Application of bioinformatics in support of precision medicine

Bioinformatics has played a major role in gene sequencing diagnostics and has been an essential tool to investigate the genetic causes of disease. With the support of new technologies and tools, bioinformatics can play an important part in the support and continued development of precision medicine.

By Mike Furness and John Wise

The European Union *in vitro* diagnostic regulation (EU IVDR) came into force in May 2017 and will come into effect in May 2022. Much work needs to be done to realign CDx R&D processes from the requirements of the old Directive 98/79/EC to the rigours of the new Regulation (EU) 2017/746 and its enhanced demands for CE Marking* (terms marked with an * are described in **Table 1**). As such, in April 2018 the Pistoia Alliance, a global, not-for-profit alliance of life science companies, technology suppliers, publishers and academic groups that work together to lower barriers to innovation in life science R&D and healthcare, formed a community of interest (CoI) on Companion Diagnostics, Next Generation Sequencing and Regulation (CDx/NGS and Regulation)[1] to consider the many challenges facing the diagnostics industry and to contribute to knowledge sharing within the community. The CoI identified three main areas where knowledge sharing needed to be enhanced viz:

- Applying NGS technologies in precision medicine.
- Application of bioinformatics in support of precision medicine.
- Aligning research standards with clinical standards for precision medicine.

This article is based on the presentations and discussion of a symposium that was held in March 2019 on the theme of 'Application of Bioinformatics in support of Precision Medicine'.

An introduction to the biology underpinning clinical bioinformatics

In the last few years there has been a rapid development in NGS sequencing technology and a substantial increase in the capacity to generate genomic sequence data, along with a significant decrease in costs. The cost of sequencing the first whole human genome*, completed in 2001, was estimated at $2.7 billion[2]. Veritas Genetics[3] is now citing the costs of whole genome sequencing to be $599 and the same company is reported to be predicting the $99 genome will become available in the next three to five years. At that price, whole genome sequencing (WGS) would become an affordable standards component in patient care.

This increase in productivity and cost-effectiveness has led to a growth in genomic sequencing projects to levels where there are now dozens of research and clinical genome projects running across the world[4]. These projects are now including more clinical data with the genomics data and these data sets are being analysed by research bioinformatics teams in pharmaceutical companies

