Biological analysis and interpretation for improved research outcomes

In the last few years, technological advancements in the life sciences have changed many ways in which we think about research. Next-generation sequencing, qPCR and microRNA offer new avenues to ask and answer research questions in more detail and in less time. However, much of the effort today centres around data gathering, and many researchers are realising that collecting massive quantities of data is not the same as biological discovery.

By Dr Douglas Bassett

Thought leaders in the life sciences industry are actively discussing data analysis approaches necessary for actionable biological insights[1,2], but the challenge of what to do once data is generated is still often not given as much of our collective research mindshare as it merits. With the increasingly broad adoption of technologies that produce large quantities of data, and more research groups leveraging multiple types of data, there is a critical need to implement solutions that enable researchers to effectively sort through and better understand their data outputs to translate data into actionable information and insights.

In academia, few are more knowledgeable about a molecular process or disease under study than the graduate students and their PIs who are at the forefront of the research. However, when large datasets are being interpreted such as whole-genome or exome resequencing studies or RNA-seq experiments, important potential causal relationships can easily be overlooked if they are not intimately tied to known players. Given the breadth and volume of data involved, manually assembling related connections to genes, diseases and other interesting biology is not only time-consuming, it is difficult to do thoroughly, given the vast quantities of published literature and information now available. Complex cause-and-effect or 'net effects' relationships of differentially expressed genes and/or deleterious or gain-of-function mutations can be difficult, if not impossible, to identify systematically without having a very high-quality curated knowledge base at hand.

While there are excellent freeware tools available, they often require 'command line' expertise to utilise and integrate to their full potential, placing them just out of reach of some of the researchers who best understand the disease and molecular processes under study. Many freeware tools also only cover one aspect of analysis – pathways, or molecular interactions, or biological processes and partial annotations – rather than truly providing an integrated view of overall biology. It can be extraordinarily time-consuming to integrate and interpret the results from various independent pieces of software. Also, it is difficult to find research tools that are built on the foundation of a high-quality, comprehensive knowledge base of findings curated by experts from the biomedical literature – critical to provide the optimal context for biological interpretation of these large, integrated studies. The results from individual tools are difficult to reconcile and interpret collectively, often resulting in overlooked insights that could have been discovered had all factors been evaluated together and manually attempting to integrate multiple tools is error prone.

On the other hand, commercial organisations frequently use a combination of in-house solutions to analyse data. This approach also presents

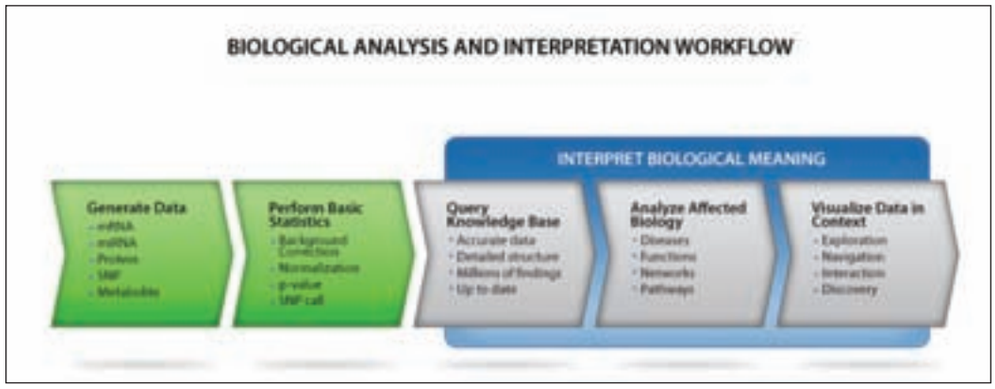

Figure 1
Biological analysis helps researchers transform statistically significant data into evidence-backed insights about affected biology by leveraging biological knowledge

challenges. Bioinformatics teams sometimes do not have the capacity to leverage the vast body of published biological knowledge in their tools. Instead, much like freeware tools, these solutions may focus on a particular subject matter or facet of biology. This makes it more difficult for researchers to obtain a systems-level understanding of the impacted biology.

