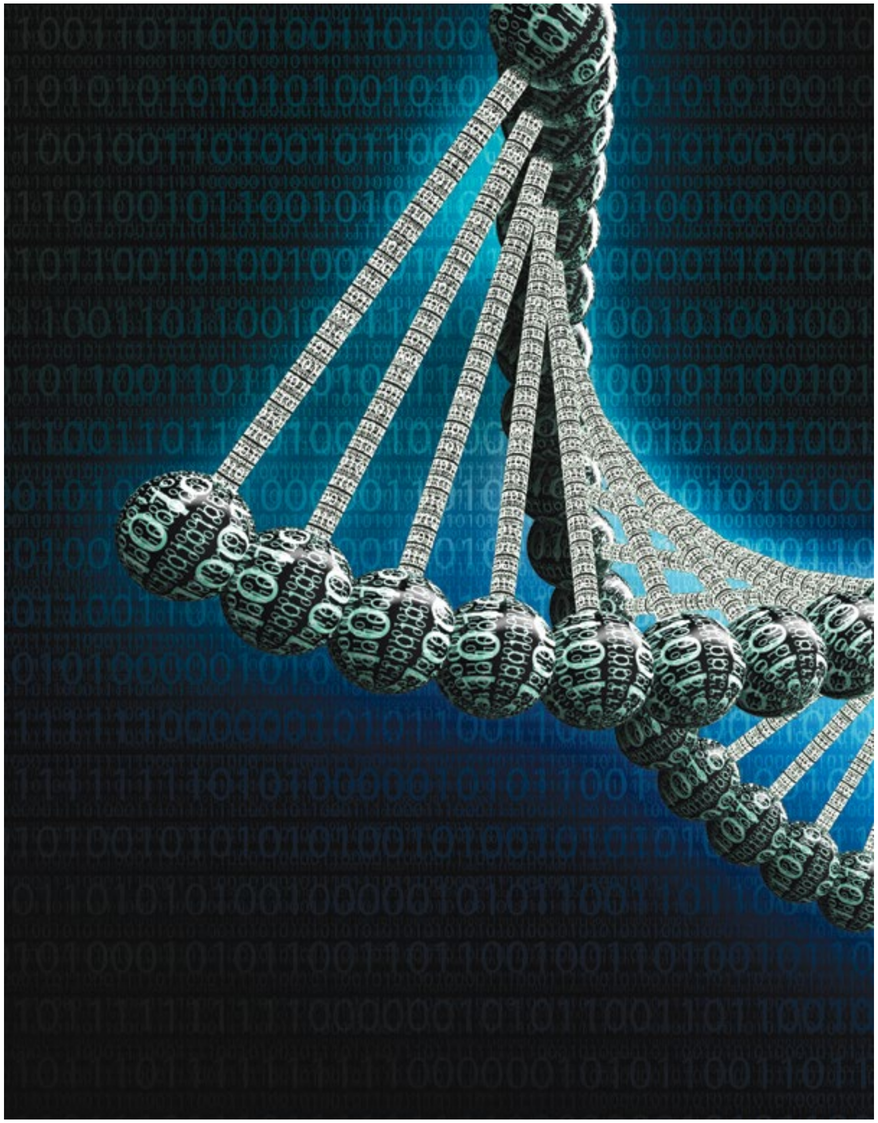
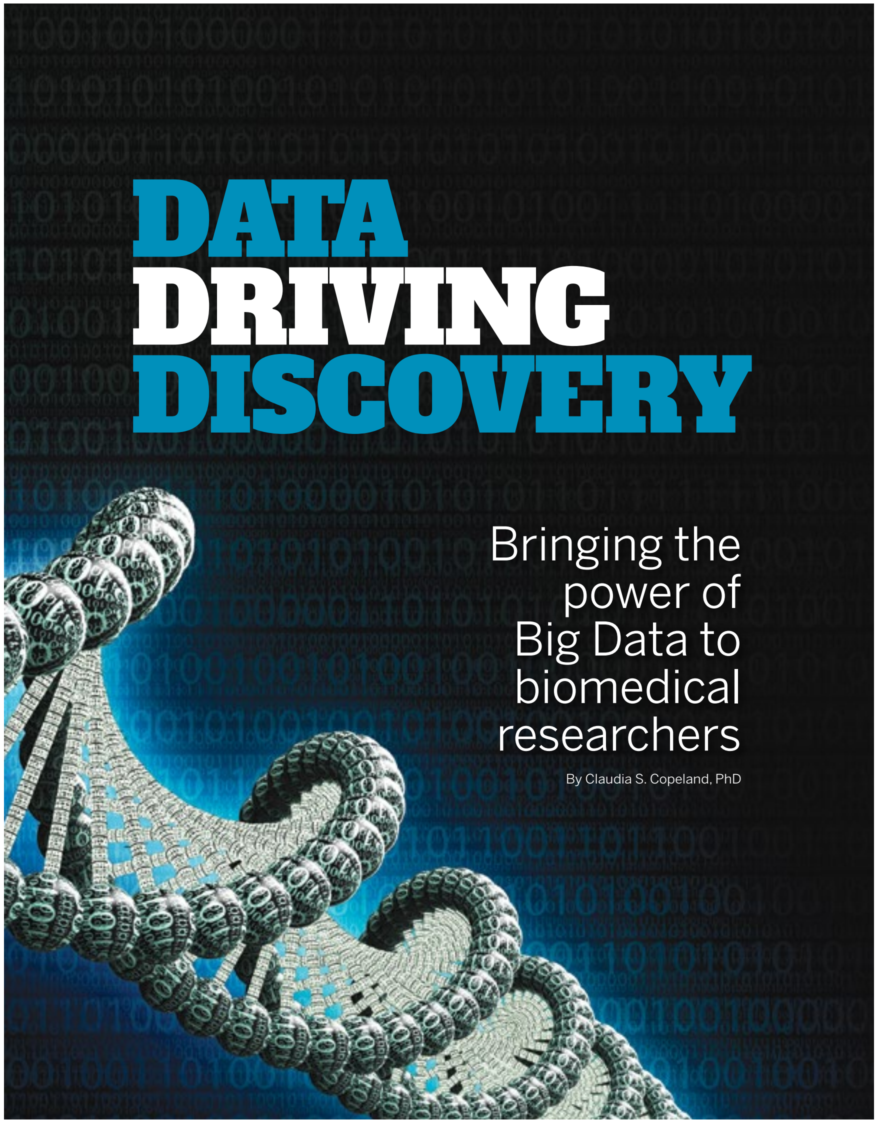




