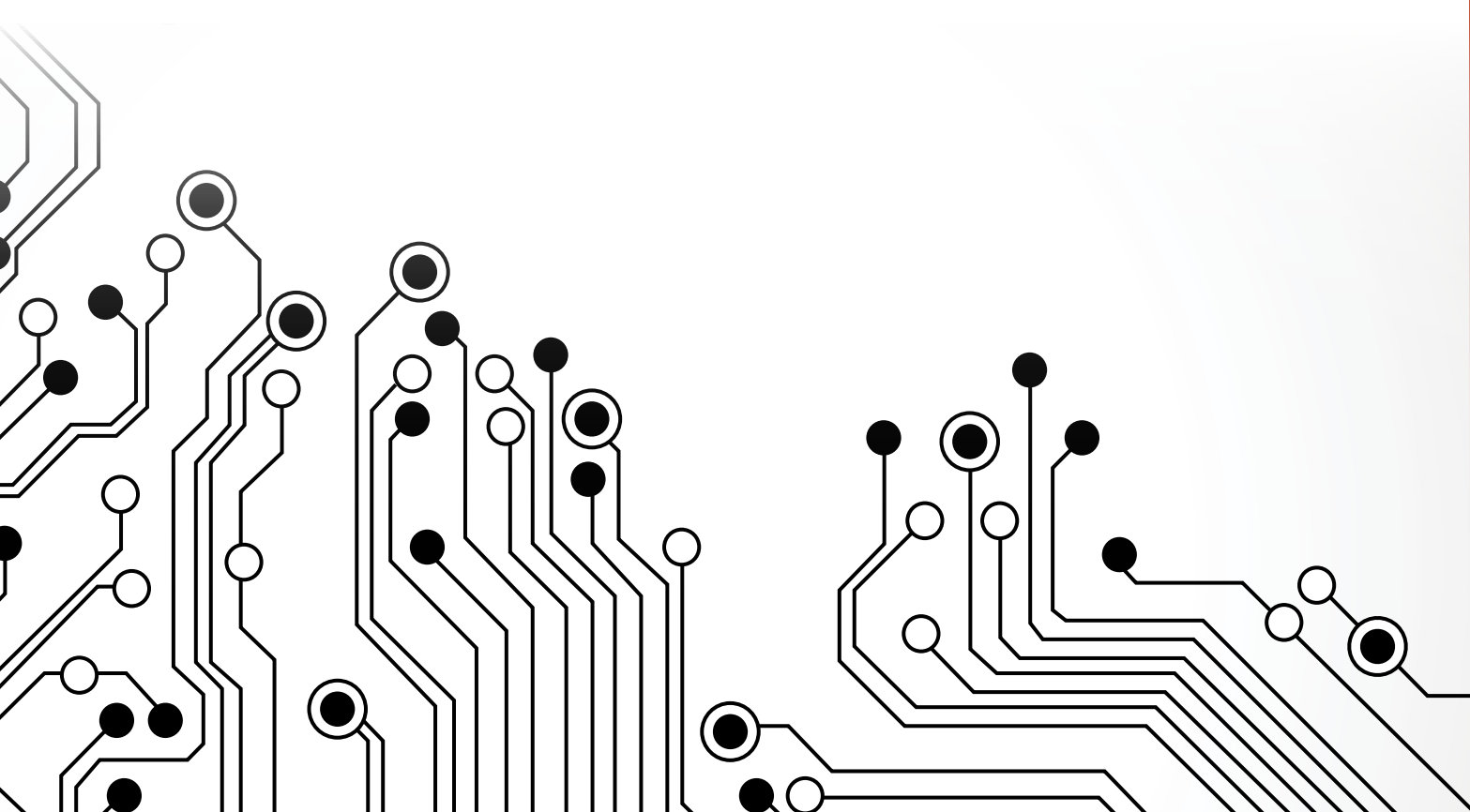
The security upgrades again expand Bina’s client pool and, Sutherland thinks, offer a distinct advantage over cloud solutions.

