Inside Bio-IT World

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

Evolutions in

Next Gen Sequencing

Part II: Genomics in the Clinic

Sponsored by

EMC® ISILON

Produced by

Cambridge Healthtech Media Group

Evolutions in Next Gen Sequencing

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

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

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

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

SCIENCE WRITER
Aaron Krol (781) 972-1341
akrol@healthtech.com

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

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

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

MANAGER, BUSINESS DEVELOPMENT
Katelin Fitzgerald (781) 972-5458
Lead Gen Companies A-K
kfitzgerald@healthtech.com

MANAGER, BUSINESS DEVELOPMENT
Elizabeth Lemelin (781) 972-1342
Lead Gen Companies L-Z
elemelin@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

Key

Look for the

Click on this icon at the end of each article 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

The Clinic Question

> "As the sequencing market has matured, there has been a rapid proliferation of new clinical sequencing startups looking to please a whole host of users."

Getting genomics to work successfully in the clinic is, as Nathan Pearson points out, "easy to discuss but hard to fix."

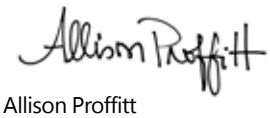
As the sequencing market has matured, there has been a rapid proliferation of new clinical sequencing startups looking to please a whole host of users: physicians, genomicists, bioinformaticians, patients and payors. Cancer and rare childhood diseases are the preferred clinical areas so far.

As an early mover, Foundation Medicine's FoundationOne assay has been validated. Other groups are making strong showings as well. The Broad Institute has launched a CLIA lab; Molecular Health has entered the US market from Germany; and Real Time Genomics has moved into Mendelian disease testing.

Partnerships are becoming increasingly important. Foundation Medicine has several including one with Memorial Sloan Kettering Cancer Center. GenoSpace out of Dana Farber, PathGroup and Thomson Reuters announced an expansion of PathGroup's SmartGenomics product to include Thomson Reuters' information on genes, variants and therapeutic implications, and GenoSpace's advanced analytical and information integration capabilities.

At the 2013 Consumer Genetics Conference, a new class of consumer genomics companies were highlighted including GENEWIZ announcing its "me better" version of the FoundationOne test and uBiome with its crowd-sourced microbiome tests. And of course the ethical issues surrounding clinical sequencing and data disclosure were debated and discussed.

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 newest players in the market and insight into some of the emerging trends.



Allison Proffitt
Editorial Director

About Our Sponsor:

EMC Isilon: A Leader in Life Science Workflows

EMC® ISILON

EMC Isilon is a global leader and trusted partner in providing scale-out storage to our Life Sciences customers. We deliver powerful, simple solutions for organizations that want to manage data, not storage.

As a leader and trusted partner at more than 250 Life Science organizations worldwide, EMC Isilon enables you to extend your analytics workflows across the entire Life Sciences R&D life cycle. The EMC Isilon single storage system provides the simplicity, high availability, and scalability needed to cost-effectively manage life sciences workflows today and in the future.

EMC Isilon is fully committed to advances in application development—including supporting the trend to incorporate Hadoop into evolving Life Sciences applications. EMC Isilon is the only scale-out NAS platform natively integrated with the Hadoop Distributed File System (HDFS). Using HDFS as an over-the-wire protocol, you can deploy a powerful, efficient, and flexible Big Data storage and analytics ecosystem.

Isilon storage and analytics solutions support multiple instances of Apache Hadoop distributions from different vendors simultaneously—including Pivotal HD, Cloudera CHD, and Hortonworks Data Platform. Our solutions also support both HDFS 1.0 and HDFS 2.0. This allows you to leverage the specific tools you need for each of your unstructured data analytics projects.

Isilon's in-place analytics approach eliminates the need to invest in a standalone Hadoop infrastructure. Our solution also allows you to eliminate the time and resources required to replicate your data into a separate infrastructure. This means that you can initiate data analytics projects faster and get results in a matter of minutes. And when your data changes, simply rerun the job with no re-ingest requirement.

Isilon's in-place analytics approach eliminates the need to invest in a standalone Hadoop infrastructure.

About EMC Corporation

EMC Corporation is a global leader in enabling businesses and service providers to transform their operations and deliver IT as a service. Fundamental to this transformation is cloud computing. Through innovative products and services, EMC accelerates the journey to cloud computing, helping IT departments to store, manage, protect and analyze their most valuable asset—information—in a more agile, trusted and cost-efficient way. Additional information about EMC can be found at www.EMC.com

LIFE SCIENCES AT RENCI

BIG DATA IT TO MANAGE, DECIPHER, AND INFORM

Learn how the Renaissance Computing Institute (RENCI) of the University of North Carolina uses EMC Isilon scale-out NAS storage, Intel processor and system technology, and iRODS-based data management to tackle Big Data processing, Hadoop-based analytics, security and privacy challenges in research and clinical genomics.

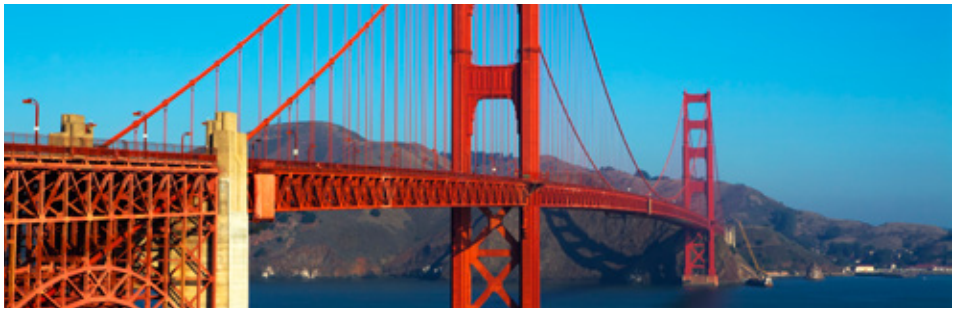
Go here to read the full story:

www.emc.com/lifesciences-unc

Three Small Steps Toward Genomically Sensible Healthcare

BY NATHANIEL PEARSON | AUGUST 27, 2013

A talk at the Clinical Genome Conference resonated with some folks, who suggested sharing it. Crucially, that crowd included doctors, who have much to both teach and learn in the brave new world of genomic medicine. With them on hand, the day's session loosely echoed a grand rounds, where the case was, soberingly, the tall order of making genomes widely useful in healthcare.



The 2013 Clinical Genomics Conference in San Francisco

In such circles, a genomicist like me is a narrow specialist. Bigger speakers — general practitioners, so to speak — were lined up to cover the chronic, integrative needs on the conference bingo card: convening stakeholders; establishing standards for reporting and payment; managing big data…a wordsquall that looms over our field, easy to discuss but hard to fix.

So rather than tackle such broad challenges, my talk stayed bite-size. Building from insights on genome structure, function, and variation, it urged three small but concrete ways to help put genome-informed healthcare on firmer footing.

- Use different reference genomes to align a person's raw data (pick reference(s) most like her/him) versus store her/his finished genome (as clear or potential differences from the human ancestral reference).
- Clinically classify genotypes, not variants.
- Filter a genome against other individuated genomes, not allele frequency tables.

Though these ideas aren't new, they would break convention — so need justifying. But even if you skip the explanations that follow, know that the proposals reflect long thought on how current convention, rooted in sparse data, will ultimately fail for millions of whole human genomes. Thus consider them early course tweaks that can save bigger tacks later, en route to genomically informed healthcare for all of us.

Small step I: The right reference genome(s)

A reference genome — we'll just say reference — is a long string of letters used as a common template for comparing the genomes of closely related organisms, such as people. As an archetype, a reference often shortens and simplifies real genomes,[1] to help read, write, and interpret them.

In teasing apart these tasks, note that today we use the same human reference for all three…and that it's right for none of them. Below we'll see why, and what we should do about it. But if you're rushed, here's the gist:

The current single reference is arbitrary and ethnocentric; inevitably misaligns most people's raw data; and is poor for writing and interpreting genomes afterward, because it includes rare and risky variants, and muddles summary insights on data quality and evolution.

An alternative made of just common or putatively healthy variants would still be unreliable for aligning raw data, and as a foil for writing and interpreting genomes.

Instead, we should read your genome by aligning raw data to references most like you (we can usually guess which). We should then write all our genomes against the human ancestral reference — a solution that's ethnically neutral, straightforwardly informative on data quality and evolution, and stabler than alternatives.

And we should give up on using any reference to proxy an idealized healthy genome. As later posts will detail, reliable health insight will instead require comparing your genome to the individuated whole genomes of many other people who, like each of us, get some diseases and not others.

Ok, now let's walk through those reference tasks in detail, to better understand why we must do them differently.

Read

To read your genome — that is, to make out the long eye chart of letters that form it — a modern sequencer streams zillions of DNA snippets, each copying a chromosome tract roughly at random. By comparing each snippet to a good reference, a computer can find where it best fits, much as we match jigsaw puzzle pieces to the picture on the box. As snippets pile up, the computer surveys what DNA letter(s) amass over each spot, to guess what letter(s) your chromosomes carry there.[2]

Conventionally, we've taken a one-size-fits-all approach to this task of aligning snippets, using the same reference, called Hg# (where today # = 19), to scaffold everyone's genomes. But Hg# wasn't carved in stone. Instead, it's quilted from several real people's genomes that were read by costly, reference-free methods. And the haphazardly picked people who contributed to it have their own ancestry, which gives Hg# their genetic quirks.

