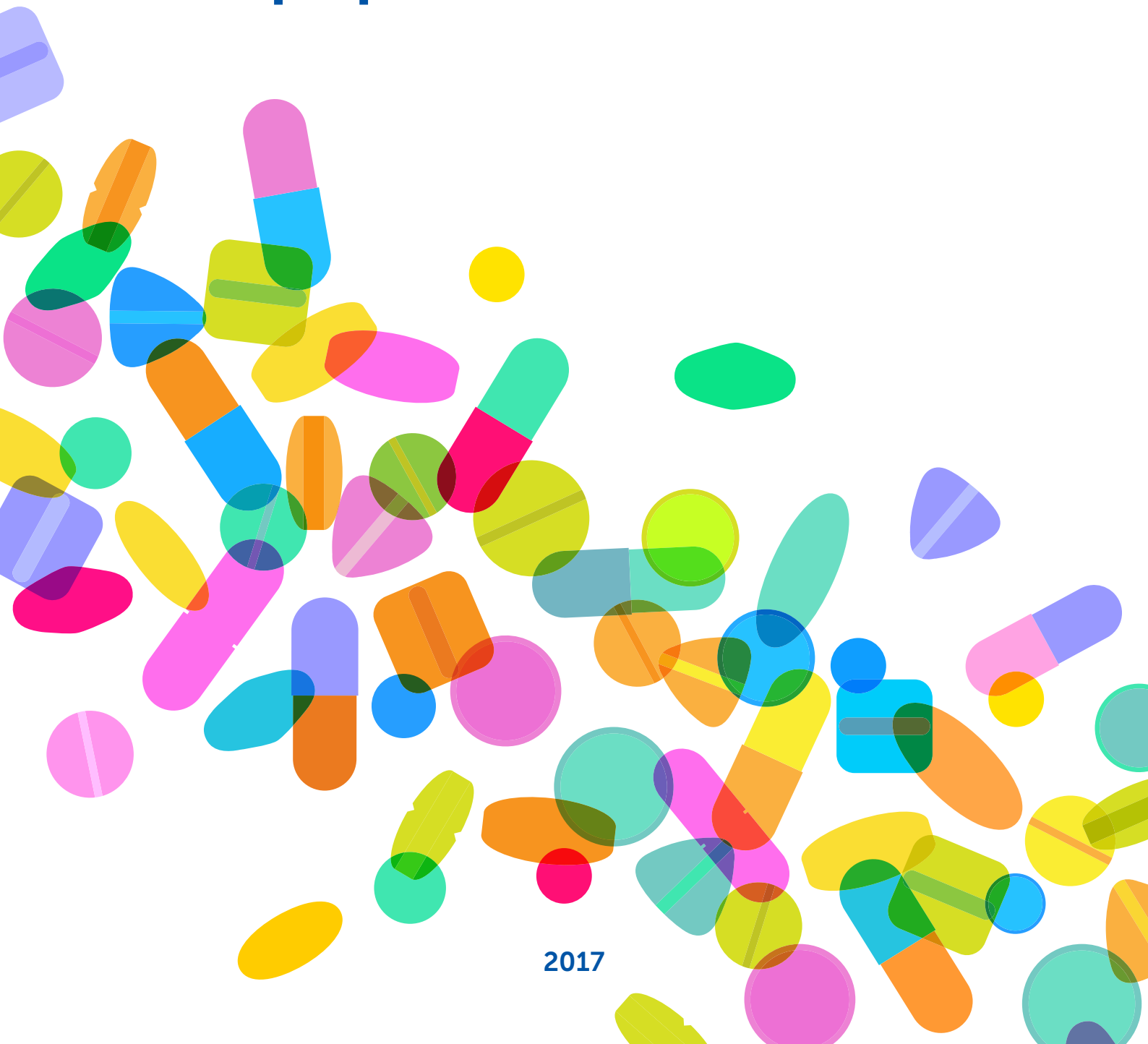
Genestack

How pharmas can extract greater value from large-scale population studies



2017

A key malaria treatment failed for the first time in the UK earlier this year. The drug combination was unable to cure four patients who had all visited Africa. This is an early sign that the parasite is evolving resistance, generating warnings of an urgent need to review the wider situation.

Drug resistance is one of the increasing pressures in the drug discovery race.

Genomics-based approaches offer promise in this area by creating a better understanding of the etiology of the disease, revealing new targets for therapies, and aiding the identification of new drug candidates.

To underpin this new approach to drug discovery and to influence better decision making, advanced data analytics are becoming embedded across the value chain by the pharma industry. However, the industry is still struggling with key issues of data management and integration.

In this white paper **Genestack CEO Dr Misha Kapushesky** dissects the big data challenge facing the pharma industry, identifies the five questions all discovery scientists are asking and proposes a strategy for turning a problem into an opportunity.

