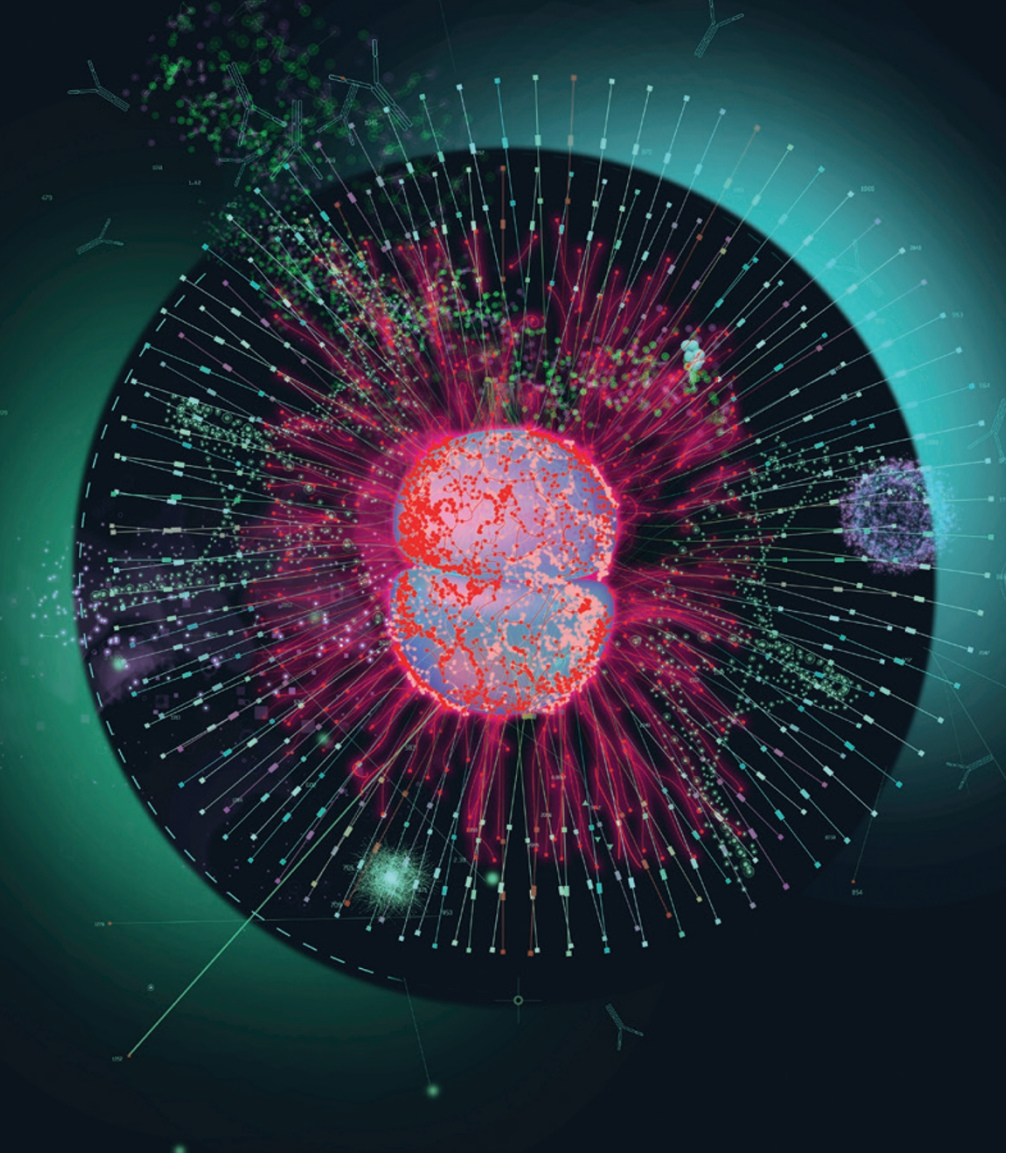

TATIANA PLAKHOVA

COLLABORATIONS

Mining the motherlodes

Collaboration and competition are spurring on major '–omic' projects.

BY KATHERINE BOURZAC

When actress Angelina Jolie announced in 2013 that she'd had a double mastectomy to reduce her chance of developing breast cancer, after testing positive for a genetic risk factor, the *BRCA* genes responsible were all over the media. These genes carry a significant risk: 55–65% of women with a harmful *BRCA1* mutation, and 45% of women with a mutation in *BRCA2*, develop the disease by the age of 70.