As a result, some human genomes are more like Hg# than others. And if my genome resembles it more than yours does, my snippets will, on average, align more reliably. Conversely, because big populations tend to be genetically diverse,

Hg# — like any single option — inevitably misaligns raw data from most people's genomes, in ways both big (mutual gaps and rearrangements wreak havoc) and small (clustered small differences leave good snippets unaligned).

In the end, this means that we can best read your genome today by first aligning to the already available genome(s) most like it.[3]

Happily, skimming your genome — or even just looking at you — strongly hints whose genome(s) might work best.4 How well we can play reference sommelier depends on what options are on hand (more and more, starting with synthetic references that proxy what's common in a particular part of the world), and how saliently mixed your recent ancestry is. But if needed we can try multiple references, and see which work best for which snippets.

And that raises two deeper points. First, for some genome segments, such as a stunningly diverse and health-relevant stretch of chromosome 6, it's hard to predict what your genome looks like regardless of where your forebears came from. For such segments, it makes sense to always align your snippets to many reference options.5 Doing so takes a few more electrons, but usefully sharpens the resulting picture of your genome.

Second, aligning your snippets to even one whole real genome would itself be like aligning them to two versions of a conventional reference (with its one copy of each chromosome that's paired in real genomes). Smartly, that could fully leverage new algorithms that track everywhere a snippet decently fits during alignment, rather than just picking one spot (often by tossup). And that, in turn, would let us read your genome more finely, without — yet — needing the compact simplicity of a conventional reference like Hg#.

Which brings us to the next use of a reference…

Write

After we read your genome in detail, a reference helps write it. Namely, because copies of a given human chromosome are all grossly alike, we can thriftily store yours by just noting where one or both mismatch a simple (single-copy) reference, or were read too poorly to tell.

Everywhere else — typically, >95% of the currently sequenceable parts of human chromosomes — we can assume that your copies match that reference. And because many of your poorly read sites will themselves clump in compressible tracts, we can shrink your genome >>20-fold in the end. That saves memory, of course, but also helps us query it — most usefully, by comparing your DNA (plus your phenotypes, ideally) to others'.

But there's a catch. Because mutation anywhere on a chromosome can make it longer than some other copy, genomes can best be compared if stored as differences from the same reference, so their mapping coordinates match. That way, like sailors agreeing to track longitude from Greenwich, we can neatly record findings like 'One of your chromosome 7's shows five more bases (ACGTA) than mine at reference site 1000; but one of mine shows three fewer bases at reference sites 2000-2002'…

Note the dilemma here: to read genomes, we should align their snippets to various most-appropriate real reference(s); but to compare them, we should write them as differences from the same simple reference.6 Bottom line, we need task-specific references.

But that still means picking one best reference for writing genomes. Given that so much work has gone into Hg#, we might ask whether it's the right one. Which leads us to the third use of references…

Interpret

After shrinking your genome to a list of differences from a reference, we'd like to understand that list — what it says about how sequencing went and, more importantly, about you. We might even hope to use the reference to proxy a healthy genome, so that anything worrisome in your genome stands out from it.

Alas Hg# makes a poor interpretive foil for real genome data, starting with quality control: because Hg# comes from a few modern people, it's poor not just for aligning, but also for writing, where it can conflate statistical signatures of lab-bench problems (sample contamination, chemistry failure, &c.) with those of ancestry. QC that first compares heterozygosity of particular genome segments, rather than just counting reference sites called with any mismatch, will help there (an issue for another day…), but the problems with the current reference go deeper.

In particular, Hg# also includes many variants already implicated in diseases — which means it won't always flag your own worrisome DNA spellings7, and that it troublesomely differs from some single-gene references familiar to clinical geneticists. Moreover, Hg# includes many other variants that, while not yet well studied, are suspiciously rare enough to be harmful too.

Given those shortcomings, many have suggested replacing Hg#'s rare and/or known risky variants with common and/or healthy alternatives, ostensibly yielding a new reference that reliably proxies a healthy, normal genome.

Alas, that won't work, for two reasons. What's common varies. And what's healthy depends.

Informative beats (un)healthy

At first glance, one of your DNA spelling variants may be rare enough on earth overall to intrigue us — but turn out to be boringly common among millions of mostly healthy people in some small patch of the planet. More profoundly, the commonest variant at a genomic site today or in five years may not be the commonest one next year or in ten years. That's evolution — and it means that a common-only reference is inherently unstable.

On the health side, meanwhile, many variants aren't simply good or bad. Their effects depend on what how many copies you have (0, 1, or 2), what disease we ask about, and what other variants lurk in your genome.

You may know a few such twists already. One or two copies of T here help avoid malaria and high cholesterol — but two copies leave you with crippling anemia. One copy of A over there can drive breast cancer, but mainly if you also lack a working copy of the SRY gene (which, on the flipside, helps you avoid testicular cancer, among other diseases…). And so forth.

Data from billions of us will unfurl more astounding complexity, where variants throughout your genome — some inevitably present in any reference we use — interact in surprising ways with each other, and with habits and other factors, to favor some diseases and disfavor others.8 Other posts will further explore how this hard truth should alter our approach to genomic healthcare. Here, it simply dooms any hope of using any reference to reliably proxy what's healthy.

And more deeply, using a reference like Hg# as an interpretive yardstick also obscures how genomes change and, by extension, how various kinds of changes tend to affect health in the first place. Hg# can't, for example, tell us whether a so-called deletion in your genome (where it's missing a tract found in Hg#) really reflects a mutation that deleted bases in you or your forebear, or instead reflects an insertion of bases in someone who contributed to Hg#.

As such, because a given letter in a reference like Hg# could itself reflect a past mutation, writing everyone's genomes as differences from Hg# makes statistical questions like 'How often does

the snippet CG mutate to TG? And how well does that TG survive, over generations, if it changed a protein's arginine to cysteine?' trickier than they should be.

Such questions matter. They can unlock basic physiology (How do mutations happen? Why do tumors correct them so poorly?); hint how a new variant may affect health (Does changing active-site arginine to cysteine often make an enzyme fail?); and clarify how variants interact with each other, and with habits, to cause disease (Why do some genetic variants, like APOE4, make us sick but leave chimpanzees healthy?).

Those big questions require the big data inside us. Even if no more than a handful of your DNA spellings alter your own healthcare, the rest of them, pooled with similar data from all of us, can shed light on many diseases to greatly refine care for our grandkids.

But using a conventional reference like Hg# needlessly hinders that effort. So while we must abandon the idea of any reference reliably proxying a healthy genome, can we at least find a sensible reference to write and compare the coming flood of genomes, to catalyze those deeper insights?

An ancestral reference

We can. The sensible yardstick for writing your genome is the human ancestral reference (HAR) — that is, a single-copy genome comprising, at each chromosomal site, the DNA letter carried by the last common ancestor of all people for that site.

In picturing the HAR, note two things. First, Suganthi Balasubramanian and colleagues have already built (and used) it, nearly site for site9, by comparing our genomes to those of other great apes. Second, two genomic sites can trace to different last common ancestors. That's because, when eggs and sperm are made, chromosomes pair up, swap segments, and move into different cells. Each copy of a chromosome thus quilts together pieces of earlier copies; so everyone's last common ancestor for site 1000 may not be our last common ancestor for site 1001 (they may have even lived eons apart). Which also means it's implausible that any person ever carried the whole HAR.10

Among reference options for writing and comparing our genomes, the HAR uniquely combines several appealing features:

- *It's neutral.* As noted, no one ever carried the whole HAR. And because the mutations that distinguish our genomes from it have struck roughly randomly among our ancestors, your genome resembles it about as much as mine does.[11]

In this important sense, the HAR belongs to none of us, and to all of us. Being roughly equidistant from everyone, it offers a uniform, non-ethnocentric baseline for assessing sequencing quality, and for reporting what's genetically distinctive about you.

- *It's stable.* The HAR actually looks a lot like a common-variants-only reference, because nearly all ancestral variants are common. But while a common-only reference would in principle need many edits each year to stay perfectly accurate, the HAR would need just one or two (as atypically rare ancestral variants go extinct).[12]

Such editing isn't urgent, because few variants with allele frequency near 50% are functionally intriguing enough, or surveyed precisely enough in the population, to day-trade anyway. But that just makes it even smarter to build a reference on the stable, reliably inferrable, and meaningful criterion that a variant be ancestral, rather than worry whether its allele frequency fell to 49.9%. That way, we get summary insights even from otherwise boring variants — and a low-maintenance reference to boot.

