FROM SANGER SEQUENCING TO GENOME DATABASES AND BEYOND

We look at how next-generation sequencing has advanced research across different disease fields, and the growing importance of open access genomic databases

N ext-generation sequencing (NGS) technologies have revolutionized genomic research. Since the completion of the Human Genome Project (HGP), the sequencing of whole genomes has become cheaper, faster and more accurate. Now, whole human genomes can be sequenced in as little as a day [1].

This is something that would never have been possible with Sanger sequencing methods – the sequencing of one entire human genome in the HGP took 13 years. Leaders in the field of NGS – including Illumina, Qiagen and ThermoFisher Scientific – continue to drive forward novel technologies and initiatives. One of the current focuses in this field is the development of large genomic databases for clinical research.

The expectations for personalized medicine are high. In oncology, personalized medicine has huge potential with both diagnoses and treatment options being driven by different factors associated with an individual, including genetic information.

The advent of genome databases is also driving research into rare and complex diseases. One such example are psychiatric disorders, which, although not necessarily rare, are often polygenetic. Therefore, comprehensive genetic information is needed to hypothesize mechanistic bases of the diseases.

The potential of NGS, and genome databases, seems to be continually expanding and developing. In this tech news piece, we look at some of the current ways in which NGS is being utilized to advance research across different disease fields, assessing the outlook for personalized medicine and determining how research can be further advanced.

> " *...personalized medicine has huge potential with both diagnoses and treatment options being driven by different factors associated with an individual.* "

CURING THE RAREST DISEASES

In 2012, then-Prime Minister of the UK, David Cameron, gave the go-ahead to the 100,000 Genomes Project. The ground-breaking program led by Genomics England in partnership with NHS England aimed to utilize whole-genome sequencing technologies to find new diagnoses and improved treatments for patients with rare inherited diseases and cancer [2].

The project was to involve the sequencing of 100,000 genomes of individuals with rare disorders and cancers. As some of the disorders did not have names nor was the cause known, the project was to be pioneering in the sense that it could begin to piece together known information into a bigger picture.

As of 5 December 2018, the team sequenced their 100,000th genome and completed the project [3], which has been referred to by news outlets as 'key to beating our rarest diseases'. As a result of this, 13 NHS Genomic Medicine Centres have been created, as well as a state-of-the-art sequencing center, run by Illumina, and an automated analytics platform for whole-genome analyses in the NHS.

"At launch the 100,000 Genomes Project was a bold ambition to corral the UK's renowned skills in genomic science and combine them with the strengths of a truly national health service in order to propel the UK into a global leadership position in population genomics. With this announcement, that ambition has been achieved. The results of this will be felt for many generations to come as the benefits of genomic medicine in the UK unfold," explained Sir John Chisholm, chair of Genomics England (UK).

The UK has become the first nation in the world to apply whole-genome sequencing at scale in

direct healthcare, solidifying its place as world leaders in the field of genomic medicine. The project has already benefited many of its participants, enabling diagnoses to be made and disease names to be assigned.

The outcomes are testament to how these kinds of programs and investment are crucial to pushing forward technology, and therefore research – something which may have begun with the HGP, but whose potential is evidently limitless.

COMPLICATED GENETIC LINKS

When it comes to determining the biological basis of many diseases, genome sequencing has been invaluable and has allowed for the identification of many genetic variants and genes that cause illness.

However, for many psychiatric disorders, the genetic link is much more complicated and, despite strong genetic associations, there is still little understanding of their mechanistic basis. Psychiatric disorders are often polygenetic and have a greater number of potential genetic variants than other disorders; however, it is likely that the presence of an individual gene has only a weak impact when it comes to increasing disease development risk.

The complexity of these disorders means that the traditional, one-dimensional genome sequencing approaches may not be as useful in developing a clear understanding of the disease.

In a recent multicenter study, the genomic data from over 2000 human brains were compiled to form the most complete picture of how regions that regulate DNA expression can influence the brain and its function [4]. With special focus on schizophrenia, autism and bipolar disorder, the PsychENCODE study combined data from DNA and RNA sequencing methods to identify the genome and the transcriptome with information regarding the DNA structure, transcription factors [5] and enhancer regions.

"It's the most comprehensive functional genomic resource ever developed for understanding the brain, and it establishes a framework for integrating different kinds of genomics data to get deep insights into the biology of brain disorders," commented co-first author Hyejung Won (University of North Carolina, USA).

The research, published as a series of 11 papers in *Science*, documents the range of epigenetic [6] and transcriptomic [7] changes that occur during human development and how these may be related to psychiatric disorders. The compiled data can be used to provide information regarding the genes, cell types, loci and co-expression molecules that may play a role as neuropsychiatric risk factors.