Jolie's case involved a single gene, *BRCA1*, that markedly increased the risk of a specific disease, but the risks of developing genetic diseases are usually much more complicated than that. These complexities are being explored by the many huge research efforts that have been launched in recent years.

Collaborations involving hundreds of scientists and computational biologists are starting to make sense of genomics, proteomics and a host of other '-omics'. Researchers are tracing the twists and turns as thousands of different forms of proteins are churned out and modified (see 'High-protein research', page S6). They are mapping the molecular pathways that flow into or away from different diseases, and are examining the effects of other factors, such as bacteria, on the human body (microbiomics). They are building and testing algorithms to predict how all these '-omic' signatures connect to human health. And they are collaborating to share their ideas and keep each other on track (see 'New eyes on the prize').

➲ NATURE.COM
You can read more about bioinformatics competitions here:
go.nature.com/ihed2h

These large studies make it possible to identify and focus on risk factors for particular diseases. This research, which should enable more personalized treatment for individual patients, is creating huge data sets. Finding rare variations in the genome — and being sure they are not missing something — means sifting through the three billion base pairs in the genomes of tens of thousands of volunteers. To make it work, clinicians from across the world are working with bioinformaticians and computer scientists on a grand scale.

In the process, these researchers also are evolving the art and science of collaboration in the era of big data.

THE QUEST FOR BURIED TREASURE

A disease-focused approach to the genome often involves so-called genome-wide association studies, which are particularly well established in cancer research. In breast cancer, for example, genome-wide association studies have revealed about 90 variants — 'typos' in the genomic code — that are associated with the disease. Of these, only five occur in parts of the genome that code for proteins, says Sara Lindström, a genetic epidemiologist at Harvard University in Boston, Massachusetts.

The other 85 breast-cancer variants are mostly a mystery. "When you see one of these signals, it's not clear if it increases disease risk, or if it's just correlated with disease," says Lindström. Sifting out the important variants requires knowledge of what all these parts of the genome do.

One of the biggest resources for computational biologists tasked with sorting genomic cause from correlation in such puzzles is the Encyclopedia of DNA Elements (ENCODE). Launched in 2003, ENCODE is a mammoth collaborative project funded by the US National Human Genome Research Institute, which maintains a publicly available, searchable genome database.

In 2012, 442 researchers in 32 labs jointly released ENCODE papers that connected more than 80% of the human genome to specific biological functions and identified more than 4 million regions where proteins hook up with DNA (see J. R. Ecker *et al. Nature* **489,** 52–55; 2012 and references therein).

"If you have a favourite gene, you can look it up in ENCODE and find out what regions are likely to regulate that gene," says Michael

Snyder, a Stanford University geneticist and one of the leaders of ENCODE. A breast-cancer researcher, for instance, might find out that a genetic variation uncovered in an association study is a target for a particular transcription factor, a protein that regulates gene expression. That regulatory protein might then be a new target for therapy.

Complementary approaches taken by researchers from 28 institutions are filling in this genome encyclopedia. Many participants study RNA, while some focus on transcription factors or on the regions of the genome where these regulatory elements attach. And still others carry out mapping and data analysis.

Sometimes the sheer size of the ENCODE project can slow things down. A postdoc's idea must be vetted by a larger group, for example, and sometimes researchers have to wait for other labs to finish their work before they can publish a paper, says Manolis Kellis, a computational biologist at the Broad Institute in Cambridge, Massachusetts.

But such problems are far outweighed by the benefits of working together, he says. When you work alone, "bugs can be introduced, and it often takes years to find them", he says. That does not happen in ENCODE — mistakes are usually swiftly spotted by one of a large group of colleagues. The collaborative structure also encourages standardization; researchers need to call a gene or regulatory element by the same name so that they can communicate, and so that the database is searchable and user-friendly.

CANCER IN SEQUENCE

This sort of standardization is essential when dealing with more complex data. The International Cancer Genome Consortium (ICGC), set up in 2008, is trying to deal with this issue at the moment.