- *It's compact but comprehensive.* Like conventional references, the HAR is a simple single-copy (haploid) genome. Real genomes, compressed against it, would yield files consistently intermediate in size between the biggest and smallest files compressed against Hg#.[13]

Nonetheless, because new chunks of DNA are usually copied from chunks elsewhere in the same genome, the HAR includes source DNA for nearly all chunks of real human genomes (missing only those recently copied from viruses or bacteria, or other oddities). Other reference options tend to be less comprehensive on these counts, which poses an ongoing dilemma of when to add a segmental copy (to make them more thorough), versus omit it (to keep them compact).[14]

That dilemma would still apply, but the HAR offers a framework for handling such segments that we choose to include in the extended HAR that Subramanian et al. proposed. For a newly arisen extra segment that some but not all people have, variation among such copies could in turn be mapped to common coordinates in the inferred earliest (nearest to universally ancestral) version of the new copy.

- *It's directly informative.* Most importantly, the HAR is the only reference option that directly shows how human genomes change. As a foil for writing all our genomes, it would thus most quickly reveal summary patterns of change that in turn shed light on basic biology and health.

Concretely, if shortening one bend in a protein makes people sick, but lengthening it — or shortening another bend — doesn't, the HAR would let clinical geneticists reliably spot this faster than other references would.

The benefits of an ancestral reference for making sense of genomes, both individually and together, have long been starkly clear for geneticists studying the first fully sequenced human chromosome: the mitochondrial genome (mtDNA).

Starting in 1981, we used the first sequence of a person's mtDNA as a reference. That sequence forms a leaf on the simple evolutionary tree that binds all our mtDNA versions. And because each of us gets only our mom's version of this short but gene-rich chromosome, with no backup from dad, that tree's branches are key foci of health research.

But using a modern person's mtDNA as a reference meant treating a leaf as if it were the treetrunk. Like a concave mirror, this flipped and warped our view of the tree, prompting epicycle-like contortions to figure out where your leaf was, and how your branch may or may not have mutated in telling ways.

In 2012, researchers cut that gordian knot, proposing the human ancestral mtDNA as a reference for writing real genomes. That new reference lets you easily a) find your mtDNA leaf, and b) see how DNA has changed throughout the tree, to better understand key biological processes.

Having learned the hard way with mtDNA, we needn't wait 31 years for our other chromosomes. In the end, by using multiple references to align raw data, and adopting the HAR to write and compare our finished genomes, we can best read, write, and learn from the millions of human genomes soon to be sequenced.

Getting smarter with references

So where are we, as a community, on human reference genomes?

There's modest reason to hope. Researchers have begun to show how much better we might read genomes by aligning snippets to similar reference(s); that an ancestral reference helps compare genomes more easily and informatively; that Hg# doesn't proxy a healthy genome; and that no alternative would reliably do so either.

On the practical side, we're accumulating diverse, ever better sequenced human genomes that can serve well as alignment references (and, as a bonus, help benchmark new sequencing methods). And we're getting better genomes from elsewhere in the great ape family tree, to refine the HAR.

Moreover, today's de facto standard, Hg#, continues to improve via thoughtful work by Deanna Church's team and others. Beyond fixing errors and filling in previously missing segments, the pending Hg20 version will include multiple versions of more segments, in part to better align raw data.

That's a sensible stopgap, until more folks start picking from multiple alignment references from the start. But adding alternate versions of more segments to Hg# requires ongoing arbitrary choices, slows the task of writing finished genomes, and tends to statistically weaken comparisons of many genomes. The latter jobs are really better served by writing genomes against the HAR.

Communal habits like using Hg# for all human reference needs are hard to break — even for open-minded scientists (and maybe moreso in famously stubborn medicine). But given the clear flaws in our current approach to reference genomes, it's likely better to break those bad habits now than let them entrench further, as we start sequencing patients' genomes by the thousands (and more).

Making all those genomes useful in healthcare, for us and future generations, will mean reading them well; writing them efficiently; and, as coming posts will explore further, interpreting them wisely.

All these goals rest on the bedrock of reference genomes. Let's get them right.

Nathaniel Pearson is Principal Genome Scientist at Ingenuity Systems. Previously he served as senior director of science and research at Knome. He blogs at genomena.com. This piece was also posted on his blog on August 26.

---

[1] Today's ~2.9 billion-letter human reference, for example, comprises just one version of each of the distinct-looking molecules (chromosomes 1-22, X, Y, and M) in the >6.5 billion-letter genome of a man's skin cell. That cell's genome comprises two copies of most such chromosomes — and those copies, in turn, differ in chemical makeup (base sequence), and include tracts that have never been seen (or successfully sequenced), so are simply missing from the reference.

[2] Importantly, many snippets from your genome differ from even their best-matching parts of the puzzlebox picture (otherwise, why bother sequencing?). But the reference template still helps piece them together faster than we otherwise could. And by piling many snippets over each site, we can tune out errors from cooking finicky chemicals under tiny image sensors — a bit like how astronomers, at the other end of the spatial scale, distinguish lasting light sources from noise by overlaying many pictures of the same part of the sky.

[3] That point was moot in 2003, when we had to use the hardwon sequence that became the current reference. But since then, we're bootstrapping our way to good sequences of many human genomes from around the world — a pool that we should tap to better align newly sequenced genomes, as some folks have already shown.

[4] Ideally, we'd use parents' genomes to align those of their kids…but when sequencing is common enough to make that practical, sequencers will likely make longer snippets that are easier to piece together from the start anyway, even without aligning to a reference.

[5] Helpfully, Hg# itself includes several options for some such segments — and those who built and refine it plan to add some more.

[6] Note that even though we write them as differences from a simple reference, which has just one copy of each chromosome, we can still keep track of which spellings go together on each copy of your chromosomes (if our sequencing method was good enough to tell in the first place).

[7] Especially if whoever compressed your genome didn't bother noting where your genome was too poorly sequenced to know what it carries — a corner that geneticists too often cut.

[8] Such insights stand to turn much of the noise that we currently sweep under the rug of partial penetrance into far better understood signal — think, for example, about how genetic insight turned the apparent random noise of why a baby was born female or male into causal signal tracing largely to the sex chromosomes.

[9] For the remaining sites, we can't reliably guess what variant our last common ancestor carried, because the state of variation we see among our copies extends to other great apes, suggesting that such variation has lasted too long to reliably unravel. In extreme cases, like the sex chromosomes, such lasting variation is already enshrined in the current reference (Hg# has one X sequence and one Y sequence, despite the fact that not everyone has the latter).

[10] Even the last common ancestors who contributed to the HAR had variants in their own genomes that aren't in it.

[11] Planetwide, we have no idea whose genome happens to differ from it most, though that person — ironically, in some sense the most evolved (geneticists would say derived) of us — is almost certainly very sick, thanks to gross genetic changes…

[12] Many thanks to Graham Coop and Justin Fay for helping think through the relevant numbers here.

[13] Average compression is the main measure that could instead be optimized by a common-only reference. But the HAR has several substantive advantages over that less stable and informative option.

[14] Note, btw, that this question matters most if we're using a reference to align snippets — which we're not proposing here. But we do need to map each of the alignment references themselves to the writing/comparing reference, which is where it helps to make sure the latter includes source DNA for the segments that those alignment references may have extra or fewer copies of.

Broad Institute Launches CLIA Lab and Exome Sequencing

BY ALLISON PROFFITT | OCTOBER 20, 2013

The Broad Institute today announced that it has passed Massachusetts state inspection, and can begin processing clinical samples as a CLIA-certified lab.

"We're trying to leverage and make best use of the infrastructure and expertise that we've established over time in sequencing and data processing and data management and trying to apply this in the clinical setting," Stacey Gabriel, director of the Broad's Genomics Platform, told Bio-IT World."

The announcement is a critical step in establishing a Clinical Research Sequencing Platform (CRSP) at the Broad.

The CLIA lab will make use of the Broad's 50 HiSeq 2500 instruments and will offer exome sequencing, returning a "technical exome report," Gabriel said. She expects the lab to process 100-200 samples per week to start, though she says the Broad is well-equipped to scale that number up if there is sufficient interest.

Gabriel said she expects the Broad's customers to be, "people who are performing clinical research studies—academic medical centers, other institutions including biotech or pharma groups who need to have this level of CLIA certification for their sequencing studies." She mentioned specifically NIH research grants now calling for CLIA-certified sequencing.

The Broad will not be offering interpretation, just, "the very best set of exomes and variant calls," Gabriel said. Partners will deliver or interpret results and inform decisions regarding patient care, including the care of patients with rare diseases. Gabriel said that the Broad is not interested in a direct-to-consumer offering.

Gabriel does expect the CRSP offerings to expand soon though. "We're developing a cancer panel, which will have 400 genes and other regions of interest for cancer genomics. We'd offer that for tumor-normal sequencing. That will probably be available early next year," she said. "We've also got to decide about activities like whole genome sequencing, or transcriptome sequencing."

"Developing and applying genomic methods that advance medicine is central to the mission of the Broad Institute," said David Altshuler, deputy director and chief academic officer of the Broad Institute in a statement. "Right now, there is a pressing need for technology development and clinical research that enable learning about genome sequencing in the clinic. Working with partners, CRSP will contribute to the efforts by the greater medical and scientific community to build the knowledgebase needed to evaluate and establish the clinical utility of genomic information."

