Cancer: smoother journeys for molecular data

Vivien Marx

Data integration and tool interoperability can ease analyses of cancer 'omics data and yield surprises.

Cancer 'omics data sets are now available in all sizes, especially XXL, thanks to large-scale tumor-genome sequencing efforts and the transcriptomic, epigenomic and proteomic characterization of thousands of tumors. These data help scientists decipher molecular mechanisms involved in tumor formation and cancer progression. Labs increasingly take on integrative analyses of these large, complex data, drawing on a wealth of bioinformatics tools and their own growing skill in using those tools. But these analyses are not yet a smooth, integrated ride.

Data integration is challenging on multiple levels: a comprehensive readout of tumor biology data involves more than the simple addition of different data types, and computational tools do not readily converse with one another. A number of labs and cross-lab initiatives are addressing these issues.

Tool interoperators

According to an informal survey in 2012 by scientists at the Broad Institute of MIT and Harvard, there are over 10,000 bioinformatics tools and more than 5,000 data sources. Jill Mesirov, the Broad's chief informatics officer and director of computational biology and informatics, says it is hard to estimate the number of tools, because they keep emerging, and platforms with assemblies of tools such as Galaxy and GenePattern keep expanding. In cancer research, data, too, keeps growing, thanks to tumor sequencing projects such as The Cancer Genome Atlas and the work of the International Cancer Genome Consortium.

Maria Nemchuk, Broad Institute of MIT and Harvard

GenomeSpace bridges widely used software tools, so scientists don't have to write scripts to glue software tools together, says Jill Mesirov.

For analysis and integration of data such as whole-genome or exome sequences, single-nucleotide polymorphisms, gene expression, and changes in epigenetic, proteomic or metabolomic processes, different tools are connected into a pipeline. Often scientists must write scripts to 'glue' software tools together, says Mesirov, and tools differ in their data format needs such that, for example, scientists have to convert output about gene variants from one tool so that they can look at the expression of these genes with another tool.

To address these issues, the Mesirov and Regev labs at the Broad, along with six other labs, developed the platform GenomeSpace, which is targeted toward the non-programming biomedical investigator.

To construct the platform, the team built a 'lightweight' computational bridge connecting the widely used software Cytoscape, Galaxy, GenePattern, Genomica, the Integrative Genomics Viewer and the University of California at Santa Cruz Table Browser. The tools are written in a variety of programming languages and have different architectures, but the platform lets the tools speak computationally to one another.

The GenomeSpace team also addressed data conversion by curating existing data converters and by developing new converters, to give researchers fewer and cleaner choices.

The developers have also created 'recipes' consisting of prearranged tool combinations for certain experimental questions. For example, a biologist might want to identify up- or downregulated pathways from expression data. The GenomeSpace recipe for that question: InSilicoDB as the source of gene expression data sets, GenePattern for identifying differentially expressed genes, and then MSigDB to find biological functions and pathways enriched in that set of genes.

A. Wong, Troyanskaya Lab, Princeton University

There are many gene expression data resources with data collected on different platforms; the search engine SEEK can help people explore them.

"The most popular recipes, by a long shot, are those that relate to RNA-seq data: preprocessing and QC, finding differentially expressed genes, and finding subnetworks of differentially expressed genes and identifying the associated biological functions," says Mesirov. To develop and test their recipes, they drew on their own work on gene regulatory networks in cancer[1,2]. Mesirov and her colleagues will keep developing and posting recipes and hope to entice the scientific community to submit recipes as well.

Cross-platform analyses

Choices and preferences will differ as existing tools are built into workflow approaches for integrated multi-omic analysis. Jorrit Boekel of the Karolinska Institute, Timothy Griffin of the University of Minnesota and their colleagues believe that the widely used platform Galaxy is especially promising for multi-omic types of analysis. The researchers say that after a period of rapid expansion

Simons Foundation

Hubby genes such as *TP53* get in the way of cancer data integration. Hubbiness correction is a way around those hindrances, says Olga Troyanskaya.

of the genomics and transcriptomics tools hosted there, Galaxy has been the home for multi-omic applications in proteomics and metabolomics since 2013[3].

