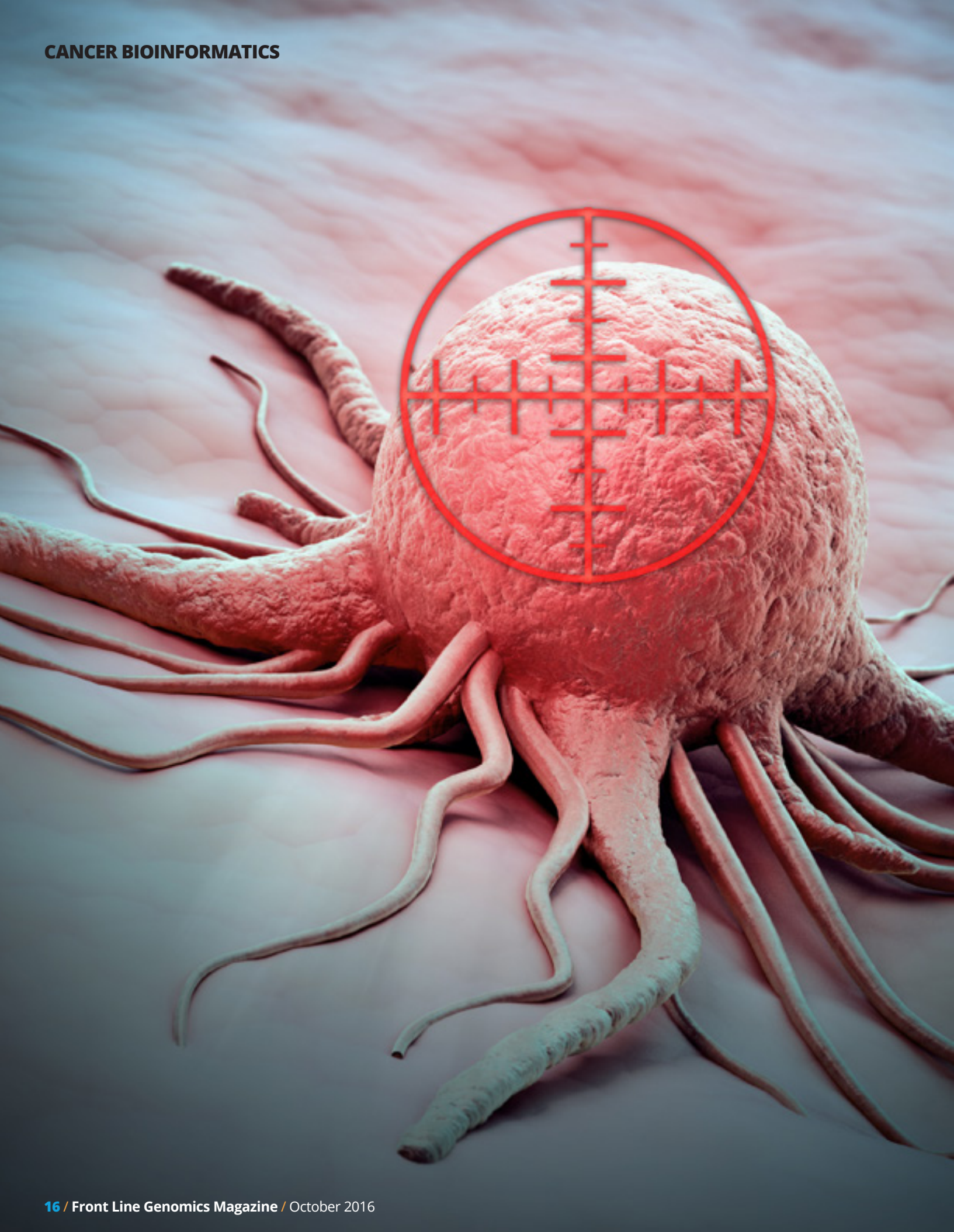

# A DISEASE OF THE GENOME

*Liz Harley*

Imagine the scenario. A patient, diagnosed with a tumour, presents to their oncologist for treatment. The oncologist takes a series of genomic profiles, from the patient and the tumour itself, and compares those to a clinical database of thousands at the click of a mouse. Computer fans whir for a minute or two, and finally a message pops up on the screen announcing a diagnosis and a course of treatment ideally suited to that patient. The patient is immediately started on the appropriate treatment, which either increases their survival or even goes on to save their life.

