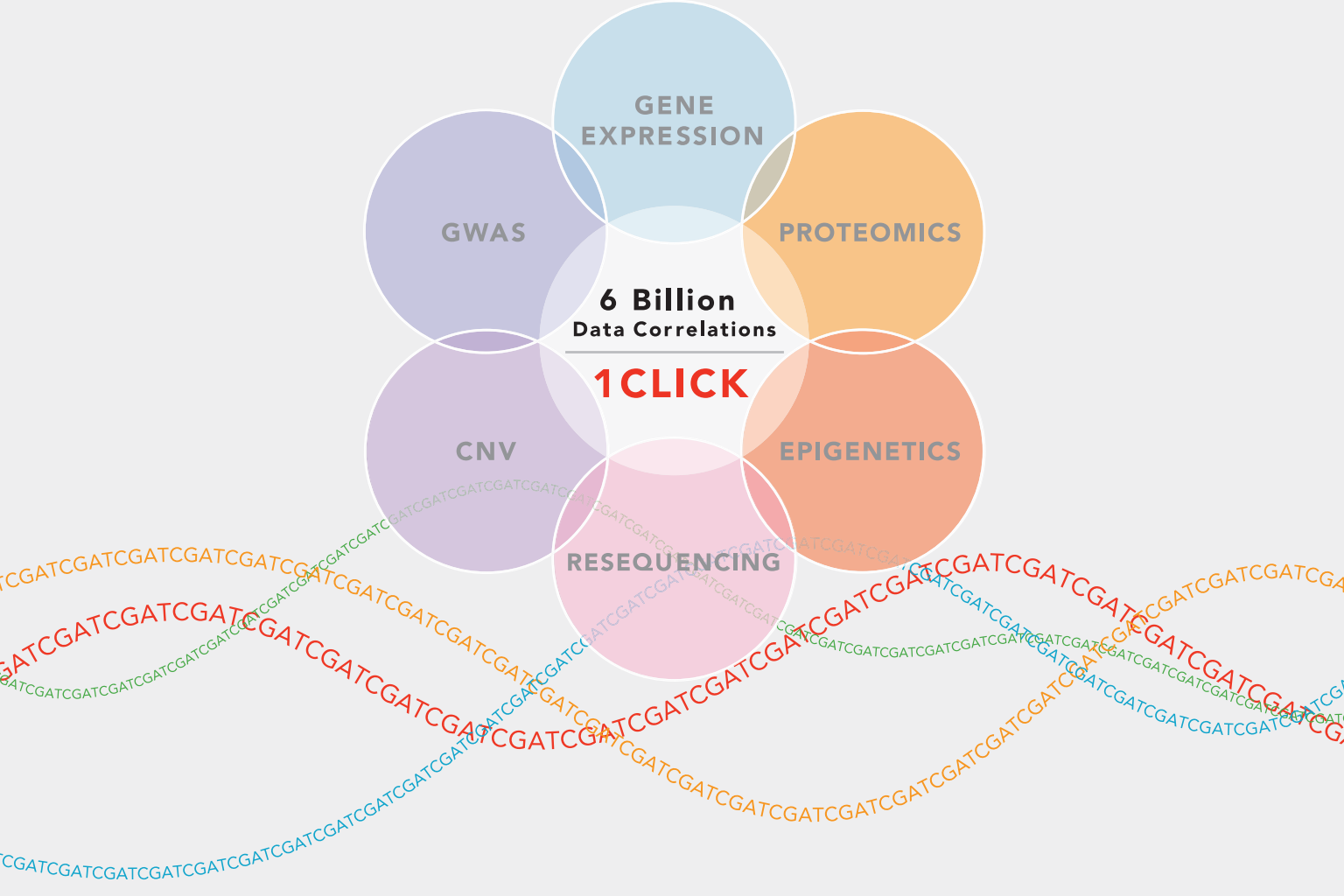
Democratizing Data

By Karen Hopkin, PhD.



Produced by Cambridge Healthtech Media Group Custom Publishing

Democratizing Data

For most biologists, the pipette is mightier than the keyboard. When faced with a scientific conundrum, their instinct is to design an experiment and hit the bench. But with the continued accumulation of untold terabytes of high-throughput data across a variety of different databases, information relevant to the problem at hand could already be out there. The question then becomes how to best extract those precious nuggets of knowledge from the public domain and put others' discoveries to good use.

