BIOINFORMATICS

2020 : A LOOK AHEAD

Part of Proventa International's U.S. Bioinformatics Strategy Meeting 2019 Le Méridien, Cambridge, MA - 12 November 2019

ATTENDEE STATISTICS - WHO WENT AND WHAT THEY'RE INVESTING IN

HIGHLIGHTS FROM ALL OUR TRACKS THIS YEAR

TOP STRATEGIC CHALLENGES FOR BIOINFORMATICS, 2020 AND BEYOND

AN EXPERT LOOK AT THE NEXT FIVE YEARS IN BIOINFORMATICS

NEXT

BIOINFORMATICS

INTRODUCTION

Proventa's U.S. Bioinformatics Meeting has just concluded for another year, showcasing the success of the company's innovative format: delegates and sponsors alike were pleased and surprised by the usefulness of Proventa's unique roundtable discussion format, the amount of connections made with peers and the seniority and experience of attendees present.

Discussions touched on the biggest topics of the moment, focusing on the problems of big data and potential solutions within the industry, the possible goldmines of data mining and NGS methods, and of course the looming presence of AI and machine learning.

THE FUTURE OF BIOINFORMATICS

This report features a wealth of information for those who attended the 2019 strategy meeting and indeed those who did not, but more importantly looks beyond the event to the future: it contains not only statistics showing job titles and investments of this year's delegates, but highlights from the event talks themselves and our facilitators' impressions of how bioinformatics will evolve and change over the next five years.

There is a wealth and variety of information packed into the pages of this report: we hope you find them of interest and use, and enjoy your time reading.