This is what many people imagine when they think about precision medicine. Perfectly personalised, tailored therapies that work in holistic harmony with the patient. We are not there yet, but how can we realise this vision for the future? Is it even a plausible vision? Over the past few months I have been learning about the complex, and often messy world, of how we can leverage the power of big genomic data against cancer.

Since 1971, when President Richard Nixon declared 'war' on the disease, cancer has been public health enemy number one in the western world. The National Institutes of Health describe cancer as "a disease of the genome," and studying the genetic profiles of different tumours, in individual patients, has become a major cornerstone of research into effective treatments.

Genomic information has already been brought to bear in the clinic, leading to the development of several new drugs that specifically target certain genetic profiles. Today, an enormous global research effort is focussed on building, analysing, and effectively leveraging large databases of patient and tumour profiles. →

**"IF YOU PRESENT TO YOUR ONCOLOGIST WITH A TUMOUR, YOU HAVE ANYWHERE BETWEEN ONE HUNDRED AND SEVERAL HUNDRED MUTATIONS. AND THAT'S JUST TUMOUR MUTATIONS, NOT THE OTHER GENOMIC CHANGES, OTHER THAN MUTATIONS, THAT CAN BE PRESENT."**

### BUILDING THE PLATFORMS FOR PERSONALISED MEDICINE

Data is always one of the big themes at any Festival of Genomics. Collecting, storing, manipulating, analysing, and leveraging genomic data always feature on the agenda and in the exhibition. This year, alongside the ever popular Enabling Data track, data-sharing takes centre stage with a series of presentations hosted by the Global Alliance for Genomic Health.

With the advent of next generation sequencing, collecting data at scale has become incredibly simple. Since 2005 no presentation on the collection of genomic data has been complete without a 'Flatley's Law' slide, showing how the rate of genomic data collection far outpaces the rate predicted by Moore's Law. But this rate of data acquisition brings with it two problems for precision medicine: centralisation and integration. Simply put, genetic data is being collected into different databases all over the world that is not always straightforward to access and query.

The Global Alliance was formed with a simple mission: to drive the development and uptake of genomic medicine. Comprising some 400 institutions working in healthcare, research, disease advocacy, life science, and information technology, their goal is to create a common framework that enables responsible, voluntary, and secure sharing of genomic and clinical data. Their data-sharing projects range from finding out how willing organisations are to share their data, to building international networks of data specifically for diagnostics.

The Matchmaker Exchange is one of the Global Alliance's flagship projects, launched in 2013, which seeks to show what can be achieved when data is liberated from its isolated database silos. The 'diagnostic odyssey' can be a protracted and ultimately frustrating process for rare disease patients, who may show no clear etiology even after exome and genome sequencing. The aim of the Matchmaker Exchange is to create a federated platform – the Exchange – that will make the matching of patients with similar profiles easier.

On the cancer front, Global Alliance have launched the Cancer Gene Trust, which similarly to the Matchmaker Exchange seeks to break existing data out of its silos, readily available to researchers, clinicians, and others.

Another organisation championing the cause of readily accessible data at the Festival is Seven Bridges, a biodata analysis company engaged in some of the boldest, and most exciting, data projects in genomics. The company's software platform powers one of the largest genomic datasets in the world, US National Cancer Institute's Cancer Genome Atlas (TCGA), a serious bioinformatics behemoth containing almost a petabyte of genome sequences.

Involvement in precision medicine projects across the globe, including Genomics England's 100,000 Genomes Project, coupled with some innovative work on data visualisation, have earned Seven Bridges a highly coveted spot on MIT Technology Review's 2016 list of the 50 Smartest Companies.

Dr. Gaurav Kaushik, a Scientific Program Manager for Seven Bridges, has a pretty unique perspective on bioinformatics. A bioengineer and molecular biologist by training, with Seven Bridges he now focusses on how to make data analysis scalable and portable with a view to 'harmonising all data underlying precision medicine' – the title of his presentation to the Enabling Data crowd.

"What we're trying to do with cancer bioinformatics is to enable as much science as possible," he explains to me, over a cup of coffee. "The data model that we use for TCGA was made so that we could extend it to incorporate additional data sets, additional data types, and additional observations."

"You can't predict the future," he chuckles, "but we can work to create systems that are scalable and that are future-proofed as much as possible."