The information on offer comes in many flavors, most notably high-throughput, nextgen sequencing data. But stir in readings from microarray experiments, proteomic studies, metabolomic profiles, and even clinical information, and you get a rich stew of data suffused with information about the molecular mechanisms of disease, the baroque meanderings of metabolic pathways and intracellular signaling cascades, and the genetic underpinnings of disease and response to treatment.

But shy of obtaining advanced training in bioinformatics, how can a hard-working human geneticist or a bench-bound biochemist access and leverage this treasure trove of information? "If you look in the public domain, there are thousands of experiments being published and very complex data sets from array and nextgen sequencing platforms being deposited into public databases. It's virtually impossible for anyone to really take advantage of that body of valuable information," says Ilya Kupershmidt, co-founder and Vice President of Products at NextBio. "It's like going to a library and browsing through the shelves to find a book of interest versus going to Google and doing a global search across all the information that exists."

To that end, NextBio is working to bring a Google-like approach to mining the myriad public databases, making that information just a few mouse-clicks away. "The idea is to bring all this high-throughput experimental data together and make it available in such a way that it can be explored in real time, based on an individual's biological interest," says Kupershmidt. Their efforts could democratize access to public data and allow scientists in industry and academia to begin to reap the rewards of the 'omics era, facilitating the design of experiments that can validate hypotheses and point the way toward the genes, molecules, and pathways that govern biology and go awry in disease (Figure 1).

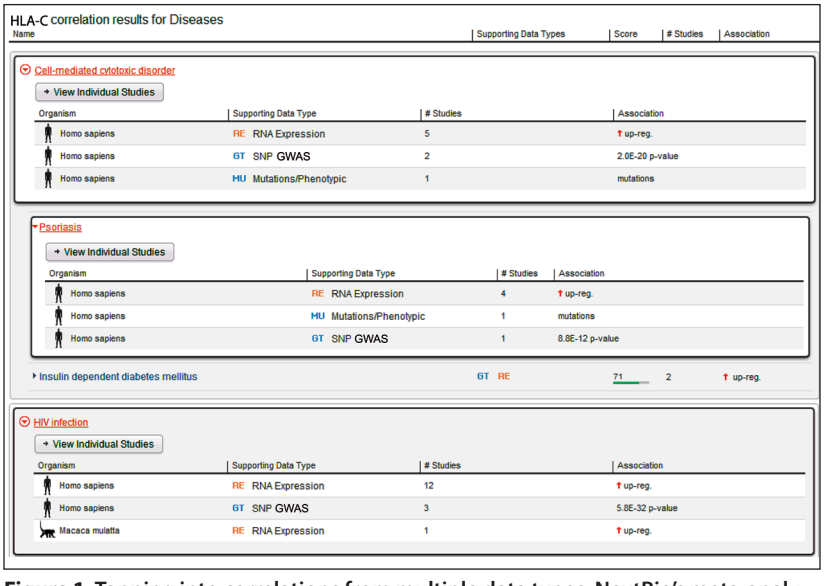


Figure 1. Tapping into correlations from multiple data types. NextBio's meta-analysis engine delivers an integrated view of studies involving RNA, DNA, or GWAS. Shown here are the results for HLA-C gene association with top-ranked diseases.

MUCHO MAS DATA

The push to mine public information springs from the sheer volume of data currently in play. "When I started my career, we were just starting to sequence genomes," notes Mark Gerstein of Yale University. "Now we have these huge sequencing machines that churn out gigabytes, if not terabytes, of data at a shot so the thing that's essentially driving bioinformatics is, of course, having large amounts of data." A spokesperson from Illumina estimates that 1500 next-generation sequencers have been installed in laboratories to date. And whole-genome sequencing is picking up steam, with thousands of whole-genome studies planned this year (compared to 600 conducted in 2009).

And biologists aren't the only ones awash with data. Physicists, for example, routinely weed through mind-boggling barrages of data in search of a certain subatomic particle or signs of dark matter. But physicists have some advantages when it comes to handling the data deluge. First, they usually have a heads-up. Take, for example, the Square Kilometer Array radio telescope slated to be built in the southern hemisphere. "This thing will generate something like a terabyte of data per second," says Paul Flicek of EBI. "But it won't be ready until around 2020. So physicists have plenty of time to determine how they can capture that much data and how they can pull information out of it as quickly as possible."

