THE CHALLENGES AND OPPORTUNITIES FACING NGS

As part of the NextGen Omics Series, we sat down with an advisory group of leading experts to hear thoughts on NGS (Next Generation Sequencing), how it is developing, its current and future applications, and what challenges face it today and tomorrow.

A revolution has taken place over the past decade and a half; with remarkable rapidity, next generation sequencing has become fundamental to the developments in molecular biology and genetics. Indeed, the name has become something of a misnomer: next generation sequencing is now just the norm for sequencing. Thousands of instruments are installed globally, and NGS has become a gigantic, billion-dollar market. Outputs are up, costs are down, both quite considerably. What is the appeal? NGS is essentially shorthand for several different modern sequencing technologies, all capable of high-throughput sequencing. These innovative technologies permit much faster – and much cheaper! – sequencing of DNA and RNA than the old Sanger sequencing method, which enjoyed a hegemony in the genomics world for forty years. Fred Sanger's sequencing technology was the foundation which the Human Genome Project built on; its completion in 2003 inaugured modern genome analysis with NGS. That year, the International Human Genome Sequencing Consortium sequenced a whole human genome after thirteen years and for €2.3 billion; but by 2008, costs had dropped to under a million euros per genome, and by 2017, it was possible to have your genomes sequenced for about €900 over a few days. Genome sequencing has never been easier: for the first time in history, up to millions of DNA molecules can be sequenced simultaneously, providing new insights at the genomic, transcriptomic and epigenomic levels, and meaning potential wonders for the future of personalised medicine, our comprehension of genetic disease, and clinical diagnostics.

Yet despite the marvel of NGS itself, its use comes with some challenging problems. It is not so much the generation of data itself that is the difficulty, as the interpretation and storage of it. Sequencing genomes is only part of the overall process. As Peter Tonellato, Professor of Biostatistics at the University of Wisconsin, has commented:

> *"Everybody talks about the $1,000 genome, but they don't talk about the $2,000 mapping problem behind the $1,000 genome."*

The reprocessing of sequenced reads and mapping to a reference genome require far more powerful computing facilities than the life sciences industry has ever had at its disposal. NGS produces vast sums of data that laboratories cannot currently keep up with. In addition, issues such as quality checking – to detect contamination – consume yet more time and effort. Such problems will not prove insurmountable, though; the accomplishment of NGS even existing is already staggering, new applications for the technologies are proliferating, and our speakers have witnessed remarkable strides taken, not least by one another.

Jane Wilkinson, Senior Director of the Broad Genomics Alliance Management at the Broad Institute, enthused:

> *"It really has been incredible… I give tours of the Broad Sequencing Facility, and it still takes my breath away to say that we are now looking at the possibility of a $100 genome. I really do think we're on a path there. Having worked on the Human Genome project, which cost billions of dollars and took years to do, it really is impressive."*

It certainly looks that way: earlier this year, the CEO of the American biotechnology colossus Illumina said that the $100 genome is well within reach. Illumina's earliest machines, first active back in 2006, sequenced genomes for a sizable $300,000, but a 2014 model did it for just $1,000. A $100 genome would mean an even greater expansion of research and drug possibilities, not to mention further data.

But therein lies the problem. As Jane Wilkinson says,

> *"I also think that we're rapidly facing the consequences of how much data storage is going to cost – and that could overtake the cost of sequencing itself."*

The sheer, staggering amount of data that NGS technologies can provide is obviously extremely helpful, but retention and handling of it all is unwieldy. Wilkinson's own Broad Institute is generating sequencing data at the rate of one 30X genome every twelve minutes, piling up to nearly four-thousand TB worth of BAM files a year. These are unprecedented levels of information for biologists, which their storage capacity cannot handle. The lack of storage space and the inadequate infrastructure necessarily means that substantial amounts of sequencing data are permanently lost. The time spent trying to solve storage space problems can take longer than the time collecting and analysing the data.