When scientists want to move beyond their favorite tools and platforms and, for example, hunt for data broadly and then integrate it, they run into some speed bumps. Some of these are being addressed by Olga Troyanskaya and her team at Princeton University, who have created SEEK (search-based exploration of expression compendia), a search engine that lets researchers explore the many gene expression data resources that house data collected with different platforms, such as microarrays and high-throughput sequencing technologies, on over 50 instrument types. Troyanskaya is at the Lewis-Sigler Institute of Integrative Genomics and the computer science department, and SEEK's co-developers are Moses Charikar and Kai Li, her colleagues in the computer science department[4].

SEEK helps scientists find associations between coexpressed genes and limit their analysis to a particular disease or pathway. It can help researchers find the knowledge that's hidden in the piles of banked expression data, says Troyanskaya.

"Biomedical labs cannot download thousands of expression data sets, analyze each of them, come up with a systematic quality assessment metric to identify relevant and accurate data sets for their area of interest, and then somehow integrate signals from any such relevant data sets," says Troyanskaya. "In fact, simple solutions that trust each data set equally fail completely in this task."

One important SEEK element is 'hubbiness correction,' which is a way of removing genes that might skew a search. For example, well-connected 'hubby' genes can dominate the list of gene coexpression results. These hubby genes are prominent because they represent global, well-coexpressed processes, says Charikar. Hubbiness correction gets to the genes that matter in a particular query. Without this correction, Troyanskaya says, hubby genes such as *TP53* would likely be among the top results of almost any query related to cancer. With hubbiness correction, these hubby genes can be side-stepped, allowing researchers to integrate data and find connections between the data and pathways that had not been previously apparent. Cancer researchers can, for example, more quickly detect processes related to faster and uncontrolled cellular growth and see which factors enable this growth in the tumor types they are studying.

Moving beyond point-and-click

Experimental biologists used to resist computational tools that went beyond point-and-click analysis, says Wolfgang Huber, an 'omics researcher at the European Molecular Biology Laboratory. But biologists have learned to edit, adapt and compose scripts.

A pre-devised query can be fine, such as a GenomeSpace recipe for RNA-seq data to find differentially expressed genes, locate subnetworks of such genes and identify function, says Huber. Though complex enough to require computational tools, the analysis is often straightforward enough for a pre-assembled analysis pipeline. But there are issues to keep in mind, because data integration is not easily standardized.

Among the issues to heed, says Huber, are batch effects: 'omics data are calibrated not in universal physical units, such as meters or kilograms, but rather in units specific to a lab. This situation is not unlike that in the Middle Ages, when towns would define their own measures in feet, inches or stones, he says. The integration of data across different cities, as was sometimes needed for commerce and trade, could be challenging.

Another hurdle to integration is that different data sets can have different rates of false positives and false negatives, says Huber. They might have different biases, which push measurements slightly off in a systematic way, such as when a bathroom scale adds two kilograms to a person's true weight. "This is not really a problem as long as one uses these data to monitor one's own weight over time, but might become a problem when integrating the data," he says.

EMBL

'Omics data are not calibrated in universal physical units, so researchers must be cautious about them, says Wolfgang Huber.

Yet another issue relates to semantics. Integrating genetic, transcriptomic and proteomic data means taking into account that genes are linked to transcripts and transcripts are linked to proteins. "This is simple at first sight, but can become arbitrarily subtle when alleles, paralogs, isoforms, post-translational modifications are important for the biology considered," says Huber.

These challenges are neither new nor particular to genomics or cancer, but they apply to all reanalyses of data. Integrating a data set from another lab or a database with one's own data to compare diseases A and B might seem straightforward, says Huber. But it also implicitly, and perhaps unbeknownst to a researcher, might compare young people with old, people of one gender with people of another, or people leading one lifestyle with people leading another, and the incidence of the disease might be different between these groups. After completing an integrated analysis, a scientist might end up reporting a molecular effect purported to relate to these diseases, when in fact the differences are explained by the underlying lifestyles.

When designing and adding to their Bioconductor data analysis platform for 'omics analyses, Huber and his colleagues keep these aspects in mind, he says[5]. They want to educate biologists about statistics and programming and entice computer scientists and physicists to learn about biology. They also want to ensure that analyses can be inspected and re-run by others.

Researchers with particular areas of expertise need to be able to add to the functionality of software tools and make them interoperable. Often these scientists with specialized expertise are not necessarily the best software engineers, but they know which applications would be helpful and which methods are currently being used, says Huber. Code produced by these scientists might need to be polished by professional programmers, but over the course of the past 14 years with Bioconductor Huber has seen users becoming active tool developers.