In either scenario, researchers face a lengthy, time-consuming process that results in limited information from disparate sources, along with a significant risk of missing novel connections or crucial insights from the data. And most importantly, the data analysis often falls short of effective biological analysis, so scientists are frequently left with unanswered questions about the larger biological context of their results, and the full potential value of the extensive dataset that has been generated remains unrealised.

This article will discuss how biological analysis addresses these challenges and provides an efficient way to get actionable biological meaning from large datasets. It will provide a definition of biological analysis, discuss the benefits of incorporating biological analysis into a research workflow, and examine what factors are necessary for efficient and accurate biological analysis.

The importance of biological analysis

Biological analysis is a scientific approach that combines analytical tools and biological content in one place, so researchers can obtain a fundamentally deeper and broader understanding of biological relationships and processes known to be connected to experimental observations, and the translation of that understanding to actionable insights and concrete hypotheses. Biological analysis can transform basic data analysis results into useful research outcomes, so researchers can leverage what they have discovered to make informed decisions, generate well-formed, testable hypotheses, design follow-up experiments, and provide compelling biological and mechanistic evidence for results (see **Figure 1**).

Biological analysis represents a very natural but extraordinarily powerful extension to improve traditional data analysis and interpretation approaches. It connects molecular information coming out of various experimental platforms to help researchers understand whether the genes from their experiment work together as molecular modules, assess their impact on higher level biological processes and phenotypes, and determine whether or not those collections of events also impact diseases. For example, if the goal is to identify molecular mechanisms that link a genotype to a phenotype, biological analysis is the crucial approach that links gene expression changes in cancer cells to the observed cellular phenotype or related disease phenotype.

Biological analysis can rapidly identify relationships already known to be involved in experimental changes. These capabilities help researchers by providing a broader biological picture when they analyse experimental results. For example, by examining a gene of interest in the context of a pathway, it becomes easier to get a sense of what is happening in an experimental model. What are the key players? What are the known interactions? What are the top pathways involved in the data set? Asking these kinds of questions and relating experimental data back to the larger biological picture is a key part of biological analysis.

In addition to providing a more relatable, high level biological picture, biological analysis can identify key findings and novel discoveries from large amounts of data. For example, a basic microRNA dataset might return 13,000 potential miRNA targets. Using biological analysis, a researcher could begin to narrow down and prioritise that list, using questions like: which of those are experimentally

demonstrated and involved in particular pathways of greatest interest to me? From those, which mRNAs are expressed in a relevant tissue, and have inverse expression from their matched microRNA? Which are known biomarkers?

These advantages all demonstrate another key benefit of biological analysis, which is that it significantly decreases the time it takes to obtain a novel discovery. The integration of a wide variety of structured biological content in one place, in combination with analysis tools that let researchers effectively use that content to narrow in on a targeted set of experimental findings or explore outward from their findings to other biological relationships, saves an immense amount of time over manual, piecemeal or overly specific tools and approaches.

Biological analysis also speeds the process of creating a validated and testable hypotheses, either at the end of an experiment using insights gained from experimental results, or prior to beginning a new experiment. Generating a hypothesis that can be interrogated and vetted against published research provides added confidence that wet lab testing makes sense. With biological analysis tools, researchers can challenge their hypothesis and examine it in the context of additional layers of biological and chemical knowledge before investing in the physical experiment.

By informing decisions throughout the experimental cycle, biological analysis decreases the time it takes to get from instrument to insight, and improves the ability to complete that process without dead ends, mistaken directions and other research obstacles (see **Figure 2**).

Considerations necessary for strong biological analysis

What are the critical enablers for biological analysis? The most crucial requirement is the tight integration of powerful analysis tools with an associated high-quality content database. When that content is combined with an analysis platform, the resulting tool can take current, accurate scientific discoveries and make them accessible within the context of a researcher's specific data analysis task. This is a powerful resource for searching for relevant, validated knowledge, and for interpreting experimental results in the context of larger biological systems.

