Genomics
Genetics, Transcriptomics and Epigenetics

Subgroup report

Didier Meulendijks, NL
Dieter Deforce, BE
Hans Ovelgönne, NL
Renate König, DE
Marjon Pasmooij, NL (subgroup lead)

See websites for contact details
Heads of Medicines Agencies www.hma.eu
European Medicines Agency www.ema.europa.eu

An agency of the
European Union

1. Summary

This report is a deliverable from the HMA/EMA Big Data Task Force, 'Genomics' subgroup, and focuses on the characterisation and mapping of a type of 'big data', namely genomics data, including epigenetics and transcriptomics data. In addition, this report describes the potential usability/applicability of genomics data in regulatory processes as well as recommendations from the HMA/EMA Big Data Task Force regarding how to optimise the future use of genomics (big) data in regulatory processes.

There are different public data sources (databases) where genomics data can be uploaded and freely accessed. Most publicly available genomics data sources are derived from investigator-initiated (non-industry-driven) initiatives and contain only genomics data, without phenotypic/clinical outcome data linked to the genomics data. Some data sources do contain phenotypic data, mainly data on the presence or absence of genetic/hereditary diseases. Other phenotypic data, in particular clinical outcome data (e.g. data on efficacy or safety of treatments) are currently only sporadically found in available public databases (e.g. PharmGKB, https://www.pharmgkb.org/), although a vast number of initiatives are now being performed which do link genomics data to clinical data. Most genomics data generated by pharmaceutical industry, which is often genomics data linked to clinical outcomes (e.g. from clinical trials), are not publicly available. These data would be of interest to regulators as the data could be used for regulatory purposes (e.g. pharmacovigilance, identification of biomarkers for efficacy, etc.).

Incentives to make genomics data available to regulators and/or publicly available could be beneficial for the regulatory system, e.g. it could facilitate more individual patient-based B/R assessment as opposed to population-based B/R assessment. Therefore, it could be considered to request companies upon a marketing authorisation application to make genomics data linked to clinical data from the pivotal clinical trials available in public databases or to the EMA. By doing so, the data would be accessible for further analyses for academic as well as potentially for regulatory purposes. To facilitate data sharing in a secure way, it could be explored whether the EMA should provide a central platform for sharing of clinical trial data, or whether the EMA could provide a portal linking to industry-owned data. To be able to optimally profit from available data, it would be important to link the most important parameters related to phenotype and/or treatment outcome to the genomics data (e.g. adverse events, primary efficacy outcomes). Importantly, various privacy, security and ethical issues need to be addressed before genomics data sharing can become common practice. For example, informed consent would have to be adequately covered in relation to data sharing. Furthermore, to be able to link data from different sources, a system that includes patient identifiers that ensures both adequate linking of data and patient privacy would be needed. Linking clinical and phenotype data across datasets would both empower precision medicine, but also introduce new privacy risks. The latter is especially of concern for rare diseases where there are sometimes only a few patients with a specific mutation worldwide.

Genomics data quality can be improved by improving standardisation (e.g. making use of standard operating procedures), by requiring raw data to be shared in addition to processed data, by requiring meta-data to be attached to the data (i.e. descriptive information about the overall study, individual samples, all protocols, and references to processed and raw data file names), by certification of the instruments used for analysis, and by setting a minimal data standard. It is important to not only standardise the genomics data, but also the clinical outcome/phenotypic data.

Genomics analyses require highly specific skills and knowledge. Therefore, although it is anticipated that regulators will not do these highly specialised analyses themselves, knowledge should be available within the regulatory network to be able to assess big data analyses as part of a marketing

authorisation application. Collaborations with skilled academic groups as well as clustering of expertise within a working party (similar to the Pharmacogenomics Working Party) and/or different regulatory agencies in the network or educating assessors using the EU NTC platform could be considered.

Lastly, in order to advance genomics-guided treatment in clinical practice, it would be advisable to make clinically meaningful information regarding genomics data more readily available in the SmPC, including the most up-to-date information. Further, it could be considered to make this information available online in a separate database, which could be searched by pharmacists, geneticists or physicians, and which would be linked to the SmPC. However, it would be important to have an adequate system in place to curate the presented information and be clear about the level of evidence available for clinical utility of the described genomics-outcome associations.

The end result of this report by the genomics subgroup of the Task Force on Big Data is a number of recommendations for future actions in relation to the use of genomics (big) data in regulatory processes. The full table with recommendations based on the mapping exercise can be found below in Table 12 in section 6.2 Specific recommendations from the analysis.

2. Background

This report is a deliverable from the HMA/EMA Big Data Task Force, 'Genomics' subgroup, and focuses on the characterisation and mapping of one type of 'big data', namely genomics data, including epigenetics and transcriptomics data. Genomics data can be expected to have major implications on regulatory processes in the near future, and therefore an overview of the regulatory challenges ahead is warranted. This report summarises the results of a mapping process to map the available relevant sources of big data and define their main formats, which was performed in the context of the HMA/EMA Big Data Task Force. In addition, this report describes the potential usability/applicability of these genomics data sources in the regulatory processes as well as recommendations from the Big Data Task Force regarding how to optimise the future use of genomics (big) data in regulatory processes.

2.1. Genomics

Genomics – the study of genes and their functions – comprises different aspects of the genome, including genetics (variations in DNA sequence and their function), transcriptomics (variations in RNA sequence and their function), and epigenetics (the study of modifications of gene expression rather than alteration of the genetic code itself; see also Appendix 1 for explanation of definitions). These different aspects of genomics already have a wide variety of applications in current medical practice, and the number of applications can be expected to further increase in the coming years.

2.2. Genetics

Genetic variants have been used for decades to diagnose a number of hereditary diseases which previously were of unknown origin, including cystic fibrosis, Duchenne muscular dystrophy, and spinal muscular atrophy (disease genomics). Genetic variants are also increasingly used to inform medical treatment decisions, i.e. as part of 'personalised medicine'. One promising area of genomic medicine is the ability to match an individual's genetic profile to the likelihood of experiencing an adverse reaction or a therapeutic response with particular drugs (pharmacogenomics). A large number of anticancer drugs are already being used specifically in patient populations selected based on certain genomic characteristics in order to achieve optimal efficacy, including vemurafenib in BRAF V600-mutated melanoma (SmPC Zelboraf®), erlotinib in epidermal growth factor receptor (EGFR)-mutated non-small cell lung cancer (SmPC Tarceva®), and cetuximab in RAS wild-type metastatic colorectal cancer (SmpC Eribitux®). The patient's genetic makeup can also be used to predict the occurrence of adverse drug

reactions, such as the severe and potentially life-threatening hypersensitivity reactions that occur in patients carrying the *HLA-B**5701 allele who receive abacavir for the treatment of human immunodeficiency virus infection (SmpC Ziagen®), or severe gastrointestinal and haematological toxicity in patients treated with fluoropyrimidine chemotherapy who carry a deficient *DPYD* gene (Dean, 2016). Furthermore, genetic variability in cytochrome P450 isoenzymes (CYP's) occurs frequently in patients and is a major source of interindividual differences in drug metabolism, with potential consequences for efficacy and safety of many small-molecule drugs (e.g. *CYP2C19* deficiency in patients treated with clopidogrel; Dean, 2015).

There are many examples of evolving genetic applications that will change the way we treat patients in the very near future. One such example is the use of circulating tumour DNA (Dawson et al., 2013; Lippman and Osborne, 2013; Han et al., 2017). Circulating tumour DNA allows for the highly sensitive and non-invasive detection of tumour DNA in blood plasma, which can be used to determine the genomics of tumours without the need for an invasive biopsy. Importantly, quantitative and qualitative changes in circulating tumour DNA can be used to determine the patients' response to treatment, and the use of circulating tumour DNA has already been shown to be superior to conventional methods of determining treatment response in subsets of patients with cancer.

2.3. Transcriptomics

The transcriptome is the set of all RNA molecules in a cell or a population of cells. It is used to refer to all RNAs, or to just messenger RNA (mRNA). It differs from the exome in that it includes only those RNA molecules found in a specified cell population, and usually includes the amount or concentration of each RNA molecule in addition to its molecular identity. Advanced clinical applications related to RNA expression are already being implemented in clinical practice, and these applications change the way patients are treated. One example is gene expression profiling of primary breast cancer after surgical resection, e.g. using the 'MammaPrint', to determine whether women need adjuvant chemotherapy after surgery, or whether adjuvant chemotherapy provides no additional benefit (Krop et al., 2017). A recent large randomised-controlled study showed that when gene expression profiling is used in combination with clinical–pathological risk stratification to determine whether chemotherapy should be administered, approximately 46% of the women with breast cancer for whom adjuvant chemotherapy used to be the standard of care, can be spared from receiving chemotherapy, as their outcome is not further improved with chemotherapy (Cardoso et al., 2016).

2.4. Epigenetics

Epigenetics – the study of changes in organisms caused by modification of gene expression rather than alteration of the genetic code itself, e.g. through DNA methylation – appears to be of critical importance in regulating gene expression. Therefore, like genomics and transcriptomics, epigenetics is likely to have an increasing impact on health care in the near future. For example, cancer is nowadays considered to be both a genetic and an epigenetic disease. The epigenome comprises the chemical changes to the DNA and histone proteins of an organism and is overlaid on DNA in the form of epigenetic traits that are heritable during cell division, but do not alter the DNA sequence itself. The pattern of these chemical tags is called the epigenome of the cell, whereas epigenetics is the study of the function of these marks that lead to alterations in gene expression (Figure 1). Epigenetic mechanisms can be divided in:

- **DNA methylation** occurs by the addition of a methyl (CH3) group to DNA, thereby often modifying the function of the genes and affecting gene expression. The most widely characterised DNA methylation process is the covalent addition of the methyl group at the 5-carbon of the cytosine ring resulting in 5-methylcytosine. These methyl groups project into the major groove of

DNA and inhibits transcription. In human DNA, 5-methylcytosine is found in approximately 1.5% of genomic DNA. Methylation in promoter regions correlates negatively with gene expression.

- A **histone modification** is a covalent post-translational modification (PTM) to histone proteins, which includes methylation, phosphorylation, acetylation, ubiquitylation, and sumoylation. The PTMs made to histones can impact gene expression by altering chromatin structure or recruiting histone modifiers. Histone modifications act in diverse biological processes such as transcriptional activation/inactivation, chromosome packaging, and DNA damage/repair.
- **Noncoding RNAs** (ncRNAs) may contribute to regulation of protein expression and therefore modulate drug effects. NcRNAs are RNA molecules that are transcribed from DNA but not translated into proteins. Their function is to regulate gene expression at the transcriptional and post-transcriptional level by playing a role in heterochromatin formation, histone modification, DNA methylation targeting and gene silencing. They can be divided into two main groups; the short ncRNAs (<30 nucleotides) and the long ncRNAs (>200 nucleotides). The short ncRNAs can be further classified into three major classes of microRNAs (miRNAs), short interfering RNAs (siRnAs), and piwi-interacting RNAs (piRNAs).

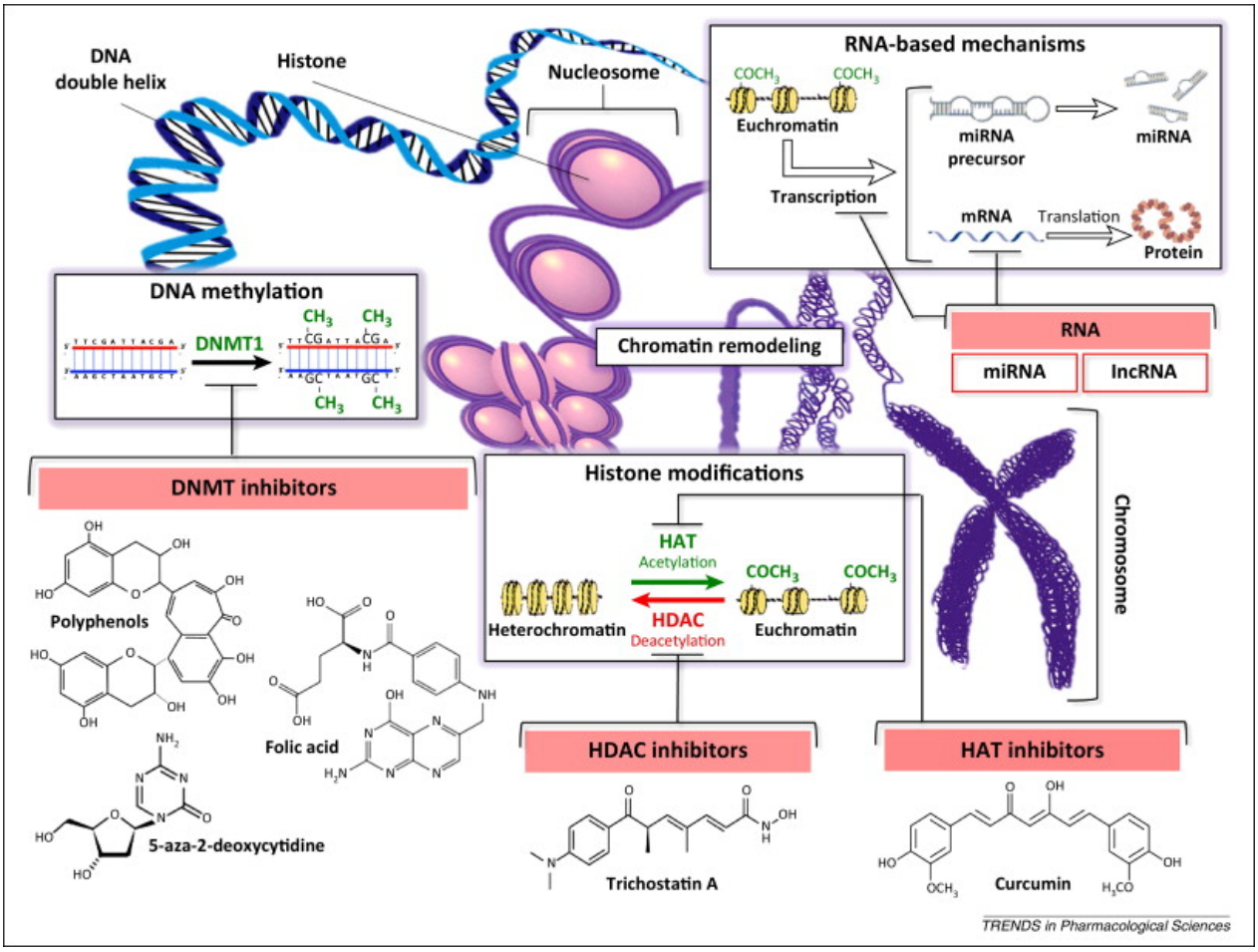

Figure1. This schematic representation illustrates epigenetic mechanisms, including DNA methylation (including cytosine methylation), histone alterations, and RNA-based transcriptional control, which can alter the cellular gene expression profile. Chemical structures of selected compounds targeting epigenetic modifications are also reported. A simplified scheme illustrating the structure of mammalian chromatin is also presented. Abbreviations: DNMT, DNA methyltransferase; HAT, histone acetyltransferase; HDAC, histone deacetylase. Source: Schiano et al., 2015.