"Biologists haven't had that luxury," he says, "because new sequencing machines just appeared on the scene and data suddenly amasses ten or a hundred or a thousand times more quickly than it used to with very little warning."

But Flicek sees cause for hope. "If there's a lesson to be learned from other fields," he says, "it's that: it can be done. There's no need to throw up one's hands and say, 'it's too hard a problem to solve.' Because clearly it can be solved."

And now is a good time to try. "We have all this information in the literature and in public databases," says Rong Chen of the Stanford Center for Biomedical Informatics Research. "For hundreds of diseases we know the changes in gene expression, and what kinds of nucleotide changes increase the risk of disease. So now we have this great opportunity to combine all this information from the public databases and public literature to potentially improve patient care."

PANNING FOR GOLD

Filtering useful information from vast repositories of data is an easy enough concept to grasp. But it can be a bit tricky to accomplish. At its core, data mining involves looking for patterns. Learning, for example, that a pair of genes show expression levels that increase or decrease in lock step across a variety of tissues could suggest that their proteins are involved in the same process, fall in the same pathway, or perhaps even physically interact. Or finding that a particular single nucleotide polymorphism (SNP) tends to associate with a disease could point the way toward a gene involved in the disorder. Patterns can also indicate that a particular nucleotide stretch might represent a methylated region affecting a gene with a particular function. Or that a cluster of amino acids folds into a particular three-dimensional structure. Searching for these links often involves pattern-recognition software that can tease out these associations.

The low-hanging fruit includes interactions that were already suspected. "Often, we can find something to substantiate our hypothesis in a big pot of data," says Nick Tsinoremas of the Miami Center for Computational Science. "But that's not true data mining. True data mining allows us to find, in an agnostic way, things we did not know before."

Chen, for example, used the approach to discover a protein that helps predict rejection of transplanted tissue. He and his colleagues pulled from public databases microarray data from patients who'd received a donated kidney, lung, or heart. And they compared the gene expression patterns of patients who later rejected the transplant to those who hadn't. "We found 45 genes that are consistently upregulated when you have rejection across these different organs," he says. Those genes make proteins that can be detected by a simple diagnostic assay. "Previously you had to stick a needle into the kidney or the heart to detect rejection," he says. "Now, based on this data from the public domain, you can simply draw blood or collect urine"—

much less invasive procedures—and determine the best course of treatment for the patient.

Working with microarray data, Kupershmidt and his NextBio colleagues discovered a previously unrecognized connection between a type of hormone replacement therapy and an aggressive form of breast cancer. The team started out by identifying a transcriptional signature characteristic of invasive versus noninvasive breast tumors. That exercise yielded a collection of 1500 genes whose expression differs between the two. Kupershmidt and company then used that data as a query, comparing it to the other publicly available data sets they had in the system. And they found that this gene expression profile matched the expression profile of patients receiving this particular hormone therapy. "This is something we determined in silico," says Kupershmidt. "So you could have potentially avoided giving this treatment to certain patients if you'd done this analysis beforehand."

ALL SYSTEMS GO

But *in silico* research is not just about grappling with large volumes of information. Integration is also key, particularly when combining different types of data—whether it's genomic, proteomic, transcriptomic, or any other –omic. "Each data source has its troublesome noise," says Stanford's Russ Altman. "However, when you combine data sources—each of which has a different source of noise—the net result is a much cleaner signal."

Take, for example, the vexing false positives. "If I have a list of hits from a proteomics experiment and a list of hits from a microarray experiment, the chances that both of those methods will both make the same mistakes is low," says Altman. So, seeing an increase in transcript levels coupled with an increase in protein concentrations can be reassuring. "Combining them doesn't get rid of the false positives, but it makes them much less likely," says Altman. "That's important, because you don't want to go ahead and spend a lot of time and money following up on something that's not true."

Some of this integration can be done in an automated fashion. NextBio uses a semi-automated process to normalize heterogeneous information from a variety of databases, tagging and curating the data so that they can be associated with data from other sources. "It's a big job," says Tsinoremas. "But it's the only way you can even attempt to compare different data sets."