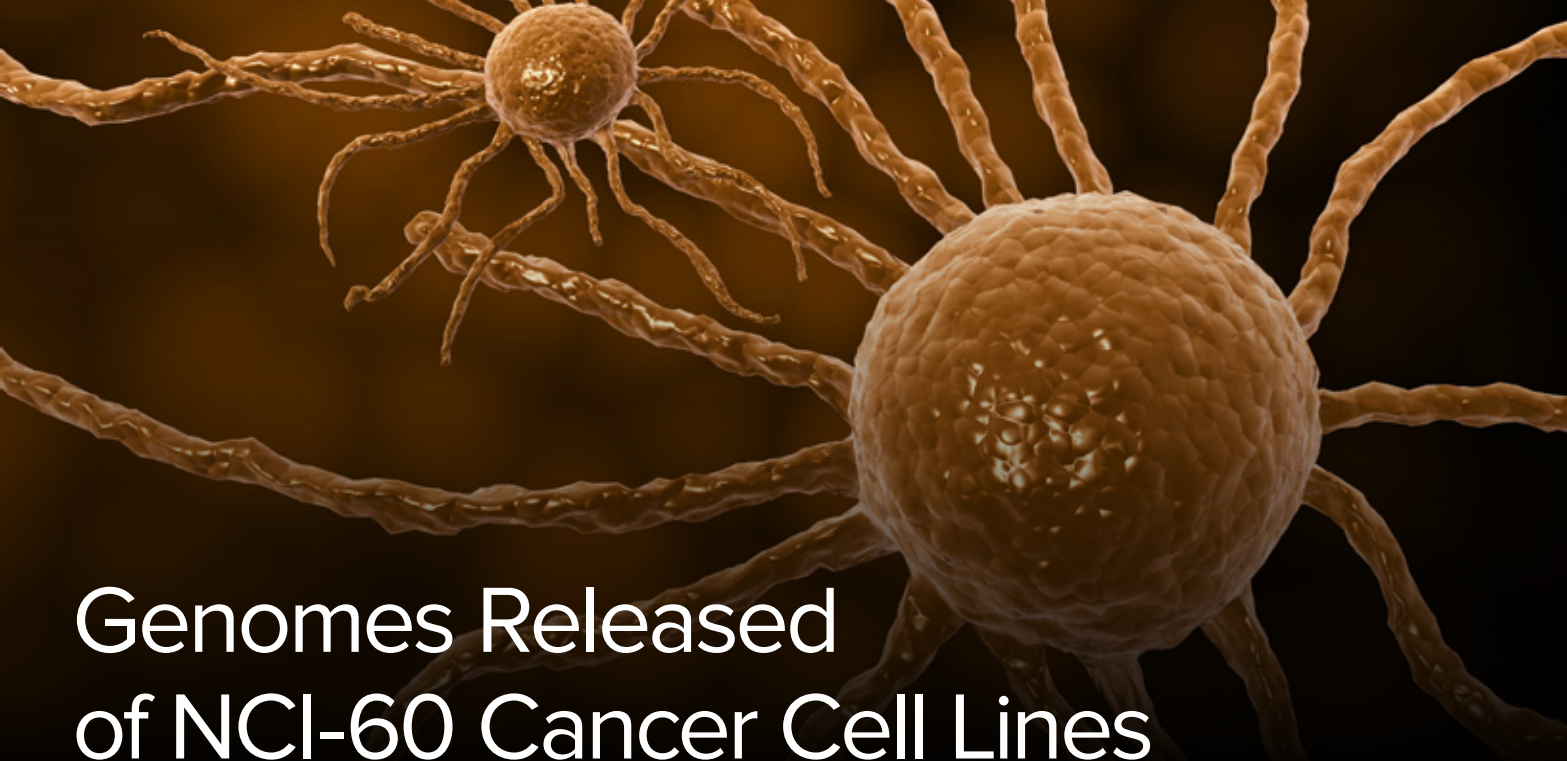
“We’re seeing more and more of these applications of interest happening in children’s hospitals, cancer centers—what we call clinical core labs,” he said. “Eventually there will be patient privacy and HIPAA issues around these datasets and many of these customers of ours feel more comfortable having an on premises solution that sits behind their firewall that they feel helps keep them consistent with best practices around patient privacy.”