# DATA DRIVING DISCOVERY

## Bringing the power of Big Data to biomedical researchers

By Claudia S. Copeland, PhD

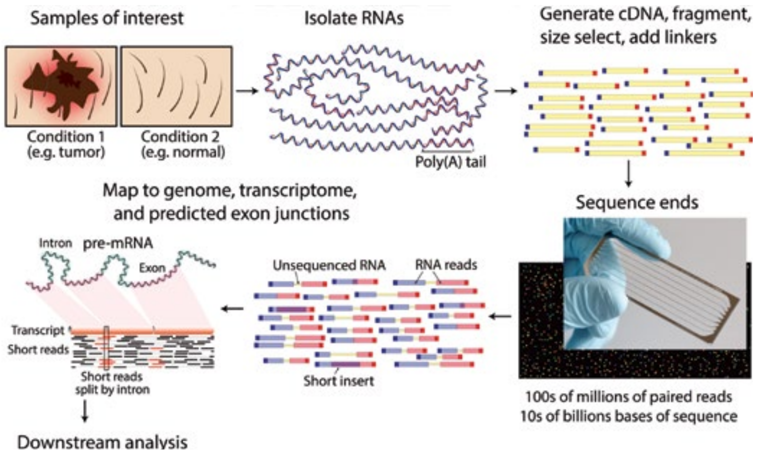
PHOTO COURTESY OF ECSEQ

Among the very first things a good science teacher emphasizes is the overall structure of the scientific method: come up with a hypothesis, design an experiment or study to test that hypothesis, and analyze the results.  Everything must be done with careful objectivity in mind, from experimental design to statistical tests for significance.  It's a great method, and hypothesis-driven research has led to some of the most solid advances in knowledge throughout history.  However, it's slow.  And expensive.