"It's the promise of integrative biological data analysis: being able to look at biological phenomena from a variety of different angles to identify what's driving a disease or a response to a particular drug," says Kupershmidt. "That's why bringing together this heterogeneous data—from , SNPs to copy number variations, epigenetic chang-

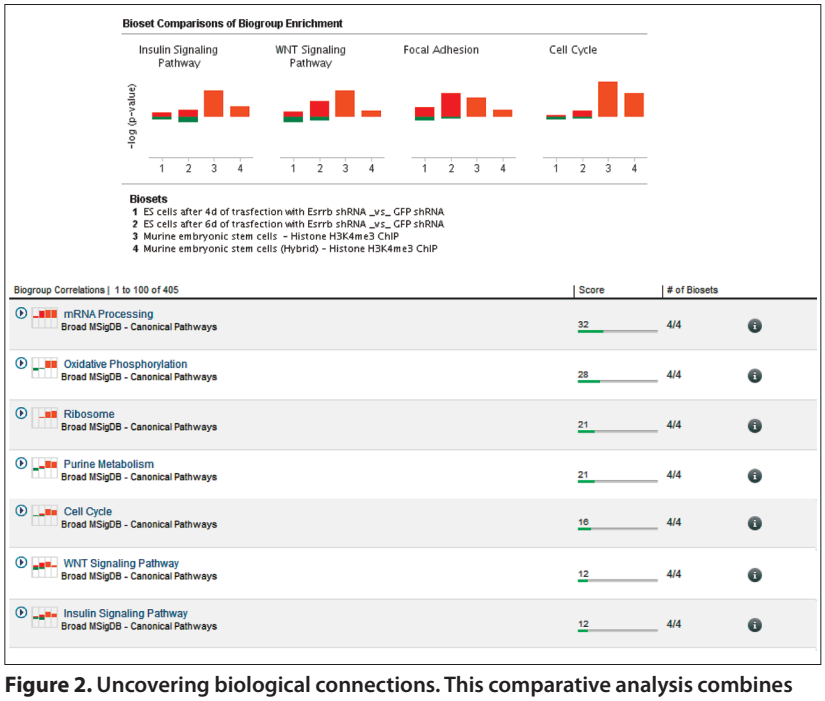


Figure 2. Uncovering biological connections. This comparative analysis combines gene-expression data with transcription factor-binding data for the estrogen-related receptor protein ESRRB generated by nextgen sequencing (ChIP-seq) in embryonic stem cells. The results present a prioritized list of pathways that are involved in ESRRB signaling during stem cell differentiation.

es, transcriptomics, and proteomics—is critical."

With the latest evolution of the NextBio platform, researchers are now able to look across tens of thousands of datasets from array and nextgen sequencing technologies to find genes, pathways, and DNA regions impacting a disease, a patient's response to a therapy, or a normal tissue development. By correlating and comparing these data sets within one integrated environment, NextBio enables researchers to seamlessly compare, for example, mutation data from resequencing studies with DNA copy-number, gene expression, and DNA methylation changes identified in individual patients, cell lines and animal models (Figure 3).

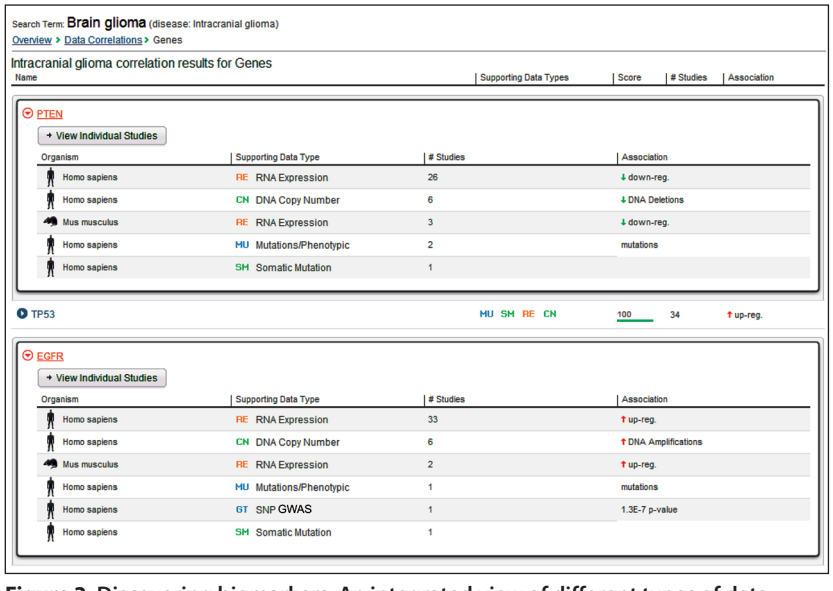


Figure 3. Discovering biomarkers. An integrated view of different types of data highlights genes that are upregulated—or downregulated—in brain gliomas across diverse types of experiments in human and mouse.