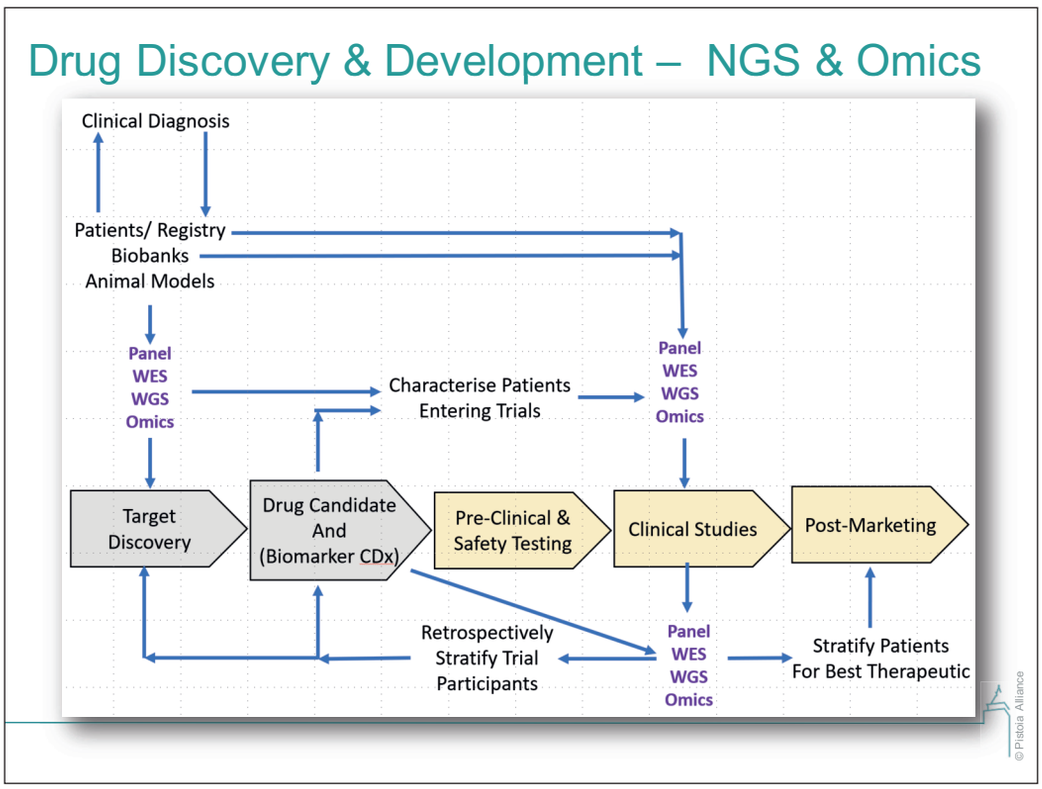

Figure 1
Looking at a very crude model of the drug discovery and development process, we can map on where omic technologies are currently having an impact. Creation of registries and with biobanks provide access to large, more standardised data sets on participants and help identify relevant participants for both research and clinical trials. With the growth of clinical trials collecting omic data, we now also have the possibility of stratifying patients with biomarkers (and ultimately CDx) todefine those most likely to respond to particular therapeutic agents

and also embedded within health services. This research is focused on a better understanding of disease mechanisms, the identification of biomarkers* of particular diseases or conditions and a characterisation of the patients' drug responses (both good and bad). Such work, and the increasing affordability of genome sequencing, provides an opportunity to develop these biomarkers into CDx to enable drugs to be targeted to the specific patient populations most likely to respond positively to treatment with a specific therapeutic agent (**Figure 1**)

Figure 2 is a schematic that shows how retrospective analysis of clinical trial results might stratify the participants into different cohorts allowing insight into the genomic profile of those patients who might well respond to a drug therapy and those patients who are unlikely to respond.

It should be noted that if such biomarkers had been identified in the comparatively unregulated research environment, then if these biomarkers were to be used in a clinically applicable CDx, the research work that had been carried out to identify the biomarkers could well need to be repeated and validated under clinical regulatory conditions, in accordance with an appropriate quality system, in order to be eligible for registration as a CDx.

An objective of the Pistoia Alliance 'CDx/NGS and Regulation' CoI was to bring together protagonists in the research, clinical and regulatory domains relevant to CDx to consider whether standards could be identified and agreed such that if the research analyses were aligned to them, then the data first used to identify the biomarker could form the basis of the regulatory filing of the CDx. Such an approach would have the benefit of minimising the duplication of effort, saving time and cost in development, and as such getting more effective therapeutics to the marketplace faster, benefitting patients, healthcare providers and the companies providing the therapeutics and CDx.

The inaugural workshop in April 2018 brought together representatives from pharma and biotech, technology companies, clinical scientists and regulators to identify the key issues that would need to be addressed. One of the key learnings was that CDx business embraced a broad range of disciplines such as genomics, NGS, bioinformatics, clinical and regulatory affairs and the workshop delegates were in large part familiar with some of those disciplines, but by no means all. As such, a first step was to help address these needs and the Pistoia Alliance CoI organised a symposium on the 'Application of NGS Technologies in Precision Medicine' in September 2018, and this symposium on the 'Application of Bioinformatics in support of

Precision Medicine' in March 2019. Both were targeted at the interested but non-specialist audience.

Use case: applying bioinformatics in drug discovery and development – Cystic Fibrosis

In the EU, Cystic Fibrosis affects one in 2,000-3,000 new-borns and in the USA one in 3,500. In Asia existing evidence indicates that the prevalence of CF is rare[5]. CF is a multisystem disease caused by one of several different mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) gene located in chromosome 7[6]. The CFTR gene provides instructions for making a protein which functions as a channel across the membrane of cells that produce mucus, sweat, saliva, tears and digestive enzymes[7].

Genetic markers can be used to understand the disease stratification in terms of symptoms and severity across populations, as well as to enable drugs to be targeted more effectively.

More than 1,700[8] genetic variants have been identified in the CFTR gene for patients with Cystic Fibrosis. Only five of these mutations have a frequency greater than 1%. The deletion of phenylalanine in position 508 of the CFTR (F508del-CFTR) is the most common mutation in CF patients[9] found in ~90% of CF patients.