NEXT

BIOINFORMATICS

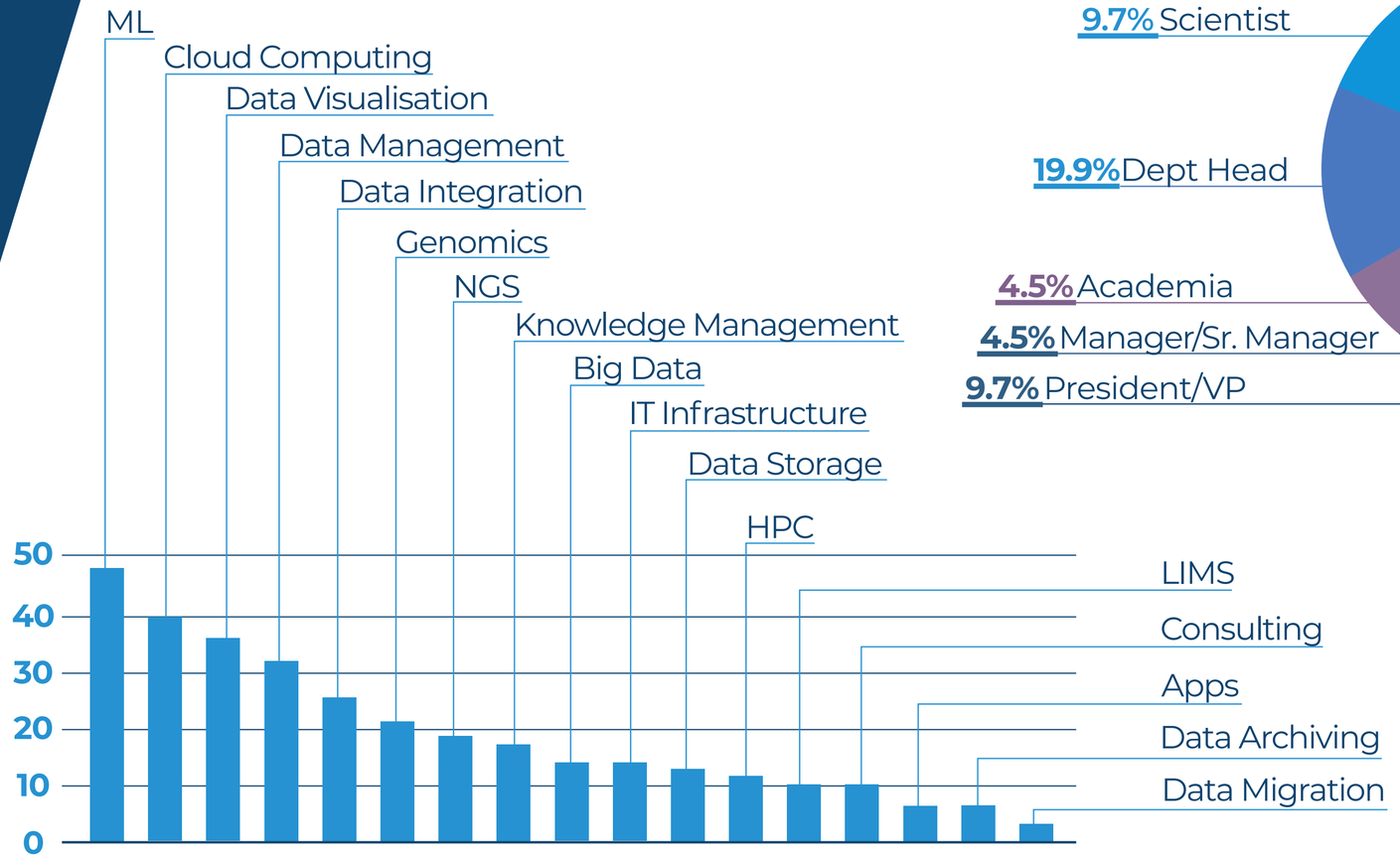
CONTENTS

NEXT

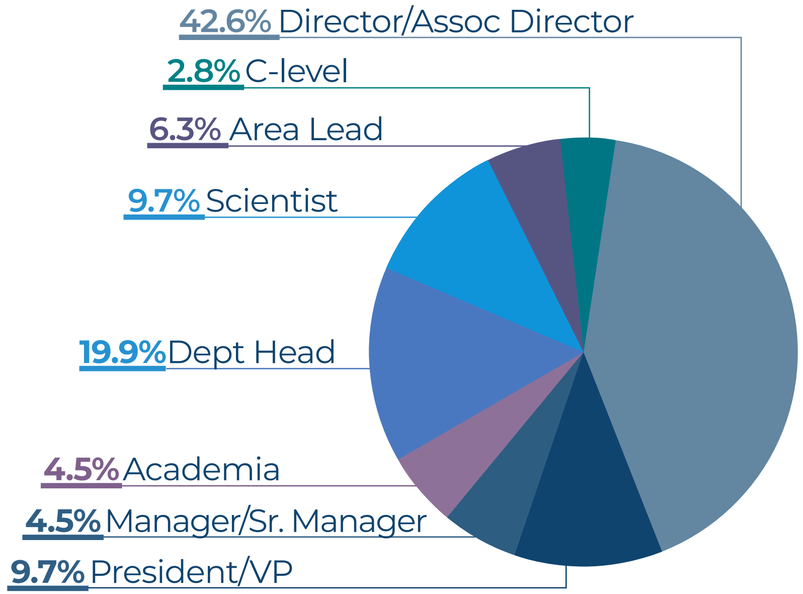
BIOINFORMATICS

1 DELEGATES BREAKDOWN

2019 DELEGATES BIGGEST INVESTMENTS



2019 ATTENDEE BREAKDOWN



BIOINFORMATICS

2 2019 EVENTS HIGHLIGHTS

The 2019 U.S. Bioinformatics Strategy Meeting saw engaging roundtables across six main tracks, with everyone who attended finding something of use. The tracks were: big data/data management; NGS; data integration/mining; data visualisation; AI and ML; and clinical research & translational informatics.

BIG DATA/DATA MANAGEMENT

The big data/data management track began with Hannah Payne's roundtable on how the tokenisation of healthcare data is providing valuable new insights into the patient journey for in-market drugs. Following this, David Fenstermacher of DNANexus led a talk on scalable infrastructure and platforms, and how they're used to integrate -omics and clinical data to power discovery in Pharma Facilitator. Following this, Adina Mangubat introduced a discussion around unique bioinformatics challenges for national-scale genome projects.

The day ended with a talk on metadata standards for documentation in selecting appropriate data storage options, facilitated by Imogene Dunn of vTv Therapeutics.

NEXT-GENERATION SEQUENCING

The NGS track began with a discussion led by Hongyue Dai, of Cygnal Therapeutics, about using NGS as a reference for molecular diagnostics. After a short break Donald Jackson, Director of Translational Bioinformatics at BMS, facilitated a fantastic discussion about the challenges of single cell NGS, from processing to interpretation.