Asadi says the On-Demand offering can even be considered a private cloud. “You should think about it as a private, on premises cloud. It scales in hardware, software infrastructure, as well as all the applications. It’s very easy to tune it to that institution’s demands.”

The “box” name can be a bit misleading, Asadi and Sutherland said. Bina Boxes can be stacked and customized to provide a scalable solution.

“You used to hear about studies for 10s or even 50 patients. Now you’re hearing about studies about 500, 1,000, 5,000 or more. We’re aware of a couple of national initiatives where over 100,000 patients will be sequenced,” Sutherland said. “One of the advantages of the Bina platform is not only its speed, but its scalability. Narges and the development team here have designed a platform that is rather extensible. People today may only have 20 or 50 samples in a month. We firmly believe these same accounts will be processing 500 a month a year from now. So by getting started with them today, we can build out to meet their needs as they expand.”





Genomes Released of NCI-60 Cancer Cell Lines

BY ALLISON PROFFITT | JULY 17, 2013

On Monday, researchers released the largest database of cancer-related genetic variations—the genomes of the 60 cancer cell lines represented by the NCI-60 list. The project was published online in *Cancer Research* (2013;73:4372-4382. Published OnlineFirst July 15, 2013).

The NCI-60 cell lines were derived from cancers from nine tissues of origin. “These tissues of origin were selected because they are hard to treat,” and include lung cancer, melanoma, ovarian cancer, renal cancer, colon cancer and others, Yves Pommier, chief of the Laboratory of Molecular Pharmacology at the National Cancer Institute, told Bio-IT World.

Pommier and his collaborators performed whole exome sequencing on the 60 cell lines using an Illumina Genome Analyzer IIx instrument and catalogued the found variants.

“A small number of genes had been sequenced by Sanger several years ago... and when we compare, the matches are beautiful, the data match perfectly with the known few genes that had been sequenced by classical methods. But now we have the whole genomes.”

The cell lines have historically been used to screen chemicals and anticancer agents for possible development into cancer therapies. Over the years there have been thousands of compounds screened against the list, Pommier said, but having full sequences for the 60 cell lines expands the possibilities.

Cancer cell lines are certainly different from tumors, but the genetic results are still very similar, Pommier said. “When we compare the gene expression profiles of these cancer cell lines to their tissues of origin, to a good fraction they retain the tissue of origin signature. If you look at the melanoma cell line, the [cells] still look like melanoma,” he said.

“The main difference is the cell lines are homogeneous—they are clonal; they are developing all the same in the tissue culture flask. A real cancer is very heterogeneous. That has pros and cons. The advantage of the cell line [is that] because they are very homogenous, it is easier to interpret the results. The gene mutations and expressions are the results of one population, where when you have a real tumor you have an average of everything.”

Public Mining

The dataset is ripe for query. The authors did some initial data mining, but are releasing the entire dataset to the research community. The data are made available through the CellMiner, NCI DTP and Ingenuity Systems’ websites.

