Betting on Genome Interpretation

By Katharine Miller

Six Startups Jockey for a Place at the Table

A handful of startups are wagering that genome interpretation is the next big thing.

Why is this business space so hot?

"Once you can produce a better faster genome, thanks to Illumina and others, the bottleneck shifts downstream to processing, making sense of, and interpreting that data," says **Jorge Conde, MBA**, co-founder and chief financial officer at Knome.

Each new startup seeks to turn whole genome sequences (or whole exome sequences—the portion of the genome that codes for proteins) into meaningful information. But each is also making a different bet about what approaches will succeed on the open market. Some companies plan to serve biotech or pharmaceutical researchers while others target clinical researchers or physicians and hospitals. Some focus on one step of the interpretive pipeline while others cover the whole shebang. Some provide cloud-based services, while others are betting on embedded platforms. And some rely on open-source algorithms and databases while others look to polish proprietary ones.

Which combination will ultimately prove successful is anyone's guess. But right now, "there's a lot of buzz," says **Nicholas Schork, PhD**, founder of Cypher Genomics and professor of molecular and experimental medicine at the Scripps Research Institute. Researchers and medical institutions that are buying sequencing machines to do genomic profiling need someone to turn to who has a clue about the data they generate. "The time is definitely right to think about interpretation," Schork says.

Sequence Crunching

Gaining knowledge from genomic data—the strings of C's, G's, T's and A's that constitute the standard output from a next-generation DNA sequencer—follows a series of fairly predictable steps, and some startup companies are putting their eggs in baskets defined by those steps.

Bina Technologies, for example, is focused on optimizing what's called secondary analysis, the essential data-crunching step that happens immediately after the DNA sequence comes off the next-gen sequencer. That step requires software that first aligns an individual's DNA sequence with a reference sequence and then picks out the differences between that individual's sequence and the reference (a process known as variant-calling).

Bina is betting that the speed and accuracy of secondary analysis will matter in a clinical context, says **Mahni Ghorashi, MBA**, director of marketing at Bina. Physicians don't want to wait two weeks to determine appropriate cancer treatments based on genetic differences between tumor cells and normal cells, he says. They want that information now.

And for newborns whose lives are at risk, "they need the information in 48 hours or less," Ghorashi says.

To meet that demand, Bina built a big-data platform for genomics. Called the Bina Genomics Platform, it pairs specialized software with specialized hardware that's designed to sit right next to the sequencer and analyze the data as it's generated. "We're able to take a process that used to take days or weeks and reduce it down to under four hours," Ghorashi says. At Stanford, which is using the Bina platform as part of a pilot program for new-

bina TECHNOLOGIES

Company launch date	June 2011
Funding	$6.25 million in Series B funding announced March 2013
Staff size	14
Target customers	Researchers and clinicians
Products	Bina Genomic Analysis Platform—an embedded system of software and hardware for secondary analytics
Publicized successes	Pilots with Stanford University and the Palo Alto Veterans Administration

born screening, it used to take 10 days to process one genome on a shared computer cluster. "They are now processing 10 genomes a day, a 100x acceleration," Ghorashi says.

Bina is only one of several companies that are focused on secondary analysis. DNANexus, Realtime Genomics and Appistry (not interviewed for this story) also focus on this area. In addition to its hardware version, Bina also offers its platform over the cloud, as does DNANexus. Realtime Genomics' new genome analytics platform for the study of early childhood disease can be embedded locally or accessed on-demand in Amazon's public cloud. And Appistry licenses GATK, the genome analysis tool kit developed by the Broad Institute. Each company makes claims of speed and accuracy akin to Bina's, and other companies described below also incorporate secondary analytics as a part of their pipelines.

From Variants to Interpretation: Predictive Modeling

The output from secondary analytics software, such as Bina's platform, is a variant file. To determine whether any of the variants are associated with disease, researchers or physicians must put those files through another computer pipeline (tertiary analysis) using a different product—one either developed in-house by the client or provided by another company—at least until Bina adds that step to its platform, which it plans to do. "We planted our flag upstream and our goal is to ultimately own the pipeline to guarantee accuracy," Ghorashi says. But several other startups begin where the Bina platform lets off. Cypher Genomics and Knome, for example, are two startups that specialize in this space.

