A White Paper for Genomics in Horizon Europe

Draft of 3 Dec 2020

Contact:
Declan Kirrane
Declan.kirrane@iscintelligence.com

ISC
Intelligence in Science

1. SUMMARY OF THE PROPOSED STRUCTURE

Medical genomics is the use of genomic data to better understand the genetic and molecular basis of health, disease and drug responses. The major applications are to improve the diagnosis, prevention and treatment of disease, and to accelerate the development of new and better diagnostics, drugs and biologicals such as antibodies and vaccines. Although the key principle of medical genomics is the use of **human genome data** for the applications listed above, it should also accommodate the genomes of important **human-associated microbes** (friend and foe) due to their impact on health and disease, including chronic disease. Furthermore, although genomics can be strictly defined as the use of **genome sequence data**, it is widely recognized that medical genomics also benefits from the broader omics disciplines (some of which use exactly the same technologies). These can be loosely grouped as **functional genomics** and they include (but are not limited to) the epigenetic regulation of genetic information (**epigenomics**), the analysis of gene expression and gene products (**transcriptomics**, **proteomics**) and the analysis of small molecules (**lipidomics**, **metabolomics**) moving towards a **systems biology** approach which incorporates multiple aspects of the (clinical) phenotype not just the disease (**clinomics** or **phenomics**). The future development of medical genomics should consider the **underlying technologies** which comprise standardized methods and operational procedures for data acquisition (DNA/RNA sequencing, protein/metabolite analysis, clinical analysis and diverse phenotyping methods such as high-content imaging) and analysis (increasingly reliant on deep learning and other forms of artificial intelligence) as well as data security, and the surrounding ethical issues.

Medical genomics arose from the human genome project and has traditionally had an impact on three key areas of clinical relevance:

1) Identifying associations between **specific genetic variants and human disease** leading to more accurate **diagnoses of genetic diseases** and more effective **genetic testing** particulary in the arena of **rare disease**.
2) Enabling the **stratification of cancer patients** based on the specific genetic characteristics of tumours to allow **more specific diagnosis** and in many cases more **targeted therapeutic intervetions.**

3) Predicting the **impact of drugs based on genetic variation**, commonly referred to as **pharmacogenomics**, allowing patients to be classified as responders/non-responders or grouping them by dose/therapeutic window.

These genetic variations are either the basis of the disease/drug response because they directly affect the corresponding genes (which are therefore likely to lead us to good drug targets) or are so tightly linked to these genes that they are reliable biomarkers for disease diagnosis or for disease tracking and risk prediction in family pedigrees. By including the broader aspects of functional genomics we can add a fourth major clinical area:

4) Identifying and validating new **drug targets** (mostly proteins, occasionally RNAs, and in a few cases actual genes) as well as **disease or drug-treatment biomarkers** that include RNAs, proteins, lipids and other metabolites, as well as extracellular vesicles (exosomes) and circulating tumour DNA (ctDNA).

The general trend in medical genomics has been to move from a generalised to a more personalised approach ultimately leading to the concept of **personalized medicine** where disease diagnosis and treatment is a bespoke process driven by the unique genomic data of each patient. The future funding of medical genomics should move beyond these four pillars to consider additional goals but should continue to support the original pillars while increasing their scope and ambition as set out below. To encourage rapid translation of scientific breakthroughs and innovations, the commitment to stimulating meaningful collaboration between industry and academia should also continue and grow within Horizon Europe.

The proposed white paper will therefore highlight **eight activities** that should be recommended for inclusion in Horizon Europe.

ACTIVITY 1 – GENETIC VARIATION AND OTHER BIOMARKERS FOR HUMAN DISEASE DIAGNOSIS/TESTING

Genetic variations were initially associated with Mendelian diseases via linkage mapping or association studies and these genetic markers or biomarkers evolved into the first genetic tests (including prenatal diagnosis). We now have the ability to

sequence and analyse the entire genome (or its most established parts, the exome) and this should be exploited *to find associations with (and develop tests for) more complex diseases including multifactorial diseases with heterogeneous manifestations*. This is especially important for diseases that benefit from early identification and treatment but are difficult to diagnose early in the disease course based solely on clinical manifestations. It is very likely that in the short to medium term genetic variation alone will not be used singularly for diagnosis and that more information, gathered through the analysis of biomarkers or combinations of biomarkers (**biomarker profiles**) as well as environmental/lifestyle data from electronic health records (a field sometimes described as "**clinomics**"), will be required for effective diagnosis and treatment. Notwithstanding this, **the accrual of data and evidence following standardised methods,** does over time, present an opportunity for detection of variation alone to play a greater first line role in diagnosis, prognosis and guiding treatment.