Given current technology and the amount of available information, it can be highly inefficient to research discoveries or genes individually and even harder to put them into the context of an existing dataset with expression changes. A more streamlined research workflow occurs when the most relevant scientific findings can be summoned at the precise time they are needed.

Combining content with analytics is not enough, however. The content must also be of sufficient

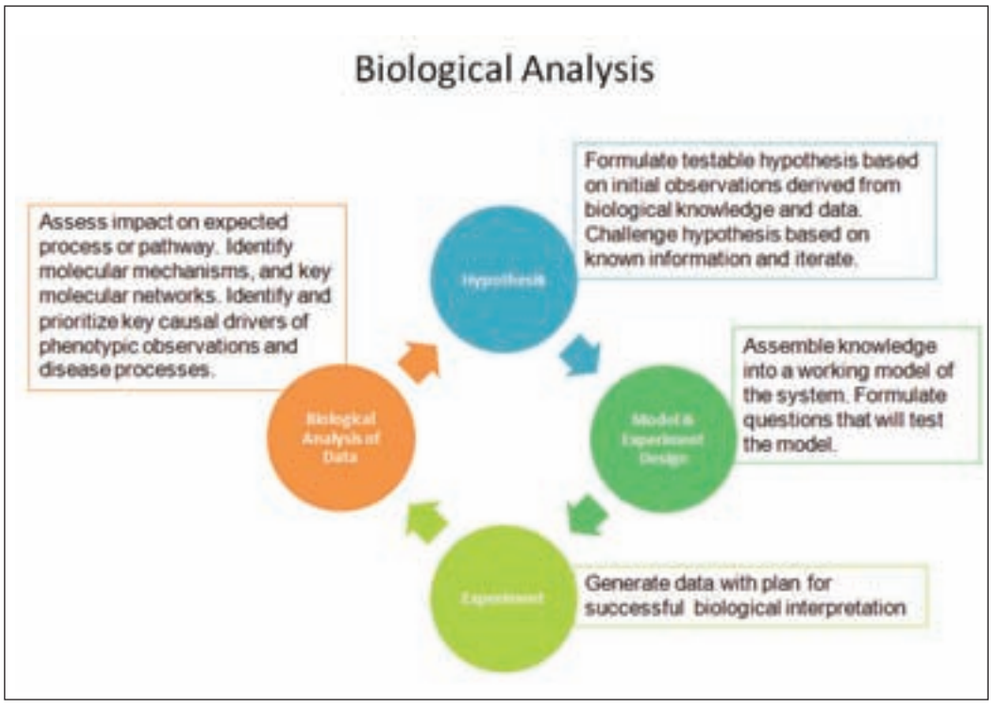

Figure 2
Biological analysis informs decision-making at all phases of the experimental cycle. Adapted from H. Kitano, Computational Systems Biology, Nature 2002 Nov 14;420(6912):206-10

breadth, quality and detail to enable sophisticated and accurate biological analysis. An effective biological analysis tool must make it easy for researchers to connect their data with that biological information through powerful analytics and an intuitive interface that encourages exploration and the generation of novel insights. The following sections explore these further.

Key analysis capabilities

Data analysis means different things to different people. Often, statistical analysis is considered the first step. Or, researchers turn to analysis software to indicate a few relevant pathways or diseases. But comprehensive biological analysis is a deeper dive. Researchers need to filter down their large amounts of initial data to focus on the most interesting and relevant information from their experimental results. They then need tools that allow them to explore outward to identify interesting connections and gather supporting evidence. Ultimately they need to understand if their experimental data holds together in a cohesive and supported biological story based on the known body of biological knowledge.