Such data cannot just be abandoned; after all, depending on the experiment, especially if it is a long and incremental process, the reads may have to be retrieved and processed again, with new data refining these earlier results. As the cost of sequencing has come down so much, some have even concluded that simply resequencing samples might be the more efficient route. For Augusto Rendon, Director of Bioinformatics at Genomics

England, it's this handling of data that is the real pressing issue:

> *"… the problem stops being – how do we sequence a genome? We know how to do that; we know how to deal with the bioinformatics. The question becomes, how do we deal with the clinical data? How do we get the medical records so we can really do research?"*

Data storage is not the only problem biologists face in sequencing genomes. For Graziano Pesole, Professor of Molecular Biology at the University of Bari, one of the big problems is *"… in our reliance on existing databases, which are contaminated, i.e. contain taxonomically mislabeled entries."* Data contamination, whether from their biological source or the experimental environment, is the other big practical challenge. Notably, the amount of reads that correctly map to the reference genome range between 70% and 90%, leaving a consistent fraction of unmapped sequences; some of these unmapped reads can be explained by downstream or upstream contamination from exogenous nucleic acids. Bacteria, fungi and viruses can also easily contaminate sequences.

> *"Virtually every genome project is contaminated when it's done by microbial sequence,"*

says Pesole. It is a major difficulty: it has been estimated that between one thousand and a hundred thousand contaminating microbial reads are detected per million host reads when sequenced by RNA-sequencing. These contaminations can result in phenotypic changes, diminishing the quality of the sequencing data. NGS is an improvement compared to earlier sequencing when it comes to carefully identifying these problems, but contamination and cross-contamination remain frequent. Of course, as Jane Wilkinson notes, *"…contamination has also been highlighted as potentially informational as well."* NGS has been used on microbial species: the identification of these pathogens in patient samples has naturally helped with the research itself.

But for Patrick Descombes, Head of Genomics at Nestlé Research, the talk of the technology, its advances and its practical challenges is less interesting and less important than considering the applications.

> *"You were saying we are close to the $100 genome. But somehow, it's a little – so what? In the sense that, we know it's going to come. The question is, what are we going to do with it?"*

It's an exciting question. NGS and bioinformatics approaches have already been used to identify the origins of disease outbreaks, investigate epidemic dynamics, and discover human pathogens, amongst much else. NGS could even become routine for the investigation and control of pathogens. One of the most prominent areas is in combatting cancer: NGS analysis of tumor genomics, transcriptomics and epigenomics is driving biomarker discovery for cancer diagnostics. Biomarkers delineated via NGS are, for instance, helping diagnose whether someone has cancer, or how they might respond to therapy – and can measure the treatment's effects. In addition, as Patrick

Descombes notes, there are other areas beyond human health where NGS could prove useful: *"…plant genetics, sustainability, there are so many things."* Regarding plant genetics, some are working on developing software via NGS that can take millions of short reads or RNA from a mixed sample and produce a list of viruses present – for the identification of virus infected crop plants in the UK. NGS is positively revolutionizing approaches to food safety, by isolating DNA either from food materials or the production environment to identify microbial populations.

However, Achim Kless, Scientific Director at Grünenthal GmbH, believes that there is more work to be done before considering further applications.

> *"I think in the very beginning, when the Human Genome Project was done, everybody thought, now we have it. In the end, we only understand roughly 5-10% of the whole story."*

One particular area that Kless wants to explore more?

> *"The three-dimensional structure behind it. There's a complete lack of work there, at least for me … there are some ideas about how this could be arranged, and you heard some talks about it today, people are speaking together. But there is no familiar kind of strategy behind how we can go and understand a sequence genome in three dimensions. That's the map that is completely neglected."*

Three-dimensional analysis, Kless contends, *"brings the last 20-30%"* for quality scores. Indeed, three-dimensional configuration of the genome, for all its complexity, is crucial for gene regulation. In eukaryotes such as humans, the genome does not exist as a linear molecule, but is instead *"hierarchically packaged"*; three-dimensional organisation shapes it. Three-dimensional analysis could enrich our understanding yet further.

There are other major questions that our speakers did not have the time to touch upon, including further obstacles: the samples can be damaged in ways other than contamination, for instance through the fixation process, or the storage conditions; NGS may have lowered costs generally, but they remain high for certain applications like sequencing bacterial genomes; and while short reads, which constitute the vast majority of data generated so far, are a good fit for many applications, they are insufficient elsewhere. Despite the advances, the challenges remain; yet so do the many possibilities. It is an exciting time for genomics. NGS has completely transformed the field within a decade. Where it goes next will surely be just as transformational.