Public data: opportunity and challenge

Increasing volume and diversity of data exceeds our capacity to process it

In all areas of R&D, there has been a massive increase in data generated. Five years ago, the leading national genomics studies were looking at thousands of samples. In a very short span of time, this has become tens of thousands (e.g. gnomAD with 16,000 whole genomes reprocessed and reanalysed), hundreds of thousands of samples (100,000 Genomes Project) and on the horizon are projects investigating millions of individual genomes (AstraZeneca, HLI and the Wellcome Trust Sanger Institute collaboration).

Additionally, there has been a massive increase in the diversity of data generated, both from a wider number of sources (clinical, translational, genomic and public health) and in the types of data (genomics, proteomics, metabolomics, as well as epigenetics). The need for an integrated view of all of these disciplines is gaining more and more traction.

We need to build ecosystems to make it possible to work with and manage data of this scale and variety.

Need to put private data into context with public data to exploit opportunities

As publicly funded national population studies begin to produce results, there are increasing opportunities to put private data into a wider context and to mine the multi-omics data for new insights.

To fully benefit from this evolving data ecosystem, organisations must deploy the newest methods of data integration, curation, management, analysis and visualisation. However, the skills required to do this are limited.

We need to think about how to extend the data lifespan to make it live longer, how to automate the analytics so that biologists can find the data they need and analyse it effectively so that the pipelines can be reproduced with different data.

Inconsistency of data makes integration problematic

With the growing public data repositories, the opportunities for translational medicine are becoming very interesting. For example, mining data to look for other therapeutic indications in already licensed products would enable re-purposing of medicines or finding new therapies without the need for toxicology trials.

The challenge here is how to aggregate data sets that were collated for different purposes and are held in different formats with inconsistent parameters.

Common challenges include data living in silos (often stuck on someone's laptop) or public data issues like insufficient metadata or no data provenance. These issues need to be addressed before truly meaningful insights can be gained from the available data assets.

Having data that are consistent, reliable and well-linked is one of the biggest bottlenecks facing pharmaceutical R&D and it is becoming more acute. With the speed at which new omics data is created, the variety of the data generated and the multitude of sources producing data, the integration of data to give a holistic picture seems like a daunting task.

The time and expense required to prepare the data ready for analysis is underestimated and this is the missing link for many pharma companies.

Private sector collaborations: generating new insights

There has always been an air of secrecy around pharmaceutical R&D efforts, with little internal and external collaboration.

This is changing, with several pharmaceutical companies making their clinical data (ranging from basic laboratory research to elaborate clinical trial programmes) available to outside researchers for further analysis.

An example is Project Data Sphere, where the research community can share, integrate and analyse historical patient-level data from academic and industry phase III cancer clinical trials.

Another example is ClinicalStudyDataRequest.com, a consortium that aims to drive scientific innovation and improve medical care by facilitating access to patient-level data from clinical studies. Study sponsors committed to use the site include: Bayer, Boehringer Ingelheim, GSK, Novartis, Roche, Sanofi, Takeda and UCB.

Effective external collaboration can broaden capabilities and insights; a good example of this is Eli Lilly's Open Innovation Drug Discovery Initiative. This enables researchers to submit their compounds for screening using Lilly's proprietary tools and data. With this it is possible to identify whether the compound is a potential drug candidate, and participation in screening does not require the researcher to give up their IP right.

Need to resolve five big questions asked by discovery scientists

The biggest asset for any pharma company is its top scientists, so in this environment of increasing complexity how do you support their innovation and make them more productive?

Discussions with our clients in the pharma and biotech industries have revealed five pain points experienced by scientists:

1 How can I access my data and that of my collaborators effectively?

All scientists working with big data are facing this problem. Most institutions lack a centralised knowledge management system, and this is slowing down the pace of scientific discovery.

2 What data is relevant to my research?

There are two aspects to this question. The first is the need to see all public or shared data that is similar to a scientist's experiments. The second is what public or shared data could help answer a particular scientific question.

3 How do I ask my data questions?

Scientists want to be able to query the data themselves and to know the provenance of the data (the way it was collected). The requirement for scientists to use an intermediary such as a bioinformatician who may not have intimate knowledge of the therapy area is one of the biggest bottlenecks in discovery.

4 How do I stop my data from residing in silos?

How to break down the data silos both internally and between organisations is a central theme of big data discussions. Integrating different types of data to identify trends, patterns and correlations in order to gain better insights is the true goal of data integration.

For this to happen, there is a need for improved metadata (information about the data) to understand what the data is, its relevance and the context in which it was collected.

5 How do I reuse public data if it doesn't have comprehensive metadata?

A common problem in the field is that even though loads of data is publicly available, it often lacks comprehensive metadata required for reanalysis or use as control or validation sets.

Industry response to data inconsistency

The lack of consistency is being addressed by some key stakeholders in the industry. They are coming together to formulate a set of principles that will serve as a guideline for researchers wanting to enhance the reusability of their data.

The FAIR principle aims to overcome data discovery and reuse obstacles by ensuring that data and metadata are (F)indable, (A)ccessible, (I)nteroperable and (R)eusable. The principles should apply not only to 'data' in the conventional sense, but also to the algorithms, tools and work-flows leading to the data. A particular emphasis has been put on making the data disc overy and reuse easier not only for humans, but also for machines.