While six common classes of the disease have been identified[10], based on molecular deficit, there is a move from just genotyping patients towards 'theratyping' (matching therapies or medications to specific types of mutations) based on lack of protein (correctors) or lack of function of the protein (potentiators). Development of drugs such as Ivacaftor initially treated specific groups of patients (primarily the G551D mutation), who account for 4-5% of cases of cystic fibrosis. Notably, Ivacaftor was the first medication approved for the management of the underlying causes of CF (abnormalities in CFTR protein function) rather than control of the symptoms of CF. But dual therapies including a combination of Ivacaftor with Lumacaftor have increased the number of patients who can benefit from drug therapy. Furthermore, work is now also under way to use a triple combination of tezacaftor, ivacaftor and an experimental drug VX 455 as well, which has the potential to treat twice as many CF patients.

This use case shows the importance of being able to stratify patients within the overall population of those suffering from CF. Understanding the details of the genetic abnormalities provides opportunities for drug therapy to address the physiological cause of the disease and not just the symptomatic relief that has been the standard-of-care until recently. Furthermore, such detailed genetic understanding of the abnormalities in the gene will pave the way for gene therapies to be developed.

Some principal concepts of bioinformatics

Bioinformatics can be used in clinical diagnostics. The bioinformatics tools can be used to detect the presence of genetic variants that act as markers for a condition or a disease. However, these bioinformatics tools when deployed in Europe must follow the stipulations of the EU IVDR and be CE marked to demonstrate that they are fit for purpose.

The EU IVDR includes a specific exemption for diagnostics that are used in the same health institution as they are made or modified but with some specific requirements as set out in Article 5, paragraph 5. Health institutions wishing to apply the exemption in the new Regulations will need to ensure that products meet the relevant General Safety and Performance Requirements. In addition, health institutions will need to have:

- An appropriate quality system in place eg ISO 15189.
- A justification for applying the exemption including that the target patient group's specific needs cannot be met, or cannot be met at the appropriate level of performance by an equivalent device available on the market.
- Appropriate technical documentation in place.

Several key issues are present in all bioinformatics and probably one of the most important is how the tools are deployed. This includes managing dependencies, eg other data associated with the analysis, software versions and version control, and operating system compatibility. Source code for the tools is generally stored in repositories (eg Github, BitBucket) and containers (eg Docker) can be used to wrap up all the source code and its dependencies into a standardised format, ready to run.

In clinical diagnostics, bioinformatics software, including the sequencer's own software, should be validated, ie shown to be robust and repeatable. So, it must be demonstrated that, given identical input (reads from the sequencer), the analysis pipeline will always produce identical output (markers identified). However, this will not be the case when stochastic analysis techniques are deployed, or AI/ML is used. As such, what needs to