For example, using SNPs that were found to be associated with psoriasis in a large genome-wide association study, the NextBio team were able to fish out gene expression signatures from patients and cell lines treated with psoriasis drugs—and identify a handful of drugs, currently used to treat other diseases, that could potentially represent novel therapies for psoriasis. "So this approach could enable drug-discovery companies to find applications for their existing drugs in new therapeutic areas, or to further explore these connections to find biomarkers predicting patient's response to a particular drug," says Kupershmidt.

"Many of these findings can only be found by doing the data correlation," adds Saeid Akhtari, co-founder, President, and CEO of NextBio. "The great thing about research in silico is that you can do these types of analyses in real time, without ever leaving your computer. We provide a framework that allows you to make discoveries by systematically correlating all these different data types," he adds. And the analyses go beyond simply looking for information about your favorite gene or disease. "Your query is often not a single gene name or protein. Your query could be list of 10,000 genes or one million SNPs, which requires a whole different paradigm."

FOR HERE OR TO GO?

Fortunately, individuals don't need to import all that raw data to be able to mine it for information. "While people working on some aspect of human genetics might have gone and downloaded the entire human genome sequence, it's unlikely that most people will download all the data from the 1000 genomes project," says Flicek. "They might be more likely to download data for a single locus or some other subset of the data."

"We're informaticians," says Altman. "We have big computers and this is our laboratory science. So the first thing we do when we see a database is we bring it in-house. But most biologists are not geared up to do that," he adds. "If a biologist has a specific protein of interest and wants to know what other proteins it interacts with or what cells its gene is expressed in, those are perfectly good questions where you don't have to download an entire database just to get an answer."

NextBio provides access to manageable amounts of processed information culled from GEO, ArrayExpress, caBIG, Sanger Institute and dbGAP, among others. "For the past six or seven years, we'd

been trying to manually import public data into our system," says Eric Muise of Merck. "It was pretty painful" But by combining his own data with the publicly available information that had been collected by NextBio, Muise was able to identify a new, insulin-resistant mouse strain with a gene expression profile similar to ones he'd seen in his own lab. "That validates what we'd found," he says. And it does so quickly. "It used to take me two days to pull up expression data on a gene for somebody," he says. "Now it takes five seconds. So that's also an immeasurable amount of time saving."

BLAZING CONNECTIONS

And speed is a big part of the equation. "There are really two aspects of data mining," says Gerstein. "One is finding subtle connections or non-obvious patterns among bits of data by integrating different data sources. Another is the speed." A typical BLAST search for a nucleotide sequence returns an instant match. "We're growing accustomed to that type of interactivity with computers," he says.

"Of course, as you store more and more information, you get into the 'needle in a haystack' problem," he adds. "So it becomes harder to locate that valuable piece of information." The solution, Gerstein says, is to build representations of the data that summarize its important features, making querying the data quicker and easier. "To some degree Google does that," he says. "So if you type 'hamburger' into Google, it doesn't search absolutely every web page, then and there. That would take a gargantually long time. Instead, long before the query, it trawls the web for occurrences of the word and then it generates and stores indices that show where references to 'hamburger' might be. So you can find what you're interested in quickly. That's why everyone is so gaga about Google. It's just so fast."

NextBio, too, does an enormous amount of tagging, indexing, and up-front, behind-the-scenes computation to make their system run smoothly—and quickly. "What we've done is actually analyzed all the data to extract the important findings, the important signatures from each experiment in a very systematic way," says Kupershmidt. This data is then integrated with all the other data within NextBio, regardless of whether it comes from a different platform, a different organism, or a whole different type of experiment. "Then we basically pre-compute all possible combinations between various data sets—between different genes, different disease studies, different sequenced regions, different compounds—and store all of that information in our computational pipeline. We have hundreds of boxes performing billions of data correlations as new data gets incorporated into the system, and these correlations are continuously recomputed."