At the most basic level, tertiary analysis involves querying whether the genome contains variants that have already been discussed in the literature and are known to be associated with disease. But that's just the beginning. Given that a human diploid genome contains 3 billion base pairs and each genome has about 10 million variants in it, according to the National Library of Medicine, most variants found in an individual's genome will not be described in the literature. Therefore, Conde says, at a bare minimum, interpretive pipelines need to include algorithms for predicting how a variant, though unknown, might be relevant for disease. Most companies, including Knome and Cypher, incorporate several open-source predictive algorithms to accomplish this goal. "If you don't do that, you'll only be looking for the keys under the streetlamp, as the story goes—because that's where the light is," Conde says.

In addition to open-source algorithms, Cypher licensed proprietary approaches developed at Scripps Research Institute by Schork and **Ali Torkamani, PhD**, a founder of Cypher Genomics and assistant professor of molecular and experimental medicine at Scripps. "Cypher Genomics has tools that can assign functional labels to virtually every variant in the genome—not just those that have been discussed in the literature," Schork says. As a simple example: If a gene variant codes for an amino acid substitution in a particular protein, Cypher's tools could predict whether that change would be likely to affect the

THE BET: Cypher Genomics

That its analytical tools for genome interpretation, including some proprietary approaches licensed from Scripps Research Institute, will prove useful to pharmaceutical and other researchers.

Cypher Genomics

Company launch date	2011
Funding	Not publicly disclosed
Staff size	5
Target customers	Pharmaceutical companies, research groups, and clinical partners
Products	Cypher Analytics—a pipeline for ranking candidate gene variants in rare diseases; conducting family studies; and performing genetic association studies. The pipeline includes variant-impact prediction and gene-phenotype prediction.
Publicized successes	Contracts with pharmaceutical companies and research groups

protein's structure or function—even if that variant has never been seen in the literature. Cypher also has tools for winnowing down from millions of variants to those that are likely responsible for a particular trait—be it a disease or drug response. And they have tools for leveraging annotations to make claims about groups of people, for example, individuals who do or don't respond to a drug.

Knome offers similar capabilities. "You need to be able to rapidly compare genomes to one another," Conde says. For example, if a family member is sick and other family members are both sick and healthy, Knome's software can ask for mutations in genes that are predicted to affect protein function or structure in sick individuals where that variant is very rare and is not present in healthy individuals. "That very quickly filters you down to the needle in the haystack," Conde says. In a study conducted on a family in British Columbia, Knome researchers used this strategy to find the sixth known genetic cause of Parkinson's disease. The same approach can work with unrelated individuals.

Knome's pipeline also includes "nifty algorithms," Conde says, that look first for identical point mutations, then for mutations in the same gene, and then ultimately for mutations in genes that are part of the same pathway. In that final step, he says, "we tend to get very interesting hits." This is important because unrelated people with the same disease or drug response are likely to have different mutations that fall within similar pathways.

For example, a pharmaceutical company asked Knome to look for gene variants that could explain why a group of unrelated people didn't respond to a particular drug. Knome's algorithms found that the nonresponders all had some level of mutation in different genes in the same network for metabolism of a particular starch. "To us it meant nothing," Conde says. But the pharmaceutical company used that starch to stabilize the drug. "People with that metabolic deficiency were excreting the drug and the body was never really seeing it."

Both Cypher and Knome provide genomics interpretation services for pharmaceutical and biotech companies as well as clinical researchers.

In addition, Knome has created a product called KnoSYS™100, an end-to-end system for interpreting human genomes and exomes. The company is essentially betting that as the cost of sequencing goes down and the resolution of data goes up, clinics will shift from ordering an occasional test for a specific gene, to sequencing and storing whole exomes or genomes and querying them *in silico* whenever a test is needed. "That's why our platform exists," Conde says.