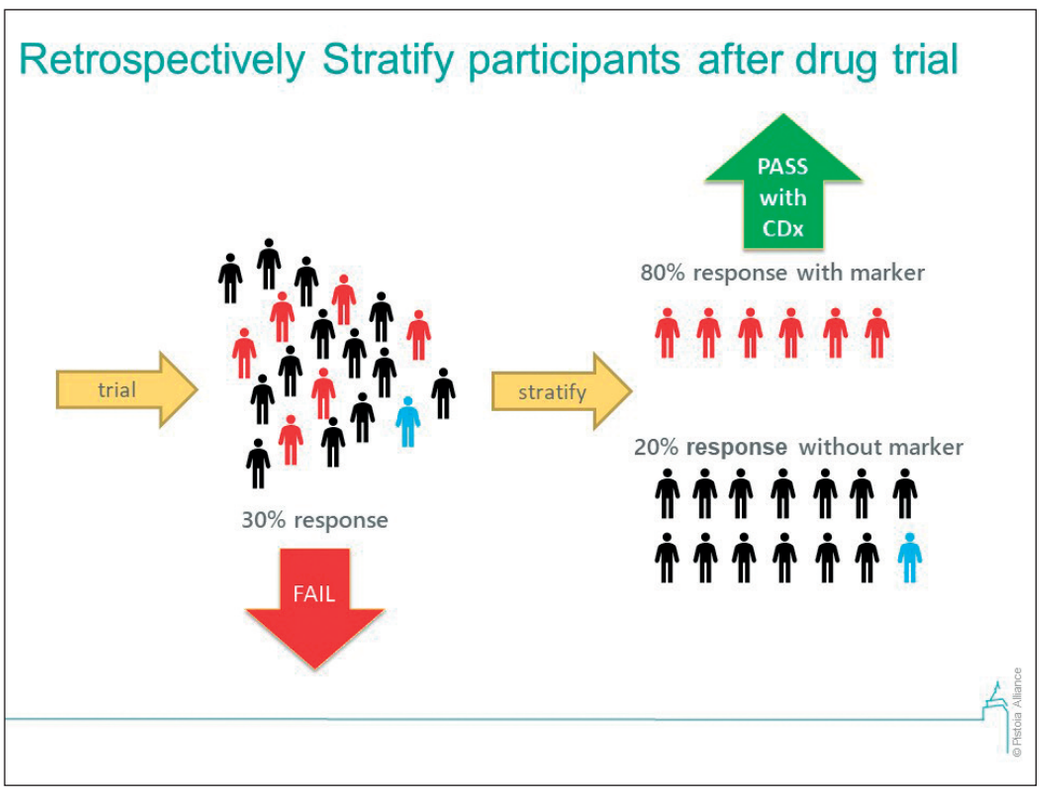

Figure 1

be demonstrated and documented is that the error rate is within acceptable limits.

Analyses and results – sharing and reproducibility – repositories, containers and workflows

Workflows can be deployed to automate analyses to enable them to run faster and more reproducibly and to scale. For example, it had been calculated that running a virtual drug docking simulation on a laptop computer would theoretically take 8.5 years (not useful), but that simulation could be run in the cloud with a workflow using 40,000 CPUs in just four hours. Workflow manager software comes in a variety of packages. A comprehensive list is available in GitHub[11].

The current emphasis on the deployment of the FAIR data principles (Findable, Accessible, Interoperable, Reproducible) in bioinformatics was noted as indeed was the Pistoia Alliance project[12] and its multi-author paper on the 'Implementation of FAIR Data Principles for Pharma and Life Sciences'[13]. The use of workflow managers helps to address the reproducibility of these analyses and sharing the code through repositories such as Github or ContainerHub allows other users to run exactly the code that was used to generate the initial results.

Provenance was a crucial issue to be addressed – who ran it, when, where did they run it, which workflow manager was used, etc. The Common Workflow Language (CWL) makes an important contribution to this challenge and the paper published in 2018 entitled 'Sharing interoperable workflow provenance: A review of best practices and their practical application in CWLProv'[14] focused specifically on provenance in this environment.

Scaling bioinformatics – high performance computing and the cloud

The ability to scale bioinformatics solutions is important. A good example can be demonstrated by Genomics England. In the 100,000 Genomes Project, DNA is sequenced by Illumina. As analyses scale, so the underlying platforms need to change, eg individual applications might require SaaS (Software as a Service) such as GATK* and Dragen*, through genomics platforms (PaaS [Platform as a Service]) such as BaseSpace Sequence Hub*, SevenBridges*, DNAnexus* and up to Infrastructure platforms (IaaS [Infrastructure as a service]) such as AWS*, Google Cloud* and Microsoft Azure*.

Key business considerations for investing in bioinformatics include:

- Scalability.
- On-premise hardware investment (capital expenditure) versus cloud-based implementation (operational expenditure).
- Workforce – bioinformaticians, software engineers, DevOps teams.
- Compute and storage costs.
- Direct instrument integration.
- Security and compliance requirements, especially in a clinical environment.
- Accuracy and reproducibility.
- Turnaround times.
- Out-of-the-box, plug-and-play solutions versus custom-built.