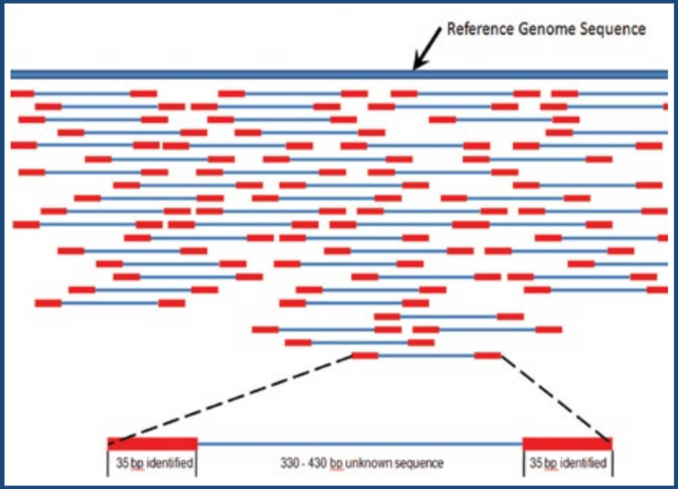
WHEN TIME IS MONEY—when a professor's next grant or a CRO's next pharma contract is dependent on current results—the slow nature of hypothesis-driven science can lead biomedical researchers to choose more conservative avenues of inquiry. This is especially true for academics in the early stages of their careers or start-ups still dependent on venture capital. When you're in the make-or-break stage, when results are a must, it's wise to choose subject matter that's sure to yield results. But this kind of "safe" subject matter tends to advance our body of knowledge slowly, and in the same general direction it's already going. Breakthroughs, on the other hand, often come from a completely different, unpredictable angle—often landing in researchers' laps by chance rather than through knowledge-based hypotheses. Penicillin was discovered not by a mycologist carefully testing a hypothesis about Penicillium mold, but by a bacteriologist who noticed that accidental mold contamination of some of his staphylococci cultures was releasing a substance that killed the bacteria. The antidepressant properties of monoamine oxidase inhibitors were discovered when anti-tuberculosis drugs had the "side effect" of improving depressive symptoms. Sildenafil was being used to treat hypertension and angina pectoris when it was also discovered to have a side effect that became a highly profitable main effect. Marketed under the brand name Viagra, it is now among Pfizer's top selling drugs. In these three cases and countless others, it was the result that drove the hypothesis—only after discovering the unexpected properties of these drugs did researchers develop hypotheses to explain their effects (which often, as in the case of the antidepressants, led to the development of better drugs through enhanced understanding of the underlying biology).

Serendipity can be a great shortcut to discovery, but it is an elusive, fanciful beast that's not easy to harness. What if there was another way to circumvent the constraints of predictions based on already-understood principles? A way to simply explore at random until bumping into a result that can then be put on the hypothesis-driven track? Well, there is such a way—harnessing not the power of serendipity, but the power of massive data analysis enabled by today's technology. With the advent of next-generation sequencing, which allows rapid sequencing of fragments making up whole genomes, and software powerful enough to reconstruct those genomes from the massive datasets NGS outputs, a road has been paved to do just that.

To get an idea of how powerful today's technology is, imagine looking at tumors from a group of patients suffering from a specific type of cancer and being able to get



Overview of RNA-Seq



**...RNA-seq, is one of the hottest subfields of computational molecular biology, and one of the most powerful ways to understand disease processes and discover new treatments.**

a readout of every single gene that was transcribed in that group, alongside every single gene transcribed in a group of healthy controls. You could then isolate the genes transcribed in the cancer group, but not in the controls, and start testing them for their functions and possible roles in the cancer. Just 10 years ago, this type of whole-transcriptome study would have been impossible. Today, this procedure, called RNA-seq, is one of the hottest subfields of computational molecular biology, and one of the most powerful ways to understand disease processes and discover new treatments. (Although a similar approach, microarray analysis, was being used earlier, it requires gene chips made from known genes rather than whole transcriptomes, including unknown genes.)