ACTIVITY 2 – GENETIC VARIATION AND OTHER BIOMARKERS FOR HUMAN DISEASE STRATIFICATION AND RISK ANALYSIS

Genomics was initially used for disease stratification in the field of oncology because cancer cells undergo overt genomic rearrangements associated with disease progression and tumour tissues can be excised by biopsy and compared to healthy tissue. With technological advances it is now possible to analyse the genome at a **single-cell level**, or indeed to compare genome level variation that arise spatially in cells and tissues distal to eachother in one individual. These technologies will allow not only *more sensitive detection and diagnosis of cancer (including haematological malignancies) but also other diseases that have been inaccessible thus far to this kind of detailed analysis*. One example is the complex area of autoimmune and inflammatory disorders, involves a subset of normal cells 'going rogue' and attacking other tissues. Autoimmune and inflammatory disorders have an overlapping and heterogeneous set of clinical phenotypes but genomic analysis of the immune cell population at the single cell level combined with the sensitive detection of inflammatory biomarkers could provide a new approach to the diagnosis and treatment of such disorders. Lipidomics is an emerging aspect of this field because lipids have been identified as important signalling molecules during inflammation.

The use of supervised learning could help to diagnose and/or stratify diseases and quantify the risk of progression, which is important for making early therapeutic decisions. Unsupervised learning could also reveal new and informative stratifications and risk categories, including the risk of comorbidities, especially when combined with other clinical data.

ACTIVITY 3 – GENETIC VARIATION AND OTHER BIOMARKERS FOR THE PREDICTION OF DRUG RESPONSES

The study of drug metabolism and responses associated with specific genetic variations is known as pharmacogenetics, and at the genomic level it is pharmacogenomics. Progress in this field has allowed the correlation of specific genetic variants (often SNPs or haplotypes in CYP genes) and more recently gene expression profiles defined by RNA-Seq analysis and proteomics with drug responses (ADME). As above, we now have the data to use this approach *in more complex scenarios such as comorbidities (drug responses in patients with multiple diseases), heterogeneous/multifactorial diseases (integration of other clinical data including environmental/lifestyle data into drug selection/dosing decisions) and polypharmacy (genomic associations of responses to multiple simultaneous drugs*. Research and innovation efforts that accelerate the translation and application of pharmacogenomics into clinical pathways would be worthwhile.

ACTIVITY 4 – DRUG TARGET IDENTIFICATION AND VALIDATION AND DRUG DEVELOPMENT

The association of genetic variation (typically SNPs) and gene expression profiles with diseases often leads to the identification of candidate drug targets, but the value of such targets depends on their frequency in the population as well as their 'druggability' (suitability for pharmacological intervention). *The combined use of genomic data, the analysis of sequence information to identify the gene product, structural analysis and modelling to predict interactions with small molecules, high-content imaging to reveal the distribution and abundance of the target, and population-wide genomics data to investigate target frequency, provides a powerful strategy to rank new drug*

target candidates and avoid the wasteful pursuit of dead ends. The development of new cell-based assays including stem cells and 3D-cultures combined with complex readouts involving panels of biomarkers will facilitate the downselection of candidates while avoiding expensive and unnecessary animal studies. Furthermore, the stratification of patients with heterogeneous and multifactorial diseases using progress from Activities 1-3 will allow the preselection of clinical trial subjects that are more likely to respond to new drug candidates.