Gabriel says that the program is a logical progression from the Broad's research work.

"I think people don't really think of us in a clinical setting or a CLIA setting. We're more of a big research operation," she said. "We are really trying to couple these things together and apply what we're learning in research directly to the clinical data generation. We hope that people are interested to work with us."

Foundation Medicine Reports Validation of FoundationOne Assay

BY BIO-IT WORLD STAFF | OCTOBER 21, 2013

Foundation Medicine has announced findings from a 24-month, multi-institution collaboration demonstrating the analytic validation of its FoundationOne cancer genomics assay. The results were published in the online edition of Nature Biotechnology.

FoundationOne characterizes all classes of molecular alterations (base substitutions, short insertions and deletions, copy number alterations and select rearrangements) across 287 cancer-related genes from routine formalin-fixed, paraffin-embedded (FFPE) clinical specimens. The publication describes clinical application of this assay across 2,221 consecutive patient cases.

The publication applied and extend the guidelines established by the Next-Generation Sequencing: Standardization of Clinical Testing (Nex-StoCT) workgroup to validate a clinical sequencing-based assay for cancer, therefore setting the standard for validation of targeted NGS in cancer.

"Clinical cancer care is undergoing a fundamental shift toward treating patients based on the specific molecular drivers of their disease, and a sequencing-based diagnostic assay that comprehensively and accurately characterizes the genomic alterations occurring within an individual's tumor is essential for the implementation of this therapeutic strategy," said Lajos Pusztai, M.D., co-director of the Cancer Genetics and Genomics Research Program at Yale Cancer Center and co-author of the study in a statement. "This study is instrumental in establishing the technical validity of next-generation sequencing in the clinic and enables the practice of precision medicine wherein the molecular characterization of a patient's tumor informs the patient's individual treatment."

Foundation Medicine assessed the accuracy and precision of FoundationOne using reference samples of pooled cell lines and hundreds of clinical cancer specimens with diagnostic testing results generated by established clinical assays. FoundationOne was found to be highly accurate in identifying genomic alterations, including sensitivity greater than 99% for detection of base substitutions, 98% for detection of insertions and deletions, and greater than 95% for detection of copy number alterations, while maintaining greater than 99% specificity, reports the company.

Application of FoundationOne to 2,221 clinical cases revealed clinically actionable alterations in 76% of tumor samples, three times the number of actionable alterations detected by other currently available diagnostic tests. Alterations are defined as clinically actionable if linked to an FDA approved targeted therapy in the tumor under study or another solid tumor, a known or suspected contraindication to a given therapy, or an open clinical trial for which the alteration confers patient eligibility.

"FoundationOne was proven to have the sensitivity and specificity required for routine clinical practice, and it identified more than three times the clinically actionable alterations that are identifiable using a collection of six commercially available and commonly used diagnostic tests, including the other most common NGS-based tests. This comprehensive approach directly translates into more treatment options for patients," said Michael J. Pellini, M.D., president and chief executive officer of Foundation Medicine. "We believe this study establishes the standard for analytic performance that is required for patients with cancer to benefit from the clinical application of next-generation sequencing of their tumors."

Read more:

Frampton, G.M. et al. Validation and clinical application of a cancer genomic profiling test using next-generation sequencing. Nature Biotechnology, 2013; DOI: 10.1038/NBT.2696.

Gargis, A.S. et al. Assuring the quality of next-generation sequencing in clinical laboratory practice. Nature Biotechnology 30, 1033-1036 (2012).

MolecularHealth Enters the American Cancer Genomics Market

BY AARON KROL | OCTOBER 11, 2013

Cancer care is growing more particularized as the disease becomes better understood. The national burden of cancer is enormous – with more than a million and a half new cases diagnosed every year – but that figure obscures a highly heterogeneous set of conditions as distinct from one another as the people they afflict.

The unique genetics of each case present a moving target for oncologists and molecular pathologists. "Anyone in this field knows that the whole genomic aspect of cancer is rapidly developing," Dr. Lloyd Everson, a former vice chairman of U.S. Oncology with four decades' experience in cancer care, told Bio-IT World. "It is the future of medicine."

Dr. Everson is part of a team of four specialists who left U.S. Oncology in September to join the management team of MolecularHealth's new North American branch in The Woodlands, Texas. MolecularHealth, founded in 2004 and based in Heidelberg, Germany, is preparing to launch a personalized, direct-to-consumer genomic service for cancer patients, sequencing the whole exomes of their tumors and recommending treatment based on the best available research. As CEO of the company's U.S. branch, Dr. Everson is charged with putting this ambitious diagnostic program into practice in the world's largest market.

Genomic cancer diagnostics is a multi-tiered enterprise. First, a representative sample must be extracted and sequenced – already a complex task due to the amount of genetic variety that may be present even in a single tumor. Still, rapid progress in lowering the timeframe and cost of sequencing makes this step far more practicable than it would have been just a few years ago. Depending on the genetic panel desired, Dr. Everson expects MolecularHealth to turn around samples in as little as 12 hours. "There's been a lot of activity, a lot of research and development, in the actual gene sequencing environment," he says. "But where the real opportunity and challenge lies is in the bioinformatics area of putting this information to use for the patient."

That's the second step: analyzing all the data generated and mining it for expected outcomes, treatment responses and drug associations. Despite extensive research on cancer genetics, a large amount of subjective judgment remains in deciding which information is actionable, and what the best course of treatment for a cancer with a given profile should be. This is perhaps the greatest opportunity for a cancer diagnostics company to distinguish itself from competitors, by creating a track record of effective recommendations, uncovering treatment options that physicians might not have found on their own – and acting responsibly in determining when an obscure study offers a viable course of action. "When you do a whole exome on [cancerous cells]," says Dr. Everson, "you find a lot of things that aren't necessarily actionable in the current sense of the term, but do find associations in preclinical, Phase I and other studies with drugs that may indeed be very targeted to those gene variations. It's a moving field." Keeping pace with that field will demand constant reevaluation of MolecularHealth's metrics for tying variants to treatments, as well as a tumor DNA bank so that results can be reinterpreted in the light of new research.

Of course, the best information in the world won't help patients and their physicians fight cancer if it isn't formatted in a useable way. The third piece of the diagnostic pipeline is delivering the right results to the right end users. MolecularHealth is developing three separate modules for reporting results: a detailed report for molecular pathologists that includes both findings and test procedures; a one-page summary for oncologists with access to the relevant research; and a tool for the patients themselves. This last represents a delicate balance of good customer service and responsible medicine. "We want to make sure that if and when our patients access their own material, they have an educational environment that explains what this all means," says Dr. Everson.

If MolecularHealth is successful, the company will be generating a huge database of genetic cancer profiles. That information could be of invaluable use to research partners like the MD Anderson Center, with whom the company intends to share its data on genetic variants, treatments and outcomes. MolecularHealth also has an exclusive partnership with the FDA to exchange information on adverse drug responses, integrating the FDA's existing Adverse Event Reporting System with the company's more detailed molecular analysis of drug interactions in the body. The resulting tool, called the Molecular Analysis of Side Effect information (MASE), will help alert MolecularHealth when an otherwise recommended pharmaceutical course could be expected to trigger an adverse response, adding another layer of personalization to the company's service. MASE will also provide targets for future research on the molecular pathways that trigger side effects. "This is an absolutely unique tool," Dr. Everson says, which will help MolecularHealth stand out when cancer patients and their families search for healthcare solutions.

MolecularHealth is currently seeking CLIA certification for its laboratory in Texas, and hopes to launch commercial services in the first quarter of 2014. Although integrated genomic cancer diagnostics is an emerging discipline, the company will already be entering a crowded field, with a competitor, Foundation Medicine, staging a successful IPO this September. Both companies have the backing of major players in the commercial IT field: MolecularHealth is supported by Dietmar Hopp, a co-founder of SAP, while Google and Bill Gates are both major investors in Foundation Medicine.

These are "visionary people," says Dr. Everson, looking for ways "to push the envelope and make these kinds of advances accessible to real patients in real time." With whole genome sequencing cheaper and easier than ever, and analytic software catching up to the volume of data involved in genomics, it seems a tipping point has been reached in cancer care, when long-awaited diagnostic tools can finally reach the patients who need them.

Gene Information Directly to Doctors

BY ALLISON PROFFITT | AUGUST 21, 2013

Last week, GenoSpace, PathGroup and Thomson Reuters announced an expansion of PathGroup's SmartGenomics product to include Thomson Reuters' information on genes, variants and therapeutic implications, and GenoSpace's advanced analytical and information integration capabilities.

PathGroup provides anatomic, clinical and molecular pathology services to medical practices and hospitals in a CLIA certified laboratory.

"A physician, presumably an oncologist or general surgeon, could order a tumor profiling testing under our product SmartGenomics," Ben Davis, PathGroup's CEO told Bio-IT World. Tumor specimens are subjected to a whole menu of tests including traditional pathology services and chromosomal cytogenetics as well as next generation profiling panels and arrays. "Basically digital karyotyping," Davis said. "We can offer a broad, comprehensive spectrum of diagnostic services."