Transcriptomics can be used in subfields spanning the whole of modern biology, but one of the most promising areas for RNA-seq is medical phenomena for which mechanisms are a mystery. For example, it has long been known that alcohol consumption is associated with head and neck squamous cell carcinoma. However, no clear mechanism underlying this association had been determined. U.C. San Diego researchers Yu et al. knew that noncoding RNAs–RNAs that exert their effects directly, rather than by coding for a protein–are important regulators of gene expression, and that ethanol can influence the activity of noncoding RNAs. However, they had no idea which specific ncRNAs, if any, might have an effect–no specific ncRNAs had been associated with alcohol-associated SCC. So, they had reason to believe there might be one or more ncRNAs involved, but finding them would be, essentially, like searching for a needle in a haystack. Fortunately, with today's bioinformatic software, needle-in-a-haystack searches are, if not easy, quite do-able. Yu et al. decided to approach this question through the winning combination of next-generation sequencing, which allows the sequencing of massive amounts of DNA, with bioinformatic algorithms capable of analyzing these massive datasets, a procedure called RNA-seq.

"Together with our biomedical customers and a team of high-level academic experts, we want to find new diagnostic approaches to considerably improve people's quality of life."

Dr. David Langenberger

To conduct an RNA-seq experiment, researchers must first fragment the entire set of RNA in the cell, which represents all of the genes being expressed by the cell at that moment in time, and then reverse-transcribe those fragments to create DNA copies that can be sequenced using NGS. The entire process–reverse transcription of the RNA transcript fragments, direct sequencing using flow cells in a massively parallel fashion, and use of advanced bioinformatic algorithms to reconstruct and analyze the transcriptome from this sequence data–constitutes RNA-seq. Yu and her colleagues used RNA-seq to look at the differential expression of 13,338 long noncoding RNAs, or lncRNAs. In this "haystack of RNA", using bioinformatic analysis, they were able to find two "needles", NETO1-1 and PSD4-1, that were differentially expressed in alcohol drinkers vs. non-drinkers among head and neck SCC patients. Further experiments indicated that these two lncRNAs are at least partially responsible–via their regulatory effects on the expression of cancer-related genes–for the pathogenesis of alcohol-associated head and neck SCC, and may be possible candidates for therapeutic targets. The power of next-generation sequencing and bioinformatics allowed these researchers to not only find "needles in a haystack," but to find them when they had little or no idea of what they were looking for; they had no prior hypothesis about NETO1-1 or PSD4-1. Countless other computational biologists are using this same approach, and examples abound of important biomedical research stemming from an initial hypothesis-free discovery using RNA-seq.

RNA-seq is clearly a powerful approach–it can be used to understand the etiology of disease, screen drug treatments, or figure out why some patients respond well to a certain drug while others do not, to name just a few applications. However, while Big Data approaches like this hold great potential for biomedical research, many biologists and clinical researchers are averse to taking advantage of it because of discomfort with an informatics mindset. Researchers investigating medical questions tend to be "biology people" not "computer people." Running bioinformatic algorithms has required a computer science mindset that many biomedical researchers simply do not have and are not interested in cultivating. Until recently, such research was therefore restricted to true bioinformaticians; engineers with the skills to write computer code and work with command-line algorithms. Many biomedical researchers, however, have never worked "in the terminal" in their lives. So, those most familiar with patients, diseases, and other "real-world" aspects of biology and medicine have not been skilled enough to take advantage of the power of Big Data. Fortunately, a budding field of bioinformatics entrepreneurs are developing ways to bridge the gap.