A THANK YOU TO OUR EXPERT SPEAKERS, CONVERSATIONS WITH WHOM INSPIRED THIS PIECE AND HELP US CONTINUE TO PROVIDE YOU WITH HIGH-QUALITY, INDUSTRY-LEADING EVENTS.

Stefano Ceri, Professor of Database Systems, Politecnico di Milano

Stefano Ceri is professor of Computer Science at Politecnico di Milano; his research covers four decades and is broadly addressed to extending data models and technology; his recent interest is in data-driven genomic data management. He is author of more than 300 research articles and of 11 international books (H index 72, 24895 citations as of March 2017). He is the recipient of the ACM-Sigmod Edward T. Codd Innovation Award (June 2013) and of two Advanced ERC Grants, on "Search Computing" (SeCo, 2008-2013) and on Data-Driven Genomic Computing (GeCo, 2016-2021, (SEE HERE); he was the coordinator of the PRIN Project "GenData 2020" (20013-2016).

Graziano Pesole, Professor of Molecular Biology and Director of the Institute of Biomembranes and Bioenergetics of the National Research Council, University of Bari

Graziano Pesole is full professor of Molecular Biology in the University of Bari "A. Moro" and past Director of the Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies (IBIOM) of the National Research Council (Bari). Bibliometric facts: h-index: 68 (Google scholar), 61 (ResearchGate), 56 (Scopus) peer-reviewed publications: >280 Sum of Times Cited without self-citations: >25,000 Graziano Pesole has since long carried out research activity in the fields of bioinformatics, comparative genomics and molecular evolution. His current research interests are focused on bioinformatics application for the management and analysis of next generation sequencing data, obtained also at single-cell resolution. He has been one of the founders and past-president of the Italian Society of Bioinformatics and is currently the head of the Italian node of ELIXIR European infrastructure for Bioinformatics.

Massimo Delledonne, Professor, Department of Biotechnology, University of Verona

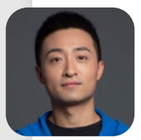
Massimo manages a vigorous research program that emphasizes interdisciplinary approaches. His main focus is on the characterization of those parts of the genome that are not shared among all genotypes of a species and that can contribute to the establishment of differences in phenotype. Comparison of the different technologies available and testing of new emerging technologies for DNA sequencing and analysis is another integral part of his mission.

Achim Kless, Scientific Director, Grünenthal GmbH

Dr.rer.nat. Achim Kless, Scientific Director at Grünenthal, holds a Ph.D. in chemistry and has more than 20 years' experience in the pharmaceutical industry. His research focus is on computational methods for drug design and unbiased omics approaches. Currently, he is working in the department of translational science and intelligence leading the genetics team for clinical trials as well as the disease understanding team for CRPS.

Charlotte Guldborg Nyvold, Professor, Odense University Hospital, Denmark

Charlotte Guldborg Nyvold is Professor at Odense University Hospital in Denmark. She holds a Master of Science in Chemistry and Biotechnology from Aarhus University and has completed her Ph.D. studies at the Department of Clinical Immunology at the National University Hospital in Copenhagen. For more than 25 years, Charlotte Guldborg Nyvold has worked with molecular diagnosis in haematological cancers. Her research group in Odense is focused on the heterogeneity of B cell neoplasms and characterization of new molecular markers directed towards personalized diagnosis and prognosis.



Grover Yu, EVP, BGI

Dr. Yu is the Executive Vice President of the International Department at BGI Genomics, a global leading genomics organization. In that role, he is in charge of the company's global development, operations, and M&As. With more than 10 years of experience in the genomics and diagnostics industry, he also leads BGI in promoting precision medicine applications internationally. Prior to taking over the Executive Vice President role, Dr. Yu served as the COO of BGI Europe and Vice President of BGI Europe.

Jane Wilkinson, Senior Director, Broad Genomics Alliance Management, Broad Institute of MIT and Harvard

Jane is a Senior Director at the Broad Institute where she leads the Broad Genomics Alliance Management team. Jane has over 20 years of high-throughput genomics experience from the Wellcome Trust Sanger Genome Center, UK where she was a key leader on the Human Genome Project and at Monsanto Company, USA where she led a new directive in plant genomics. Jane has been at the Broad Institute for fourteen years and has worked on various initiatives including Comparitive organisms, Cancer, Mendelian, Infectious and Common Diseases.