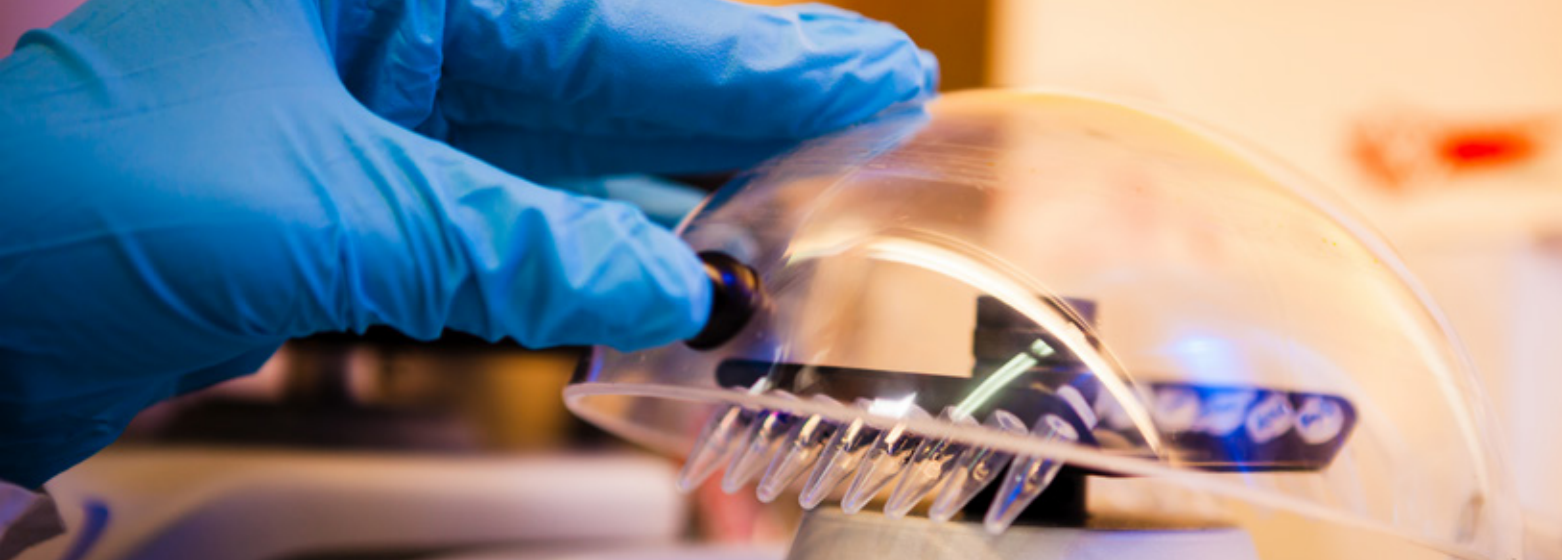
“We’re well aware that there are so many questions to be asked,” Pommier said. “You can enter the data in different ways. You could be a drug-minded person and look at a specific drug and see how the drug activity matches different mutations and gene expression. You can be a gene person and you can enter from the gene side. There are 21,000 genes! How many possibilities are there?”

“That’s why these things need to be publically available, so people can use it as a platform to ask their own questions and find their own discoveries. It’s essential that this is publically available, because it’s a tool.”

The dataset is a tool that should be accessible for all, the authors believe—not just bioinformaticians.

“One of the intents here is to enable people who don’t have a bioinformatics team to look at these data and look at biological and drug insight,” Pommier said. The website is for “regular people,” he contends. All of the data are delivered in Excel spreadsheets and can be stored and manipulated according to the user’s needs. Next steps for the team are to create tools to enable more mining by anyone.

On Monday afternoon, “in the early afternoon already 230 queries of the whole dataset had been put in and that was just a few hours from the release of the paper,” Pommier said. “It’s extremely active.”



Arpeggi Adds Genome in a Bottle Consortium Data to GCAT

BY MATTHEW DUBLIN | JULY 15, 2013

Texas-based bioinformatics startup Arpeggi (see, “Arpeggi’s Harmonious Approach to NGS Data Analysis”) has announced the addition of data from the Genome in a Bottle Consortium to its online Genome Comparison & Analytical Testing (GCAT) toolkit. GCAT is a freely available cloud-based platform for evaluating the accuracy of next-generation sequencing (NGS) analysis pipelines that provides performance reports which users can share and discuss with the community.