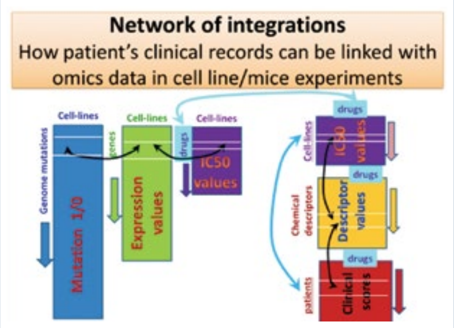
One example is the start-up ecSeq which has been offering nucleic acid bioinformatic analysis services for genomics, transcriptomics (RNA-seq), and epigenetics research since 2012. Intimately linked with the University of Leipzig, the company employs four full-time bioinformaticians (all with PhDs and publications) and an advisory network of over three dozen PhD-level researchers from the university. More akin to a creative outshoot from a university lab than a traditional corporation, this structure has laid the groundwork for a company that can both understand the way science works and quickly adapt to the rapidly changing needs of researchers. Recently, in response to demand for software that biologists could easily learn to use themselves, they rolled out a user-friendly computing platform called Seamless NGS. This software solution, designed for diagnostic labs and other practical biomedical labs, consists of easy-to-use, push-button analysis software with output that can be customized to customers' wishes. In addition to providing NGS analysis services and developing software, ecSeq also conducts training workshops in NGS analysis applications that draw an international audience of students. (Their last workshop, held in Munich, Germany, is a snapshot of the global nature of this field, with students from Germany, USA, Saudi Arabia, UK, Romania, Belgium, Slovenia, and Turkey.) With clients from universities, hospitals, biotech firms, and pharmaceutical companies, ecSeq's goal is to bring the power of computational biology to all areas of health-related research. In order to tap the great potential of NGS, the worlds of academic computational biology and translational biology/clinical applications need

Dr. Leonid Brodsky

> “One of his (Brodsky’s) foremost goals, however, is integration of heterogeneous omics analyses, through clustering and an algorithm called BiAssociation.”

> “Our collaboration with Pine Biotech will involve testing innovative big data approaches to integrate clinical and genomic information to support precision medicine clinical decisions in cancer.”

Dr. Lucio Miele



Network of integrations

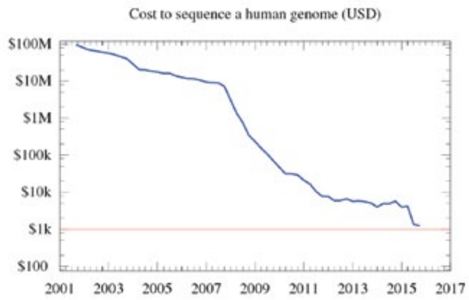
How patient’s clinical records can be linked with omics data in cell line/mice experiments

PHOTO COURTESY OF TAUBER BIOINFORMATICS RESEARCH CENTER

to meet, and that exchange needs to be an active, two-way relationship, explains CEO David Langenberger. “It is our goal to build bridges between academia and industry in order to continuously improve the technical possibilities of Next-Generation Sequencing for the health market. Together with our biomedical customers and a team of high-level academic experts, we want to find new diagnostic approaches to considerably improve people’s quality of life.”

While NGS has catapulted nucleic acid-based analyses like RNA-seq forward, parallel advances in technology have pulled other “omics” areas into the realm of Big Data. Just 15 years ago, proteomics was done using 2D electrophoresis gels, which allowed the discovery of a couple of hundred proteins via a 3-day procedure. Today’s mass spectrometer can quantify about 4,000 proteins in just one hour. Following this increase in capacity, an increasing number of proteomics datasets are becoming available in the public domain, allowing for more comprehensive comparative studies. Metabolomics, the study of small-molecule metabolites in cells or biofluids, is rapidly taking its place alongside proteomics, transcriptomics, and genomics in the pantheon of analyses dubbed multi-omics. One local company, Pine Biotech, is developing user-friendly software to allow biomedical researchers to do not just genomics and transcriptomics, but also proteomics and metabolomics, as well as specialized big data analyses such as CirSeq, which utilizes rolling-circle reverse transcription to generate multiple copies of individual viruses, allowing the detection of rare variants and fine characterization of viral populations.