Data integration gives scientists a more systems-level view of their data—for example, a way to query pathways involved in tumor progression. But surprises await those integrating 'omics data.

Data integration surprises

Data integration can enable prediction. For example, researchers might integrate data to see whether genomic, epigenomic and proteomic signatures of tumor cells predict reactions to perturbations. This knowledge can help determine whether a cancer patient might react favorably to an approved drug or one still in clinical trials. However, such integrative prediction is often difficult.

In cooperation with the National Cancer Institute, a team of scientists—the Dialogue for Reverse Engineering Assessment and Methods (DREAM)—recently compared the ability of almost four dozen drug-sensitivity prediction algorithms to predict the effects of over two dozen compounds on over 50 breast cancer cell lines for which genomic, proteomic and epigenomic data had been collected[6]. For example, the cell lines had been profiled for mutations, copy-number variation, methylation, gene expression and protein abundance.

IBM

Data integration can deliver interesting, unexpected results, says Gustavo Stolovitzky.

The team ranked the results and found that the predictions were robust for the majority of tested compounds. But there was also an interesting, unexpected result, says Gustavo Stolovitzky, the IBM researcher leading the DREAM project: most of the signal came from the gene expression data. Adding the other data types did not boost the performance of the algorithms as much as gene expression data did.

There are many ways to explain this result, says Stolovitzky. Gene expression experiments are probably the most mature of the technologies in this data integration and prediction comparison. The outcomes likely have the least amount of technical noise, and labs are perhaps most familiar with analyzing these types of data. But he also wonders whether the scientific community just might not yet know how to use and integrate data from different types of molecular characterizations, such as copy-number variation or proteomics data.

Perhaps, says Huber, researchers are not yet as adept at reading the genome as they are at reading transcriptomes. Exome-seq data do not cover potentially important regulatory regions, and many tumor-driving mutations are individually rare, because tumors arise by random mutation and selection. Different mutations in the same pathway can cause the same tumor phenotype, but, Huber says, “we currently don’t understand these groupings.” Another complication is that proteomics methods are currently limited by sensitivity.

The transcriptome offers an important readout of the genome. Anything that is important for a cell—its genome, its metabolic and proliferative state, its epigenome and other factors—tends to be reflected in the transcriptome, says Huber. This may be particularly true for cell lines, whereas in multicellular organisms with plenty of tissue heterogeneity, this aspect might be more nuanced.

Cancer biologist Stephen Friend, too, sees gene expression readout as important for integrating data. Friend is the former vice president of cancer research at Merck, and he now directs Sage Bionetworks, a nonprofit that builds platforms that enable collaboration and data sharing.

At Rosetta Inpharmatics, a genome-analysis company Friend cofounded and which Merck acquired, he and colleagues did large-scale gene expression experiments. He likens the

The same tumor type can have utterly different mutational patterns. That makes comparing and integrating such profiles tough, says Trey Ideker.

experiments to hanging 20,000 microphones in a cell in order to listen to each gene's activity. Such types of experiments, although perhaps not practiced on the same scale, were plentiful in labs between 2002 and 2010, he says. Then, sequencing emerged, prices dropped and labs began finding 'actionable' mutations that suggest a clinical course of action and offer a view of tumor biology. At the time, many scientists belittled arrays for their indirect and inferior approach, whereas the exact readout of changes in the DNA sequence was seen as the "absolute truth."

What ensued, says Friend, was a conflict between two schools: one that looks to mutations in an altered component list in order to see what is amiss in disease, and another that looks at expressed genes as a readout of the integrated state of the cell. "Eventually the two of them are going to have to meet," he says. That meeting will require an understanding of genomic alterations and integration of these data in a network context, he says.

That kind of data integration calls for mutational data, expression data and more. But for now, the abilities to capture many aspects of the altered component list do not match the aggregated benefit that comes from expression profiling, says Friend. Scientists need a much better understanding of the consequences of the altered component list to use it as a direct map of disease.

For now, the "20,000 microphones" in expression analysis deliver a better readout of the state of the cell, Friend says, which explains the DREAM result highlighting the power of gene expression data. "Over time I would be surprised if that continues to be true," he says.