Released in April at the Bio-IT World Conference & Expo held in Boston, Arpeggi’s GCAT has been steadily adopted by both academic investigators and commercial vendors to compare the performance of their NGS alignment and variant calling software tools. Users can

simply download and analyze data on their local resource and then upload the results to GCAT to compare against various performance metrics. Benchmarks include coverage depth, correct to incorrect read mapping ratio, and transition/transversion ratios. After analysis is complete, GCAT produces visual reports on results and performance that then can be compared to reports uploaded by other users.

GCAT, which is hosted on Amazon Web Services’ cloud, was originally developed to benchmark and evaluate the Arpeggi engine, the company’s

proprietary variant caller. In addition to the Genome in a Bottle Consortium data, GCAT contains four real exome datasets produced by Life Technologies and Illumina sequencing platforms.

Now that GCAT has been fortified with the Genome in a Bottle Consortium data, researchers have access to highly confident genotype calls

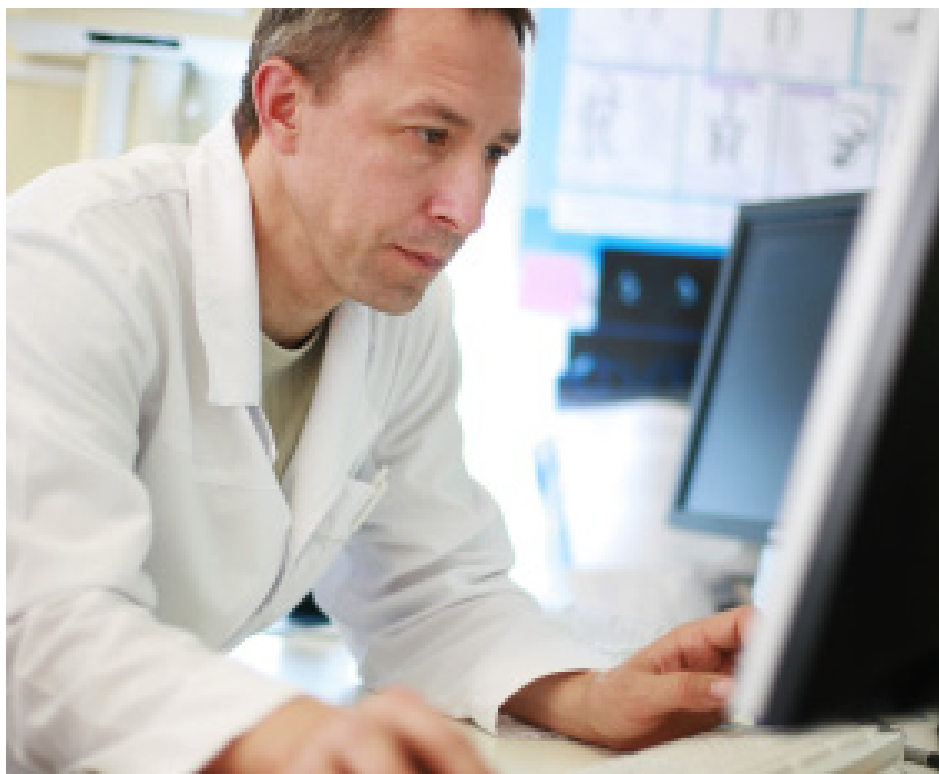
that can be easily employed to score the performance of analysis pipelines.

“The value of the Genome in Bottle data is that it’s a ‘truth set,’ they have areas of this genome for which they have developed highly confident genotype calls,” said David Mittelman, an associate professor at Virginia Bioinformatics Institute and Arpeggi’s chief scientific advisor. “So by integrating that data into GCAT, we’re offering a great metric you can use to evaluate your performance.”