Beginning with a discussion about the many things "single cell" can now refer to - CITE-Seq, RNA-Seq, and NGS among others - the delegates moved on to talk about how such work is accomplished across pipelines or in individual labs. Delegates agreed that the work was hard across pipelines, with better cell signatures easing the process but still containing issues that require subject matter experts in clusters across the pipeline. When discussing the preferred data analysis methods, delegates agreed that Seurat was the general choice to perform checks; some mentioned however that when dealing with clustering it is occasionally necessary to check Seurat against other algorithms.

Other questions posed to the delegates included:

- Who's generally involved in single cell interpretation? - largely biology teams, according to the assembled delegates
- How is analysis usually shared? - generally through internal apps, which allow for selecting and sharing analysis through distributed curation
- What is the difference between single cell transcriptome data and bulk RNA-Seq data? - process time is generally faster in the latter, with the latest version of Callisto up to 100 times faster than the alternative
- How do you ensure cell integrity when working with non-blood human samples? - nuclear sequencing methods offer some hope, though are as yet limited

The track ended with a final discussion hosted by Jeremy Jenkins of Novartis, who aided a discussion around NGS' impact on genomic research and the entire biological field.

BIOINFORMATICS

DATA INTEGRATION/MINING

The data integration and mining track began with a discussion about the rise of data mining in the modern era, looking particularly at capturing, storing, and processing tremendous amounts of data and led by Ed Trautman, VP of Science and Clinical Analytics and Informatics at Pfizer.

The session began with a broad question on what companies are searching for, as well as what data they're holding onto: the importance of quality as opposed to quantity was discussed, with a general agreement that some companies are holding on to too much data following the mining process.

Quality was defined as the expensive part of the capture and storage process, and raised the question: when so much compression is being used, does quality really matter?

Suggestions to improve the situation included creating categories of data to define how long to store them before they become useless, and throwing out any data without a source (which is functionally useless to a company).

Metadata was posited as a necessary element of stored data: delegates advocated for a persistent identifier, storage as a file type that will always be usable, as well as the use of key search terms.

The topic then moved on to nano- and micro- journalistic publications, in which all data/code/models must be published to ensure reproducibility. It was pointed out that without correct stewardship to ensure data is correct, there can be little trust in the industry. To encourage stewardship, crediting the author was suggested as a powerful incentive.

The track ended with two further excellent discussions around data integration and mining: a roundtable on the verification of experiments through the data mining process, led by Sigilon Therapeutics' Hozefa Bandukwala, and one on developments in analysis methods as they are applied in large databases, led by Takeda's Elaine Hoffman.

DATA VISUALISATION

The data visualisation track had two main discussions throughout the day: the first, on diagrams versus colours in statistical data analysis considering user-centred design for good visualisation in understanding biological data, was facilitated by Christian Airiau of Sanofi; while in the latter delegates spoke about high-content imaging on the visualisation of single-cell RNA and DNA, as led by Arthur Liberzon of Alkermes.

AI AND MACHINE LEARNING

The AI/ML track began with a discussion on prediction-making through ML, facilitated by Diane Joseph-McCarthy of EnBiotix. This was followed by a talk on going beyond tSNE, and specifically the ML and statistical challenges when dealing with single cell RNA sequencing, led by Gregory Ryslik of Celsius Therapeutics.

Following this, Transcode Therapeutics' Oliver Steinbach, VP of Research and Development, facilitated a roundtable on AI in medical imaging informatics. The talk began with a discussion of what regulators and tumour boards look for: with classical interpretations done by eye, and the whole process still largely hands-on, it was determined that there is still great diversity in how different experts observe and annotate, something that can be hard for tumour boards to understand.

It was then agreed, when discussing how many image annotations are really needed to generate a solid ML model, that the process is very laborious - often separate groups will conduct imaging before sending results back to the oncologist. Anywhere from tens of annotations to thousands are needed for a solid model, which in areas such as CT can be enormously difficult, where samples are hard to come by.

Another point raised was that if all involved parties are looking at the information on different types of display, it can distort the information available: as such it is important to have identically calibrated pictures and displays. After this, a question of whether in-house or outsourced labour was more desirable was raised, though delegates had no preference.

The discussion ended with the point that human expertise is still very useful for flagging uncertain things for the algorithm to pick up - but the delegates were divided on whether it is arrogance at this stage to say that machines will never be better than doctors in future.