Diagnostic Odysseys

Some of the splashiest genomics news in recent years involved diagnostic odysseys—cases where whole genome sequencing was used to diagnose and treat patients with unique or very rare diseases. Both Knome and Cypher offer interpretation pipelines for diagnostic odyssey patients—ways to sift through genetic variants

Knome®

Company launch date	2007
Funding	~$12 million
Staff size	"Under 50"
Target customers	Pharmaceutical, medical, and academic researchers
Products	KnoSys™100—a fully integrated, locally installed, hardware and software system for the interpretation of human genome sequence data. KnomeDiscovery—an end-to-end solution for interpreting large numbers of human whole genomes and exomes—starting with sequencing and ending with a interpretation findings report.
Publicized successes	Discovered a new gene for Parkinson's disease

to find the likely culprit.

SVBio offers a combination of secondary and tertiary analytics to that same end. But SVBio differs from Knome and Cypher in its clinical rather than research focus. "Clinical companies come from a different mindset," says **Dietrich Stephan, PhD**, SVBio's CEO. "Rigor levels are much higher than for a research product."

Because data that comes off next-gen sequencers is not in a form that can be used in the clinic, SVBio does a lot of massaging of the primary data in the alignment and variant calling step, Stephan says. And when assigning pathogenicity to a variant, they have to make sure they aren't relying on a polluted public database.

In addition to accuracy, SVBio wants to be comprehensive. It's not helpful to tell someone "you have a variant of unknown significance." Instead, according to Stephan, SVBio can say with 99.5 percent precision, whether the variant is a mutation or polymorphism, based on classifiers that are trained on all the historical data. Many labs do this, including Knome and Cypher, but according to Stephan, "Few have gone to the level we've gone to in terms of training complex classifiers on hundreds of attributes across 300,000 variants with publicly stated precision metrics around pathogenicity."

In January 2013, the Mayo Clinic's Center for Individualized Medicine teamed up with SVBio to build a robust software pipeline for interpreting a patient's exome sequence—the portion of human DNA that codes for proteins—in a clinical setting. The system will go live in June.

What attracted the Mayo Clinic to SVBio? "We'd talk about sensitivity and specificity on a per patient basis," Stephan says. "And we'd talk about the low probability of missing a diagnosis across a specific number of patients. They liked that."

One thing's for sure: Getting the contract with Mayo didn't hurt business. "This stuff is really complicated and nuanced and multifaceted," Stephan says. "Being able to say Mayo Clinic is using it makes things a lot easier."

SVBio

Company launch date	2011
Funding	Undisclosed funding by Sequoia Capital
Staff size	20
Target customers	Hospitals and physicians
Products	Cloud-based genome diagnostic services
Publicized successes	Contract with Mayo Clinic

Soup-to-Nuts Research and Diagnostic Services

Personalis[1] is one company that's gone full bore into clinical research and diagnostics, with the goal of enabling accurate clinical grade insights into genomic data. They start with a DNA sample, sequence it in-house, do the alignment and variant calling, and analyze the variants to identify those with potential to cause disease. "It's a kind of soup-to-nuts offering," says **John West, MBA**, the company's CEO. "We allow a customer to go from sample to insight."[2]

By owning the whole process, West says, Personalis can innovate every step of the way. "We're doing something novel in each area to achieve higher accuracy."

So, for example, exome sequencers don't actually catch all the genes. Coverage can be inadequate for many reasons, including sequencer bias against regions rich in guanines and cytosines (GCs), or because repeats are difficult to sequence, says **Richard Chen, MD**, chief science officer at Personalis. Because there could be something medically important in those gaps, Personalis has innovated to fill those holes—creating what they call ACE Technology™. "You don't have to wonder if the variant you're looking for isn't listed because it wasn't covered by the sequencing," Chen says.

Similarly, Personalis has taken a close look at secondary analysis. "There are so many details there, and mastery of the process is not trivial," Chen says. Many companies use standard tools and align against a standard reference that itself includes rare alleles. So Personalis has created its own reference genomes that contain the most common alleles for people of different ethnic backgrounds. "It gives us better alignment and variant calling," he says. The company is also improving on public tools that are good for calling certain types of variants but do poorly on others, such as inserts/deletions and structural variants, Chen says. "For case-control analysis, accuracy in sequencing and alignment really matters so that the real biology can be dissected from the noise in the data."