What GenoSpace and Thomson Reuters bring to the table is context for the results.

"With next generation sequencing and array CGA panels, what we find today—as the number of genes that are tested expands up to 100s or 1000s, exomes and whole genomes… —here's a massive amount of data that must be processed," Davis said. "GenoSpace provides us with a highly sophisticated algorithmic approach to arriving at a specific variant calls as well as confirming them and putting them in the form of a reporting process that becomes the product itself. In addition, they provide a linkage to the Thomson Reuters Genomic Knowledgebase that allows us then to put the results that we've found through the various testing platforms in context with the individual patient and the global knowledgebase about drugs and therapeutics that are currently available."

"[PathGroup] has worked with GenoSpace and implemented a system that pulls that data together along with reference data that's been provided by Thomson Reuters," explained Joe Donahue, Senior vice president at Thomson Reuters Life Sciences "Out of that system comes the analysis report that will then be approved by pathologists at PathGroup, and sent back to the clinicians"

The resulting report will include not only the patient's test results, but information about drugs and therapeutics that are currently available, actionable targets in an off-label approach, or clinical trials.

Davis said the service will initially be offered to existing clients in PathGroup's market in the Midwest and Southeast, but he hopes to expand to new customers soon. "Depending upon the market response we could offer this on a national or even global basis," he said.

"This is exciting because it really is all about how we can provide more options to the clinicians and the patients with drugs that are on the market, with things that we know are in the pipeline, or existing clinical trials or trials that are shortly to get underway," said Donahue. "Effectively it's going to provide more options to the both the clinicians and the patient going forward… From our standpoint, it's a great fit because of the quality and the depth that the [Thomson Reuters] data has. We think it's a great partnership with both GenoSpace and PathGroup."

Gene By Gene to Acquire Arpeggi

BY ALLISON PROFFITT | AUGUST 8, 2013

Arpeggi announced yesterday its impending acquisition by Gene By Gene, a Houston-based consumer genomics company that develops products for ancestry and genealogy applications. The new company hopes to make available an innovative suite of more affordable genetics testing and diagnostics services to consumers, researchers and healthcare providers.

The new company will keep the Gene By Gene name; financial details were not disclosed.

Arpeggi develops solutions for genome sequencing, data management and computational analysis. In April, the company released GCAT, the Genome Comparison and Analytic Testing, a free community driven platform for evaluating the performance of nextgeneration sequencing (NGS) data analysis methods.

The company is funded by the Startup Health and GE Entrepreneurship Program, and develops proprietary sequencing tools, designed for scale, that enable accurate, fast, and costeffective analysis of genomes.

Gene by Gene is a private company that has developed a cash flow positive business, said David Mittelman of Arpeggi. Gene By Gene's full-feature lab is state of the art with the latest Illumina sequencers and fully robotic sample handling capabilities.

"We were building the best model of the geneome we could build from next gen data and we'd done quite a bit, but the next growth opportunity for us would be a full vertical," Mittleman told Bio-IT World. "We'd raised a seed round, and thought of building our own lab, but that takes money and time. We fell into discussion with Gene by Gene, originally looking for a way to use their data generation capabilities. And it occurred to us that if we merged we'd be more than the sum of the parts."

"The acquisition of Arpeggi's technology and worldclass team of data and technology experts will enable us to accelerate Gene By Gene's plan to make nextgeneration DNA sequencing and clinical genomics accessible and affordable to all," said Max Blankfeld, Managing Partner of Gene by Gene in a statement. "We are on a mission to transform health care by dramatically speeding up the process, and reducing the costs of genetic tests, which today are often far too expensive for the average consumer."

"By forging Gene by Gene's state of the art lab and existing customer reach with Arpeggi's NGS analytics platform we can bring maximum value to the customer," said Jason Wang, CTO of Arpeggi and now Gene By Gene. "We start with blood or saliva as an input and can produce extremely accurate assessments on genomic variation by owning the entire process and optimizing every step along the way. Our goal is to make personal genomics available to everyone by making it more affordable and easier to understand."

Arpeggi was previously based in Austin, Texas, but the entire Arpeggi team and technology platform will be incorporated into Gene By Gene in Houston. Arpeggi's founders will join Gene By Gene's management team, effective immediately. Arpeggi's Nir Leibovich was named Gene By Gene's Chief Business Officer, Jason Wang was named Chief Technology Officer and David Mittelman, Ph.D was named Chief Scientific Officer.

Mittelman stressed that Arpeggi's open data commitment "remains as strong as ever." "We were interested in engaging the community and [with the Gene By Gene acquisition], we can do more now." GCAT will remain a free, fully-supported resource.

Fast Science: Real Time Genomics Moves to Mendelian Diseases

BY ALLISON PROFFITT | AUGUST 7, 2013

Real Time Genomics chose cows for their first proof of principle, but a lot has changed in the 18 months since I sat down in their Hamilton, New Zealand, offices and toured a dairy farmer's cooperative down the road. Today RTG's extremely fast genomics analytics platform is proving itself faster, cheaper and more efficient than the competition for tackling Mendelian genetics.

Real Time Genomics first traced trait heritability in dairy cows

The company now calls San Francisco home, and is headlined by a completely new leadership team and commercial focus. The founding CEO of Real Time Genomics, Graham Gaylard, left the company to attend to family matters and the company began looking for an "industry insider" to steer it into commercial viability, said current CEO Steve Lombardi. With an established biotech pedigree, having served as president and CEO at Helicos BioSciences, senior VP at Affymetrix, and VP of genetic analysis at Applied Biosystems, Lombardi was ready when Real Time Genomics approached him.

"I saw your article on RTG right when they were contacting me," Lombardi told me. "Your article [aligned] very much with my thinking of them, which is that these are a bunch of really, really smart computer scientists, and with a little bit of pointing them in the right direction, their work could be really interesting."

Real Time Genomics first vetted their technology with a local bovine genetics project, but the opportunities were much broader, Lombardi said. "I thought, boy, if this thing is good enough, we could apply it to the Mendelian genetics world. We could really have something!"

Lombardi took the reins in April 2012 (see, Lombardi Lays Out His Vision for Real Time Genomics) and started building his core team. He asked Francisco De La Vega, an Assistant Professor of genetics at Stanford University and previous colleague of Lombardi's, to look at the platform from a computational biology perspective.

"I remember [Francisco] said, 'Steve, this thing looks really good, but there's no validation. We don't know if it's going to work.' And I said, Francisco! That's why we need you!"

De La Vega signed on as Vice President, Genome Science in August. Pam Morley, a former sales exec at Applied Biosystems and Fluidigm, became head of sales in June. Jason Blue-Smith, previously a senior product manager responsible for BaseSpace at Illumina, joined the team in September as Head of Product.

The New Zealand computer scientists now work as part of a wholly-owned subsidiary of RTG. "The team in New Zealand has been absolutely fantastic," Lombardi says. "These guys are real seasoned applied mathematicians and professional computer scientists. They share that wonderful sort of collective ego and sense of urgency that are needed to win."

In the US, Lombardi and his new team have turned their sales, product, and research resources to human applications. RTG has two product lines; the first released to market is a shotgun sequencing-based metagenomics platform, primarily used by customers to "estimate species frequency composition and protein function, to understand the biology at play in a given sample," Blue-Smith says. "The metagenomics was pretty much done before I got here," Lombardi adds. "We have continued to work with a small number of customers doing some amazing things, but we just don't think the market is ready yet for RTG to make a bigger investment in growing that application right now."

Instead the company's area of focus is the genome analysis platform—"RTG Variant" for now—which includes the Family Caller, Population Caller, Singleton Caller, and other potential products. RTG Variant technology delivers variants from raw sequence data or BAM files from large, complex pedigrees, or family trios for highly penetrant single-gene diseases to more complex, adult-onset diseases.

"A unique aspect of the technology is that the data from all available individuals can be jointly analyzed to not simply improve accuracy, but to detect types of variants that you can't get by looking at a single individual," Blue-Smith said. "But as importantly, this simultaneous joint analysis allows RTG to deliver comparable results when less sequence coverage is available. The implication this has on reducing sequencing costs for family and population studies is considerable."

The first human validation performed with RTG using Illumina's Platinum Genomes dataset revealed some impressive numbers. The analysis pipeline proved to be about 7 times faster than a comparable BWA mapping," Lombardi says. "But where the rubber meets the road is in RTG's ability to quickly determine variants. We're 65 times

faster… and we've got patented technology that allows us to do innovative things in the context of identifying actionable variants for different types of human disease."

The newest validation is especially exciting for the field of early childhood disease, Lombardi says.

"Our technology can not only reduce the cost of analysis, but it also can reduce the actual cost of sequencing. The novelty of [the Family Caller] is that you bring all three aligned and mapped genomes into a single caller. What everybody else does is do each person in the trio—the mother, father, and offspring—separately… Because we do all three simultaneously—you're using the mother and father to get as accurate view as you can off the offspring—you can reduce the coverage—i.e. the amount of sequencing you do on the parents—by half and still get the same results."