The National Cancer Institute launched TCGA in 2005, while next generation sequencing was still in its infancy, as an effort to catalogue the genetic mutations responsible for cancer. In 2007 Illumina made their fateful acquisition of Solexa, triggering an avalanche of sequencing data. "Next generation sequencing revolutionised the field," Gaurav recalls, "and provided opportunities for scientists and researchers to collect data that hadn't been considered before."

Gaurav describes the current state of play, in particular around the TCGA, as academic research that will pave the way for eventual diagnostics. "We have seen studies look at particular molecular alterations in cancer, and what that means for the patient," he says. "For future datasets, the more patient data that we have to couple with genomic data, the more we can understand how specific alterations are going to affect patients."

"That data should be easy for the community to find, and easy for the community to use, so that they can focus on the science. That would be my ideal vision."

### EVERY TUMOUR IS UNIQUE: A CHALLENGE FOR PERSONALISED MEDICINE

Gaurav's vision is ideal. In fact it is the dream scenario for precision medicine. But as I am about to discover, biology is never that simple.

"Every patient's tumour genome is different and complex. That makes interpretation a hard problem," explains Dexter Pratt. Dexter and his colleague Rudolf Pillich are, respectively, Director and Project Coordinator of the Network Data Exchange [1,2] (or NDEx for short), an open-source framework for sharing, storing, manipulating, and publishing biological network data, based out of Trey Ideker's lab at UC San Diego, California.

As systems biologists, Dexter and Rudolf have a very particular view of cancer, the genomic changes that cause cancer, and how we can leverage big data to treat cancer. On a conference line covering the thousands of miles between a soggy London afternoon, and (what I believe is) a beautiful morning in La Jolla, I'm about to find out just how deep this particular rabbit hole goes.

Dexter began by telling me how every tumour is remarkably unique: "When a patient presents to an oncologist with a malignant, metastatic cancer, those cancer cells are the ones that have survived, proliferated, and spread in the hostile environment of the human body. But in each cancer, there is a unique combination of genomic changes that enable evasion or sabotage of the many safeguards of our cells, tissues, and immune system."

"Depending on the tumour type," he continues, "the cells in a particular tumour may have 30, 60, or even 200 somatic mutations that can significantly alter the proteins the mutated genes produce. Other changes, such as copy number variations, further increase the complexity of the tumour genome."

"Say you find 50 somatic mutations in a tumour genome,"

Rudolf adds. "A small number - perhaps only 1 or even none - will be in known oncogenes and tumour-suppressors, cancer drivers characteristic of that tumour type. But what about the rest? Are they just "passengers," mutations that have accumulated in the development of the cancer but which do not contribute to its success? Or are they crucial for the survival and spread of that cancer?"

This feature of cancer has a name: tumour heterogeneity. Two patients may present with lung cancer, but just about the only thing their tumours have in common is growing in lung tissue. Genetically, these two tumours may be wildly different, and may respond very differently to the standard treatments for lung cancer.

By way of illustration, Rudolf points me towards a recent study of hairy cell leukaemia. Hairy cell is a rare blood cancer, a slow-growing condition that causes bone marrow to create defective B white blood cells. Under a microscope these cells have a distinctly "hairy" appearance, leading to their deceptively charming name.

"About 90% of the patients with hairy cell leukaemia have acquired a mutation (V600E) in the BRAF gene of some of their B cells," Rudolf explains. "This mutation is frequently observed in some other cancers, such as melanomas, and its mechanism as an oncogene has been extensively investigated. But the remaining 10 % of patients do not have the BRAF V600E mutation. This is referred to as the variant form of HCL, or HCLv."

Since the 1980s, treatment for HCL has involved purine analogue drugs and has been highly successful. Patients treated with pentostatin or cladribine typically expect a complete remission rate of between 80 and 95%, and progression-free survival of nine to 11 years[3].

"However, in patients with HCLv, treatment with purine analogues is less effective, and other forms of treatment are necessary," says Rudolf. "Consistently, the percentage of patients with HCLv that achieve complete remission is lower and the median progression-free survival is much shorter. Other genetic mutations have been identified in patients with hairy cell leukaemia although their implications in disease treatment and outcome have not been defined."

**A NETWORK APPROACH TO TUMOUR HETEROGENEITY**

Why can't we explain each cancer as a straightforward combination of known cancer genes? Why is it so difficult? Dexter and Rudolf described to me a widely accepted view of cancer as a disease of required and contributing "hallmark" capabilities[4]. The population of cells in each successful cancer must acquire these capabilities that enable them to survive their host's immune system, to keep on growing, and to keep on spreading through different organ systems.