At present, a major area of interest in the clinical use of epigenetics is that of biomarkers, which are prognostic and/or predictive of response to therapeutics. Epigenetic changes have been shown to be a key aspect of cancer development, and epigenetic variability has been shown to be prognostic and/or predictive in many cancers, including haematological malignancies (e.g. multiple myeloma and myelodysplastic syndrome) and many solid tumour types (e.g. colon cancer, prostate cancer, and glioblastoma) (Blute et al., 2015; Glavey et al., 2016; Lao et al., 2011). A number of diagnostic tests for epigenetic changes are under development as prognostic and/or predictive biomarkers for different types of cancer. One epigenetic biomarker, which is already being used in routine clinical practice, is MGMT promoter methylation, a marker that is predictive of the patient's response to treatment with temozolomide chemotherapy in glioblastoma (Seystahl et al., 2016; Herceg et al., 2017).

In addition, there is increasing evidence that individual differences in drug response may also result from epigenetic alterations such as histone-acetylation or DNA-methylation (Cascorbi and Schwab, 2016). Fisel et al. (2016) outlined the influence of DNA methylation on genes involved in the absorption, distribution, metabolism and excretion (ADME) of drugs, showing that over 60 ADME genes have been considered to be influenced by epigenetics, including via histone modifications, DNA methylation, and miRNAs. An extensive summary regarding the available knowledge on regulation of ADME gene expression by DNA methylation is provided in Fisel et al., 2016.

MiRNAs are frequently dysregulated in malignancies and involved in tumour cell drug resistance (Fanini and Fabbri, 2016). Depending on which genes or pathways are regulated by specific miRNAs in a specific cancer type, miRNAs can act as onco-miRNAs, or suppressor-miRNAs. A new field of drug research is the modulation of specific miRNAs deficiencies, by either antagonists or mimics, with the aim to improve treatment outcome by restoring the network of gene regulation associated with pathways such as drug resistance.

Acknowledging the importance of epigenetics, large-scale studies of human disease-associated epigenetic variation, specifically variation in DNA methylation have been performed, e.g. to determine links between epigenetics and development of human diseases (EWAS, Rakyan et al., 2011), similar to how genome wide association studies (GWAS) have been performed. An important aspect will be to integrate the EWAS with GWAS data to allow better functional analysis.

2.5. Microbiomics

The human microbiota (the total of all microorganisms present in/on humans) consists of the 10-100 trillion symbiotic microbial cells harboured by each person, primarily bacteria in the gut; the human microbiome consists of the genes these cells harbour. The microbiome thus refers to the overall collection of genes of all the microbes comprising a human microbiota. While an individual's genome is fixed for life, the microbiome changes over time. More and more studies highlight the potential utilisation of the microbiome as a potential therapeutic option, although the large majority of studies on the role of the microbiome in the pathogenesis of disease are correlative and preclinical (Lynch and Pedersen, 2016).

3. Objectives

- To map relevant sources of genomics data (i.e. genetics, transcriptomics, and epigenetics) and genomics data formats.
- To discuss issues related to data quality.
- To discuss issues on access to data (data sharing) and privacy/ethical issues.

- To identify regulatory challenges related to the use of big data sources within regulatory processes
- To make recommendations on the usability and potential applications of genomics data in regulatory processes across the product life cycle.

4. Methods

For the mapping exercise, a focused internet search was conducted for genomics data, i.e. genetics, transcriptomics and epigenetics data. Use was made of peer-reviewed publications by searching PubMed. Furthermore, informal input was received from a member of the Pharmacogenomics Working Party (Marc Maliepaard, NL) and several external experts (Hanns Lochmüller, RD-connect; Lude Franke, associate professor, Department of Genetics, University Medical Center Groningen, NL). The focus was on providing an overview that especially mapped those data sources that could be of value in regulatory processes.

Microbiome data was not mapped in detail in this report, given that the methodology for analysing microbiome data resembles that of genomics data, and because microbiomic applications are far from mature and will likely be highly specific. Therefore, it is difficult at this stage to provide specific recommendations for microbiome data. However, the recommendations regarding genomics given in this document will generally also apply to microbiomics.

5. Results of the data characterisation

5.1. General overview and history

In Figure 2 an overview is provided on the different aspects of collecting and analysing genomics data. For obtaining information on the individual's genetic makeup, usually a blood sample or a buccal swap will be required. In case of a specific disease, for instance in the field of oncology, a biopsy is usually necessary to study somatic mutations in tumour tissue.

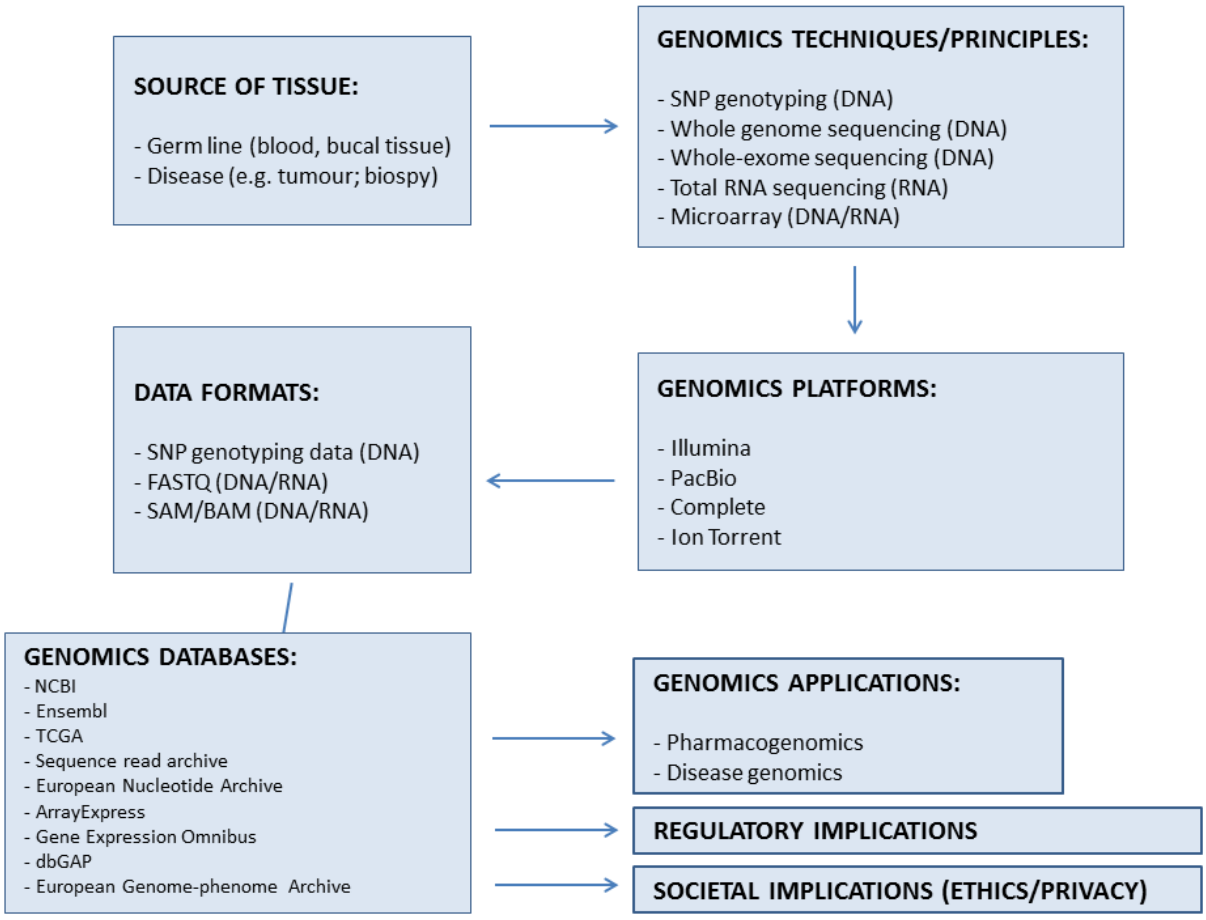

Figure 2. Overview of the different aspects of genomics data, from source of tissue to applications and implications.

There is a wide variety of genomic techniques that can be performed (Table 1). Sanger sequencing, which typically focusses on sequencing of a single gene, has been used traditionally. More recently, however, targeted gene panels were introduced. Targeted gene panels have been optimised to capture key genes or regions of interest. For instance, cardiopanels have been developed by different hospitals, which include known cardiogenes (e.g. http://biosb.nl/wp-content/uploads/2014/10/Day-2-Jongbloed-Cardio-Gene-Panel.pdf). With such a targeted gene panel all cardiogenes can be screened in one run, leading to a significant reduction in the time required to establish a diagnosis compared to Sanger sequencing (i.e. separate sequencing of individual genes).

Nowadays, next-generation sequencing (NGS) is increasingly performed, and different NGS systems have been introduced in the past decade. The critical difference between traditional Sanger sequencing and NGS is that NGS extends the process of sequencing of a single DNA fragment to sequencing of millions, or even billions of sequencing reactions at the same time. Although different machines have been developed, with varying technical details, they all share several common features, i.e.: library preparation, cluster generation, sequencing, and data analysis.

Because the quality of whole-exome sequencing (WES) and whole-genome sequencing (WGS) has improved, and the costs have been reduced (Figure 3), this technology is increasingly applied in clinical practice to inform patient care. Further, in some hospitals and in clinical trials it is already becoming the standard (e.g. specialised cancer centres and academic hospitals).

Whole-genome sequencing allows for analysis of both the exonic (i.e. coding) regions as well as the intronic (non-coding) regions of the DNA. With WES, on the other hand, only the exons in the genome are captured and analysed (approximately 30 million base-pairs, instead of the whole genome, composed of roughly 3 billion base-pairs). The focus is on the exons, because these are translated into functional proteins, in which mutations are most likely to have a direct phenotypic consequence. The pros and cons of different sequencing techniques are summarised in Table 1.

Table 1. Pros and cons of different sequencing techniques

GENOMICS TECHNIQUE	PROS	CONS
Sanger sequencing	- Simple technique, widely available. - Cost-efficient and time-efficient when there is an indication, which gene/mutation to investigate.	- Sequencing restricted to one or several DNA fragments. - Mutations in non-coding regions (e.g. intronic variants) will be missed. - No discovery of new genes involved in a specific disease. - In case multiple genes could be involved, then it costs a lot of time compared to WES/WGS.
Whole-exome sequencing (WES)	- High coverage in targeted regions. - Reduced costs for large genomes compared to WGS*.	- Mutations in non-coding regions (e.g. intronic variants) will be missed. - Genetic markers can only be genotyped if they are in the targeted regions. - Requires information about targeted regions and enrichment kits. - Risk of incidental findings, i.e. previously undiagnosed medical or psychiatric conditions that are discovered unintentionally and are unrelated to the current medical or psychiatric condition which is being treated or for which tests are being performed. - Data interpretation: analysis of found variants can be time-consuming.
Whole-genome sequencing (WGS)	- Most comprehensive technique, i.e., whole genome analysed, including the identification of non-coding mutations (e.g. intronic variants). - Access to the whole genomic sequence.	- Expensive for large genomes*. - Risk of incidental findings. - Data interpretation: analysis of found variants can be time-consuming.
RNA-sequencing (RNA-seq)	- Simultaneous analysis of expression (differences) possible. - Complexity reduction of large genomes without prior knowledge about genes. - Effects of splice-site mutations are readily identifiable.	- Mutations in regulatory regions or non-expressed genes will be missed. - Genetic markers can only be genotyped if they are expressed.
Epigenetic techniques (e.g. promoter methylation)	- Analysis of epigenetic changes, which cannot be assessed by the other techniques mentioned above.	- Wide variety of techniques depending on the specific type of modification, i.e. DNA methylation or histone modification including methylation, acetylation, phosphorylation and sumoylation. - Epigenetics may change over time and vary between different organs/samples.