Like ecSeq, Pine Biotech is closely connected with a university–the University of Haifa in Israel, where Pine’s bioinformatics analysis platform, called t-BioInfo, was created by bioinformatics professor Leonid Brodsky. Dr. Brodsky, originally a mathematician and bioinformatician at Moscow State University, moved to Israel in 1998, where he conducted bioinformatics research at Quark Biotech and Tel Aviv University before becoming a scientist at the University of Haifa. Today, he is the director of the Tauber Bioinformatics Research Center, founded and supported by Dr. Alfred Tauber, where the t-BioInfo platform was developed. With t-BioInfo, Dr. Brodsky is aiming for a platform with broader analytic capabilities than open-source platforms like Galaxy, which, in addition to being user-friendly, “is mostly about RNA-seq; our platform is much wider, covering many heterogeneous omics issues and data integration.” He goes on to explain that t-BioInfo can conduct analyses in “almost any omics research direction: transcriptomics, RNA-editing, genomics (including somatic mutations), epigenetics (DNA methylation and histone modification), mass-spec proteomics and metabolomics, structural biology, machine learning methods, and evolution in virology.” In addition, several other capabilities are under development, including “image analysis, extra-structural biology, extra-mass spectroscopy, and systems biology modeling.” One of his foremost goals, however, is integration of heterogeneous omics analyses, through clustering and an algorithm called BiAssociation.



Pine Biotech CEO Elia Brodsky is working to bring the capabilities of the t-BioInfo platform, alongside the use of Pine’s powerful servers (necessary for conducting Big Data analyses) and guidance from bioinformaticians at Haifa, to American researchers. This includes researchers close to Pine’s home in the New Orleans BioInnovation Center. “We are working with Dr. Lucio Miele at LSU. We are in the early stages of the project, [which is focused on] triple negative breast cancer.” Triple negative breast cancer is a form of breast cancer that is negative for estrogen, progesterone, and human epidermal growth factor (HER2) receptors. Since specific therapies targeting these receptors are not effective for the triple negative subtype, it has a particularly poor prognosis. Dr. Miele, Chair of the Department of Genetics at LSU and Cancer Crusaders Professor, as well as the Director for Inter-Institutional Programs at the Stanley S. Scott Cancer Center, explains the project: “Our collaboration with Pine Biotech will involve testing innovative big data approaches to integrate clinical and genomic information to support precision medicine clinical decisions in cancer. For example, is someone who has mutation X in gene Y, but also gene expression profile Z, and other mutations in specific regions of the chromosome where gene Y is AND specific clinical parameters in addition to the genomics more or less likely to respond to targeted agent XX? Information integration for clinical decision support is the key. “Specifically, with Pine, we are focusing on analyzing a large dataset including both clinical and genomic information. From this dataset we will develop hypotheses that we will validate in a clinical study here in New Orleans, focusing on predicting response to neoadjuvant therapy in triple-negative breast cancer.”

Dr. Miele is ready to take full advantage of Big Data to advance healthcare outcomes. Coming from the practical angle of treating cancer patients, he is impatient with the slow pace of integration of genomics and other omics data into the healthcare system. “We already do offer [genomics-based] tests and therapies. The main obstacle is not availability but insurance coverage. Insurance coverage has not yet kept up with the progress of medicine, and obtaining approval for high-end genomic tests is difficult, effectively restricting access to precision medicine. With new anti-cancer agents being approved on the basis of genomic tests, we anticipate changes.” He is also heartened by some good news from the FDA that gives hope that the wider healthcare system is in fact moving forward: “In a momentous first for precision medicine and immunotherapy, the FDA granted accelerated approval for pembrolizumab (Keytruda) for patients carrying mismatch repair defects irrespective of anatomical tumor site. Importantly, the approval was based on a relatively small “basket” trial, in which patients with solid tumors in diverse organs were enrolled based on microsatellite instability and DNA mismatch repair. This is very different from a standard phase 3 clinical trial, and we can look forward to more trials like these. This is the result of 30 years of molecular genetics and immunology. Without scientists dissecting DNA repair genetics and immune checkpoint receptor biology, these treatments would have never been developed. The future of cancer therapy is here.” ■