Three key scientific capabilities can enable this approach: data filtration, which allows researchers to dynamically interact with the analysis to quickly focus on relevant results; data exploration tools that widen and narrow the scope of an inquiry along different criteria as needed; and finally visualisation tools that allow insights to quickly emerge from today's large datasets and help researchers visualise networks of interacting molecules, understand relationships to larger biological processes, and effectively sort and group interesting genes. These capabilities enable researchers to approach a single biological problem from multiple angles and result in substantially cleaner, more relevant, more accurate and more verifiable data.

Data filtering

One of the most powerful and time-saving aspects of biological analysis is its ability to help researchers rapidly narrow in on what is most relevant in those experimental results and identify a small body of information, such as a subset of relevant genes, for follow-up and investigation. It allows researchers to quickly discard information that won't be helpful to them, and focus on examining the biological functions, diseases, molecular interactions and biological pathways that are significantly affected in their experiment. For example, removing uninteresting molecules from complex molecular networks reduces the networks' size to make them more tractable, reduces the number of molecular relationships to consider, and produces results that are more accurate and more relevant to a particular experiment.

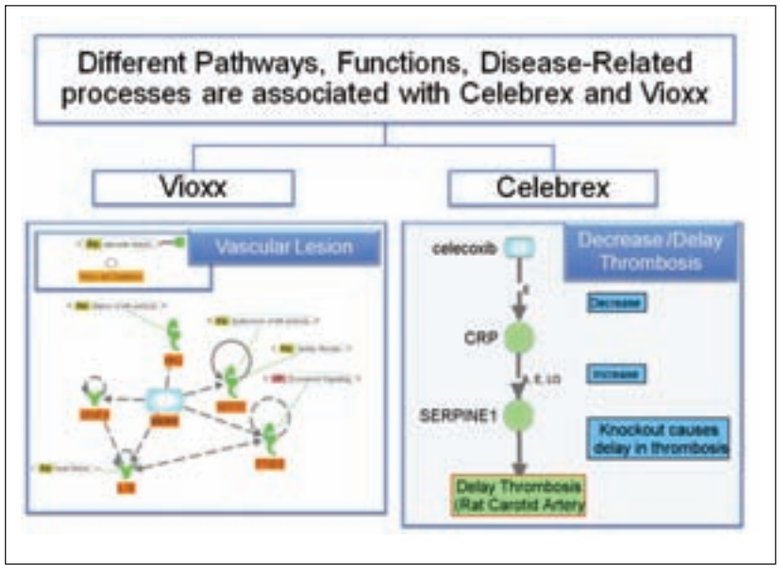

Figure 3
Clear mechanistic differences between Celebrex and Vioxx

As an example, a set of 300 significant genes is an interesting initial result from a data analysis, but the real power of biological analysis comes when you consider the utility of narrowing that down to only 30 cancer-pathway specific genes, and from there identifying four known to be biomarkers of a particular form of cancer. Or, from those same 300 significant genes, you could identify the ones known to be highly upregulated in your experimental sample, and filter down to those known to be downstream of a particular target molecule, or those known to be present in liver tissue.

A good example of this is a 2009 paper published in the Journal of the American Medical Association, where researchers integrated different types of molecular data to generate a network model of genes interacting in the promotion of gliomagenesis. From their data, they identified 214 genes affected by dosage effects in the glioma genome that were connected by protein-protein or functional interactions. From this, they organised genes and proteins into functional networks and examined genes with high connectivity as a proxy for good potential therapeutic targets. They filtered down to 11 highly connected hub genes with tumour-promoting functions and several of these had a known biological role in gliomagenesis. Using a tool that allowed them to quickly filter based on

biological information, such as functional role or level of connectivity, they identified a compelling subset of genes for further investigation.

Data exploration

Producing novel insights depends on an exploratory approach to data which can synthesise multiple levels of biology in a unified, efficient approach. Biological analysis approaches produce comprehensive, high-level summaries of the biology most significantly affected in an experiment. These can include molecular networks, disease processes and biological pathways. However, just as statistical analysis is not enough, a limited biological analysis is not enough. Discovering implicated processes can only take the researcher so far. With an exploratory approach, researchers can investigate interesting neighbouring molecular interactions, related biology and possible avenues of connection. When exploration is applied to a small set of interesting genes that are the result of a filtering exercise, the effort spent is minimal and the amount of relevant knowledge returned is maximised.