Integrating the data siloes – analysing separate data sets together

Sharing healthcare genomics data across multinational sites creates a range of challenges. Interesting work to address some of these is being carried out by Elixir and the Genomic Alliance for Global Health (GA4GH). It has been estimated that in 2012, roughly 1% of all genome sequencing was funded by healthcare; by 2022 that is expected to increase to 80%. As clinical data requires additional security and compliance requirements over research data, most clinical data sets need to remain in defined geographical locations. However, to benefit from these large volumes of data distributed globally, especially when looking at rare diseases, there needs to be a means to protect patient confidentiality while allowing researchers to search these data sets to identify where specific patient populations can be found. Federating databases is the best practice identified to date, but it requires common data models and tools to allow interoperability across sites.

This becomes even more pressing with initiatives being announced such as the MEGA (Million European Genomes Available) Initiative[15]. Elixir is working closely with other national organisations and global initiatives, such as GA4GH to address some of these issues with projects such as the Beacons initiative, which allows users to identify the locations of patients with specific genetic characteristics across multiple sites globally, and creating a Tools Platform[16] to provide easy access to bioinformatics tools.

NGS bioinformatics: challenges and solutions

For more diagnostics to be developed to identify more diseases, more biomarkers need to be discovered. The Clara^{T} assay from Almac can analyse gene expression data from microarrays or RNAseq data for biomarker discovery. The data generated can also be integrated with clinical outcomes data to perform patient survival analyses. Currently Clara^{T} covers three of the 10 biologies identified to have a role in cancer, with plans to ultimately cover all 10 areas. However, this assay is designated as RUO (Research Use Only) and as such cannot be used for diagnostic or prognostic purposes, including predicting responsiveness to a particular therapy.

Machine Learning (ML) has been deployed on sequencing data derived from FFPE (Formalin Fixed, Paraffin Embedded) tissues to automate removal of artefacts introduced by the formalin-fixing process. This approach has reduced the incidence of artefacts from 42% down to just over 1% in some cases. One of the key advantages in doing this is to allow archival samples to be screened more accurately, allowing the possibility of increasing the numbers of available clinical trial participants.

Emerging technologies – Blockchain and AI/ML

Looking forward, Blockchain and Artificial Intelligence/Machine Learning (AI/ML) could play an increasingly important role. Blockchain utilises distributed ledger technology and provides an irrevocable audit trail for all data handling without the requirement of a trusted party. Such an approach could contribute strongly to making sequence data available to researchers in a controlled manner, the partnership between Shivom and Lifebit providing one example of such capability[17]. Furthermore, the application of blockchain technology could support the regulatory compliance of diagnostic analyses. It was noted that the Pistoia Alliance had a blockchain project 'Blockchain supporting Life Science & Health'[18] to explore the capabilities of this exciting technology in life sciences.

It was anticipated that if AI were optimally exploited it could make a strong contribution to biopharma and healthcare. Some examples were put forward, ie AI could:

- Predict patient drug response.
- Support patient stratification to optimise clinical trials or to personalise treatments.
- Predict disease progression.
- Diagnose disease.
- Discover biomarkers (thereby improving diagnostics).
- Optimise drug design *in silico* to increase efficacy and decrease toxicity.
- Support multi-omics data analysis optimisation.