As Stolovitzky explains, the machine learning algorithms from the DREAM competition all calculate the rules governing cancer cell behavior in smart and sophisticated ways. But, he likes to say, when it comes to machine learning, the machine does learn, but we don't.

For machines and scientists to learn how to better integrate data, it will take increased data riches from a variety of resources and experiments. Researchers will still need to fit the pieces together into an integrated narrative to explain, for example, why a cell turns cancerous. That narrative can often be incomplete, as it is, for example, in the case of triple-negative breast cancer, which a genetic test might reveal. Physicians know this diagnosis means a patient faces an aggressive tumor. The answer to what makes the tumor tick so aggressively stands to come from data integration of all pieces of information: gene expression, proteomic data and clinical data. Despite the many efforts underway, the best way to interrogate these data in an integrated way remains elusive, says Stolovitzky.

An example of a different sequence of integration steps is one from the lab of Trey Ideker at the University of California at San Diego[7]. The analysis starts with mutational profiling, then groups genes according to pathways in such a way that molecular patterns appear with more prominence. These patterns are then used to group tumor subtypes.

The subtypes grouped according to mutation data were different than the subtypes defined by expression data, says Ideker. While the reason for this difference still has to be resolved, he guesses that the mutations are upstream and causal, whereas expression changes indicate downstream responses and effects from the tumor microenvironment.

From genes to networks

Ideker and his colleagues note that stratifying tumor data according to expression profiles is challenging for a number of reasons, such as the ample opportunities for overfitting data. Comparing and integrating mutational profiles is tough because the same tumor type can have utterly different mutational patterns, a situation that hinders research and makes treatment choices hard. Ideker and his team used mutation profiles to integrate heterogeneous data from a group of ovarian, uterine and lung cancer patients whose genomes were characterized as part of The Tumor Genome Atlas.

Even though individual mutations from one tumor genome did not match those from the next, the researchers found a way to group the data according to gene pathways such as cell proliferation and functional networks connected to such aspects as protein transport, beta-catenin signaling and fibroblast growth factor signaling.

Many of the 350 ovarian tumors they looked at, for example, had mutations in the tumor suppressor gene *TP53* (*p53*), which are well-known cancer drivers. But, says Ideker, they sought an integrated result of molecular events that were exclusive to a subset of samples—and they found it.

Ideker hopes that the success in stratifying these 'omics data with a pathway-based integrative view is a reason to continue with tumor sequencing, and also a clue that will lead researchers to place at least equal emphasis on understanding how these sequenced genes connect to one another in networks and pathways. The network perspective made it much more apparent which functional areas the mutations in these tumors were hitting. "It's the same damn networks," he says of the finding that made it possible to stratify the data into delineated tumor subtypes.

To hunt for network-based clues in another group of genomes, Ideker is continuing with this approach. These data are from ovarian cancer patients whose tumors are unperturbed by chemotherapy. Studying the DNA sequences and the mutations yields no pattern, he says, but a pattern emerges when the team integrates the mutations of these 'non-responders' in a network context. He is currently validating these results in cancer cell lines, which will help clarify the molecular mechanisms that lead the cells to resist chemotherapy.

A networked view of data offers some answers, and it yields more questions. "The network gives us a high-level clue as to what those mechanisms could be, but then you have to delve into the biochemistry," says Ideker.

Researchers must remember that data integration in network-based analysis does not automatically deliver conclusions about causality. "I believe that we simply don't know enough about the connections to add that linear alignment so far," Ideker says. In order to expand the abilities of network-based analysis, whether for tumors or for other biological systems, researchers will need more data, as well as more tools that are able to communicate with one another.

1. Wong, D.J. *et al. Cell Stem Cell* **2**, 333–344 (2008).
2. Ben-Porath, I. *et al. Nat. Genet.* **40**, 499–507 (2008).
3. Boekel, J. *et al. Nat. Biotechnol.* **33**, 137–139 (2015).
4. Zhu, Q. *Nat. Methods* **12**, 211–214 (2015).
5. Huber, W. *et al. Nat. Methods* **12**, 115–121 (2015).
6. Costello, J. *et al. Nat. Biotechnol.* **32**, 1202–1212 (2015).
7. Hofree, M. *et al. Nat. Methods* **10**, 1108–1114 (2013).

Vivien Marx is technology editor for *Nature* and *Nature Methods* (v.marx@us.nature.com).