For example, biological analysis can be applied to understand and identify the key mechanistic differences between drugs (see **Figure 3**). Celebrex (celecoxib) is on the market today; however, Vioxx (rofecoxib) has been withdrawn due to cardiovascular side-effects. Both drugs inhibit the same target, COX-2 (or PTSG2), and are NSAIDs (nonsteroidal anti-inflammatory drugs) used for the treatment of rheumatoid arthritis. Using biological analysis to explore outward from their target molecule can help identify where these drugs differ in terms of the pathways they impact, the cellular and toxicity phenotypes they play a role in and the downstream genes they impact. The resulting comparisons can be used to generate testable hypotheses around drug mechanism of action and mechanism of toxicity.

Data visualisation

Visualisation tools, such as interactive pathways or molecular network building tools, help researchers intuitively understand their data at a glance. Researchers can easily redefine the unit of study from a single gene to a more functionally comprehensible pathway, cellular function or phenotype. The visualisations highlight relevant information, and bring multiple levels of biology together within a single view. Visualisation has been widely recognised for its benefits within the field of biological research.

Pathways provide a wealth of information about molecular interactions, including upstream and downstream effects, and allow researchers to make

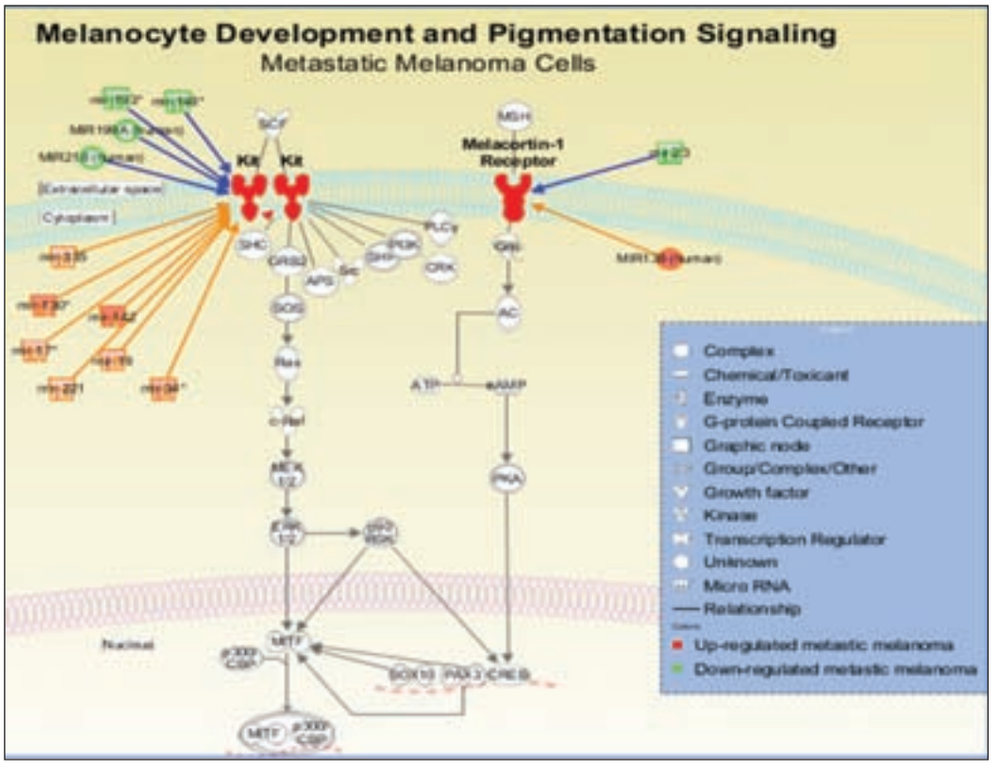

Figure 4
Pathways provide an intuitive visual representation of key biological relationships causing experimental observations