The Genome in a Bottle Consortium is an international effort spearheaded by the National Institute of Standards and Technology (NIST) aimed at developing reference materials for human genome sequencing in order to assess the performance of NGS platforms. The consortium recently released an early set of highly confident calls based on publicly-available genome data

“The value of the Genome in Bottle data is that it’s a ‘truth set,’ they have areas of this genome for which they have developed highly confident genotype calls, so by integrating that data into GCAT, we’re offering a great metric you can use to evaluate your performance.”

David Mittelman, Associate Professor, Virginia Bioinformatics Institute



»» Click [here](#) to read a related article on Arpeggi

“So far, running our datasets on GCAT has shown us that some tools are being very aggressive but they lose a lot in terms of false positives, while other tools are very conservative but you don’t get to see how much stuff is missing.”

Justin Zook, Consortium Leader, National Institute of Standards and Technology (NIST)

“I thought the GCAT tool kit would be a good platform for people to understand the performance of their sequencing image by comparing their variant calls to our highly-confident genotype calls we’re developing as part of the Genome in a Bottle Consortium,” said Justin Zook, consortium leader and a biomedical engineer at NIST. “It’s the first tool that I’ve seen that allows you to compare lots of different methods in the same way, as opposed to everyone doing their own validation of their own methods, so it’s nice to have a centralized resource like this.”

Early next year, the consortium is planning to release a highly characterized HapMap sample together with a set of highly confident genotype calls. In a collaboration with Coriell Cell Repositories, the consortium was recently able to grow a large quantity of NA12878 cells from the Hap

Map sample set, totaling roughly 8,000 vials of 10 micrograms each. Consortium members will be able to request these samples next year.

But it’s not just GCAT that benefits from integrating the Genome in a Bottle Consortium data. Participants in consortium are using GCAT to add another dimension to their published research as well. Authors can include links in published papers to their GCAT results in order to publicly share information on their pipelines and performance results. Zook and his Genome in a Bottle Consortium colleagues currently have a paper under review at Nature Biotech that links to data taken directly from GCAT. The authors leveraged CGAT to generate figures comparing different sequencing and bioinformatics methods to the highly confident genotype calls they generated for their pilot candidate NIST Reference Material.

In one example, Novoalign, BWA, and BWAMEM, and Bowtie2 are compared to determine performance in terms of the amount of false negative and false positive results that are returned. The team found that Novoalign, which uses the Needleman-Wunsch algorithm to find optimum alignments, has the lowest number of false positives while BWA-MEM, a version of the Burrows-Wheeler Aligner, is better at eliminating false negatives. Bowtie2 seemed to lag behind the other three tools, reporting a high number of both false negative and false positives results.

In 2013 BWA-MEM and Novoalign3 for example show continued improvement. In the above, you want the plot to be most to the right and to the bottom.

Another example looked at the various results provided by two versions of the Broad’s Genome Analysis Toolkit (GATK) in combination with Novoalign against Illumina’s ISAAC genome alignment tool. According to performance reports from GCAT, GATK with Novoalign demonstrated a much lower rate of reporting false negatives and positives compared to Illumina’s tool.

Zook and his colleagues have already observed significant disparities among tools operating on the consortium’s datasets.

“So far, running our datasets on GCAT has shown us that some tools are being very aggressive but they lose a lot in terms of false positives, while other tools are very conservative but you don’t get to see how much stuff is missing,” said Zook. “So the ability to see those differences in performance is a real key feature that is critical and missing in our community discussion of tools.”

Future plans for GCAT include the integration of other well-characterized genome data sequenced at exceptional depth, such as Illumina’s Platinum Genome datasets, as well as more comparison functionalities which they hope will eventually improve the lack of continuity among variant callers and some mappers.

“GCAT is a also template for new ways to do science,” says Mittelman. “Not just the openness and transparency; but also the collaborations between academic labs, government institutes (like NIST) and for-profit companies. Especially in today’s economy, we need to be more creative in how we approach scientific problems (and how we fund them). I like to think that with GCAT we were also experimenting with how to experiment.”