Table 1

TOOL	DESCRIPTION	URL
AWS	A leading provider of cloud-based Infrastructure-as-a-Service	https://aws.amazon.com/
AWS Batch	AWS Batch enables developers, scientists, and engineers to easily and efficiently run hundreds of thousands of batch computing jobs on AWS	https://aws.amazon.com/batch/
AWS Lambda	AWS Lambda lets you run code without provisioning or managing servers. You pay only for the compute time you consume – there is no charge when your code is not running	https://aws.amazon.com/lambda/
BaseSpace Sequence Hub	Data management and analysis that is simple enough for labs getting started, or powerful enough for rapidly scaling up next-generation sequencing (NGS) operations	https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_basespace.pdf
Biomarker	A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention	https://link.springer.com/referenceworkentry/10.1007%2F978-1-4419-9863-7_211
CloudKnot	A Python Library to Run your Existing Code on AWS Batch	http://conference.scipy.org/proceedings/scipy2018/pdfs/adam_richie-halford.pdf
DNAnexus	A secure, trusted cloud platform and global network for scientific collaboration and accelerated discovery	https://www.dnanexus.com/
Dragen	Dynamic Read Analysis for GENomics. A Bio-IT Platform providing accurate, ultra-rapid secondary analysis of sequencing data	https://emea.illumina.com/products/by-type/informatics-products/dragen-bio-it-platform.html
Galaxy	An open source, web-based platform for data intensive biomedical research	https://usegalaxy.org/
GATK	A genomic analysis toolkit focused on variant discovery	https://software.broadinstitute.org/gatk/
Google Cloud	A leading provider of cloud-based Infrastructure-as-a-Service	https://cloud.google.com/
Lifebit	Automates multi-omics & big data HPC/Cloud deployment. Leverages AI for breakthrough insights generation	https://lifebit.ai/
Microsoft Azure	A leading provider of cloud-based Infrastructure-as-a-Service	https://azure.microsoft.com/en-gb/
Nextflow	Nextflow enables scalable and reproducible scientific workflows using software containers	https://www.nextflow.io/
pywren	Pywren runs existing python code at massive scale via AWS Lambda	http://pywren.io/
SevenBridges	A biomedical data company, specialising in software and data analytics to drive public and private healthcare research.	https://www.sevenbridges.com/
CE Marking	CE marking proves that your product has been assessed and meets EU safety, health and environmental protection requirements	https://europa.eu/youreurope/business/product/ce-mark/index_en.htm
Genome	A genome is an organism's complete set of DNA, including all of its genes	https://ghr.nlm.nih.gov/primer/hgp/genome

While AI/ML was subject to much hype, there are a broad range of areas where this technology might USEFULLY be applied.

What's on the horizon that will impact bioinformatics?

The future will:

- Be driven by companies that either are not well known or do not yet exist.
- Be secure, for cloud-based systems offer advanced security features and alerts (eg UK-OFFICIAL security classification is supported by AWS[19]).
- Require more experiments executed more quickly.
- Demand ease-of-use, eg design thinking and user experience engineering will be increasingly important[20].
- Require more cloud-accessible, software components, datasets, tools and techniques to build sophisticated applications.
- Require cloud-based scalability of applications.
- Be about data sets, tools and techniques, eg AWS is supporting public-hosted data sets and underpins many tools (eg Lifebit*) and techniques (eg pywren*, CloudKnot* and Nextflow*).

Conclusion

CDx are the *sine qua non* of precision medicine and CDx needs to conform to the rigorous quality requirements imposed by the EU IVDR, whether by obtaining CE marking or exercising Health Institute Exemption. The capabilities of gene sequencing and its bioinformatics analysis were increasing rapidly, while the associated time and costs were decreasing. When bioinformatics was involved in diagnostics then the bioinformatics systems needed to be validated in accordance with a recognised quality system to demonstrate that their results were robust and repeatable. Bioinformatics was an essential tool to investigate the genetic causes of disease. Data standards and federated approaches to healthcare genetic data needed to be developed and deployed to allow research access to data that was geographically distributed. The cloud was playing an increasingly important role in bioinformatics analyses by enabling the scalability of systems needing to keep up with increased workload. The cloud also made available a wide variety of tools, including AI/ML-based tools, to increase the capability of bioinformatics analyses. Finally, blockchain technology could contribute strongly to the management, availability and analysis of genomics data allowing the individual to own their data and to make it available for research as and when they choose. DDW

Mike Furness *was Founder of TheFirstNuomics and currently works at Qiagen in the Bioinformatics Customer Services Team. He has spent more than 30 years working in genomics and bioinformatics, developing and applying new technologies to understanding disease and drug R&D. He has previously worked for Life Technologies, Cancer Research UK, Pfizer, Incyte Genomics, DNAnexus, Congenica and Lifebit, as well as consulting widely for pharmaceutical and technology companies and investors and the Pistoia Alliance.*

John Wise *specialises in precompetitive collaboration in the life science R&D information ecosystem. He is a consultant to the Pistoia Alliance, a not-for-profit organisation committed to lowering the barriers to innovation in life science R&D, and also serves as the programme co-ordinator for the PRISME Forum, a not-for-profit biopharma R&D IT/Informatics leadership group focused on the sharing of best practices. John has worked in life science R&D informatics in a variety of organisations, including academia, the pharmaceutical industry and a cancer research charity, as well as in the technology supply side of the industry. John graduated in physiology before obtaining a postgraduate certificate in education.*