RTG Family allows the actual sequencing coverage—and reagent costs—to be cut by 1/3. The team believes that's enough to tip the scales for physicians who are stalled by the cost of exome sequencing for their patients.

"When we talk to pediatric clinicians and researchers about the ability to add parents without having to triple the cost, their eyes light up in the hope they'll be able to provide better treatment options for these sick children," said Blue-Smith.

The speed and accuracy both turn on the underlying mathematics. The platform has two components, explains Francisco De La Vega: proprietary algorithms for searching and a Bayesian infrastructure. The searching algorithms drive "really fast" alignments, and the Bayesian infrastructure is key because it deals well with prior information. "In the case of a pedigree, that would be the information that we know about the relationships between individuals and the expectations, for example, of Mendelian segregation."

The infrastructure can integrate the sequencing platform error rates as well, producing a probabilistic quality score. "At the end of the day, the game of variant identification is about producing the right score," De La Vega says. "We are constantly refining the balance between finding the true positives and avoiding the false positives."

The platform is extensible and multi-threaded, De La Vega says. "It allows us to do variant calling with single individuals, allows us to do variant calling in populations, allows us to do pedigree variant calling, and, in fact, allows us to do variant calling in related samples such as tumor/normal samples." Blue-Smith adds that the company soon plans to introduce separate products to deliver on the unique needs of each of these three use cases.

On a commodity server—"maybe $4,000 or $5,000"—Lombardi says a fully-mapped 30x human genome would take GATK about 65 hours to do variant calling. The RTG Variant Platform takes under an hour.

RTG's goal is to be where the data is, says Jason Blue-Smith; the platform is currently deployed on the cloud, on appliances, and on customer's local infrastructure. "It can run on a single server, on a distributed grid, in a public cloud; the point being that whatever IT infrastructure a customer uses we have deployed our products on and are able to integrate seamlessly into their ecosystem. No special hardware needed."

The pricing is similarly flexible, Blue-Smith says. RTG offers pay-as-you-play pricing that scales with volume, requiring no up-front cost.

Lombardi's vision is ultimately for the platform to be cloud-hosted. "My belief is that it's all going to be moving to the cloud. As these NIH budgets get squeezed and health care costs get squeezed, people are going to get out of buying proprietary hardware in labs and they're going to utilize the cloud."

"The idea is to sell this not as a piece of enterprise-wide software—not even really think of it as software—just think of it as a consumable," Lombardi continues. "If you've got a hundred exomes to run, you come to us and buy a hundred exomes. If you're got 500 whole genomes, we'll sell you that. If you've got a long-term idea of what your business is, you can come to us and we'll give you more of a subscription-type of business. So we're trying to be flexible."

Lombardi sees nearly endless commercial opportunities for the RTG platform, but also sees a need for specificity. "The key thing you've got to do with a company is find a market where your technology can really be a winner, and then focus on it," he says.

Lombardi has set his sights clearly on analysis. "Analysis is the new consumable," he says. "Our 100% focus right now is to be a platform that transforms FASTQs into VCFs. When you look at sort of the sequencing value chain that way, you've got the sequencing companies whose main core competency is producing FASTQs, then you've got a lot of companies who are building either interpretation engines or full service CLIA labs to do the whole thing. But there're just a few people who are trying to be the best at analytics."

Lombardi mentions BINA, CLC Bio, and Novalign as having competitive pieces of the puzzle, but reiterates the same market position RTG claimed a year and a half ago.

"Our main competition, not surprisingly, is open source. Thirty-five years ago I was making DNA by hand in a lab; across the hall from us were people who were making enzymes by hand. Because no one would ever think of buying oligos or enzymes; we can do it best! Now you wouldn't even think of it.

"What we're coming now with is a value proposition, just like the DNA synthesis companies and the reagent companies and the sequencing companies and the microarray companies before, with a commercial product that is better," said Lombardi. "We are as accurate or more accurate than the academic software and we're much, much faster, and we're easier to use. We're bringing a professional approach to it."

Lombardi's confidence is supported by the product's reception earlier this year. "We launched the product for the technologists at AGBT in February and got a great response. We went to the American College of Medical Genetics meeting in March and go an even better response."

The buzz has borne out in partnerships as well. In a one month span from mid-April to mid-May, RTG announced partnerships or collaborations with Knome (to integrate the RTG Variant platform on the knoSYS 100 system); the J. Craig Venter Institute (a long-term study looking at the genetic changes that induced pluripotent stem cells may acquire during the process of differentiation); and Omicia (integrating the two platforms into a seamless workflow).

In May the company also announced a $5 million investment to further expand commercial operations. (It is currently backed by funding from Catamount Ventures, Lightspeed Venture Partners, and GeneValue Ltd.)

The funding, "gives us a nice stretch of runway," Lombardi said. "We'll continue to do what we're doing, make further investments in our commercial franchise,"—the company is hiring in marketing, sales, and bioinformatics—"it's only a matter of time until we breakthrough and really get going on this."

GeneInsight: Genetic Knowledge to Action

BY ALLISON PROFFITT | JUNE 6, 2013

Today's biotech grail is surely genomics in the clinic—using sequencing to inform care, treatment, and disease prevention. Implementation is easier said than done, but Partners Healthcare has been doing it since 2005. Its GeneInsight suite of applications was awarded the 2013 Bio-IT World Best Practices Editors' Prize.

Heidi Rehm of Brigham and Women's Hospital and Director of the Laboratory for Molecular Medicine, Partners Healthcare Center for Personalized Genetic Medicine (PCPGM) in Boston has been running a clinical genetics lab for over 10 years. For years the lab used Sanger sequencing, Rehm said, but was able to make major leaps in the volume of testing when it shifted to next generation sequencing a few years ago.

Thankfully, Rehm had been working closely with an IT team led by Sandy Aronson, Executive Director of IT of PCPGM to develop a platform designed to assist labs in, "storing genetic knowledge across genes and diseases and variants and tests in a way that allows data to be structured more efficiently," Rehm says.

The problem isn't a new one, and GeneInsight isn't a new solution.

GeneInsight has been in, "full production clinical use since 2005," says Aronson. "Our Laboratory for Molecular Medicine—[Rehm's lab]—began providing sequence-based tests very quickly after it opened," he says. "When you do sequencing-based tests you start finding these variants of unknown significance on a regular basis and you need mechanisms for dealing with that, and that really was the impetus for building GeneInsight and tracking the data and the knowledge lifecycle around each one of these variants."

The platform has grown with the genetic data. The goal, Rehm says, has always been a platform that can effectively analyze data and automatically generate patient reports. Her lab has been using GeneInsight for over eight years and has generated 30,000 reports.

Two Sides, One Solution

The clinical genomics problem has always been two-sided, says Aronson.

"You have a physician that is treating patients, and you need to be able to both communicate results effectively to them, give them the ability to manage those results, and then also keep those clinicians up to date as more is learned about their patients over time," he explains.

"From the laboratory perspective, what goes into that is you begin running genetic tests on patients, you start sequencing genes, and you find more and more variants of uncertain significance in those genes. And one of your objectives becomes to do as good a job as possible at re-classifying those variants… into pathogenic categories or benign categories."

Building a platform to address those challenges needed to be multi-faceted.

"GeneInsight consists of a clinician-facing application that can be integrated with electronic health records or stand alone, a laboratory-focused application that manages knowledge, and facilitates reporting. Those applications can be federated either lab- to-clinic or lab-to-lab," Aronson says.

The clinician-facing application—GeneInsight Clinic—simplifies genetic testing reports, while also staying dynamic. GeneInsight, "uses a lot of sophisticated rules-based logic to enable the auto-drafting of patient reports using patient-specific and disease-specific information," explained Rehm. The platform delivers Web-based reports to physicians and can be integrated into several electronic health records (EHRs). But keeping the reports connected to the system, "allows the variant database to be connected to patient reports, so if knowledge changes in variants, it can be delivered in real time to physicians," Rehm says.

Partners' Partners

Early on, Partners Healthcare knew that this wasn't a task to tackle alone. "Even a place with the scope of Partners will not be able to curate the genome by themselves for every indication that could be seen in one of our patients. Achieving our goal required working with others," Aronson said.

First, GeneInsight was registered as a Class 1 exempt medical device with FDA, so it could be shared with other labs and clinics across the country. Later, GeneInsight LLC was set up to facilitate that distribution.

Aronson says Partners is working with Mount Sinai Medical Center, the New York Genome Center, Illumina's CLiA laboratory, Rehm's lab, and ARUP Laboratories in Utah to define how "share and share alike" networks could work and what the governance surrounding that should be.

Aronson wants to encourage, "more and more places to operate under a model where in exchange for contributing your data… [labs] can benefit from the data that are contributed by other places."

Rehm agrees that interpretation is the major bottleneck in clinical sequencing, and believes that as a community, "[we] can evolve and improve that process over time through widespread data sharing."

Moving Forward

Even after almost eight years, Aronson still has a GeneInsight wishlist. He plans to provide deeper support for kinds of variants that are becoming more and more important, such as structural variants and other types of omics data. He also hopes to develop deeper integration with clinical process to take advantage of the "clinical context" that clinicians can bring.