Augusto Rendon, Director of Bioinformatics, Genomics England

Augusto undertook his PhD at the University of Toronto in the Department of Medical Biophysics. He then moved to the University of Cambridge to pursue postdoctoral work in computational biology and statistical genomics. Augusto currently holds a Principal Research Associate position at the University of Cambridge in the Department of Haematology. Since 2014 he had been seconded to Genomics England as Director of Bioinformatics. There, he leads a team of over 50 developers, analysts, bioinformaticians and curators. Augusto's team is responsible for establishing pipelines to analyse and manage all genomic data for the UK 100,000 genomes project and the future NHS Genomic Medicine Service. The team designed and implemented the clinical interpretation pipelines that return findings to patients, while ensuring that knowledge accumulated through this process is best exploited to improve patient care and enhance discovery.

Matt Hall, Senior Scientific Officer, European Bioinformatics Institute (EMBL-EBI)

Matt Hall works at the European Bioinformatics Institute (EMBL-EBI) in the Strategic Partnership Office. His primary responsibility is for managing the EMBL-EBI Industry Programme. Matt has a background in bioinformatics, molecular biology, cell biology, genomics, gene expression technology and protein biochemistry and previously worked in the pharmaceutical industry. Specialties: computational biology, bioinformatics, genomics, molecular and cellular biology.

Patrick Descombes, Head of Genomics, Senior R&D Expert, Nestlé Research

Patrick Descombes is the head of Functional Genomics at the Nestle Institute of Health Sciences (NIHS), located in Lausanne Switzerland. His responsibilities in this role include supporting, managing and executing projects involving genomics at NIHS and across the Nestle R&D. The cutting-edge genomics technologies routinely used include Next Generation Sequencing, Third Generation Sequencing and microarrays for discovery, and real-time PCR and nanoString nCounter for validation. The major applications cover transcriptomics, genotyping, DNA (re)sequencing, miRNA profiling. Prior to joining NIHS, Dr. Descombes did set up and managed for ten years the Genomics Platform of the University of Geneva. He holds a Ph.D. in Molecular Biology from the University of Geneva.

John Weightman, Senior Scientific Officer, Cancer Research UK, Manchester Institute

John is a Senior Scientific Officer at the Cancer Research UK Manchester Institute. He moved from Newcastle upon Tyne to complete a degree in Cell and Molecular Biology at Manchester Metropolitan University before joining the institute in 2012. John worked on the Cancer Research UK Microarray Service and is now a senior scientist for NGS in the CRUK MI Molecular Biology Core Facility.

Yasushi Okazaki, Professor and Director of Diagnostics and Therapeutics of Intractable Disease Center/Team Leader, Laboratory for Comprehensive Genome Analysis, Juntendo University, School of Medicine/RIKEN Center for Integrative Medical Sciences

Prof. Yasushi Okazaki, M.D., Ph.D. is a Professor and Director of Diagnostics and Therapeutics of Intractable Disease / Intractable Disease Research Center in Juntendo University, School of Medicine in Japan. He also serves as a Team Leader at Laboratory for Comprehensive Genome Analysis, RIKEN Center for Integrative Medical Science in Japan. After obtaining his M.D. degree from Okayama University in 1986 and Ph.D., degree from Osaka University in 1995, he led the worldwide collaborative project, FANTOM, as a Team Leader at RIKEN and published several Nature papers (Okazaki Y et al, Nature. 420:520-62, 2002). Prof. Okazaki is now leading the part of genome analysis at several national projects, including mitochondrial disease cohort and hereditary colon cancer syndrome cohort.

Ed Schuuring, Full Professor in Molecular Oncological Pathology and Senior Clinical Scientist in Molecular Pathology, University Medical Center Groningen

Prof Dr Ed Schuuring, PhD, is a senior clinical scientist in molecular pathology, University Medical Center Groningen, Groningen, The Netherlands. His research focuses on the identification of prognostic/predictive epigenetic and molecular markers for clinical outcome, response to chemo-radiotherapy, gene-targeted therapy and treatment-resistance in lung, GIST, head and neck cancer, as well as the early detection of cervical cancer in scrapings. He is heading the laboratory of Molecular Pathology and acts as a principal investigator for several postdoctoral, clinical and PhD-student projects.