Genomics and the Healthcare Revolution

BY AMANDA GOH | JUNE 10, 2013

SINGAPORE—At the recent 2013 Bio-IT World Asia conference, scientists and clinicians discussed how they could better coordinate their efforts in clinical genomics and “make genomic information usable in the clinic”.



Marcel Dinger of the Garvan Institute for Medical Research at the Kinghorn Cancer Center enthused, clinical genomics could “revolutionize our understanding of our genetic programming”.

Starting with Basics

Professor Patrick Tan, from the Genome Institute of Singapore and the Institute for Genome Sciences and Policy at Duke-NUS Medical School, began the session by describing how he has used exome sequencing to investigate the molecular mechanisms underlying various cancers endemic to Asia.

These studies have led to the identification of new genes and pathways implicated in tumorigenesis. For instance, Tan showed that the cell adhesion and chromatin modification pathways contribute to gastric cancer development. Such information may enable the identification of new drug targets and increase treatment options for patients.

Exome sequencing data may also show how existing drugs may be applied to new indications. Tan’s work has also revealed that peripheral T cell and NK/T cell lymphoma may be treated with a drug already in use for rheumatoid arthritis, and patients with cholangiocarcinoma may respond to therapeutic strategies already in use for pancreatic ductal adenocarcinoma.

Moving to the Clinic

Clinical genomics may be used to diagnose inherited diseases, especially rare ones. Jimmy Lin founded the nonprofit Rare Genomics Institute (RGI) to help children with rare genetic diseases. He described how more than 300 million patients are afflicted by rare diseases. There are more than 7,000 such rare diseases, but therapy is available for less than 5% of them. Helped by an international network of collaborators, RGI performs exome sequencing to help these patients identify their illness and seek appropriate treatment.

Another major application is cancer genomics, which enables the classification of cancers based on their pathological mechanisms and thus facilitates the design of treatment regimens. One would be able to “treat cancer by molecular stratification instead of by tissue type,” said Dinger, who set up a clinical genomics sequencing center.

A National Effort

Taking genomics a step further, Bogi Eliassen from the Faroe Islands’ Ministry of Health described their vision to make genomics “the cornerstone for optimal individualized healthcare with emphasis on prevention as well as treatment, cost effectiveness and democratic implementation”.

Eliassen is the program director of FarGen, which aims to incorporate whole genome sequencing into healthcare for all 50,000 Faroese people. The small size, isolation and transparency of this close-knit community are all advantages that permit mining of their rich genetic data.

The Faroese spent eight years developing an ethical framework, including the Biobank Act to protect individual rights. The FarGen project now receives strong political and public support, so the public health, administrative and education systems are integrated into the effort, as are multiple international academic institutions.

The FarGen project is still in its pilot phase as they optimize and implement a prototype workflow while aiming to have 1000 genomes sequenced by this year.

A priority for both the Faroe Islands and RGI is patient empowerment, to enable people to take responsibility for their own health. However, it is important to manage expectations and to allow people the right not to know.

“All in the Learning Phase”

Marcel Dinger identified decision-making as the single biggest challenge for clinical genomics, particularly when the information available is new or incomplete. There are many decisions to be made, ranging from participant selection and study design to the process of data analysis to the mode of reporting. Different groups have different goals, which adds to the difficulty of process standardization. Also, clinical testing protocols are subject to rigorous regulation but constantly evolve as technology advances.

Most speakers cited cost as a major problem. Although the cost of sequencing has decreased significantly, the subsequent computational and manual analyses as well as data storage are time-consuming and expensive.

To re-evaluate and improve treatment strategies, a genotype-phenotype database that records clinical phenotypes is necessary. But information must be obtained from the patients’ medical records, entries into which are not standardized in terms of language, rendering analysis difficult.