* The price of WGS is dropping rapidly (see Figure 3), and it is expected that costs will not be a limiting factor in the near future.

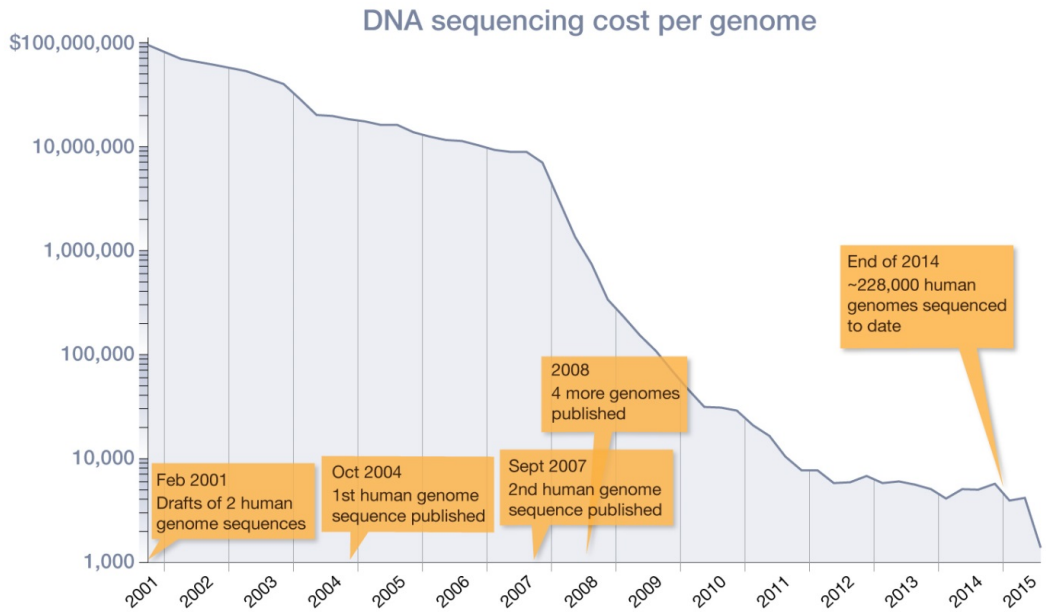

Figure 3. DNA sequencing cost per genome over time.

Source: http://learn.genetics.utah.edu/content/precision/time/. References: 1) National Human Genome Research Institute (updated October 2, 2015). *DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP)*. Retrieved November 2, 2015, from http://www.genome.gov/sequencingcosts/.; 2) Nature editorial staff (2010). Human genome at ten: The sequence explosion. *Nature, 464*, 670-671.

5.2. European regulatory scientific guidelines

Several regulatory guidance documents on genomics have been made available by the EMA. An overview is provided in Table 2. In addition, the Pharmacogenomics Working Party (PGWP) provides recommendations to the CHMP on matters directly and indirectly related to pharmacogenomics (http://www.ema.europa.eu/ema/index.jsp?curl=pages/contacts/CHMP/people_listing_000018.jsp&mid=WC0b01ac0580028d91). The PGWP has expertise on genomics in the regulatory process and consists of up to 14 experts nominated by the CHMP.

The "Guideline on good pharmacogenomics practice" (22/02/2018; http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2018/03/WC500245944.pdf) points out that although pharmacogenomics research has revealed a number of variable genetic loci that influence drug response, some clinical studies on pharmacogenomics have resulted in "ambiguous findings", which highlights the importance of correctly measuring, interpreting and translating pharmacogenomics data into clinical treatment. Important pitfalls that were identified in published studies are:

- Poor quality of the employed analytics.
- Analyses of non-relevant Single Nucleotide Variations (SNVs).
- Analysing somatic instead of germline DNA when germline DNA analysis is intended and vice versa.
- Lack of appropriate patient selection.
- Lack of appropriate phenotype identification.
- Lack of power in relation to the frequency of the genetic variation studied.
- Non relevant endpoints selected for the basis of the study.
- Failure to take into account the pharmacology of the drug in the design of the study.

Further, this guideline lays out its guidance on pharmacogenomic variants: phenotyping and genotyping; important issues to consider when analysing the tumour genome, including information on the hot topic of liquid biopsies and identifying circulating tumour DNA; DNA sequencing design; quality aspects of pharmacogenomic analyses, including guidance on analytics; study design; pharmacogenomic biomarkers and translation in the clinic today; and the future dynamics of drug labels.

The EMA does not assess (companion) diagnostics and genetic testing platforms; this responsibility lies with the notified bodies. However, the EMA and national agencies can be requested by a notified body to provide input on the medical device in a consultation procedure. Of note, currently a guideline is being drafted by the PGWP, outlining recommendations on developing predictive biomarker-based assays including companion diagnostics (CDx). Moreover, there are new European regulations on medical devices ((EU)2017/745) and *in-vitro* diagnostics ((EU) 2017/746), which will go into effect from May 2020 and May 2022, respectively.

No specific guidelines currently exist on the use of epigenetics; however, much of the guidance described above may also apply to epigenetics data.

Table 2. EMA scientific guidelines on genomics ordered by date

Title	**Status**	**Date**
Position paper on terminology in pharmacogenetics EMEA/CPMP/3070/01	adopted	21/11/2002
Guideline on pharmacogenetics briefing meetings EMEA/CHMP/PGxWP/20227/2004	adopted	27/04/2006
General principles processing joint FDA EMEA Voluntary Genomic Data Submissions (VGDSs) within the framework of the confidentiality arrangement	adopted	01/04/2007
Reflection paper on pharmacogenomics samples, testing and data handling EMEA/CHMP/PGxWP/201914/2006	adopted	15/11/2007
ICH E 15: Definitions for genomic biomarkers, pharmacogenomics, pharmacogenetics, genomic data and sample coding categories - Step 5 EMEA/CHMP/ICH/437986/2006	adopted	01/11/2007
Reflection paper on the use of genomics in cardiovascular clinical intervention trials EMEA/CHMP/PGxWP/278789/2006	adopted	01/11/2007
Reflection paper on pharmacogenomics in oncology EMEA/CHMP/PGxWP/128435/2006	draft: consultation closed	01/04/2008
ICH: E 16: Note for guidance on genomic biomarkers related to drug response: context, structure and format of qualification submissions - Step 3 EMEA/CHMP/ICH/380636/2009	draft: consultation closed	01/06/2009
Reflection paper on co-development of pharmacogenomic biomarkers and assays in the context of drug development EMA/CHMP/641298/2008	draft: consultation closed	30/11/2010
Reflection paper on methodological issues associated with pharmacogenomic biomarkers in relation to clinical development and patient selection EMA/446337/2011	draft: consultation closed	25/11/2011
Guideline on the use of pharmacogenetic methodologies in the pharmacokinetic evaluation of medicinal products EMA/CHMP/37646/2009	adopted	19/01/2012
International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use guideline E16: Genomic biomarkers related to drug response: context, structure and format of qualification submissions - Step 5 EMA/CHMP/ICH/380636/2009	adopted	11/02/2013
Guideline on key aspects for the use of pharmacogenomics in the pharmacovigilance of medicinal products EMA/CHMP/281371/2013	adopted	20/11/2015
ICH: E 18: Guideline on genomic sampling and management of genomic	adopted	06/10/2017

data EMA/CHMP/ICH/11623/2016		
Concept paper on an addendum on terms and concepts of pharmacogenomic features related to metabolism to the Guideline on the use of pharmacogenetic methodologies in the pharmacokinetic evaluation of medicinal products (EMA/CHMP/37646/2009) EMA/CHMP/644998/2016	draft: consultation closed	07/07/2017
Concept paper on predictive biomarker-based assay development in the context of drug development and lifecycle EMA/CHMP/800914/2016	draft: consultation closed	28/07/2017
Guideline on good pharmacogenomic practice EMA/CHMP/268544/2016	adopted	22/02/2018

5.3. U.S. Food and Drug Administration (FDA)

Several regulatory guidance documents on Genomics have been made available by the FDA (Table 3). In contrast to the EMA, the FDA assesses and grants approval for (companion) diagnostics and genetic testing platforms, whereas in Europe this responsibility lies with the notified bodies. This difference is reflected by a larger number of FDA guidances related to *in vitro* (companion) diagnostics than in the EU.

Table 3. FDA guidances related to pharmacogenomics

Year	Status	Guidance Title
2018	Final	Use of Public Human Genetic Variant Databases to Support Clinical Validity for Next Generation Sequencing (NGS)-Based In Vitro Diagnostics (PDF, 499KB)
2018	Final	Use of Standards in FDA Regulatory Oversight of Next Generation Sequencing (NGS)-Based In Vitro Diagnostics (IVDs) Used for Diagnosing Germline Diseases (PDF, 708 KB)
2018	Final	E18 Guideline on Genomic Sampling and Management of Genomic Data (PDF, 170.5KB)
2017	Discussion paper	Discussion Paper on Laboratory Developed Tests (LDTs) (NB: not a guidance document)
2016	Draft	Principles for Codevelopment of an In Vitro Companion Diagnostic Device with a Therapeutic Product (PDF, 1.1 MB)
2016	Final	Clinical Pharmacology Section of Labeling for Human Prescription Drug and Biological Products — Content and Format (PDF, 143.8 KB)
2014	Final	Qualification Process for Drug Development Tools (PDF, 498.8 KB)
2014	Final	In Vitro Companion Diagnostic Devices (PDF, 159.2 KB)
2014	Draft	Framework for Regulatory Oversight of Laboratory Developed Tests (LDTs) (PDF, 312.5 KB)
2013	Final	Clinical Pharmacogenomics: Premarketing Evaluation in Early-Phase Clinical Studies and Recommendations for Labeling (PDF, 130.6 KB)
2012	Draft	Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products (PDF, 996.7 KB)
2012	Draft	Drug Interaction Studies — Study Design, Data Analysis, Implications for Dosing, and Labelling Recommendations (PDF, 827 KB)
2011	Final	E16 Biomarkers Related to Drug or Biotechnology Product Development: Context, Structure, and Format of Qualification Submissions (PDF, 708 KB)
2010	Draft	Adaptive Design Clinical Trials for Drugs and Biologics (PDF, 423.1 KB)
2008	Final	E15 Definitions for Genomic Biomarkers, Pharmacogenomics, Pharmacogenetics, Genomic Data and Sample Coding Categories (PDF, 1.1 MB)
2005	Final	Pharmacogenomic Data Submissions (PDF, 306.9 KB)

The FDA has recently assessed and approved the **MiSeqDx platform**, which is a sequencing instrument that measures fluorescence signals of labelled nucleotides through the use of instrument specific reagents and flow cells, imaging hardware, and data analysis software. The MiSeqDx Platform is intended for targeted sequencing of human genomic DNA from peripheral whole blood samples. The MiSeqDx Platform is not intended for whole genome or *de novo* sequencing. https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN130011.pdf.

In addition, in Table 4 a list of nucleic acid-based tests that have been cleared or approved by the FDA are provided. These tests analyse variations in the sequence, structure, or expression of DNA and RNA in order to diagnose disease or medical conditions, infection with an identifiable pathogen, and determine genetic carrier status.

Interesting to point out in table 4 is the first consumer-oriented genetic service of 23andME that was recently approved (www.fda.gov/newsevents/newsroom/pressannouncements/ucm551185.htm). These are the first direct-to-consumer tests authorised by the FDA that provide information on an individual's genetic predisposition to certain medical diseases or conditions, which may help to make decisions about lifestyle choices or to inform discussions with health care professionals.