"Right now we're talking about thousands of experiments, and within the next few years we'll be looking at hundreds of thousands, if not more, large-scale data sets that will be made available in the public domain," adds Kupershmidt. "There's simply no other way to deal with this information—except to start building indices and making correlations so you can do your work in real time. That way, when you go into the system and want to explore your own data sets, you can immediately see important correlations. And a user-friendly interface lets you query all this information in real time."

GOOD EYE

And the interface is an important part of the package. "I think there's a lot of value in simply presenting data in a way that allows humans to do the integration," says William Noble of the University of Washington in Seattle. "Because the reality is that the human visual system is way better than most computational methods in finding patterns in data sets."

Kupershmidt agrees. "Visual exploration is very critical when it comes to sequence-centric data," he says. Currently available genome browsers are geared toward experts, so they serve the computational biology community as opposed to clinicians or lab scientists." In an effort to make its own data—which is increasingly sequence-centric—more visually accessible, NextBio has also launched a genome browser. "It's a nice interface that gives you a genomic view of different types of data," says Ku-

pershmidt. Over the past decade, high-throughput data has tended to be more or less gene-centric: microarrays, for example, provide a readout of levels of gene expression. Now, the types of data that can be collected via large-scale, high-throughput methods have expanded. And in the NextBio genome browser, each type of data is tagged to its relevant location in the genome. "So you can see, within the context of a particular cancer, which regions are amplified, where structural variations occur, and where mutations identified through next-gen sequencing efforts fall," says Kupershmidt. "Its flexible design allows you to select, visualize and navigate data of interest out of thousands of different experiments. So really all you need is a mouse to navigate through the entire genome" (Figures 4,5).

DATA FIRST

Going straight to the data may also require a paradigm shift in mindset of the average biomedical researcher. "The traditional paradigm involves going to the literature and doing some research, reading what other people have published," says Kupershmidt. "But today's high-throughput platforms are producing millions and millions of measurements that are not actually captured in publications. So we're seeing a new research paradigm in life science today - doing exploratory analysis by going directly to the data and looking for meaningful biological and clinical correlations *in silico*."

And the data contain more information than any publication ever could. "One publication might share at most maybe ten findings," says Akhtari. "A large-scale data set produces on the order of thousands of potential hypotheses. Or even millions if you correlate it with other types of information. So the number of potential hypotheses that can be derived from the data

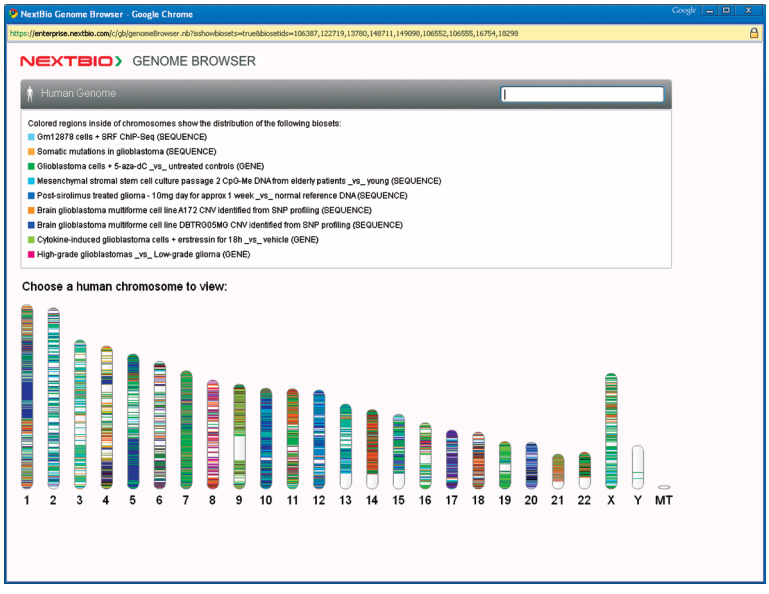


Figure 4. A genome-wide view. Whole genome view of multiple datasets from gene expression, copy-number variation analysis (CNV), DNA methylation and mutation data from glioblastoma primary tumors and cell lines. Sections of chromosomes with significant enrichment of impacted DNA regions from different types of data are clearly visible and can be further analyzed by zooming in.