ACTIVITY 5 – DATA ANALYSIS MOVING FROM GENOMICS TO PHENOMICS

Medical genomics associates genetic variation with disease and drug responses and incorporates other biomarkers (ctDNA, RNA, protein, metabolites) that can be regarded as "molecular phenotypes" linked to the actual phenotype (disease or drug response). *There is also great value in approaching the problem from the other direction because there is a great deal more phenotypic data than genotypic data*. **Direct phenotypes** are clearly associated with the disease, but other **indirect or abstract phenotypes** can be equally informative even if their relevance to the disease is unclear. Some of these phenotypes are discovered by chance, but the combination of electronic health records (which record a wide range of health parameters) and now **multimodal image analysis**, particularly **high-content image analysis at the cellular and tissue levels** as well as traditional imaging modalities such as MRI and CT, allow the extraction of large amounts of visual phenotypic data that can be highly informative. These data can be analysed by machine learning algorithms either in a supervised manner (using known diseases as categories, allowing new phenotypes to be associated with particular genotypes and diseases). Furthermore, unsupervised (deep) learning algorithms can be used to seek patterns in the data without the prior imposition of categories, allowing the use of abstract phenotypes and combinatorial phenotypes (e.g. images combined with biomarkers) for stratification and monitoring clinical outcomes/efficacy. The sucessful utilisation of advanced analysis techniques such as unsupervised machine learning are highly dependent on widespread availability of large well curated and harmonised datasets. Enabling Activities summarised in section 8 will assist in the creation of such utilities.

Activity 6 – Optimising the development, production and delivery of cell-based therapies & biologic drugs

The production of cell-based therapies, monoclonal antibodies and other biologic drugs can benefit from production enhancements that arise through innovations based on the application of genomics in the development and supply chain lifecycle. Production yields and batch qualities are heavily dependent on the unique characterisitics of the cell systems employed in the production process. Genomics can play an increasing role in optimising these production pathways. Similarly, and related to activity 3, in the future the targeted production and prescription of cell-based therapies will rely heavily on genomics in both a cell donar and recipient context.

Activity 7 – genetic variation and interactions with microbes and the environment

Humans are associated with microbes from birth and nurture a microbiome of commensal species that helps to establish a healthy gut and healthy skin. The role of the microbiome in health and disease has been overlooked, but the gut-brain axis (interactions between enteric microbes and the central and enteric nervous systems) is now known to play a key role in health and disease, especially chronic disease. The makeup of the microbiome is known to have both genetic and environmental components but the role of human genome variation in this context has not been explored in detail and it would be valuable to investigate the role of human genetic variation and nutrition on the microbiome and health. This is a gateway to the wider aspects of genomics in society, where the human genome and associated microbial genomes interact with the "nutritional genome" (the genomes of our staple food crops) and the microbes in our environment.

Activity 8 – enabling activities: data access, analysis, Development of standards, privacy, security and ethics, public awareness

The collection of large amounts of heterogeneous data (genomics, functional genomics, phenomics and clinomics) and their analysis along with environmental and lifestyle factors will continue raise numerous operational, societal, ethical and legal

issues. The evolution of privacy law globally particularly as it relates to human health research will continue to be an important feature of the field. As the technology and the field advance, questions surrounding data management and access, data ownership, the identifiability of individuals based on nucleic acid genomic analysis and so on, will continue to straddle the domains of ethicists, privacy legal experts and society researchers.

This research lens of ethics, privacy law and societal attitudes will need to be applied to the multitude of new technologies and methods associated with medical genomics that will arise, especially those that utilise artificial intelligence for analysis of datasets, and those relating to the modification of the genome (gene therapy/cell therapy, CRISPR and other genome editing technologies) and the provision of genomic information to patients (DTC genetic testing, risk evaluation, genetic counselling and more direct consumer orientated 'services' such as ancestry profiles).

The greatest value can be derived from heterogeneous data by standardizing the means of collecting, annotating (including metadata), storing and distributing data, heading towards "Genomics 4.0" following the model of Industry 4.0 and, more recently, Pharma 4.0. This **development of standards** would fold the current accepted standards for sequence data, proteomics/metabolomics and clinical records with real-world data collected from wearable or implanted devices, and would improve the performance of new deep learning algorithms designed to identify objects in these data, assign data to categories, or self-learn new categories (representing disease types/stages or drug responses).

We also need to develop a more robust process for the commercial utilisation of genomic resources within the ethical limitations discussed above. The development of **public-private partnership and collaboration models** that can accelerate the realisation of the promise of precision medicine for European citizens will also flow naturally from such processes.