Table 4. List of FDA approved Human Genetic Tests

Acute Myeloid Leukaemia	Vysis D7S486/CEP 7 FISH Probe Kit	Abbott Molecular Inc.	K131508
	Vysis EGR1 FISH Probe Kit	Abbott Molecular Inc.	K123951,K091960
	LeukoStrat CDx FLT3 Mutation Assay	INVIVOSCRIBE TECHNOLOGIES, INC	P160040
	Abbott RealTime IDH2	ABBOTT MOLECULAR, INC.	P170005
Acute Myeloid Leukaemia or Myelodysplastic Syndrome	VYSIS EGR1 FISH PROBE KIT - SC (SPECIMEN CHARACTERIZATION	ABBOTT MOLECULAR, INC.	DEN130010
Aggressive Systemic Mastocytosis	KIT D816V ASSAY	ARUP LABARATORIES	H140006
Autosomal Recessive Carrier Screening	23ANDME PERSONAL GENOME SERVICE	23andMe	DEN140044
B-cell chronic lymphocytic leukaemia	VYSIS CLL FISH PROBE KIT	ABBOTT MOLECULAR, INC	K100015
	VYSIS CLL FISH PROBE KIT	ABBOTT MOLECULAR, INC	P150041
	CEP 12 SpectrumOrange Direct Labeled Chromosome Enumeration DNA Probe	Vysis	K962873
Bladder Cancer	Vysis UroVysion Bladder Cancer Recurrence Kit	Vysis	K033982, K013785, K011031
Breast Cancer	Prosigna Breast Cancer Prognostic Gene Signature Assay	Nanostring Technologies	K130010
	MammaPrint	Agendia BV	K101454, K081092, K080252, K070675, K062694
	INFORM HER2 Dual ISH DNA Probe Cocktail	Ventana Medical Systems, Inc.	P100027
	HER2 CISH pharmDxTM Kit	Dako Denmark A/S	P100024
	GeneSearch Breast Lymph Node (BLN) Test Kit	Veridex, LLC.	P060017 S001-S004
	Dako TOP2A FISH PharmDx Kit	Dako Denmark A/S	P050045 S001-S004
	HER2 IQFISH PHARMDX	DAKO DENMARK A/S	P040005
	INSITE HER-2/NEU KIT	BIOGENEX LABORATORIES, INC.	P040030
	SPOT-LIGHT HER2 CISH KIT	INVITROGEN CORPORATION	P050040

	INFORM HER-2/NEU	VENTANA MEDICAL SYSTEMS, INC.	P940004
	DAKO HERCEPTEST	DAKO A/S	P980018
	PATH VYSION HER-2 DNA PROBE KIT	ABBOTT MOLECULAR, INC.	P980024
	DakoCytomation Her2 FISH pharmDx™ Kit	DakoCytomation Denmark A/S	P040005
Colorectal Cancer	*Cologuard*	Exact Sciences Corporation	P130017
	Therascreen KRAS RGQ PCR Kit	QIAGEN MANCHESTER LTD	P110027/P110030
	Epi ProColon®	Epigenomics AG	P130001
	Cobas KRAS MUTATION TEST	Roche Molecular Systems, Inc.	P140023
	Praxis Extended RAS Panel	Illumina, Inc.	P160038
Chronic Myeloid Leukaemia	Quantidex qPCR BCR-ABL IS Kit	ASURAGEN, INC.	DEN160003
Coagulation, late-onset alheimer's disease, parkinson's disease, celiac disease, alpha-1 antitrypsin deficiency, early-onset primary dystonia, factor ix deficiency, gaucher disease type 1, glucose-6-phophate dehydrogenase deficiency, herditary hemochromatosis, hereditary thrombophilia	Personal genome service (pgs) genetic health risk test for hereditary thrombophilia	23andME	DEN160026
Cystic Fibrosis	Illumina MiSeqDx Cystic Fibrosis Clinical Sequencing Assay	Illumina, Inc	K132750
	Illumina MiSeqDx Cystic Fibrosis 139-Variant Assay	Illumina, Inc	K124006
	xTAG Cystic Fibrosis 60 Kit v2, xTAG Data Analysis Software (TDAS) CFTR	Luminex Molecular Diagnostics, Inc.	K163336
	xTAG Cystic Fibrosis 39 Kit v2	Luminex Molecular Diagnostics, Inc.	K163347
	eSensor CF Genotyping Test	Osmetech Molecular Diagnostics	K090901
	xTAG Cystic Fibrosis 60 Kit v2	Luminex Molecular Diagnostics Inc.	K083845
	xTAG Cystic Fibrosis 39 Kit v2	Luminex Molecular Diagnostics Inc.	K083846
	Verigene CFTR and Verigene CFTR PolyT Nucleic Acid Tests	Nanosphere, Inc	K083294
	InPlex CF Molecular Test	Third Wave Technology, Inc.	K063787
	Cystic Fibrosis Genotyping Assay	Celera Diagnostics	K062028
	Tag-It Cystic Fibrosis Kit	Tm Bioscience Corporation	K060627, K043011
	eSensor Cystic Fibrosis Carrier Detection System	Clinical Micro Sensors, Inc.	K060543, K051435
Coagulation Factors	Invader Factor V	Hologic, Inc.	K100980
	Invader Factor II	Hologic, Inc.	K100943

	Illumina VeraCode Genotyping Test for Factor V and Factor II	Illumina, Inc.	K093129
	eSensor Thrombophilia Risk Test, eSensor FII-FV Genotyping Test, eSensor FII Genotyping Test, eSensor FV Genotyping Test, eSensor MTHFR Genotyping Test	Osmetech Molecular Diagnostics	K093974
	Xpert HemosIL FII & FV	Cepheid	K082118
	Verigene F5 Nucleic Acid Test Verigene F2 Nucleic Acid Test Verigene MTHFR Nucleic Acid Test	Nanosphere, Inc.	K070597
	INFINITI System	Autogenomics, Inc.	K060564
	Factor II (Prothrombin) G20210A Kit	Roche Diagnostics Corporation	K033612
	Factor V leiden Kit	Roche Diagnostics Corporation	K033607
	Invader MTHFR 677	Hologic, Inc.	K100987
	Invader MTHFR 1298	Hologic, Inc.	K100496
Chromosome abnormalities	Affymetrix CytoScan Dx Assay	Affymetrix, Inc.	K130313
	AneuVysion	Vysis	K010288, K972200
	CYTOSCAN(R) DX	Affymetrix, Inc.	DEN130018
	GenetiSure Dx Postnatal Assay	Agilent Technologies, Inc.	K163367
	CEP 8 Spectrumorange DNA Probe Kit	Vysis	K953591
	CEP X SpectrumOrange/ Y SpectrumGreen DNA Probe Kit	Vysis	K954214
Drug metabolizing enzymes	xTAG CYP2D6 Kit v3	Luminex Molecular Diagnostics, Inc.	K130189, K093420
	xTAG CYP2D6 Kit v3	Luminex Molecular Diagnostics, Inc.	K130189, K131565
	Spartan RX CYP2C19 Test System	Spartan Bioscience, Inc.	K123891
	Verigene CYP2C 19 Nucleic Acid Test	Nanosphere, Inc.	K120466
	INFINITI CYP2C19 Assay	AutoGenomics, Inc.	K101683
	Invader UGT1A1 Molecular Assay	Third Wave Technologies Inc.	K051824
	Roche AmpliChip CYP450 microarray	Roche Molecular Systems, Inc.	K043576, K042259
	eSensor Warfarin Sensitivity Saliva Test	GenMark Diagnostics	K110786
	eQ-PCR LC Warfarin Genotyping kit	TrimGen Corporation	K073071
	eSensor Warfarin Sensitivity Test and XT-8 Instrument	Osmetech Molecular Diagnostics	K073720
	Gentris Rapid Genotyping	ParagonDx, LLC	K071867

	Assay - CYP2C9 & VKORCI		
	INFINITI 2C9 & VKORC1 Multiplex Assay for Warfarin	AutoGenomics, Inc.	K073014
	Verigene Warfarin Metabolism Nucleic Acid Test and Verigene System	Nanosphere, Inc.	K070804
Heart Transplant	AlloMap Molecular Expression Testing	xDx	K073482
Hereditary thrombophilia	Impact dx factor v Leiden and factor ii genotyping test	SEQUENOM, INC./AGENA Bioscience	K132978
Melanoma	Roche cobas DNA Sample Preparation Kit, COBAS 4800 BRAF V600 MUTATION TEST	Roche Molecular Systems, Inc.	P110020
	THXID-BRAF KIT	BioMerieux, Inc.	P120014
Myelodysplastic syndrome/myeloproliferative disease	Fluorescence in situ hybridization, platelet-derived growth factor receptor, beta polypeptide (pdgfrb), rearrangement	ARUP LABARATORIES	H140005
Non-Small Cell Lung Cancer	VYSIS ALK BREAK APART FISH PROBE KIT	ABBOTT MOLECULAR, INC.	P110012
	Cobas EGFR MUTATION TEST v2	QIAGEN MANCHESTER LTD	P120019
	THERASCREEN EGFR RGQ PCR KIT	ROCHE	P120022
	Cobas EGFR MUTATION TEST v2	Roche Molecular Systems, Inc.	P150044
	Cobas EGFR MUTATION TEST v2	Roche Molecular Systems, Inc.	P150047
	Oncomine Dx Target Test	LIFE TECHNOLOGIES CORPORATION	P160045
Ovarian Cancer	BRACAnalysis CDx	Myriad Genetic Laboratories, Inc.	P140020
	FoundationFocus CDxBRCA	FOUNDATION MEDICINE, INC	P160018
Platforms, Imaging Systems, and Reagents	ILLUMINA MISEQDX PLATFORM	Illumina, Inc	DEN130011
	MISEQDX UNIVERSAL KIT 1.0	Illumina, Inc.	DEN130042
	DUET SYSTEM	BIOVIEW LTD.	K130775
Polycythaemia Vera	Ipsogen JAK2 RGQ PCR Kit	QIAGEN INC	DEN160028
Prostate Cancer	NADiA ProsVue	Illumina, Inc.	K101185
	MISEQDX UNIVERSAL KIT 1.0	Gen-Probe, Inc.	P100033
Severe Combined Immunodeficiency Disorder (SCID)	PerkinElmer Enlite TREC Test System	Wallac OY	DEN140010
Tissue of Origin	Pathwork Tissue of Origin Test Kit – FFPE	Pathwork Diagnostics Inc.	K120489, K092967
	Pathwork Tissue of Origin Test	Pathwork Diagnostics Inc.	K080896

Source: https://www.fda.gov/MedicalDevices/ucm330711.htm#human

5.4. Data sources

There are different public data sources (databases) where genomics data can be uploaded and freely accessed, described in section 5.4.1. Furthermore, a great number of genomics initiatives have been initiated in the past decade, described in section 5.4.2.

5.4.1. Public data sources

A large number of public genomics databases have been established. With regard to these databases a differentiation can be made between primary databases, which are databases that contain experimentally derived data such as nucleotide sequence data (section 5.4.1.1). Experimental results are submitted directly into the database by researchers, and the data are essentially archival in nature. Once given a database-accession number, the data in primary databases are never changed: they form part of the scientific record.

On the other hand, there are secondary databases, which comprise data derived from the results of analysing primary data (section 5.4.1.2). Secondary databases often draw upon information from numerous sources, including other databases (primary and secondary) and the scientific literature. They are highly curated, often using a complex combination of computational algorithms and manual analysis and interpretation to derive new knowledge from the public record of science. Of note, many data resources have both primary and secondary characteristics. Data sources that link gene function to disease are for instance OMIM and NCBI, whereas IHEC Data Portal links epigenomes to disease. The data are, however, usually not linked to treatment outcome.

It requires specialised knowledge and skills to know when and how to use which data resource. Many of the websites provide more background information, or courses, e.g. the European Bioinformatics Institute (EMBL-EBI), https://www.ebi.ac.uk/training/online/course/bioinformatics-terrified.

Several examples of public genomics databases are discussed in the next sections.

5.4.1.1. Genomics data – primary databases

GenBank at the National Centre of Biotechnology Information (NCBI), the DNA DataBank of Japan (DDBJ), and the European Nucleotide Archive (ENA) are part of the International Nucleotide Sequence Database Collaboration (Table 5). These three organisations exchange data on a daily basis. Researchers, or institutions, can submit their own data in these primary databases. Most journals require DNA and amino acid sequences that are cited in articles to be submitted to a public sequence repository (DDBJ/ENA/Genbank - INSDC) as part of the publication process. No restrictions apply on the use or distribution of the INSDC data. Further, it is indicated that when submitting human sequences, data that could reveal the personal identity of the source should not be included.

Table 5. Publicly available primary genomics databases

Name	Organisation	Started	Website
GenBank	The National Center for Biotechnology Information (NCBI)	Established in 1988	www.ncbi.nlm.nih.gov/genbank/
DNA DataBank of Japan	National Institute of Genetics (NIG) in Mishima, Japan	Established in 1986	www.ddbj.nig.ac.jp
European Nucleotide Archive	The European Bioinformatics Institute	Established in 1982	www.ebi.ac.uk/ena

Different submission types are accepted (https://www.ncbi.nlm.nih.gov/genbank/submit_types/). Data can also be submitted on input information (sample, experimental setup, machine configuration), output machine data (sequence traces, reads and quality scores) and interpreted information (assembly, mapping, functional annotation).

It has to be pointed out that some of the patient's sequencing information is only available through controlled access for privacy protection reasons and held in sub-partition of ENA, i.e. the European Genome-phenome Archive (EGA). The decisions of who will be granted access to data resides with the submitter nominated Data Access Committee. In the USA, an equivalent exists called DbGap. Information about submitted studies, summary level data, and documents related to studies can be accessed freely on the DbGaP website (http://www.ncbi.nlm.nih.gov/gap). Individual-level data can be accessed only after a Controlled Access application has been approved, stating research objectives and demonstrating the ability to adequately protect the data (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login).

From June 2017 GenBank contained 201,663,568 sequences and Whole Genome Shotgun (WGS) 487,891,767 sequences. Figure 4 shows how the amount of available genomics data grew exponentially over time.