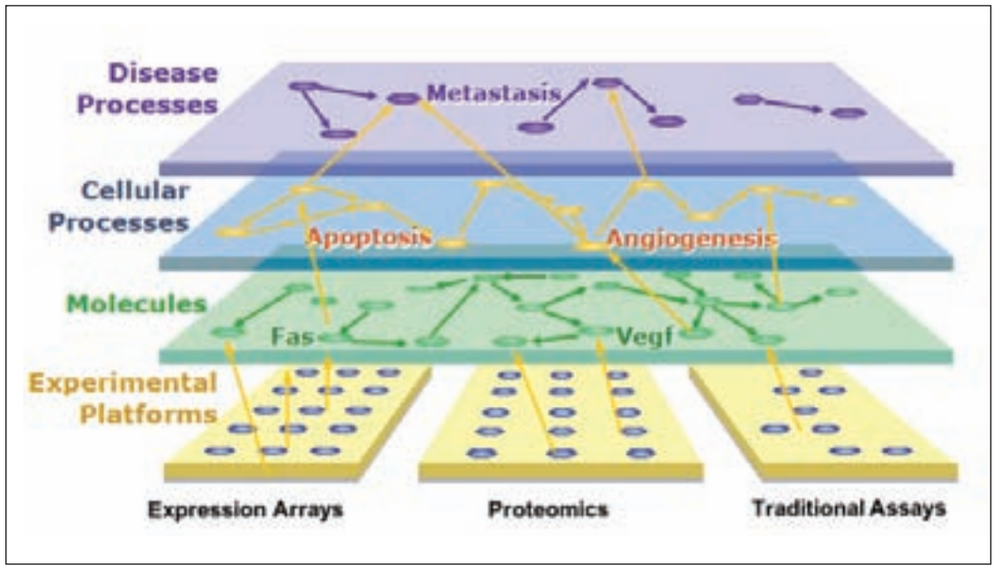

Figure 5
A broad and comprehensive knowledge base with information on multiple levels of biology helps researchers move fluidly from expression changes to an understanding of implicated biological processes and effect on disease progression

biologically-informed decisions about next steps. However, with an exploratory approach to biological analysis, a pathway becomes more than a visual representation of fixed information. The most critical aspect is the dynamic nature of these explorations. Unlike static pathways used for reference, dynamic pathways backed by strong biological knowledge bases help researchers quickly associate their results with relevant biology. When combined with filtering tools, researchers can exclude irrelevant information, include additional information most relevant to their models and verify facts. This exploration amounts to enormous time-savings in data analysis and results in more meaningful research outcomes.

A single tool that provides an intuitive interface and advanced visualisation techniques does more than speed up research productivity. It actually improves research efficacy by facilitating better insights that might otherwise be missed. According to a recently published article, "…techniques that allow the researcher to explore data interactively (rather than delivering a static view) will facilitate exploration and increase the chance of finding interesting observations or patterns"[3].

Using a knowledge base

Strong data filtering, exploration and analysis are made possible by leveraging a detailed knowledge base. For example, it is not possible to narrow down genes in a dataset known to be present in liver tissue, unless reference is made to a knowledge base that includes information about which tissues your genes are located in. The more detail a reference knowledge base contains, the more powerful filtering and exploration tools can be.

In other words, biological analysis can only be as good as the content being leveraged. Content quality is about more than just accuracy or breadth. Biological analysis must use a comprehensive, accurate and up-to-date knowledge base in order for researchers to accurately interpret biological data within the context of molecular mechanisms, and relate a wide variety of molecular events to higher-order cellular and disease processes, organismal physiology and pathophysiology.

Comprehensive

To relate more specific molecular events to larger-scale biological processes and diseases, the knowledge base must include information about a wide range of biological objects and their relationships including proteins, genes, metabolites, protein complexes, cells, cellular components, biomarkers, tissues, organs, small molecules, cellular phenotypes, pathways and disease processes. Missing any one of these components will result in the inability for the researcher to fluidly infer novel biological connections and benefit from related biological discoveries that might be relevant but which occur on a slightly different biological level (see **Figure 5**).