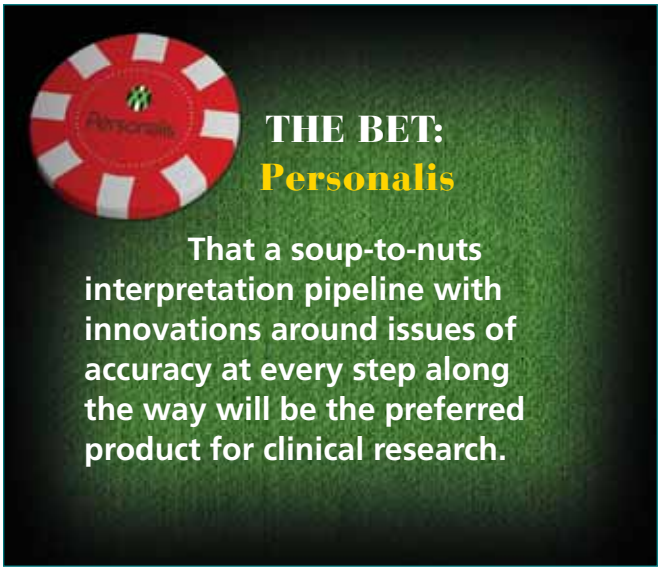
When it comes to the tertiary analytics—bringing biological meaning to genomic datasets—Personalis has also exclusively licensed and extended several large high-quality, manually curated databases in-

[1] Russ Altman is principal investigator for Simbios, which funds this magazine. He is also a founder and scientific advisor to Personalis as well as a personal friend of this author. He did not, however, play a role in the writing or editing of this story.

[2] Knome and SVBio also offer sequencing but they do not, so far, do it in-house.

cluding PharmGKB™, a pharmacogenomics database; the Personalis Variant Database, a large database of disease-related variants; and Regulome software, from the ENCODE project. All three of these were originally licensed from Stanford for exclusive commercial use. They have also built an annotation engine that integrates over 30 different databases. "In doing so we've reached a level of accuracy and comprehensiveness that is beyond what others are doing," Chen claims.

Personalis also runs sophisticated analytics (not unlike those run by Cypher and Knome) to identify differences between cases and controls at the variant level as well as at the gene and pathway level. And their tools can analyze genomic data from an individual or family with the aim of discovering the genetic cause of a particular disease or characteristic. "We have developed a detailed process to apply what we know about the family and the biology to make the most likely variant stand out from the crowd," Chen says.



Initially, Personalis is focusing on the clinical research market. In February, the Veteran's Administration contracted with Personalis to analyze 1000 genomes this year, with an option to do the same for thousands more patients in the coming years. It's a piece of the VA's Million Veterans Program—an effort to build a vast DNA data repository and correlate it with the VA's extensive electronic medical records.

Moving forward, the company would like to start working with hospitals and clinics. "It's hard work to get to the level of accuracy required for clinical decision-making," Chen says. "It's a higher standard we're holding ourselves to, but a necessary standard, whether you are doing research or using sequencing results to make life or death decisions for people."

What about Open-Source?

All of the companies described here make use of some open-source tools. But they are betting there will nevertheless be a market for their services.

Ghorashi says that although the open-source movement has been "really good" at developing algorithms, Bina adds value by making sure the open-source tools interoperate optimally. "These algorithms are written by biologists who don't necessarily take that extra step," he says.

Schork agrees. "At the end of the day, the delivery of the information is just as important as the accuracy. And the open-source tools don't do as good a job with delivery." Companies are set up to make it easy to sift through the data and present results in an effective way, he says. "That is not typically in the domain of the academic or weekend scientist."

But at least one company is making a different bet. "A lot of people starting companies are spinning them out of academic lab efforts and replicating what's available in the open-source community," says **Jonathan Hirsch**, CEO of Syapse. "The useful thing a company can do is make the use and delivery of those easier, not replace the algorithms."

In secondary analytics, for example, Hirsch thinks the bioinformatics community wants to directly use open-source tools like GATK (from the Broad Institute) and Bowtie (out of Johns Hopkins University). "If there's a choice between proprietary algorithms and the open-source algorithms, usually it's the open-source algorithm that's going to win," he says. He points to Spiral Genetics and Seven Bridges Genomics as companies that are focusing on helping customers run the open-source algorithms more efficiently by offering delivery mechanisms and a distributed computing platform. There's also GnuBIO, he says, which integrates the secondary analytics on the sequencer. "In the future, you won't need a separate secondary analytics process," he says. "The machine will do the work, just as it performs the primary analytics today." He says the same thing about knowledge bases: He predicts that the public ones will dominate, such as ClinVar.