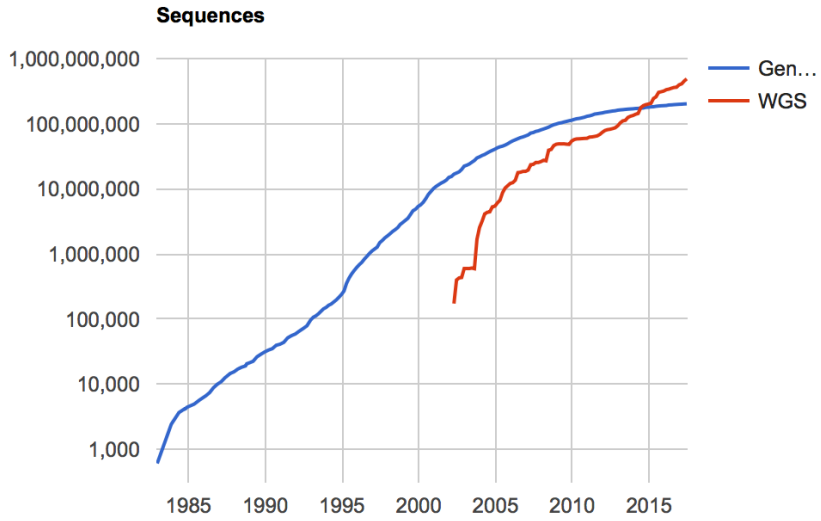

Figure 4. Number of sequences in GenBank and Whole Genome Shotgun (WGS) over time. Source: https://www.ncbi.nlm.nih.gov/genbank/statistics/.

5.4.1.2. Genomics data – secondary databases

Secondary databases comprise data derived from analysing results of primary data. In Table 6 several examples of secondary databases are summarised, as well as different resources for interpreting genomics data, e.g. in the context of pharmacogenomics.

Table 6. Publicly available secondary genomics databases and resources for interpreting genomics data.

Name	Type	Additional Details	Cohort Size	Cohort Description	Type of Data	Started	Website
The National Center for Biotechnology Information (NCBI)	Governmental institution	Part of the National Institutes of Health (NIH). The NCBI contain different databases, such as "Genome" and "GenBank".				Established in 1988.	On NCBI a list of resources with description is provided (https://www.ncbi.nlm.nih.gov/guide/all/
The International Genome Sample Resource (IGSR)	European Bioinformatics Institute	Data from the 1000 genomes project, such as variant calls (VCF format), alignments (BAM or CRAM format) and raw sequence files.	Pilot: 179 individuals; Phase 1 1092 individuals; Phase 3 2504 individuals	26 populations. The IGSR samples do not reflect all populations.	Low coverage and exome sequence data are present for all of these individuals, 24 individuals were also sequenced to high coverage for validation purposes.	1000 Genomes Project ran from 2008-2015	http://www.internationalgenome.org/data/
Online Mendelian Inheritance in Man (OMIM)	McKusick-Nathans Institute of Genetic Medicine, John Hopkins University School	A database of human genes, genetic diseases and disorders. It is updated daily, and the entries contain copious links to other genetics resources.			The full-text, referenced overviews in OMIM contain information on all known mendelian disorders and over 15,000 genes. OMIM focuses on the relationship between phenotype and genotype.	Established in 1966. Online version from 1985.	https://www.omim.org
The Pharmacogenomics Knowledgebase (PharmGKB)		Comprehensive resource that curates knowledge about the impact of genetic variation on drug response for clinicians and researchers					https://www.pharmgkb.org/

Name	Type	Additional Details	Cohort Size	Cohort Description	Type of Data	Started	Website
The UCSC Genome Browser	University of California, Santa Cruz (UCSC)	Online genome browser. Web-based tool for quickly displaying a requested portion of a genome at any scale, accompanied by a series of aligned annotation "tracks". The annotations can display gene predictions, mRNA and expressed sequence tag alignments, simple nucleotide polymorphisms, expression and regulatory data, phenotype and variation data, and pairwise and multiple-species comparative genomics data.		Genomes of 46 species.	Links to other databases, such as dbSNP from NCBI.	Established in 2000.	https://genome.ucsc.edu https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2834533/
Ensembl	The European Bioinformatics Institute	A genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation.			Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data.	Started in 1999.	http://www.ensembl.org/index.html
Expression Atlas	The European Bioinformatics Institute	Open science resource to find information about gene and protein expression across species and biological conditions such as different tissues, cell types, developmental stages and diseases among others.		Gene expression results on more than 3,000 experiments from 40 different organisms	Microarray and RNA-seq data.		https://www.ebi.ac.uk/gxa/home

Name	Type	Additional Details	Cohort Size	Cohort Description	Type of Data	Started	Website
The Pharmacogene Variation Consortium (PharmVar)	Collaboration between Children's Mercy, PharmGkb, Pharmacogenomics Research Network and Clinical Pharmacogenetics Implementation Consortium	Central repository for pharmacogene (PGx) variation that focuses on haplotype structure and allelic variation. The information in this resource facilitates the interpretation of pharmacogenetic test results to guide precision medicine.				The Human Cytochrome P450 (CYP) Allele Nomenclature Database formerly hosted at http://www.cypalleles.ki.se/ has transitioned to Children's Mercy in Kansas City, USA and will be integrated into the new PharmVar database that will launch in early 2018.	https://www.pharmvar.org/

5.4.1.3. Epigenetics data

Several national and international consortia have been organised to identify epigenomic alterations across primary human tissues and cell lines. Some examples include:

- CEEHRC Platform: A reference epigenome project for human cells.
- Classification of Human Transcription Factors: The mother list of transcription factors and their binding sites.
- DeepBlue: Store and work with genomic and epigenomic data from a number of international consortiums.
- Ensembl, featuring ENCODE: Encyclopedia of DNA elements.
- EpiDenovo, http://www.epidenovo.biols.ac.cn/, a database that provides the associations between embryonic epigenomes and *de novo* mutations in developmental disorders, including several neuropsychiatric disorders and congenital heart disease (Mao et al., 2018).
- GenExp: A web-based visualisation tool to interactively explore a genomic database.
- Human Epigenome Project (HEP).
- International Cancer Genome Consortium (ICGC).
- International Human Epigenome Consortium (IHEC). The IHEC Data Portal brings forth reference epigenomes relevant to health and disease. There is the possibility to view, search, and download all the data. Their goal is to map 1000 epigenomes.
- NIH ROADMAP Epigenomics: The NIH Roadmap Epigenomics Mapping Consortium offers maps of histone modifications, chromatin accessibility, DNA methylation, and mRNA expression across 100s of human cell types and tissues.
- Stand Up to Cancer (SU2C).
- Structural Genomics Consortium (SGC).
- The Cancer Genome Atlas (TCGA).
- The Epigenome Atlas: Human reference epigenomes.
- Therapeutically Applicable Research to Generate Effective Treatments (TARGET).

5.4.1.4. Microbiome data

Also, databases designed specifically for bringing together microbiomics data are available. One example is the NIH Human Microbiome Project (http://www.hmpdacc.org). The Human Microbiome Project Data Analysis and Coordinating Center (DACC) Portal provides access to all publicly available Human Microbiome Project (HMP) data sets, generated from healthy human subjects and demonstration project subjects - http://hmpdacc.org/resources/data_browser.php. On the following website examples of studies that have used the HMP data can be found: https://commonfund.nih.gov/hmp/databases.

5.4.2. Genomics initiatives

Many genomics initiatives have been initiated in the past decade. These initiatives were started by private/public companies, non-profit organisations (such as international consortia of academic researchers), government, and pharmaceutical companies. The focus of most of these projects is on determining associations between genomic traits and development of disease, including e.g. cancer, rare diseases, and neurological diseases. The types of data that are generated within these projects vary, and include whole-genome or exome sequences, RNA sequences (transcriptomics), gene panels and single variants. Most of these initiatives focus on genomics data, however epigenetics is expected

to become more and more important, and in the future also more initiatives on epigenetics are expected to start combining genomics/transcriptomics with epigenetics data. Data sharing is a hallmark of many of these initiatives, but the extent to which data are shared (e.g. fully open source vs. selective sharing) varies. A selection of genomics initiatives is summarised in Appendix 2.

Several examples of different types of genomics initiatives are described in more detail below:

The Genomics Evidence Neoplasia Information Exchange project

The Genomics Evidence Neoplasia Information Exchange project, supported by the American Association for Cancer Research (AACR), is a transatlantic data-sharing cooperative. During the project's first year, clinical-grade genomic data and baseline clinical information from about 19,000 patients at eight major cancer centres in the United States, Canada, and Europe were harmonised using a common data dictionary for recording tumour subtypes, and the data were then made publicly available. Longitudinal data from subgroups of these patients are being collected in order to establish genotype-specific disease registries for use in clinical care. Among the aims of the project are: validating biomarkers, drug repositioning/repurposing, adding new mutations to existing drug labels, and identifying new drug targets. In this project the AACR is working closely with the FDA with the intention of building a regulatory-grade database such that the data could be accepted as the necessary evidence to gain regulatory approval.

RD-Connect

RD-Connect is an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. To help researchers study rare diseases, RD-Connect links different data types - omics (e.g. genomics), clinical information, patient registries and biobanks - into a common resource. It is one of the examples of genomics initiatives that link genomics data to relevant clinical information. RD-Connect enable scientists and clinicians around the world to analyse genomics data and share them with other researchers. By making data accessible beyond the usual institutional and national boundaries, the aim of RD-Connect is to speed up research, diagnosis and therapy development to improve the lives of patients with rare diseases. RD-Connect contains exomes and genomes of patients with rare neuromuscular, neurodegenerative and kidney diseases thanks to the collaborations with research projects NeurOmics and EURenOmics. The number of other disease areas is also increasing. In total the number of samples is 2478 (25 October 2017).

AstraZeneca's integrated genomics initiative

AstraZeneca and its global biologics research and development arm, MedImmune, have started an integrated genomics initiative with the aim of transforming their drug discovery and development process across the entire research and development pipeline. This is an example of an initiative driven by the pharmaceutical industry. The initiative includes collaborations with Human Longevity, Inc., US; the Wellcome Trust Sanger Institute, UK, and The Institute for Molecular Medicine, Finland. AstraZeneca will also establish an in-house Centre for Genomics Research which will develop a database comprising genome sequences from samples donated by patients in its clinical trials together with associated clinical and drug response data.

AstraZeneca will generate genome sequences of up to 2 million patients, including over 500,000 patients from clinical trials. It is believed that embedding genomics across its research and development platforms will deliver novel insights into the biology of diseases, enabling the identification of new targets for medicines, supporting selection of patients for clinical trials, and allowing patients to be matched with treatments more likely to benefit them.

U.K. Biobank collaboration with Regeneron Genetics Center / FinnGen study

In March 2017 collaboration was started with U.K. Biobank and the pharmaceutical companies GSK and Regeneron to enable sequencing of the first 50,000 samples from volunteer participants in the U.K. Biobank, to be completed before the end of 2017. The goal of the initiative is to generate whole exome sequencing data and extensive phenotype information from 500,000 volunteer participants (http://www.ukbiobank.ac.uk/2017/03/gsk-regeneron-initiative-to-develop-better-treatments-more-quickly/). Sequencing of the full 500,000 samples was expected to take three to five years. The new genetic data sequenced at the Regeneron Genetics Center will be returned to the U.K. Biobank and made available to approved researchers following an exclusive period for GSK and Regeneron, in this case, 9 months. This period is in line with the exclusive period granted to other researchers conducting comparable analyses.

Subsequently, in January 2018, it was announced that a pre-competitive consortium was formed by Regeneron with AbbVie, Alnylam Pharmaceuticals, AstraZeneca, Biogen, and Pfizer to sequence the rest of the 450,000 samples. Further, the timeline was shortened, and it is now expected that at the end of 2019 sequencing of the 500,000 samples is completed.

More of these initiatives have been started, such as the FinnGen study, launched in December 2017, which will analyse 500,000 unique blood samples collected by a network of Finnish biobanks (https://www.fimm.fi/en/press-release/1513666806). The aim is to match genomic information with digital health care data from national registries. This study is expected to continue for six years. Like in the U.K. Biobank collaboration, pharmaceutical companies are involved in the FinnGen study: Abbvie, AstraZeneca, Biogen, Celgene, Genentech, MSD and Pfizer.

BLUEPRINT Epigenome

This EU-funded research project that involved 42 European universities, research institutes and industry aimed to further understand how genes are activated or repressed in both healthy and diseased human cells. The project ran from 2011 to 2016. BLUEPRINT focussed on haematopoietic cells from healthy individuals and on their malignant leukaemic counterparts. It generated at least 100 reference epigenomes. The project contributed to the overall objective of the International Human Epigenome Consortium (IHEC). Reference epigenomes were generated using state-of-the-art technologies from highly purified cells for a comprehensive set of epigenetic marks in accordance with quality standards set by IHEC.

Commercial companies that offer ancestry analyses

There are several commercial companies that offer ancestry analyses. The data generated within these projects can be used by individuals to find out more about their ancestry, or to identify relatives. Examples of companies that offer ancestry analyses include:

- **23andme** (https://www.23andme.com/) is a company that provides a DNA analysis service. The testing is performed in a certified laboratory in the United States. On it's website it is indicated that over 1,000,000 people worldwide are in their database. One of the options is to identify relatives. Further, people are offered to contribute with their data to 23andMe Research. Their data can subsequently be used by the research team of 23andMe, or by one of their collaborators at research universities or pharmaceutical companies. For instance, 23andMe has partnered with Genentech on Parkinson's and Pfizer on inflammatory bowel disease. In addition to ancestry analyses, individuals can also request a genetic test to determine their genetic predisposition to certain medical diseases or conditions (see also section 5.3).