"We think it will have a big impact in terms of risk assessment and diagnosis for patients," continued Won.

In one study regarding schizophrenia, the

> **" They also concluded that schizophrenia was primarily a neuronal disorder rather than glial. "**

team used their newly developed framework to analyze the 142 'risk loci' that have been identified in previous genomic studies and determine how they have an effect and whether they are penetrant. As few of the loci contain coding genes, it was hypothesized that they act as regulatory factors, controlling the expression of other genes. Several genes (321) were found to be potential targets for the regulators and have roles in various neurological functions such as synaptic activity and ion channel regulation. They also concluded that schizophrenia was primarily a neuronal disorder rather than glial [8].

The backbone for the new database was information gained from NGS. Although unable to give a complete understanding of psychiatric disorders, whole-genome sequencing formed a foundation to be built upon, allowing for the development of new hypotheses and more detailed analysis techniques that have helped to drive the progress of neuropsychiatric research.

The team are now working to improve their model, incorporating more genomic data and expanding to other neurological and psychiatric disorders.

SEARCHING FOR A LACK OF VARIATION

A new map detailing genetic variation or its absence, is helping to identify mutations responsible for developmental disorders. It is believed that the map will provide a much-needed resource to study genes that previously had no disease association.

A team of researchers led by Aaron Quinlan from University of Utah Health (UT, USA) has developed a detailed map of the human genome, documenting regions of DNA that experience very little heterogeneity between different genomes [9]. These regions of low genetic variation are referred to as being highly constrained and the team believes that they could indicate genes involved in the pathogenesis of developmental disorders.

Using data from 123,126 human genomes taken from the Genome Aggregation Database, Quinlan and his team observed variation, specifically searching for ▶
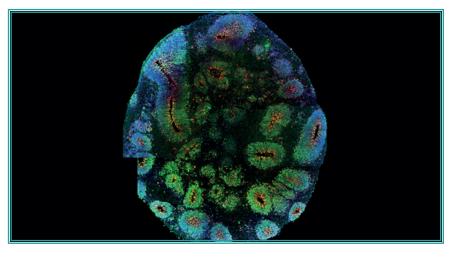

Figure 1. A brain organoid, developed as part of the PsychENCODE project, was utilized to compare the gene activity of cells to those taken from actual brains, discovering they mimic each other in the early stages of development. Credit: Vaccarino Lab, Yale University (CT, USA).

Figure 2. Francis S Collins, NIH director, presents at the The Genomics Landscape, a decade after he led the HGP to completion. Reproduced with permission from [14].

SHARING IS CARING

Open access, public databases are becoming more recognized throughout the scientific community as key tools for the advancement of both personalized medicine and novel diagnostic technologies.

For the first time, the US FDA has formally recognized a public database containing genomic information. More specifically, the agency is recognizing the genetic variant information in the Clinical Genome Resource (ClinGen) consortium's ClinGen Expert Curated Human Genetic Data, a database funded by the NIH [11].

"A major current challenge for precision medicine is the need to translate new discoveries and data from the HGP so that this information can be used by physicians and other health care providers to improve health," explained NIH Director Francis S Collins.

"ClinGen provides a standard curated data reference of genetic variants to facilitate the development and implementation of genetic tests for use by health care professionals, which is critically important for moving science into practice."

In April 2018, the FDA issued final guidance to encourage data sharing and outline how test developers can rely on evidence in FDA-recognized public databases. Going forward, the new recognition of the ClinGen database means that developers will not have to demonstrate the reliability of the database or the information within the database before proceeding with their research.

Through the utilization of publicly available databases like ClinGen, technological and clinical advances in genetic tests will help researchers develop more targeted therapeutics aimed at previously unknown subsets of disease.

Similarly, in the UK, the recent publication of the Personal Genome Project UK (PGP-UK), a group of research studies creating freely available scientific resources that bring together genomic, environmental and human trait data donated by volunteers, also demonstrates high promise for such databases [12].

Director of the PGP-UK, Stephan Beck (UCL, UK) explained: *"The PGP was founded in 2005 by George Church at Harvard University to aid the interpretation and sharing of human genomes. To facilitate more open data sharing, PGP introduced the concept of open consent and was the first project to provide human genome and trait data under open access. Since then the project has grown into a global network. All of the PGPs in the network operate independently of each other but jointly advocate for and practice open data access and collaborate on creating publicly shared genome, health and trait data. We believe that this information sharing is critical to scientific progress."*

It is well known that open access data play an important role in the public awareness and engagement of genomic research,

▶ regions that lacked it. Their findings were then used to create a map detailing the locations of these 'constrained coding regions' (CCRs).