These views have influenced

Personalis

Company launch date	2011
Funding	$20 million
Staff size	25
Target customers	Clinical researchers but moving toward hospitals/physicians
Products	Soup-to-nuts diagnostics. From DNA sample to analytical report.
Publicized successes	Contract with VA to sequence and interpret 1000 genomes as part of the VA's Million Veterans Program

Syapse's business model. "We don't do content," he says. Instead, Syapse is building what he calls "the software infrastructure for omics medicine." For content, Syapse hopes to partner with any or all of the companies described above. "We are not going to be the ones to choose the winner, so we want to partner with all of them."

Syapse is essentially a semantic computing company that builds a graph data structure that can leverage open-source ontologies for structuring biomedical terms and relationships. The goal is to make it easy to query the data to get a useful result. The company has two target audiences: data generators and clinics. For the data generators (hospital labs, diagnostic companies), Hirsch says Syapse will provide off-the-shelf software for managing, structuring, querying and reporting omics results. And for clinics, they will build an omics medical record. It will allow the clinical site to connect omics data with the electronic medical record in a clinical decision support system that can recommend appropriate courses of action to physicians.

"Content is something that will eventually be free and open," Hirsch says. "So to me, the most important thing becomes the off-the-shelf software that enables users to make use of their data."

Syapse

Company launch date	2012
Funding	$3 million
Staff size	3
Target customers	Hospitals
Products	Semantic infrastructure for omics generation and clinical reporting—agnostic as to knowledge base or analytical tools
Publicized successes	Provides infrastructure for several diagnostic companies such as InVitae and Foundation Medicine, and for the Stanford Center for Genomics and Personalized Medicine

Where are the Big Dogs?

If genome interpretation is a hot niche, it seems reasonable to ask why the sequencing companies (such as Illumina and Life Technologies) aren't jumping in.

For secondary analytics, Ghorashi agrees that Illumina is interested. "Their customers don't want the raw data files. They want the alignment and call files," he says. But at the same time, he notes, "quite a bit of innovation needs to happen," and startups are better poised to move quickly.

For tertiary analytics, Schork says that the sequencing companies have plenty of activity just staying at the top of their own market without branching out into interpretation. "They see themselves as the iPad and these other companies are the apps," he says.

If the clinical space develops more, Conde says, "it's pretty clear that Illumina would want to dominate." But right now, "there's plenty of evidence to suggest that there's a role for new companies like ours."

How will it play out?

While it's anyone's guess which business model will prevail, one thing's clear: The appearance of multiple startups with an interest in genome interpretation foreshadows a potential sea-change toward personalized medicine. Torkamani says the time is ripe for making sense of the data in certain areas, such as pharmacogenomics, diagnosis of rare conditions, and cancer. "There's plenty of actionable information there," he says. And although there's still work to be done before genomics will make a dent in chronic common diseases, "even there, a few bits and pieces of information are starting to appear," Torkamani says. He points to ApoE for Alzheimer's disease and various risk markers for macular degeneration.

There remains a risk that the hype cycle for genome interpretation is only just starting. If patients go through testing and are told "you have these variants but we don't know what they mean" or, worse, are told predicted meanings that turn out to be false, companies could unintentionally cause harm to patients—and also to the entire industry.

Many of the things the current batch of startups hope to accomplish are entirely reasonable goals, says **Mark Gerstein, PhD**, professor of biomedical informatics, molecular biophysics and biochemistry, and computer science at Yale University, who is not personally involved in any genomics startups. But, he says, "There's a long way from an idea to having evidence that it's proven." For example, connecting variation to disease is still an area of intense research, he says. And being able to find different variants in the same pathway is not as straightforward as it sounds. Still, he says, "I think this business area is a good thing." The research community doesn't create production scale products that are ready for the clinic. "There's a lot of chaos in normal research," he says, "And extracting from that chaos hardened tested workflows that people can use is very valuable."

Torkamani hopes so. "We went into genetic research with the hope that it will impact peoples' lives," he says. "Now it's really possible to make that happen." □