Dexter elaborates: "Each of these hallmarks can be potentially affected by many genes, interacting with each other through complex networks. A network biology view of this problem is that cancer is a disease of altered pathways and that a successful tumour has subverted a combination of pathways to achieve these hallmarks."

"In a given tumour type, a specific mutation may be particularly effective in undermining a network and therefore it occurs frequently, such as BRAF V600E in HCL and melanoma. But in other networks, there may be many points of weakness and so the genomic alterations are spread over many genes, each one so rare that the patterns across patients are hard to distinguish from the random "passengers"."

Within this highly variable landscape, not only do we want to identify all of the essential alterations in a given patient's cancer, we want to select effective treatments for that patient. And to select a treatment, it must have been created, tested, and approved. "This is significant", says Dexter, "because our system for developing targeted therapies requires hypotheses of mechanism and cohorts of patients to participate in trials." →



"ABOUT 90%

of the patients with hairy cell leukaemia have acquired a mutation in their B cells in the BRAF V600E gene" he explains. "This mutation is also observed in other cancers like melanomas. The remaining 10 % of patients do not have the BRAF V600E mutation. This is referred to as the variant form of HCL, or HCLv."


**This rate of data acquisition brings with it two problems for precision medicine: centralisation and integration.**

In the Ideker lab, recent work on network-based stratification (NBS) begins to address this issue[5]. Dexter continues, "NBS combines prior knowledge of networks of gene and protein interactions with patient mutation data to develop classifiers to stratify patients into groups. But critically, those groups are associated with networks of interacting genes, a potential starting point for planning therapies, either from existing drugs or in guiding next-generation drug development."

"In creating these complex network models, information on any single individual may have limited importance, but when you bring many patients and multiple data sources together – sequencing data, proteomics, metabolomics – then you have the chance to build something actionable in the lab and perhaps the clinic," says Rudolf. Working with biological information as network data facilitates methods to compare not just gene variants between individual tumours, but also how those variants behave as part of an individual tumour's 'solution' to its hallmark needs for survival, growth, and metastasis. The tumours of two patients could present in different organ systems, or with different histology, but might be related by alterations in a specific genetic network related to the hallmarks of invasion or metastatic growth. Examining patient similarity through the lens of network analysis may become an important component of truly personalised therapy.

Dexter and Rudolf's work at the NDEx Project complements the theory and analytic methods created in efforts like NBS. NDEx is a backbone infrastructure, a means of sharing network data in collaborative research efforts such as the Cancer Cell Map Initiative[6] and also as a broader publication mechanism where computable network data is transparently integrated with online journal articles. The goal of NDEx is to enable integration of diverse research by providing a commons for knowledge in network form, including protein binding networks, diagrammed pathways, patient similarity graphs, and systematically computed networks like the output of NBS.

## WHAT'S NEXT?

The ideal of truly precise, individual therapy for a cancer patient presenting to their oncologist is, clearly, still an aspiration for the future. For today, the best case scenario for a patient with a refractory cancer is that they will be assigned to a clinical trial targeting a particular mutation in their cancer. But although the challenges facing personalised cancer therapy are significant, a future where big data and biological network analysis drive drug development, diagnostics and therapy seems like a possibility.

There seems to be a lot of agreement in the sector about where we need to get to. Matching a patient to an exact therapy based on their genetics is certainly idealised, and even reductive, given the complexity of cancer as a disease. But the data that we need to realise a form of this vision is coming in, and is becoming easier to collect all the time. There are some significant challenges in the way, and many of them I have not had the space to cover. How can we protect patient privacy in massive global genomic databases? How do we overcome the physical limitations of data storage?

But what I have found over the past few months is an amazing alignment of vision in the community. We know where we need to get to, and we are starting to build the tools to make a precision medicine future happen. ■

**Reference:**

1. NDEx – The Network Data Exchange www.ndexbio.org
2. Cell Systems, Vol. 1, Issue 4, 302-305. (2015)
3. Best Practice & Research Clinical Haematology 28, 269-272. (2015)
4. Cell, Volume 144, Issue 5, 646–674. (2011)
5. Nat Methods (2013) Nov;10(11):1108-15
6. Mol Cell. (2015) May 21;58(4):690-8