- **Family Tree DNA** (https://www.familytreedna.com) is a genetic testing company based in Houston, United States. Family Tree DNA offers analysis of autosomal DNA, Y-DNA, and mitochondrial DNA to individuals for genealogical purpose. The goal of the company is to make it possible for their clients to find their family, and lineage through time.
- **DNA Ancestry** (https://www.ancestry.com) is a genetic genealogy testing partnership between Family Tree DNA and Eastern Biotech & Life Sciences. Their website indicates that over 3,000,000 people have their data stored in their database.

These ancestry databases can have an impact on the privacy of individuals. As an example, in 2017 in the Netherlands a woman tracked down her donor father by using these commercial data sources (https://nltimes.nl/2017/05/30/sperm-donor-anonymity-disappearing-commercial-dna-databases-grow). This example illustrates that potentially privacy issues could arise as a result of making genomics data publicly available.

5.4.3. Conclusions on Data Sources

A number of publicly available data sources containing genomics data is available, and there is a large number of genomics initiatives ongoing or being initiated. Based on the mapping of these data sources, the following conclusions can be drawn:

1. **Most publicly available genomics data sources are derived from investigator-initiated (non-industry-driven) initiatives.** Most pharmaceutical industry-driven genomics data, which is often genomics data linked to clinical outcomes (e.g. response to treatment), is not publicly available. These data would be of interest to regulators as it could be used for regulatory purposes.
2. **Most publicly available databases contain only genomics data, without phenotypic/clinical outcome data linked to the genomics data.** Some data sources do contain phenotypic data, mainly data on presence or absence of genetic/hereditary diseases.
3. **In the ongoing genomics initiatives, genomics data are often coupled to phenotypic data on disease.** This coupling will yield a large amount of information in the near future on the genetic origins of disease.
4. **Clinical outcome data, e.g. response to drug therapy, is not often coupled to genomics data in the ongoing genomics initiatives nor in the public databases.** It is this coupling of genomics data and clinical outcome data (i.e. data on efficacy or safety of treatments), which would be of most interest to regulatory agencies.

5.5. Volume

5.5.1. Size of the data source

The volume of data contained in the available public databases has grown exponentially since the start of the first publicly available databases such as GenBank in the 1980's (Lathe et al., 2008). The huge volume of the raw sequence data in these repositories has led to attempts to reorganise the information into smaller, specialised databases. Such databases include various genome browsers, model organism databases, molecule- or process-specific databases, and others. There are more than 3,000 distinct genomic resources, tools, and databases publicly available on the internet of which a selection is mentioned in section 5.4.

5.5.2. Structure (terminology, structured vs. unstructured data)

Genomics data are structured, and although the exact data format can vary, the basic structure of the data is similar for different genomics data formats.

Sequence data

The basis of genomics data is the sequence data. This can refer to the nucleotide sequence of a chromosome, a contig (a set of overlapping DNA segments that together represent a consensus region of DNA), a transcript, or a set of these. Sequence data are stored in different formats, including e.g. the FASTQ, SAM or BAM format. Reference sequences of genes can be found in public databases, e.g. the NCBI database. In addition to sequence data, a genomics dataset can contain additional data.

Annotations

Annotations are descriptions of features – e.g. genes, transcripts, SNPs, start codons – that appear in genomes or transcripts. Annotations typically include coordinates (chromosome name, chromosome positions, and a chromosome strand), as well as properties (gene name, function, GO terms, etc.) of a given feature. Annotations are maintained by the same public databases that maintain sequence information, because the coordinates in each annotation are specific to the genome build upon which it is based. In other words, annotations and sequences must be matched. Thus, the interpretation of the annotations is dependent on the data source that the genomics data originates from.

Quantitative data

The quantitative data refers to any kind of numerical value associated with a chromosomal position. For example, the strength of transcription factor binding to a chromosomal position in a ChIP-seq dataset. Because quantitative data associates values with chromosomal coordinates, it can be considered an annotation of sorts. It is therefore important to make sure that the coordinates in a particular data file match the genome build used by the annotation and/or read alignments.

Read alignments

Read alignments refer to a record matching a short sequence of DNA to a region of identical or similar sequence in a genome. In a high-throughput sequencing experiment, alignment of short reads identifies the genomic coordinates from which each read is derived. Read alignments can be produced by running sequencing data through alignment programs, such as Bowtie, Tophat, or BWA.

Read alignments can be converted to quantitative data by applying a mapping rule, that uses various properties of the read to assign genomic position(s) at which the read should be counted. For example, one could map reads to their 5’ ends, or to sites within the read where nucleotides mismatch the reference genome.

Table 7 summarises commonly used file formats, of which some are indexed, and others are not. Indexed files are memory-efficient, because computer programs do not need to read the entire file to find the data of interest; instead, they read the index and just fetch the desired portion of the data.

Table 7. Commonly used file formats for different types of genomics data

Data type	Unindexed formats	Indexed formats
Sequence	FASTA	2bit
Annotations	BED, GTF2, GFF3, PSL	BigBed
Quantitative data	bedGraph, wiggle	BigWig
Read alignments	bowtie, SAM, PSL	BAM

Compatibility of different data formats

To ensure that data from different sequencing providers is comparable, the raw data need to be available in FASTQ (or BAM) format and the data have to be processed through the same standard pipeline. This ensures that data from different sequencing providers are comparable.

Nomenclature

The description/nomenclature of genetic variants can lead to ambiguity, as nomenclature might have changed over time in peer-reviewed publications. The Human Genome Variation Society (HGVS) has formulated Guidelines & Recommendations on nomenclature of gene variations and guidelines on variation databases. More information can be found on the website: http://www.hgvs.org/content/guidelines.

5.6. Veracity

5.6.1. Data provenance

Publicly available data come mostly from research initiatives. Peer-reviewed scientific journals usually require that before publication of the study the genomics data are added to publicly available databases. Once added, the information will remain publicly available, and is part of the scientific record with a unique identifying number. In addition, several initiatives that include biobanks are expected to gather a large amount of genomics data, which can be made available to approved researchers. Data derived from routine diagnostic procedures will likely not be made available for big data purposes.

As described in previous sections, most genomics data derived from industry-sponsored clinical trials, which is often genomics data linked to clinical outcomes, are currently not publicly available. These data would be of interest to regulators, as it could be used for regulatory purposes, e.g. identifying subgroups of patients who would benefit more (or less) from treatments.

5.6.2. Data Quality

Quality issues related to genomics data can broadly be divided into two categories – sample quality and data quality.

Sample quality

Several aspects play a role in determining the sample quality of genomics data, such as sample collection, handling, storage and processing. Issues related to sample quality are broadly outlined in the recently adopted ICH guideline "ICH E18 Guideline on genomic sampling and management of genomic data" (EMA/CHMP/ICH/11623/2016, 06/10/2017), which provides guidance amongst others on timing of collection, preservation conditions, sample stability and degradation, specimen volume and composition, and parameters influencing genomic sample quality and quantity.

Data quality

Other features that influence data quality are: sequencing and alignment steps, appropriate thresholds and coverage that may vary between different systems. Another aspect is storage of data, which is challenging, because of the enormous amount of raw data that needs to be stored, and the accompanying costs for this.

Guidance has been provided on these aspects by the EMA as well as by the European Society of Human Genetics:

- Reflection paper on pharmacogenomics samples, testing and data handling (EMEA/CHMP/PGxWP/201914/2006).
- The recommendations of the European Society of Human Genetics on Whole Genome Sequencing in the European Journal of Human Genetics (2013) 21, 580–584 (Van El et al., 2013).
- The Recommendations of the ESHG on Genetic testing and common disorders in a public health framework' have been published in the European Journal of Human Genetics 2011;19:377-81 (Van El and Cornel, 2011).

How could data quality be improved?

Quality of data can be improved by improving standardisation (making use of standard operating procedures), by requiring attached meta-data (i.e. descriptive information about the overall study, individual samples, all protocols, and references to processed and raw data file names), by providing scripts/codes from bioinformatic analyses which allow to repeat the process in a same way, by having a certification of the instruments used and considering the importance of a minimal data standard. Harmonisation efforts are in place, e.g. by performing a ringtest, an interlaboratory external quality assurance where identical samples usually of a reference institute are sending to different laboratories for analysis. Also "EuroGentest", a project funded by the European Commission aims to harmonise the process of genetic testing, from sampling to counselling, across Europe (http://www.eurogentest.org/index.php?id=160). They also organise interactive workshops and e-courses to aid laboratories in the process of implementing and developing a quality system, in improving existing quality schemes and working towards accreditation (ISO 15189).

Regarding standardisation, this would not only apply to the genomics part of the data, but also to the clinical outcome data, which is linked to the genomics data, so that genomic information can be linked to clinical data across data sources.

Addition of meta-data

In the previous paragraph the addition of meta-data is described in the context of clinical trials, e.g. including descriptive information of the overall study, protocols, etc.. The addition of meta-data related to how the sequence data were obtained is at least as important. With these meta-data the quality of the data can be assessed, and it can consequently be determined for what type of analyses the sequencing data can be used, and/or when to be careful with using the sequencing data. Important quality parameters to have in the meta-data that accompanies the sequence data are:

1) Average coverage

Coverage (or depth) in DNA sequencing is the number of unique reads that includes a given nucleotide in the reconstructed sequence. As such coverage gives an indication of the accuracy of the generated data. For instance, the data of whole genome sequencing of the 1000 genomes project has an average coverage of 4x. This is a low coverage compared with current standards. The choice was made in the 1000 genomics project to have this low coverage, as at that time whole genome sequencing was much more expensive than today. The average coverage in research settings is now at least 30x, and for a diagnostic setting at least 50x (personal communication Professor Richard Sinke, University Medical Center Groningen, the Netherlands). When using data of the 1000 genomes project, one has therefore to keep in mind that in case a nucleotide variant was not called in the data, it can still be that the nucleotide variant was present, but that it was not covered/sequenced.

For exome data, as an enrichment step is included during the process, the average coverage in diagnostic settings is required to be even higher, 80-100x, whereas in a research setting an average coverage of at least 30x will provide a reasonable level of assurance.

2) Platform

Every sequencing platform makes systematic mistakes in sequencing samples. Further, every different type of sequencer makes use of different chemistries for sequencing, which can result in different systematic mistakes. An example of such a systematic mistake is base calling of stretches of G's.

3) The type of sample that was used (blood, tumour material)

In case DNA is isolated from blood, the quality of the DNA sample is expected to be high. However, in case tumour tissue has been investigated this could have been taken from different types of tissue of differing quality. Not only the method (snap-frozen, fixed, et cetera) plays a role, but also how much necrosis, lymphocyte infiltration there is, and also the size of the biopsy. The risk of facing challenges is higher when for instance a micro-needle-biopsy is taken from a lung tumour in comparison to larger biopsies that are taken when a surgical resection of a larger tumour is performed.

4) How DNA/RNA was isolated

This factor is especially important for RNA isolation, where differences in isolation method can for example result in differences in quantities of RNA. Big data analyses performed by combining RNA sequencing data of different data sources, have shown clear batch differences between data sources for RNA sequencing. The type of isolation is one of the aspects that could have contributed to these differences.

5.6.3. Completeness (opportunity to capture the data)

There are several aspects to completeness of genomics data (sets).

In the case of germ-line genomics data, there is the issue of completeness of the genome in question, i.e. do the data describe a whole genome, a whole exome, or only variants? Furthermore – and this is closely linked to the issue of representativeness as discussed below – do the data describe a whole patient population with all its variants or is it a sample, which is not fully representative for the whole population? Moreover, what is the ethnic background of the population, as differences may exist between ethnicity and/or geographical location?

For somatic genomics data, completeness and representativeness would pertain to the capturing of all mutations that are present and accumulating. In oncology, variability in genomic measurements in tumour tissue can also be the result of the tissue source. For example, there may be variability in the presence of the mutation from one location in the tumour lesion to another, as well as from one tumour lesion to other tumour lesions.

An aspect that is crucial for transcriptomics and epigenetics for the representativeness is the tissue from which the sample is taken. Different genes are differently expressed in different tissues, as the transcriptome captures a snapshot in time of the total transcripts present in a cell/tissue. Also, for epigenetics the tissue location is essential. For example, Byun et al. (2009) analysed DNA methylation across 11 different tissues and across six individuals. The authors concluded that DNA methylation patterns were more consistent between the same tissues from different people than between different tissues from the same individual, though the difference was subtle.

A last aspect of completeness is the possibility of linking the genomics data to phenotypic data and/or clinical outcome data. An example where linkage is made with health care data is the U.K. Biobank, which routinely links to national death and cancer registries and to national hospital data electronic record systems for all its participants since 2010. In addition, U.K. Biobank established linkages to primary care records from over 50% of its participants. They are able to do this with the support of the Royal College of General Practitioners and by working with companies that already provide data management systems to general practice for a wide range of activities.

5.6.4. Representativeness

Once the accuracy of data (data quality) has been verified, the problem becomes to define how representative an individual's genome is for a population (e.g. the target population of the drug), or how representative a set of genomes is for the population. This can be done e.g. by comparing the obtained data with genomics data in the different available databases.