Editor's Note: Heidi Rehm will be keynoting the 2nd annual The Clinical Genome Conference in San Francisco later this month, where she will address the evolution of clinical sequencing from targeting disease testing into whole genome and exome approaches, and compare and contrast and the benefits of one vs another. For more information, see http://www.clinicalgenomeconference.com.

Foundation Medicine Partners with Memorial Sloan-Kettering on Genomic Diagnostic for Blood Cancers

BY ALLISON PROFFITT | MAY 2, 2013

Foundation Medicine and Memorial Sloan-Kettering Cancer Center today announced a partnership to release a molecular diagnostic product designed to match patients with hematologic cancers (leukemia, lymphoma or myeloma) with the most rational targeted therapies or clinical trials for their cancer.

This new product will complement FoundationOne, Foundation Medicine's first product launched last year (see, "Laying the Foundation for Next-Gen Cancer Diagnostics"), which offers a similar genomic profile for solid tumors.

"The test for hematologic malignancies is something that we have been quietly working on for some time now, and we see the relationship with Memorial Sloan-Kettering as an opportunity for us to continue the test development and to—frankly—work with the leading hematologists and oncologists in the country to make sure we are then bringing a test to the market that is going to have the maximum impact on patients diagnosed with these diseases," CEO Michael Pellini, told Bio-IT World.

The test will be based on technology, methods, and computational algorithms developed by Foundation Medicine, and Foundation Medicine will commercialize the test both in the United States and internationally. Pellini says he expects the test to be available by the end of the year. Memorial Sloan-Kettering will help accelerate the development and optimization of the product by contributing their clinical and genomic expertise in hematologic malignancies.

The test for hematologic malignancies will obviously deal with new sample types—blood samples and bone marrow. This new test is also being developed using RNA sequencing in addition to DNA sequencing to better enable identification of the unique genes and classes of genomic alterations that are characteristic of hematologic malignancies.

Michael Pellini, CEO, Foundation Medicine

"We have to make sure that the genes that are being covered in this test are appropriate for all types of hematologic malignancies as well," Pellini explained. "In many cases there's a tremendous overlap with those for solid tumors, but in some cases they do not overlap with solid tumors. Just as we made sure that FoundationOne for solid tumors was a comprehensive, state of the art test for such tumors, we also need to make sure that this test for hematologic malignancies is comprehensive, is state of the art, to make sure that oncologists get the maximum insight into the patient's cancer, and that we maximize the benefit of this approach for patients with hematological malignancies."

Similar to FoundationOne, the new hematologic malignancy test will assist physicians by matching these alterations with targeted treatment options that may be relevant to the patient's genomic profile based on a comprehensive review of published literature, the company said in the announcement today.

FoundationOne results are delivered to physicians and oncologists in user-friendly reports that include mutation analysis, relevant targeted drugs, and potentially-appropriate clinical trials. The final report will have a very strong human element, Pellini told Bio-IT World last year about the FoundationOne test, with reports reviewed by a staff oncologist and pathologist with expertise in genomics before it is sent out.

"Even if we jump ahead a year and populating the report is largely automated, our medical director will be responsible for the accuracy of the information," Pellini said then. "There will always be the involvement of an oncologist and pathologist in the final report."

Foundation Medicine is not yet revealing the output plans for the hematologic test in such detail.

A Field Maturing: The 2013 Consumer Genetics Conference

BY AARON KROL AND ALLISON PROFFITT | SEPTEMBER 30,2013

The 5th annual Consumer Genetics Conference (www.consumergeneticsconference.com) wrapped up last Friday in Boston, after three days of rich discussion, new ideas and products, and a bit of debate.

Hugh Rienhoff kicked off the event with a summary of his personal odyssey to identify his daughter's rare genetic disease (see, Hugh Rienhoff Cops a Candidate Gene in His Daughter's DNA). Though Beatrice's exact condition has not been named, but she carries a TGF-beta 3 mutation that, with Rienhoff's prodding, has prompted more research. Rienhoff's journey has been a long one, but every step has added a bit more information to the broader picture. Rienhoff says that he made progress not because of who he knew, but because of a willingness to talk to people, and to endure the snubs and polite rejections along the way. Rienhoff champions patient (and parent) involvement. "It's very unrealistic that we can rely on geneticists going forward," he said. "We're all going to have to become geneticists." Today, there is a mouse model with Beatrice's specific mutation under study—and one of its line lives a pampered life in Bea's bedroom.

> "It's very unrealistic that we can rely on geneticists going forward... We're all going to have to become geneticists."
>
> **Hugh Reinhoff, , Jr., M.D.**, Director, MyDaughtersDNA.org

Rienhoff continues his search looking for "other Beatrices." He says: "My hope is to find a group of octogenarians with TGF-beta-3 mutations."

Dr. Robert Green, director of the genomes2people research program at Brigham and Women's Hospital and Harvard Medical School, presented the initial findings of the Impact of Personal Genomics (PGen) Study, a large-scale survey of genetic testing consumers led by Dr. Green and Scott Roberts of the University of Michigan School of Public Health. PGen gathered participants from users of the 23andMe and Pathway Genomics genetic testing services, and asked them questions at three different intervals: before they received their test results, 1-2 weeks after, and 6 months after. Questions were designed to explore both the motivations that led consumers to seek out personalized genomics services, and how they perceived and acted on their results afterward. Anxiety, perception of disease risks, and changes in personal health behaviors were all examined in the survey, which also delved into the demographics of those who choose to undergo genetic testing. In addition to his work with PGen, Dr. Green is also head of the MedSeq Project, an NIH-funded program that delivers the results of genetic tests direct to clinicians in a readable one-page format, and is involved in the spinoff BabySeq, which will soon do the same with neonatal sequencing.

Heidi Rehm, Director of the Laboratory for Molecular Medicine at Partners HealthCare Center for Personalized Genetic Medicine, discussed her ongoing work with the ClinVar database to standardize how the results of genetic tests are

> "I put a device in a guy's brain. Now that's serious. Getting our genomes sequenced? This is what we're worried about?"
>
> **Gholson Lyon, M.D.**, Weill Cornell Medical College

presented to patients. As genomic testing grows more routine, rare variants are regularly discovered and flagged as potentially pathogenic, leading to conflicting or tenuous conclusions about pathogenicity that may be passed on to patients. ClinVar, administered through the National Center for Biotechnology Information at the National Library of Medicine, seeks to centralize the evidence for the pathogenicity of these variants, assemble expert evaluations of the evidence, and standardize how that information is presented. Variants in the database are described as either pathogenic, likely pathogenic, uncertain, likely benign or benign, with links that allow users to view the relevant literature and curation process. Relying on large quantities of data, ClinVar takes a middle ground between being open source and relying on the gatekeeping of expert analysis: anyone can submit data, but that information will be assigned to different levels of curation depending on how many studies have examined a particular variant, and whether an expert panel has been assembled to reach consensus on its pathogenicity. ClinVar has so far assembled information on around 50,000 variants, about half from sources that were not previously publicly available. On September 25, the NIH announced an $8.25 million grant to Rehm's team to continue their work with ClinVar.

Taking on the issues of scale, Daniel MacArthur, co-founder of Genomes Unzipped, said, "We are dominated by artifacts of variant calling." On one hand, we don't have enough data. MacArthur said that the industry is only just beginning to fully grasp the scope of differentiating between a true rare variant and artifacts in samples that are simply not large enough to know. On the other hand, of course, storage is expensive and in some cases computation is simply impossible using current methods because we cannot hold enough data in memory at once. The Broad Institute led work to create reduced BAMs, "a new way of storing that raw sequencing read data in a format that allows us to actually actively pull the variants in a very large number of samples." MacArthur's lab is using a joint calling approach that combines data generated with different software. The pilot run included 25,000 samples, generated more than 350 Tb of raw data, and required more than 150,000 CPU hours of processing and joint variant calling. But it worked. MacArthur gained the largest ever catalogue of human protein-coding genetic variants.

In a session on the pros and cons of consumer testing, Ellen Matloff, Director of the Cancer Genetic Counseling Center at Yale Cancer Center, presented a GAO report from 2006 highlighting issues with direct to consumer tests. Matloff called some practices "criminal" (specifically the 23andMe recommendations included in the report for patients with positive BRCA findings), and questioned the practice of ever providing genetics results without counseling.

Gholson Lyon disagreed. After recounting a successful deep brain implant in a patient with severe mental illness, Lyon says it puts sequencing in perspective. "I put a device in a guy's brain. Now that's serious. Getting our genomes sequenced? This is what we're worried about?" The medical community has lost its way, he said. We don't have to go with them.

The New Class

A host of new players and products were presented covering the gamut of sequencing options.

Jessica Richman, CEO and co-founder of uBiome (ubiome.com), presented a different consumer genetics engagement model, raising $350,000

with an indiGogo crowd-funding drive. The financing model is just the first difference, said Richman. The company, which sequences patient microbiome from stool samples, is "throwing out" the old assumptions that patients get most of their information from doctors and that the customer is an insurance company, and is instead embracing a paradigm where patients are bright, creative, and engaged.