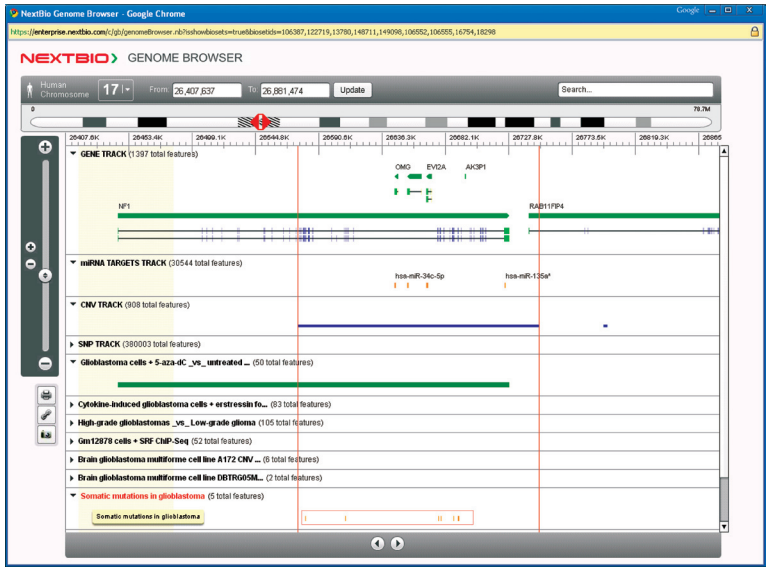


Figure 5. A sequence-centric approach. Region of chromosome 17 indicating a number of mutations across NF1 gene in glioblastoma resequencing experiment (somatic mutations track). Adjacent tracks show additional glioblastoma related information associated with this region. For example, users can see that NF1 is down-regulated in glioblastoma samples treated with 5-azadeoxycytidine ("Glioblastoma cells + 5-aza-dC" track).

is several orders of magnitude higher than that captured in the published literature. We need a very different paradigm to make sense of that amount of information."

Until now, that kind of investigative power was in the hands of bioinformaticists. "But anyone who wants to do it should be able to do it," says Altman. The folks at NextBio heartily agree. "Bioinformatics experts will always be a critical part of the research process," says Kupershmidt. "But it's not a scalable model for everyone with a biological question to always have to go through a computational expert to be able to extract value from large-scale data set. So ultimately the goal is to really put this data into the hands of all biologists and clinicians."

Such democratization would allow individual scientists to take advantage of public data sets from the comfort of their own lab bench. "I'm not a bioinformaticist. I don't know how to manipulate large data sets," says Sarah Sague of Johnson and Johnson. But she uses NextBio to search public datasets to see whether her target protein—or its receptors—are upregulated in disease, a first step toward designing therapeutic inhibitors. "It's a simple query to do in NextBio," she says. "And within seconds or minutes you have data to start working with." The process, she says, "allows biologists to poke around and formulate new questions on the fly." And it frees up the card-carrying bioinformaticists to tackle more complicated queries.

"If my colleagues all came to me with every query, I'd never be able to keep everybody happy," says Altman. "Then they'd think I'm not a nice person and I don't collaborate well. Instead I can say, 'Oh, there's this great company, they do a nice user interface, you can query for your favorite genes. Go give them a try.'"

Because using these tools can be a good place to start. "I say let people try stuff," says Altman. "And if they find something incredibly amazing, then they have to do the hard work to try to validate it. Because this is really about hypothesis generation." Which is a big part of laboratory science. "In some ways it doesn't matter where a hypothesis comes from. An apple could hit you on the head and you get an idea. That's perfectly valid. So these bioinformatics tools help you generate hypotheses. Which is great. I say, let people go wild with them. As long as they realize that these tools are not proving anything. They're just generating some ideas, providing some interesting associations. But then you have to look at those results with a skeptical eye and go back to the lab and find some way to show that your hypothesis is statistically and biologically valid."

"Access to this data allows you to perform in silico research, where you explore data and correlations from thousands of large-scale, genome-wide studies in essence in real time," says Kupershmidt. "So before you run a real experiment, why not investigate what's already been done by other companies or other academic centers? Why not test your hypotheses *in silico*, based on this incredible collection of data—and then go on to design a better experiment?"

The approach could change the way science is done—or at least accelerate the pace of discovery. "This wealth of data is producing unimaginable opportunities for generating hypotheses, designing experiments, and, ultimately, making major discoveries," says Akhtari. "With the right tools, life scientists will be able to revolutionize drug R&D, clinical research and patient care."