Content should be derived from a wide range of peer-reviewed literature. Additionally, it is important that a comprehensive knowledge base integrate the benefits of information found in databases that are ubiquitously used and familiar to researchers. The more sources that are included,

References

1 Batchelder, Keith, as quoted in http://rna-seqblog.com/data-analysis/determining-standards-for-rna-seq-data-analysis-biological-interpretation/.

2 Krohs, Ulrich. Convenience experimentation. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, Available online 28 October 2011.

3 Kelder, T, Conklin, BR, Evelo, CT, Pico, AR (2010). Finding the Right Questions: Exploratory Pathway Analysis to Enhance Biological Discovery in Large Datasets. PLoS Biol 8(8): e1000472. doi:10.1371/journal.pbio.1000472.

the better and more rich the returned findings will be, and the easier it will be to understand implicated biology.

Accurate

Data and text mining technologies provide speed, but there is a greater likelihood for inaccuracy such as false positives or even unrelated results. To maintain the needed level of accuracy for reliable biological research, content should be manually reviewed by scientific experts for accuracy and for correct structuring, to ensure that it is mapped to other terms correctly and that detailed context is included. While manual review is time-consuming, and requires a significant investment, this type of effort is hugely impactful and is more reasonably implemented in a commercially available database, where the ongoing time and investment can be subsidised and maintained.

Up-to-date

The content in the referenced knowledge base should be updated frequently enough to allow researchers to keep pace with the latest discoveries. Without constantly refreshed information, the usefulness of the analysis platform will rapidly decline over time. Ideally, new information should be added weekly, to avoid a noticeable lag time and to maintain the timeliness of the knowledge base.

Conclusions

With advances in data-generation technologies, the generation of high-quality data no longer represents the fundamental research challenge. Instead, the ability to meaningfully understand that data is the critical success factor. Explaining experimental observations in terms of complex biological relationships and teasing apart upstream causes and downstream effects requires that researchers plan for this step prior to the selection of instrumentation and reagents. As experiments are designed, researchers should ask: "How can I prepare to fully understand this data? How can I ensure I'm not wasting my investment in instrumentation and reagents, and how can I maximise the biological meaning I can extract from this data?"

Biological analysis yields results that are biologically familiar, grounded in accurate and recent biological knowledge, visually oriented, and that support exploratory, interactive analysis and display. These approaches yield accurate, direct and clear connections of phenotypes with the current physiological and disease states, genotypes and clinical observations. They enable a big picture of the affected biology and help researchers quickly and effectively narrow in on new information within their experimental data that provides valuable insights into the biological systems under investigation.

Effective biological analysis tools streamline workflows and introduce more efficient and accurate approaches for unlocking the maximum biological meaning from 'omics data. Examination of the scientific literature indicates that thousands of researchers are drawing upon these approaches to fuel advancements in different disciplines and therapeutic areas. Backed by biological analysis strategies, they are making well-informed decisions and generating groundbreaking discoveries from their research data. DDW

Dr Douglas Bassett *has a PhD in Genetics from Johns Hopkins and an MBA from the University of Washington. After completing a fellowship in bioinformatics at the National Center for Biotechnology Information (NCBI), he joined Rosetta Informatics in 1997 as head of its computational biology program and later became head of Rosetta's software products and services team. In 2001, Rosetta was acquired by Merck where Doug moved from General Manager of the Rosetta Biosoftware business unit to become Merck's Executive Director of Informatics & Molecular Profiling, and in 2008 became the site head for Merck's Seattle research facility. Doug brings to Ingenuity a deep scientific, software and bioinformatics expertise and combines it with a strong General Management perspective and experience. He currently serves as the Chief Scientific Officer and Chief Technology Officer of Ingenuity.*