On the heels of Foundation Medicine's IPO, GENEWIZ (www.genewiz.com) announced its forthcoming PGxOne and OncoGxOne tests. Guanghui Hu, VP of translational genomics, described the tests as not "me too" solutions, but "me better" versions of the FoundationOne test. The tests can detect a full range of mutations: copy number variations, SNPs, insertions, deletions, gene fusions in introns, and low-frequency aberrations.

parents are carriers. But Morriss believes that as the consumer genetics industry becomes more crowded, novel services like GenePeeks can't rely on innovation alone. "We decided we were unambiguously in a service business, and service had to be one of our advantages," said Morriss. To that end, education in the genetics of recessive disorders, and offering genetic counseling at every stage of customers' decision making, will be an integral part of GenePeeks' strategy. Morriss hopes to launch GenePeeks' commercial services in the fourth quarter of 2013, and is already thinking of expansion into the couples market, although she acknowledges the ethical considerations are more complex. You can read more about GenePeeks, and Morriss' own story, at Bio-IT World.

Focusing on the digital health market, Julio Oh and Anish Sebastian of 1EQ (www.1eq.

> "We've surpassed a million [customers purchasing genetic testing kits]... The cool thing is, it's [only] going to take a year to hit the next million."
>
> **Spencer Wells,** Director of the Genographic Project, National Geographic

Ian Curry, president of DNA Genotek (www.dnagenotek.com) gave insight into DNA sampling and collections options—crucial questions as sequencing moves every closer to the consumer and medicine highlights home care. DNA Genotek samples are stable at room temperature and can be sent through standard mail. They can be banked for years at room temperature without even a freezer.

GnuBio (gnubio.com) presented its new sequencer—as yet unnamed—built on fluidics. Using emulsions and 10,000 uniplex reactions per second, the GnuBio system needs 1/1000 of sample compared to other sequencers, said John Boyce, President and CEO of GnuBio. Boyce says GnuBio aims to be "the K-cup of sequencing," eliminating sample prep. The $50,000 GnuBio system uses $200 disposable cartridges, can handle 50 genes, 5% allele frequency, and weighs 80 lbs. Boyce reported 1 error per million drops across all experiments using raw data.

Anne Morriss, the founder and CEO of GenePeeks (www.genepeeks), spoke about her company's business and corporate culture. GenePeeks offers prospective mothers who use sperm banks the opportunity to screen donors for the possibility of dangerous recessive diseases for which both

me) presented their vision for consolidating genomic data with the health record with data from personal health devices like the FitBit and Jawbone Up. The FDA released guidelines this week saying they would not review a lot of health and wellness apps, and Martin Mendiola of Happtique (www.happtique.com) presented their vision for stepping into that space. The company was spun out of the Greater New York Hospital Association, and is offering certification for health and wellness apps for $3,000 each for content, patient data security, and more.

Spencer Wells, Director of the Genographic Project at National Geographic and a luminary in the field of using genetics to illuminate human origins and migration patterns, delivered the closing keynote. In addition to reviewing the success of the Genographic Project, which since 2005 has performed over 600,000 genetic tests on people around the world and distributed $1.9 million in Legacy Funds to preserve traditional cultures, Wells also offered his predictions for the near future of consumer genetics. "The big inflection point has happened this year," said Wells. "We've surpassed a million [customers purchasing genetic testing kits]... The cool thing is, it's [only] going to take a year to hit the next million." To continue driving growth, Wells encouraged the industry to make ancestry and genealogy its primary focus, pivot to international markets, and offer a "citizen science" experience to users. "Harnessing the community as part of the scientific process is something I hadn't anticipated in 2005," he said, recounting the story of a woman of Hungarian ancestry who insisted that Central Asian traces in her genome had to be the result of a faulty test. Her questions led the Genographic Project to finally confirm the genetic impact of the Magyar people on present-day Hungarians. This sort of result, Wells believes, shows the promise of collaborations between scientists and consumers as the industry continues to expand. You can read more about the Genographic Project at Bio-IT World.

5 Ways Technology Is Changing Personalized Medicine

BY THOMAS HEYDLER | OCTOBER 18, 2013

In today's doctor's office, when a physician diagnoses a patient, a number of tests are consulted and the best possible course of treatment is prescribed. Unfortunately there is often limited data that allows the doctor to tailor and customize treatment specifically to a patient's biology and lifestyle. But there are five ways technology will change that over the next decade, bringing personalized medicine to fruition.

1) Correlations and Data Science. As consumers we first realized the power of correlation with e-commerce. Amazon's "people like you also bought" feature introduced algorithms to look at our online buying profile and match us to others so we could easily find new products we might enjoy. These commerce algorithms are in fact the foundational technology for creating medical algorithms to segment populations for clinical trials. Ultimately, physicians will use biomarkers and genetics to correlate a patient to a population "like him" and thus match him to the most efficacious treatment. At the current moment, a handful of diseases with simple and direct markers have been found, but the power of correlation will truly come to fruition in approaches like those used by researchers Nigam Shaw and Russ Altman, who have been able to use data mining to identify potential rare side effects and segment the population into those at risk of experiencing those side effects. By understanding a person's biology and how he will react to a particular therapy, researchers will be able to develop more targeted and effective treatment options and physicians will more accurately prescribe those treatments.

2) Advancing Clinical Utility of Genomics. Obtaining sequencing data has gotten faster and less expensive, but bottlenecks exist not just in regulatory process but also in correlating DNA sequence with clinical outcomes. Great examples of sequences with clinical utility exist, such as BRCA1, BRCA2 in breast cancer or the CFTR gene for cystic fibrosis. A key driver for the future is advancement of clinical utility for other genes with advances in the bioinformatics pipelines and data management. Major players in sequencing technologies are already offering data analysis and data storage cloud services in addition to just the instrumentation. New technologies that break the bottleneck in analysis and drive clinical utility of additional genes will be crucial to advancing the translation of sequencing to the clinic.

Thomas Heydler, CEO, Definiens

3) "Datafication" of Tissue. To date, much of the buzz in personalized medicine has been focused on the increasing possibility to easily extract data from DNA. The reality is that diagnoses today and in the future will be made of multiple types of diagnostic data. It will be essential for scientists and clinicians to be able to mine not just DNA, but also extract quantifiable data from images. At Definiens, we've termed the datafication of tissue images and its correlation with clinical outcomes "phenomics". Although genomic data can give clues to the ideal therapy, tissue images typically are more highly correlated to stage and presentation of disease, making the correlation of both types of data essential to the future of personalized medicine.

4) Telemedicine and Biosensors. At September's TedMed, Eric Topol dazzled audiences by using a cell phone to remotely monitor vital signs. While the term personalized medicine originally applied to tailored therapies, many like Topol believe that personalized medicine will also entail the use of devices and sensors for physicians to continuously monitor their patients remotely and tailor treatments on the go. Today's sensors are as small as a dime, but advances in nanotechnology could shrink sensors to allow for implantation in the body. With this miniaturization, you can imagine a day in which not only could glucose levels be monitored effortlessly in diabetics, but biomarkers of response to prescribed treatments could be continuously monitored via small sensors to alert physicians if threshold levels were reached.

5) Engineering Cells and Printing Organs. Within the next few decades, 3D printing will come to medicine. With over ninety thousand Americans awaiting organs, nothing will become more personal than the ability to "print" an organ from your own cells. Regenerative medicine pioneer Tony Atala has already printed the first 3-D kidneys and San Diego-based start-up Organovo is working on the 3-D printing of a liver. Initially 3-D tissue prints will be used as models for drug action and safety, but many believe that in 10-15 years 3D printing will enable tissue and organ construction from cells harvested from the patient, providing the ability to produce custom and personalized organs on demand.

While some of these technologies like DNA sequencing and tissue datafication exist today, others such as 3-D printing of organs are still in proof of priniciple phases. Nonetheless, as we look to the future of personalized healthcare, technology is poised to be a major driver in how we get there.

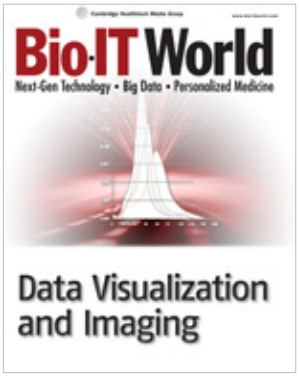
Thomas Heydler is CEO of Definiens, the leading provider of image analysis and data mining solutions for quantitative digital pathology in the life sciences, diagnostic biomarkers and healthcare industries. Heydler has more than 20 years of entrepreneurial expertise and in-depth knowledge of global software and IT, having served in prior executive roles at Barcelona Design, InterPro Business Solutions, Documentum, Cadence Design Systems and Siemens AG. He can be reached at theydler@definiens.com.

View all the Inside Bio-IT World eBooks

Evolutions in Next Gen Sequencing Part I: What's Next for NGS

Data Management and the Cloud

Clinical Genomics & Diagnostics



Data Visualization and Imagng

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