Since many biomarkers have a dynamic temporal aspect, i.e., expression may change over time, biomarker expression at the time of sampling may not always be representative of the expression at the time of interest (e.g. at start of treatment).

Lastly, gene expression and DNA methylation profiles of a number of genes involved in drug metabolism and transport are considerably different between cell lines and primary tissues. Consequently, the extrapolation from findings from *in vitro* studies conducted in cell lines to the *in vivo* situation is limited, since expression and DNA methylation profiles do not necessarily resemble the profiles found *in vivo* (Fisel et al., 2016).

5.6.5. Analytical tests / variability

A limitation of standard NGS is the high frequency with which bases are scored incorrectly due to artefacts introduced during sample preparation and sequencing (Fox et al., 2014). For example, amplification bias during PCR of heterogeneous mixtures can result in skewed populations. Additionally, polymerase mistakes, such as base misincorporations and rearrangements due to template switching, can result in incorrect variant calls. Furthermore, errors arise during cluster amplification, sequencing cycles, and image analysis result in approximately 0.1–1% of bases being called incorrectly (Fox et al., 2014). This has to be taken into account when assessing the data.

To be able to share genomics data, several Application Programming Interfaces (APIs) have been created for secure, modular, interoperable access to genomic data from different applications, platforms and organisations. Such APIs will be also important for connecting Electronic Health Care Records to genomics data. A challenge is the variety of types of data (e.g. gene expression, or sequencing archives) and variety of file formats (e.g. FASTQ, SAM, BAM, VCF). Also, some databases will contain raw data, whereas others will have only more processed data. Three currently available Genomics APIs are 1) Google Genomics (https://cloud.google.com/genomics/reference/rest/), 2) SMART Genomics, and 3) 23andMe (Table 8; see for mini-review Swaminathan et al., Comput Struct Biotechnol J 2015;14:8-15).

Table 8. Comparative view across the three genomics Application Programming Interfaces for a list of features

Features	**Google Genomics**	**SMART Genomics**	**23andMe**
Input data to API	Currently, limited to only sequencing information in the form of reads, variants, and annotation	Some of the Genomics resources are extensions of previously existing Clinical resources	Capability to use both genomics and clinical resources
Location of data	Data need to reside within Google Cloud Storage	Data available within EHR's and other genomics	Data from 23andMe database

		data sources	
API response	Returns only JSON formatted response	Returns either JSON or XML formatted response	Returns JSON formatted response
Ability to import data	Can import both reads and variant data from BAM and VCF files	Create call available for certain resources	API only used for data retrieval through GET calls
Range search for variants in a given individual	Available	Available	Not available
Identify risk for a disease in an individual	Not available	Available	Available
Availability of reference applications using the API	Client libraries and interactive API Explorer through Google Console	Some application like Genomics Advisor, Variant Mapper currently using the API	Not available
Authentication	Uses OAuth2.0	Uses OAuth2.	Uses OAuth2.

Source: Swaminathan et al., Comput Struct Biotechnol J 2015;14:8-15

Also, important to point out here are the FAIR guiding principles for scientific data management (Wilkinson et al., 2016). FAIR stands for Findable, Accessible, Interoperable, Reusable. These principles have been laid down with the intent that these may act as a guideline for those wishing to enhance the reusability of their data holdings (Table 9).

Table 9. The FAIR guiding principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource
To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardised communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorisation procedure, where necessary
A2. metadata are accessible, even when the data are no longer available
To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data
To be Reusable:
R1. (meta)data are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

5.6.6. Validation

Several types of validation are important: analytical method validation, clinical validation of biomarkers and validation of genetic variations.

Analytical method validation

To assess the performance of analytical methods in genomics, similar measures of assay performance are used as in validation of other bioanalytical methods, e.g. sensitivity, specificity, lower limit of detection, accuracy, robustness, and reproducibility. Regarding analytical method validation the following example illustrates the importance of adequate method validation in genomics: the Stanford University reported that a cluster generation called exclusion amplification (ExAmp) resulted in problems, i.e. 5-10% of sequencing reads (or signals) were incorrectly assigned from a given sample to other samples in a multiplexed pool (http://biorxiv.org/content/biorxiv/early/2017/04/09/125724.full.pdf).

Clinical validation of biomarkers

There are many guidelines regarding genomic biomarker validation, including those established by the Standards for Reporting of Diagnostic Accuracy (STARD), Evaluation of Genomic Applications in Practice and Prevention (EGAPP), NCCN Task Force recommendations and the recent Next-generation Sequencing: Standardization of Clinical Testing (Nex-StoCT) and American College of Medical Genetics and Genomics (ACMG) biomarker guidelines.

Regarding clinical validation of biomarkers, guidance is available in the EMA guideline "Guideline on good pharmacogenomic practice" (https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-good-pharmacogenomic-practice-first-version_en.pdf). For a biomarker to be suitable for use in clinical practice, diagnostic accuracy (including e.g. sensitivity, specificity, and positive predictive value) needs to be thoroughly assessed and validated in different clinical studies, i.e. in order to demonstrate clinical validity. To demonstrate that use of a biomarker test in clinical practice actually leads to better patient outcomes, a study aiming to demonstrate clinical utility has to be performed. The clinical trial design needed for clinical validation of a biomarker will depend on the context of use. Historically, randomised controlled trials have been the mainstay to demonstrate clinical utility.

Validation of genetic variations

A challenge of whole genome/exome sequencing is the interpretation of the phenotypic consequences of genetic variants. Whole-exome sequencing and whole-genome sequencing usually generate a long list of mutations, a large number of which have no known significance (variants of unknown significance; VUS). The majority of variants identified represent VUS. The assessment of each VUS for pathogenicity is time-consuming. It was indicated by Bertier et al., 2016 that there is a need to share WGS/WES results and to develop more complete, less biased databases containing fewer false positive and false negative variant-phenotype associations. However, this does not fully resolve the issue of coupling VUS to phenotypic consequences, which will require large datasets of genomic data coupled to information on phenotype.

5.7. Variability

5.7.1. Data heterogeneity

Data heterogeneity between genomics datasets can occur on different levels.

Heterogeneity in type of genomics data

The first aspect to data heterogeneity in genomics is that the type of genomics data and variables included in the dataset can vary. For examples, an epigenetics dataset looks different from a genetics dataset, and a dataset of single variants looks different than a whole-genome sequence dataset. It is possible to combine datasets, even if the type of genomics data is different. For example, there are numerous publications on combining different types of DNA modifications with gene expression, e.g. Kendziorski et al., 2006. However, merging of data from different types of genomics data is not straightforward and often requires complex statistical modelling.

Heterogeneity in data formats and dataset content

Even with the same type of genomics data, the format and content of the dataset can vary depending on data format and the variables included in the dataset, as described in section 5.5.2. This does, however, not prevent the combination of genomics data from different data formats in most cases.

Tissue heterogeneity

Tissue can be truly heterogeneous with regard to genomic material. This can occur due to contamination, e.g. during sampling. In oncology, when biopsies are taken from the primary tumour or metastases, in most cases the biopsy is composed of tumour tissue and healthy tissue. Thus, both the germ line genome and the genome of the tumour is included in the sample. In most cases this does not have to be an issue, since the genomic information in the biopsy can be compared with a sample of healthy tissue, and thereby the tumour genome can be derived. Another important source of tissue heterogeneity in oncology occurs due to heterogeneity among tumour cells themselves. Tumours are known to be heterogeneous, or clonal, meaning that the tumour is composed of a variety of tumour cells with different genomes. The composition of the tumour can even be different for different metastases. For example, it can occur that the primary tumour does not have a certain driver mutation (e.g. RAS), but that some of the metastases do have this mutation. This poses relevant problems, e.g. with regard to accuracy of genomic testing to determine whether a certain treatment should be given.

5.7.2. Data standards

There is a number of initiatives which aim to standardise genomics data from different sources to make genomics data better accessible. In addition, there are initiatives to standardise the clinical/phenotype data that can be associated with genomics data (Table 10). A few examples are described in more detail below.

Genomics data standards initiatives

One example of a data standardisation initiative is the Genomic Standards Consortium (GSC). The GSC was established in September 2005, and is an international initiative, which includes representatives from a range of major sequencing and bioinformatics centres (including NCBI, EMBL, DDBJ, JCVI, JGI, EBI, Sanger, FIG) and research institutions. The goal of the GSC is to promote mechanisms for standardising the description of (meta)genomes, including the exchange and integration of (meta)genomic data. The number and pace of genomic and metagenomic sequencing projects has increased with the use of ultra-high-throughput methods, and therefore data standards are vital to scientific progress and data sharing. A second example is the National Cancer Institute's (NCI's) Genomic Data Commons (GDC), a data-sharing platform that promotes precision medicine in oncology. It is a network supporting the import and standardisation of genomic and clinical data from cancer research programs.

Clinical/phenotype data standards initiatives

- The Human Phenotype Ontology (HPO) aims to provide a standardised vocabulary of phenotypic abnormalities encountered in human disease. Each term in the HPO describes a phenotypic abnormality, such as atrial septal defect. The HPO makes use of the medical literature, Orphanet, DECIPHER, and OMIM. HPO contains approximately 11,000 terms (still growing) and over 115,000 annotations to hereditary diseases. The HPO also provides a large set of HPO annotations to approximately 4000 common diseases.
- Different diagnose codes are used in different countries, such as READ in the United Kingdom, and ICD-9.
- Also, different medical dictionaries are used, e.g. SNOMED and MedDRA. MedDRA or Medical Dictionary for Regulatory Activities is the international medical terminology dictionary (and thesaurus) used by regulatory authorities. SNOMED was started in 1973 by the College of American Pathologists and is now international. Bodenreiser (2009) investigated the feasibility of using SNOMED as an entry point for coding adverse drug reactions and map them automatically to

MedDRA for reporting purposes and interoperability with legacy repositories. Bodenreiser concluded that it was feasible to map SNOMED concepts automatically to MedDRA, however, the quality of the mapping still needed evaluation.

Table 10. Genomic data standards resources and initiatives and clinical data standards initiatives

Genomics data standards initiatives	
Name and website	Description
Genomic Standards Consortium (GSC) http://gensc.org/	GSC is an open membership working body formed in September 2005. The goal of this International community is to promote mechanisms that standardize the description of genomes and the exchange and integration of genomic data.
The Global Alliance for Genomics and Health (GA4GH) http://oicr.on.ca/oicr-programs-and-platforms/global-alliance-genomics-and-health-ga4gh	Data Working Group concentrates on data representation, storage, and analysis, including working with platform development partners and industry leaders to develop standards that will facilitate interoperability.
Human Genome Variation Society (HGVS) http://www.hgvs.org/rec.html	Members of the Society have formulated guidelines and recommendations on a number of topics, particularly for the nomenclature of gene variations and guidelines for variation databases.
Clinical/phenotype data standards initiatives	
Name and website	Description
American College of Medical Genetics (ACMG) http://pathology.ucla.edu/workfiles/News/ACMG-NGS-Guidelines-2013.pdf	ACMG has developed professional standards and guidelines to assist clinical laboratories with the validation of next-generation sequencing methods and platforms, the ongoing monitoring of next-generation sequencing testing to ensure quality results, and the interpretation and reporting of variants found using these technologies.
Health Level 7 International (HL7) http://www.hl7.org/index.cfm?ref=nav	HL7 is dedicated to providing a comprehensive framework and related standards for the exchange, integration, sharing, and retrieval of electronic health information that supports clinical practice and the management, delivery and evaluation of health services.
Healthcare Information Technology Standards Panel (HITSP) http://hitsp.org/	HITSP is a cooperative partnership between the public and private sectors. The Panel was developed for the purpose of harmonising and integrating standards that will meet clinical and business needs for sharing information among organisations and systems.
PhenX https://www.phenx.org/	PhenX provides the scientific community with recommended, standard high-priority measures of phenotypes and exposures for use in genome-wide association studies and more generally, epidemiological and biomedical research.

Data reporting standards

In the past years a number of data reporting standards have been developed. A reporting standard pertains to how a researcher should record the information required to unambiguously communicate experimental designs, treatments and analyses, to contextualise the data generated and underpin the conclusions drawn. Such standards are also known as data content or minimum information standards because they usually have an acronym beginning with "MI" standing for "minimum information" (Table

11). The motivation behind reporting standards is to enable an experiment to be interpreted by other scientists and to be reproducible. When an experiment is submitted to a journal for publication, compliance with a reporting standards is often required. A reporting specification does not normally mandate a particular format in which the data are captured, but simply delineates the data and meta-data that the community considers appropriate to sufficiently describe how a particular investigation was carried out.

Table 11. Selection of existing reporting standards for Omics data

Acronym	**Full name**	**Domain**	**Organisation**
MIAME	Minimum Information about a Microarray Experiment	Transcriptomics	MGED
MIGS-MIMS	Minimum Information about a Genome/Metagenome Sequence	Genomics	GSC
MINIMESS	Minimal Metagenome Sequence Analysis Standard	Metagenomics	GSC
MINSEQE	Minimum Information about a high-throughput Nucleotide Sequencing Experiment	Genomics, Transcriptomics (UHTS)	MGED
MISFISHIE	Minimum Information Specification For In Situ Hybridisation and Immunohistochemistry Experiments	Transcriptomics	MGED

5.7.3. Data processing

As indicated above, some databases will contain raw data, whereas others will have only more processed data. Processing and analyses of whole-genome sequence data is complex. As an example, a typical workflow of WES analysis is illustrated in Figure 5. Whole-exome sequencing analysis consists of the following steps: raw data QC, pre-processing, mapping, post-alignment processing, variant calling, annotation, and prioritisation.