The original goal of the project was to sequence the healthy and cancer genomes of 25,000 people. The initial sequencing efforts were performed only on the protein-coding parts of the genome. But consortium leader Tom Hudson, scientific director of the Ontario Institute for Cancer Research in Toronto, Canada, says that now the ICGC has collected about 2 petabytes (2 million gigabytes) of data it plans to go much broader and deeper.

The ICGC will now sequence the non-protein-coding parts of the genome that ENCODE specializes in, and include more clinical information about the patients. This Pan Cancer Analysis of Whole Genomes project will also bring in data from more people — the target is 250,000 — and sequence both their normal and cancer genomes.

This scaling up in the size and scope of the project will be no mean logistical feat. So far the ICGC has brought together leaders from 78 projects in 16 countries. In a pilot of the larger whole-genome comparison project, researchers are analysing paired tumour and normal genomes from 2,600 people, which amounts to about 0.7 petabytes, says Jan Korbel, a computational biologist at the European Molecular Biology Laboratory in Heidelberg, Germany. This is large, but it is still possible to use academic computer centres to process the data.

But the group is at a crossroads. They either need "vast investment" in academic data-centre infrastructure for 250,000 genomes, says Korbel, or they must figure out how to use cloud computing for data sharing and analysis. "You could have several clouds, each specific to a country, as long as those clouds can 'talk' with one another — that is, as long as comparative analyses of data in one cloud with data from another cloud are possible," says Korbel.

"Big data is not particularly useful if you don't have analytics that you can trust."

ANOTHER VIEWPOINT

In efforts like these, standardizing data so that results from different groups are comparable and searchable maximizes the pool of information. This is important when hunting for rare variations that can only be spotted by analysing genomic data from tens of thousands or hundreds of thousands of samples. Working together also helps researchers to strengthen their analyses, says Gustavo Stolovitzky, a computational biologist at IBM's Thomas J. Watson Research Center in Yorktown Heights, New York.

Although big-data analytics can reveal patterns and connections that are otherwise invisible, they can also support a researcher's pre-existing assumptions, thereby obscuring the truth.

One common mistake is 'overfitting'. Stolovitzky likens this to preparing for a university entrance exam by memorizing a big stack of difficult vocabulary flashcards. You can study hard and memorize all the words and their definitions, but that does not mean those words will be on the test — and if they are, the test may use a different wording that throws you off.

Similarly, researchers who devise a predictive algorithm based on their own data set tend to make an algorithm that is good at predicting the results of their own study but fails to work on different data.

Another problem is simply human nature. "When we analyse our own work, we are very benign," says Stolovitzky. It is more useful to involve others, who may have ideas that would never have occurred to someone staring at the same data set all day.

"Big data is not particularly useful if you don't have analytics that you can trust," says Stolovitzky. "We've seen that if you aggregate the results of several algorithms — as long as none of them are bad — the whole is greater than the sum of the parts." That's just one more example of how, when researchers want to get the best results from biomedical big data, working together is crucial. ■

Katherine Bourzac *is a science journalist based in San Francisco, California.*

NEW EYES ON THE PRIZE

Competitions find different ways to solve problems.

Tough problems often benefit from a fresh pair of eyes. That was the thinking in June, when the US National Cancer Institute (NCI) launched a competition called 'Up for a Challenge' to find new ways of analysing breast-cancer data sets. The NCI gathered data from several research groups and is supplying them to teams that present a reasonable proposal, agree to uphold privacy standards, and meet other criteria. The NCI has offered the winner a $30,000 prize and the opportunity to publish a paper in *PLoS Genetics*.

Judges will score entries according to how well groups use innovative methods to find new genetic variants associated with breast cancer, whether the findings can be replicated, and whether they are consistent with known cancer biology. The competition will give extra points to competing groups who formed new collaborations to work on the problem. "We want to reach beyond the usual suspects, and encourage a greater diversity of people to work on these problems," says Elizabeth Gillanders, a genetic epidemiologist at the NCI.

This is one of many competition-based biomedical data projects. Among the others is the DREAM Challenges programme, set up to improve algorithm development in systems biology by Gustavo Stolovitzky, a computational biologist at IBM in Yorktown Heights, New York. The programme has expanded to ask researchers to, for example, predict disease progression and the effectiveness of drug combinations in people with amyotrophic lateral sclerosis.

In many cases, the best performers do not have a background in the specific biology involved. "Presented with a new data set, they shine," says Stolovitzky. K.B.