All the speakers agreed with Lin, who said that we are “all in the learning phase”. Dinger emphasized the need for international collaboration and engagement with other consortia to provide “strength in numbers”. The unifying goal would be the incorporation of genomics in routine clinical care, which he predicts would enable a shift to personal and precision medicine and “a system that is primarily geared toward health optimization rather than crisis management”.

Too Much to Ignore: Anne Wojcicki's Plan for Health Care and Big Data

BY ALLISON PROFFITT | MAY 23, 2013

STANFORD, CA—The challenge in healthcare is to change what is—and what isn't—a billable question, said 23andMe founder Ann Wojcicki, giving the opening keynote yesterday at the Big Data in Biomedicine conference at Stanford University.

Formerly, Wojcicki spent time on Wall Street making investments in healthcare and the formula was simple: disease makes money. But that presented a "real disconnect" with "my moral compass", Wojcicki said. She says 23andMe was built with the "rebellious attitude" of serving first the individual.

The response to personal genomics toggles between, "There's nothing meaningful yet!" and "Oh it's all so meaningful, how could you give this information out without an incredible amount of supervision!" Wojcicki said. And in fact, both are right. "Genetics won't be everything, but it is a useful tool," she said.

23andMe was built to capitalize on two human trends: the inclination to share quite a lot about ourselves via Facebook and YouTube and the desire to help others, to be active in causes like the Livestrong movement and the Susan G. Komen events.

It was a good bet. 23andMe offers its customers genotyping, not sequencing. "Sequencing is incredibly important," Wojcicki acknowledged, "but not for the customer. Genotyping is phenomenally accurate, and it's enough for the customer."

When customers visit the site, they are also offered plenty of opportunities to participate in research via surveys. 81% of our customers answer more than 10 research questions, Wojcicki said. Users don't visit the site daily, but when they do, most interact with the research opportunities available.

"We are collecting more than 1 million phenotypic data points each week," she said. And that data store, Wojcicki said, is the future of healthcare. "You can just run a query."



Ann Wojcicki, Founder, 23AndMe

The result is real time research, and access to an engaged patient population. For example, a Parkinson's researcher stumbled across a non symptomatic patient with both a LRRK2 genetic variant (associated with Parkinson's disease) and a GBA variant (associated with Gaucher's Disease). The researcher wanted to study what the combination might mean, and turned to 23andMe to find other individuals with both variants. The database indicated that 17 23andMe customers carried both variants. Within 36 hours, 8 had agreed to participate in the research and to having a punch biopsy.

"Where else could you find these incredibly rare people who are willing to do an invasive procedure for free, and say 'Yes, I want to participate in research!'" Wojcicki said. "I think it taps into that Susan G. Komen-like enthusiasm. People want to help!"

“We say, ‘Hey, you actually did this! That’s great!’ and that’s the kind of cycle we want. Instead of having research kind of hidden from people, we want to really engage them.”

What people don't want, Wojcicki emphasized, is to be treated like a human subject. Wojcicki said that 23andMe is completely transparent in its research practices. Patients who participate in research are sent any resulting scientific paper.

"We say, 'Hey, you actually did this! That's great!' and that's the kind of cycle we want. Instead of having research kind of hidden from people, we want to really engage them."

The company dropped the price of genotyping to \$99 in hopes of getting one million customers by the end of the year. The one million mark is the point Wojcicki expects to be "really truly disruptive."

"If I have a million people walking around town with their genetic data, it creates a little bit of chaos."

Wojcicki's goal is to, "create so much data that you cannot ignore this anymore."



View all the **Inside BioIT World** eBooks



Data Management and the Cloud



Clinical Genomics & Diagnostics



Data Visualization and Imaging



NGS Updates



Open Science



The New Clinical Trial



Knowledge Management

Visit: www.Bio-ITWorld.com



◀ Inside
Bio·IT World

The Quarterly eBook of Bio-IT World's Most Trending Articles

Stay tuned for:

Evolutions in Next
Gen Sequencing

▶ Part II: Genomics in the Clinic