CLINICAL RESEARCH & TRANSLATIONAL INFORMATICS

The first session of the day in the clinical research and translational informatics track was led by Gajanan Bhat, VP of Clinical Science/ Biostatistics at Spectrum Pharmaceuticals. The discussion centred around modelling of clinical and translational research workflow.

The first question that arose asked how companies perform successful due diligence. At this stage F.A.I.R data was brought up, alongside the importance of data being born F.A.I.R, to ensure an easier change of processes. Delegates did suggest, however, that at present F.A.I.R. is more of a goal than a realisation.

One delegate noted that for F.A.I.R data to work effectively, companies must have an initial model and know the initial questions they want to ask. Translating from pre-clinical to clinical stage, companies must turn their databases into knowledge graphs, which can confederate data without moving them from their initial locations.

A further problem arose around access and usage rights. Delegates asked how they could best borrow data from biostatisticians, when so much of it is often private or siloed. Others answered that the first step was to go use due diligence and go through company policies and governance first, before attempting to negotiate for that which could not otherwise be easily accessed.

The session ended with a discussion around AI: now so many processes have been transformed by the innovative technology, one delegate questioned how analysis should be performed. The answer was to assess data needs and create training programs to best utilise the computational power available. The delegate was also advised to look at patient demographics in order to better predict results.

After this, PerkinElmer's Alexandra Vamvakidou-Thomas led a talk on the challenges of applying ML applications in translational and clinical research. He was followed by Kirk Sudheimer, of DNAnexus, who facilitated a discussion on accessing and extracting insight from publically-available UK Biobank data.

The track ended with a discussion around use of real world data to improve efficiency of clinical trials, run expertly by Takeda's Yan Ge.

BIOINFORMATICS

3 KEY DELEGATE CHALLENGES

One of the most important resources available to any senior figure in biotech or pharmaceuticals is an understanding not only of the field at present but where the field is going, and the key obstacles that any company in the sector faces.

Proventa International surveyed a number of major players in the field prior to our 2019 event, using expert opinion and insider knowledge to uncover out of the many obstacles on the horizon the major challenges to overcome in the next few years.

MAJOR CHALLENGES - 2020 AND BEYOND

DATA

Data was by far the biggest challenge facing attending delegates, with almost every attendee voicing concern over some area within data. This ranged from management and integration to reducing the complexity of genomics data to the use of data in the wake of GDPR. It seems that in all forms, data is providing a considerable strain for industry professionals.

AI AND NEW TECHNOLOGY

What could be a possible solution to the data problem seems instead to be an additional thorn in the side of pharmaceutical players: new and innovative technologies, from AI to machine learning, were mentioned as a major challenge in the coming years. Alongside innovative algorithms and programs, delegates announced concerns around technology scaling-up, assessment of new technology and using AI to accelerate genomics strategy.

ANALYTICS

Linked with the first two concerns, a word that recurred frequently in experts' worries was analysis: analysis-ready visualisations, increasing analytic capacity with numerous integrated data types and maturing to predictive analysis were all particular points of concern for delegates.

-OMICS FIELDS

Integrating and working within the numerous -omics fields was mentioned as one major challenge for the coming few years, with delegates noting that they had concerns around integrating the data from different -omics studies and even in some cases simply working in the fields at all.

HIRING AND TRAINING STAFF

Staff training and development were mentioned as of concern in the survey, with many employers unable to find individuals with the right skills - and a particular deficit when looking for individuals well-versed in both the scientific and bioinformatics fields.

STORAGE

Storage capacity and facilities were another issue cited by many of Proventa's delegates in the survey. Particularly, cloud storage integration and determining appropriate storage solutions that align with company requirements were priorities for some delegates in the search for a solution to the overarching problem of big data.

BIOINFORMATICS

INDUSTRY AND PATIENT COMMUNICATION

Another of the major issues facing professionals was what can loosely be termed industry communication: a mix of issues around sharing information between companies and ensuring they remain fully connected to the wider pharma community. Specific challenges cited include mobile engagement, and driving innovative ways to ensure the customer is involved and active.

HIGH-PERFORMANCE COMPUTING

One of the more niche challenges facing several delegates was high-performance computing, a type of computer system containing a number of nodes which can process in parallel with others to solve complex problems faster. This challenge was cited by several big names in the industry, giving the impression that the next few years will see an increase in interest around the technology and perhaps an uptake in its implementation in the pharma space.