The public data landscape is changing, with many repositories already showing some degree of "FAIRness", each with its own technology implementation for the different aspects of FAIR, e.g. Dataverse, FAIRDOM and UniProt. There are even emerging projects in which FAIR is a key objective, e.g. bioCADDIE – a project to index biomedical data across data repositories and ag-gregators – and CEDAR – tools enabling creation of metainfo templates that follow community standards with vocabularies and ontologies integration.

However, each displays immaturity in its "FAIRness" trajectory in one aspect or another, particu-larly in genomics research, where a core challenge is the integration of external and internal data and computational pipelines.

Not only is there a need to access data easily but also to process it and to combine private/public datasets using a mix of private/public tools, whilst all the time keeping track of their complete provenance.

The provenance of processed data not only facilitates reproducibility and reusability, but it also lends itself naturally to the aggregation of knowledge, also known as meta-analysis.

Good response, but not fast enough to win the race

Until now data inconsistency and lack of accessibility have been less of an issue for researchers in academia since they operate on a scale that can tolerate some inefficient integration. However, as data volumes and collaboration requirements increase this will become a greater pinch-point.

For industrial R&D departments the bottleneck is painful. The need for bioinformaticians to focus on maintaining existing tools and creating point solutions means they do not have the time and resources to fully integrate all of the resources or to look for more innovative solutions.

Genestack's response: a powerful metadata engine to put scientists in control

At Genestack we are creating an ecosystem that is bringing together data, tools and knowledge. This is offering a platform where discovery scientists can organise, search, collect and aggregate data across different sources.

This includes data from public domain, data that is produced internally within an organisation and data that is shared with collaborators.

The result is that biologists are able to search across various metadata attributes, without the need for specialist bioinformatics skills, and put data together for meta-analysis, while at the same time imposing strict access controls.

Security and confidentiality is of utmost importance, so the platform has tools that allow the instigator to control who has access to what data, maintain the permissions and keep track of consent.

To achieve this, we have invested significant effort into managing data and metadata.

The whole system is accessible via the web or via a scripting interface, and provides a powerful, flexible mechanism for describing data, ranging from describing large projects, experiments and studies all the way down to describing individual data types.

Our platform is built so that each of the different data types have their own descriptor. So, if it is a sequencing assay, you can decide what attributes it has and from what private or public ontologies it has come from. If you're using arrays or vcf data, all of these data types can be described and annotated.

These templates control how data can be described, and templates can be shared with partners and collaborators. This ensures that if a data producer is creating data then this data will be described in the way the discovery scientist wants it.

The platform is also compatible with legacy infrastructure. For example, if a large enterprise is using Laboratory Information Management System (LIMS) then Genestack's tools can integrate with LIMS to populate it or retrieve data from it.

Once these templates are set up data can be imported and then control validation reports produced, ensuring the data has been described as expected.

We have developed lots of tools to help people create data sensibly; this means that when you want to import data into Genestack, the system recognises automatically what kind of data you're looking at and tries to extract metadata from the files as the data is entered. If the data is a dose, unit or time point, this is identified and the technology autocompletes.

This approach has enabled Genestack to rapidly upload public repositories and make this information available in a consistent format ready for meta-analysis with private data.

In conclusion

Multi-omics research has long been the domain of bioinformaticians. The number of genomes sequenced was not overwhelming, and sophisticated computational skills were needed to analyse the data. Now the pace of change has intensified and this is creating pressure points within companies trying to extract value from big data.

With a shortage of bioinformaticians and growing numbers of multi-omics data available, the industry needs to focus on enabling biologists to analyse their data, without requiring them to learn how to write code or use the Linux command line. This will free the bioinformaticians from the more routine aspects of their work, allowing them to look at the bigger picture.

I believe that Genestack has addressed many of these issues created by big data by developing a flexible platform that works with clients' existing investment in hardware, data storage and analysis.

Through working with clients and research institutes an ecosystem has grown that can integrate with public data on the cloud, private data within an in-house deployment, and everything in between.

To support the scientists, the Genestack platform incorporates a powerful user interface and metadata management. This makes it possible for them to quickly gauge how much data exists for different omics types, create virtual cohorts, see what kind of analyses are available and what kind of analysis can be done. These are tasks that would normally require the intervention of a bioinformatician.

We recognise that the situation is evolving so rapidly that a partnership approach is needed to respond to the challenges. We are working closely with our clients – leading innovators in pharma – to support them in extracting the greatest benefit from multi-omics data.

What five questions are your discovery scientists asking?

Dr Misha Kapushesky, CEO of Genestack, will be talking at 12.20 on day one of the Oxford Global Pharmaceutical IT Congress on 27th September 2017 in London, UK.

OXFORD GLOBAL

Talk to us

For more information, or to discuss your projects and challenges, get in touch with us at genestack.com/contact or email our team at info@genestack.com