Analyzing the CCRs, the research team found that the most highly constrained were enriched for mutations that result in developmental disorders such as developmental delay, seizure disorders and congenital heart defects. This information supports the researchers' hypothesis that these CCRs, so resistant to variation, are indicative of regions that, when they are mutated, can be the root cause of severe developmental diseases and other pathogenesis.

"A gene as a whole might be able to tolerate variation, but variation in one critical section could have serious developmental consequences," explained first author James Havrilla (University of Utah Health) [10].

Quinlan states that, as well as established pathogenic genes, genes that have yet to be associated with disease often harbor one or more CCRs and he believes that a mutation in these regions could result in disease. *"We are confident that these genes play a role in development of disease, but we currently know little about their role. That's where the exciting potential for discovery is."*

While Quinlan warns that this map is best suited for identifying dramatic phenotypes such as facial dimorphisms, developmental disease or congenital heart defects, the map could still prove a particularly handy tool for researchers in finding new genes and regions of DNA responsible for the pathogenesis of diseases.

"The map we created will provide the community with a resource to study genes that heretofore had no disease association," Quinlan said. *"The beauty and power of this approach is that, as we obtain more data from ever more human genomes, we can continue to improve the resolution of this map to pinpoint areas to study for disease."*

> ❝*For the first time, the US FDA has formally recognized a public database containing genomic information.*❞

and PGP-UK is no different: *"Like all PGPs, PGP-UK is classified as a research project, so our findings are not suitable or accredited for clinical use. As already demonstrated, PGP data will, however, contribute to advancing both technology and our understanding of genome function in health and disease,"* concluded Beck.

FUTURE PERSPECTIVE

While we are heading in the right direction in terms of the availability of open access databases, there is still a lack of representation of certain populations. A recent publication reported significantly fewer studies of African, Latin American and Asian ancestral populations in comparison to European populations. Moreover, these patterns were consistent across both data types and disease areas. It is important that the number of genomic studies that include non-European populations continues to improve to ensure that the promise of personalized medicine is applicable to all [13].

Ultimately, it is the advancements in the abovementioned NGS technologies that enable the construction of large open access databases. The availability of these databases will allow patients to gain access to more sophisticated tests that provide important genetic information, allowing for more targeted medical care.

With the continued development of technologies comes the continued expansion of genomic databases. Research will be accelerated even further as more recognition by agencies such as the FDA is provided.

REFERENCES

1. Behjati S, Tarpey PS. What is next generation sequencing? *Arch. Dis. Child. Educ. Pract. Ed.* 98, 236−238 (2013).
2. www.theguardian.com/science/2018/feb/03/1000-genomes-project-genetics-dna-database-key-beating-rarest-disorders
3. www.genomicsengland.co.uk/the-uk-has-sequenced-100000-whole-genomes-in-the-nhs/
4. Li M, Santpere G, Kawasawa YI *et al.* Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science* 362(6429), eaat7615 (2018).
5. Chen C, Meng Q, Xia Y *et al.* The transcription factor POU3F2 regulates a gene coexpression network in brain tissue from patients with psychiatric disorders. *Sci. Transl. Med.* doi:10.1126/scitranslmed.aat8178 (2018).
6. Kozlenkov A, Li J, Apontes P *et al.* A unique role for DNA (hydroxy)methylation in epigenetic regulation of human inhibitory neurons. *Sci. Adv.* 4(9), eaau6190 (2018).
7. Gandal MJ, Zhang P, Hadjimichael E *et al.* Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* 362(6420), eaat8127 (2018).
8. Rajarajan P, Borrman T, Liao W *et al.* Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. *Science* 362(6420), eaat4311 (2018).
9. Havrilla J, Pederson B, Layer R, Quinlan A. A map of constrained coding regions in the human genome. *Nat. Gen.* doi:10.1038/s41588-018-0294-6 (2019).
10. https://medicalxpress.com/news/2018-12-big-datasets-regions-explore-genome.html
11. https://healthitanalytics.com/news/fda-recognizes-genomic-database-to-advance-precision-medicine
12. www.ucl.ac.uk/cancer/news/2018/nov/advancing-person-alised-medicine-research-open-access-genomic-data
13. Landry LG, Ali N, Williams DR, Rehm HL, Bonham VL. Lack of diversity in genomic databases is a barrier to translating precision medicine research into practice. *Health Affairs* 37(5), 780-785 (2018).
14. www.genome.gov/27553745/nhgri-symposium-commemorates-10th-anniversary-of-the-human-genome-project/