KNOWLEDGE

Finally, many experts noted that they were looking to focus on increasing company knowledge, with knowledge management cited several times as an important challenge for the coming years

BIOINFORMATICS

4 A LOOK AHEAD

EXPERT OPINION ON THE FUTURE OF BIOINFORMATICS

The future of bioinformatics is one many could guess at, even with a limited knowledge of the subject: an expanding deluge of big data will flood into pharma companies, its sheer size matched - hopefully - by the increasing speed and storage capacity of modern computers.

Much of this big data comes from the relatively new fields of genetics and genomics, an area intrinsically linked to the future of bioinformatics - individuals who can classify, analyse and generally work with such huge datasets are more important now than ever.

Because of the increasing importance of the digital sphere, a bioinformatics background could well become a necessity for most researchers or experts in the field, and perhaps even the standard knowledge-set for anyone entering the pharmaceutical area. We asked some of the expert facilitators at our recent event what they think the future of bioinformatics will look like.

STRUCTURING AND ACCESSING DATA

Speaking about the future of bioinformatics, one delegate working as a head of data science suggested that progress would almost certainly be made in structuring and accessing complex and big data: he noted that the field is currently still in an expanding phase of tools, solutions and software, and hoped that in the near future this situation would slow and the current variety of tools used would reduce.

Regarding structure, he said that the field is currently still in a phase akin to MS-DOS in computing, before more accessible programs such as Word and Excel improved the situation. Given the ever-evolving nature of coding and applications today, he suggested that soon a more structured workflow would be seen that would clarify how scientists approach data analysis.

Availability of datasets was also on the mind of another facilitator, a director of translational bioinformatics at a leading pharma company. He suggested it was increasingly important to ensure datasets are available not only for the immediate users but as a resource for the company as a whole, or even the entire field. The generation of single-cell data, including the accessing of patient information, is extremely difficult: many datasets end up being used over and over for information. Information on what each dataset contains, which questions it is appropriate for, and how it can be easily accessed are vital to ensure as much data is disseminated as possible.

INCREASED DATA USAGE

Similarly, the head of data science suggested that the future will see a better mix of internal and external data sources, once scientists have found how best to integrate them. He said companies will also likely determine how to pool data between organisations better, with huge rewards waiting for companies that can master this concept.

Another facilitator, involved in biometrics and regulatory affairs, also pointed out quality as an important factor to be addressed, with no current solution available.

CLOUD COMPUTING

Cloud computing solutions were hailed by facilitators as an important part of the next five years, coming more to the fore as the years progress. One facilitator mentioned that during its introduction many experts were skeptical, but that “progress has been staggering.” He said fierce competition between key players over the technology would be likely as a next phase in the evolution of Cloud computing.

BIOINFORMATICS

CELLULAR IMAGING

According to one facilitator the field of cellular imaging is “just exploding”, though many challenges still remain: scientists will soon have the ability to look at potentially hundreds of markers across much wider datasets. Managing the datasets in particular will be an issue, with many thousands of cells in each sample.

AI AND MACHINE LEARNING

Most of the facilitators interviewed speculated that AI and its surrounding fields were vital, but overhyped at present, taking potentially longer than currently projected timelines to fully integrate into current processes. Some said this applied even to machine learning, which is based on well-known mathematics and is more easily applicable than some more innovative technologies.

Another facilitator said that while AI will have a huge impact, many expect that they can throw large quantities of data in one end and expect a magic answer afterwards. He said what was important was to think clearly about what types of data are present, and how best to tackle them: while AI is a valuable tool, it is not an end in and of itself.

A third facilitator agreed, suggesting that relying on presuppositions that data entered into an algorithm is clean and won’t mislead must always be thoroughly checked. Untrustworthy data can give deceitful results, potentially leading to a loss of enthusiasm for the technology, even if the methodology is sound.

DNA PLATFORMS

As mentioned earlier, the field of genomics is closely linked to that of bioinformatics. One facilitator suggested that DNA-based platforms will break new boundaries of science, already being seen with DNA-directed chemistry and using DNA molecules to modify protein levels. The trend, he said, would only accelerate, as scientists begin to read out DNA sequences at a scale and cost unmatched by anything at the moment.

BIOINFORMATICS

5 OUR SPONSORS

LEAD SPONSOR

DNAnexus

CO-HOST SPONSORS

thehyve

Benchling

dotmatics
knowledge solutions

SPIRAL GENETICS

PerkinElmer
For the Better

PREVIOUS