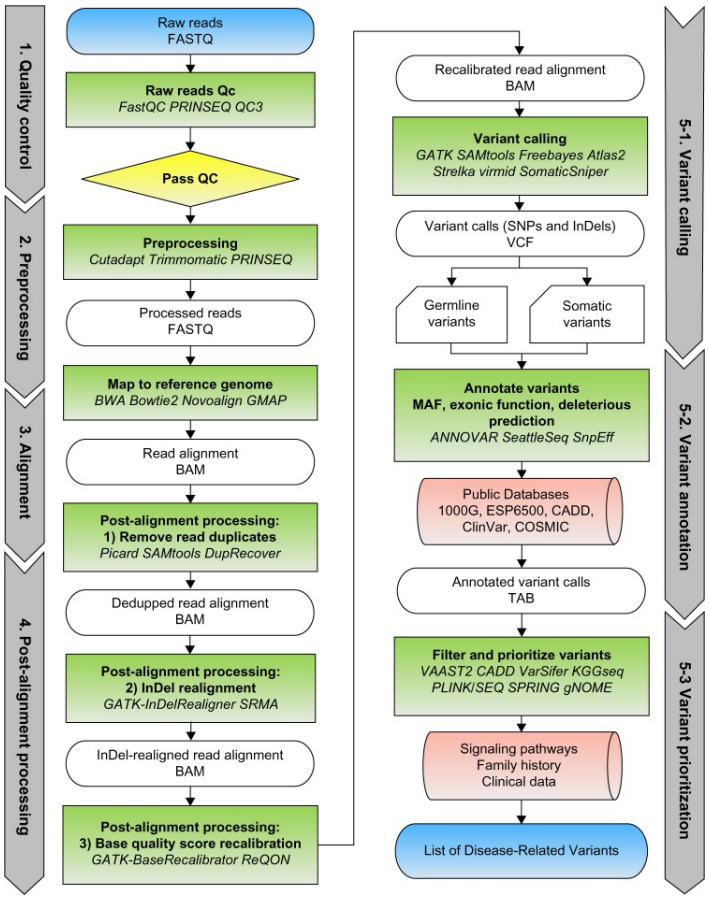

Figure 5. A general framework of WES data analysis.

Five major steps are shown: raw reads QC, pre-processing, alignment, post-processing, and variant analysis (variant calling, annotation, and prioritisation). Notes: FASTQ, BAM, variant call format (VCF), and TAB (tab-delimited) refer to the standard file format of raw data, alignment, variant calls, and annotated variants, respectively. A selection of tools supporting each analysis step is shown in italic. Source: Bao et al. Cancer Inform. 2014; 13(Suppl 2): 67–82.

A high-level overview of the processing steps is provided here, adapted from Bao et al. Cancer Inform. 2014; 13(Suppl 2): 67–82. Refer to the full publication for detailed information.

Raw data

FASTQ and FASTA are standard formats for representing raw sequence data. The FASTA format is a text-based representation of sequences, which begins with the sequence name followed by lines of single-letter coded nucleotides or amino acids.

Pre-processing

Standard pre-processing procedure includes 3'end adapter removal and trimming of low-quality bases at the ends of the reads. Depending on the study design and use of the data, redundant reads and undesired sequences such as contamination from primers, adaptors, or other species may be removed at this point.

Sequence alignment

After raw data QC and pre-processing, the next step is to map the reads to the reference genome and with high efficiency and accuracy. Alignment mapping is a classical "string match" task in computer

science. For example, most web browsers and text editors provide a "Find" function to search for the perfect matching string with a given query. However, finding the optimal alignment for a sequence read requires an alignment algorithm that is tolerant to imperfect matches, where genomic variations may occur. Moreover, the algorithm needs to be able to align millions of reads at a reasonable speed. As a first step to address this challenge, the reference genome is usually indexed in a hash table for efficient querying.

Post-alignment processing

After mapping reads to the reference genome, a three-step post-alignment processing procedure is recommended to minimise the artefacts that may affect the quality of downstream variant calling. It consists of read duplicate removal, indel realignment, and base quality score recalibration (BQSR).

In the alignment, reads aligned with exact mapping coordinates are considered "read duplicates," which represent either true DNA materials or PCR artefacts. The two cases, however, cannot be distinguished solely based on sequence or alignment information. Before sequencing, a library of DNA fragments from genomic regions of interest is prepared. Those fragments are amplified via certain amount of PCR cycles to provide a sufficient amount of DNA materials for sequencing, while limiting the duplication level of templates introduced by rounds of amplifications. For WES analysis, it is recommended to remove duplicates before variant calling, with the purpose of eliminating PCR-introduced bias due to uneven amplification of DNA fragments.

After duplicate removal, the second step is to identify genomic regions that contain indels and improve the alignment quality in the target region. Compared to reads that contain only SNVs, mapping reads composed of indels requires gapped alignment which is more prone to noise. When aligning reads to the genome (discussed in the previous section), most short-read aligners walk through the reads one by one and the optimal alignment is determined for each read independently. As a result, the introduction of gaps in each alignment may be different among overlapping reads. The quality of the resulting gapped alignment can be improved by considering all aligned reads in the same region after mapping.

In the sequencing reads, each base is assigned with a quality score generated by the sequencer, which represents the confidence of a base call. Base quality is a critical factor for accurate variant detection in the downstream analysis. However, the machine-generated scores are often inaccurate and systematically biased. Therefore, BQSR is recommended to improve the accuracy of confidence scores before variant calling. It takes into account all reads per lane and analyses covariation among the raw quality score, machine cycle, and dinucleotide content of adjacent bases. A corrected Phred-scaled quality score is reported for each base in the read alignment, assuming that all observed differences between the aligned reads and the reference genome are sequencing errors.

Variant analysis

Variant analysis consists of genotyping, variant calling, annotation, and prioritisation. Variant calling is the process by which variants are identified from the sequence data by comparing the data with the reference genome (after alignment). Annotation refers to coupling of additional information to the variants identified. Variants can be annotated in different ways, e.g. based on their genomic locations or predicted coding effects. After variant calls and annotations are generated, prioritisation analysis is performed with the aim of understanding the functional effect of variants.

5.8. Velocity

5.8.1. Speed of change

Speed of change in genomics techniques

Since the development of Sanger sequencing in 1977, DNA-sequencing technology has evolved at a rapid pace and the landscape continues to change. First generation (Sanger) sequencing was followed by second generation sequencing, also known as 'massively parallel' or 'next-generation' DNA sequencing, which was a major advance that allowed more rapid sequencing of larger parts of genomes and has now almost completely replaced Sanger sequencing. In recent years, third generation sequencing techniques have been developed, which refers to single-molecule, real-time sequencing techniques, which are further advancing the field. Despite these major advances in sequencing techniques, which have greatly increased the number of possible applications and the accuracy of the techniques, the data derived from different sequencing techniques, e.g. a DNA sequence, have remained fairly constant. Therefore, if raw data are available, it should be able to combine data derived from older techniques with data from newer techniques. When processed data is concerned, however, this can in some cases be more difficult, which emphasises the value of raw data to allow back-compatibility of techniques and merging of datasets.

Speed of change in genomics data

Another aspect related to change of genomics data is change of the genome itself. A differentiation can be made here between genomics data that remain constant (e.g. germ-line DNA), and genomic data that change over time, as is the case for example with RNA expression, and as can also be the case in oncology when tumour DNA is concerned. Important properties of epigenetic marks is that they are highly stable (methylated DNA and miRNA) in multiple biospecimens (i.e. urine, blood), in contrast to mRNA and proteins (Mitchell et al., 2008; Volinia et al., 2006). However, it has to be kept in mind that epigenetic changes may also change over time (Sierra et al., 2015), and are influenced by age, environmental and lifestyle factors representing a major challenge for the integration of the knowledge into clinical practice. DNA methylation is reversible, and therefore use of DNA methylation profiles as contributors to interindividual variability of drug response would require repeated investigation of biomarker gene DNA methylation profiles depending on patient age, treated target organ, or during long-term treatment.

Furthermore, a distinction can be made between the germ line genome of the subject, the genome of diseased tissue (e.g. tumour), and the genome of a pathogen (bacterium, virus, et cetera). The germ line genome of the subject can yield information on (susceptibility to) genetic diseases and also on variations in the rate of drug metabolism or susceptibility to experience certain adverse drug reactions. A one-time sequence determination is useful life long in predicting which drug could be effective and also what the effective dose for the individual could be.

The disease or pathogen genome is different since it evolves: tumour cells acquire resistance to anti-cancer drugs, bacteria to antimicrobial agents and viruses to antiviral drugs. As a result, a spectrum of genomes can arise, a quasispecies, which may be monitored regularly to predict the development of resistance. In many cases, viral or bacterial genomes, or circulating tumour DNA can be sampled from blood.

5.8.2. Rate of accumulation

The rate of accumulation of genomics data worldwide is very high (e.g. refer to Figure 4) and is expected to remain high and/or increase in the future as a result of decreasing costs, increasing

technical feasibility, and increased use of genomics data clinically. What is less clear is how genomics data linked to clinical data is accumulating. This requires coupling of electronic health records with genomics data and there is little oversight with regard to the extent to which these kind of linked datasets are accumulating. However, there are several good examples, such as U.K. Biobank and the FinnGen study where can be learned from.

5.9. Value

5.9.1. Usability of genomics big data

Current use of genomics data across the product life cycle

Currently, clinical genomics data are submitted in a proportion of the marketing authorisation applications, e.g. genomics data used for patient selection (such as in many oncology indications, cystic fibrosis), for investigating potential genomic causes of interindividual variability in efficacy and safety (e.g. molecular subsets within a specific cancer), or for pharmacogenomics/dose individualisation purposes (e.g. CYP2D6 genotyping to determine starting dose; SmPC Cerdelga®).

In oncology, the submission of genomics data is relatively frequent compared to other therapeutic areas. However, submission of genomics data is currently not standard in the process of applying for marketing authorisation. Also, in the post-authorisation setting, submission of genomics data or analyses is not a requirement; although post-authorisation measures to address genomics-clinical outcome associations are sometimes imposed.

Future impact of 'genomics big data' on the regulatory process?

The impact of 'genomics big data' on the future regulatory process will be dependent on different factors, including the availability of genomics data linked to clinical outcome data to regulators, and whether regulatory bodies will be involved in analysing such data in addition to assessing them.

A key question, therefore, is whether regulatory bodies will actively and systematically stimulate/request genomics data submission and/or be involved in 'genomics big data' analyses. In the current process genomics data are not systematically collected/analysed for regulatory purposes, and therefore the value/impact of genomics data is dependent on whether MAHs perform analyses on genomics data (in relation to clinical outcomes) or not. The value/impact of genomics data on regulatory processes could potentially be increased if a more proactive role is taken by the regulatory bodies. It is therefore important to consider which future role regulatory agencies should have in relation to the use of 'big data' to improve the regulatory system (or, more broadly, health care in general).

Different levels of involvement could be foreseen:

1. *Limited* active involvement in the collection and analysis of big data (status quo).
2. *Limited* active involvement in the collection and analysis of big data *but* actively stimulating applicants and MAHs to make use of available big data for defined/recommended purposes.
3. Active involvement, by stimulating or requesting applicants to submit their data for big data purposes, used by regulators for in-house analyses in order to improve medicinal product regulation.

Another aspect to be considered is whether applicants/MAHs should be stimulated to publicly share data to facilitate analyses by academia/third parties.

Considerations in relation to the agencies' future role in using big data

If a more proactive approach to the involvement of regulatory bodies in requesting and/or analysing genomics (big) data is developed – i.e. with active involvement in the collection and analysis of big data – there is potential for more rapid innovation in medicine e.g. by using genomics (big) data to better target medicines to patients who are likely to benefit.

However, active involvement by regulatory agencies in genomics (big) data analyses will also imply drastic changes in the regulatory process. It could for example be questioned whether this will lead to a shift in responsibility for the (results of the) analyses from MAHs to the regulatory agencies, and it is difficult at present to oversee the consequences of such a drastic change in processes. In addition, the regulatory network would need to acquire substantial resources to be able to make these changes in the process.

Developments in genomics

Machine Learning / Deep Learning

The human genome is now investigated through high-throughput functional assays, and through the generation of population genomic data, such as for instance in the previously mentioned U.K. Biobank collaboration and the FinnGen study. These initiatives will generate an enormous amount of genomics data that is expected to be combined with other 'omics' data (Figure 6). Artificial intelligence / deep learning strategies will help to analyse this enormous amount of data and will be more and more applied in the future. See for reviews also Telenti et al., 2018 and Yue and Wang, 2018. Also, in the clinical setting it is foreseen that machine learning will facilitate diagnosis setting and analysis of images. An interesting example in this respect is the publication of Haenssle et al. (2018) in which the performance of deep learning convolutional neural networks (CNN) were compared to a large international group of 58 dermatologists, including 30 experts. Most dermatologists were outperformed by the CNN.

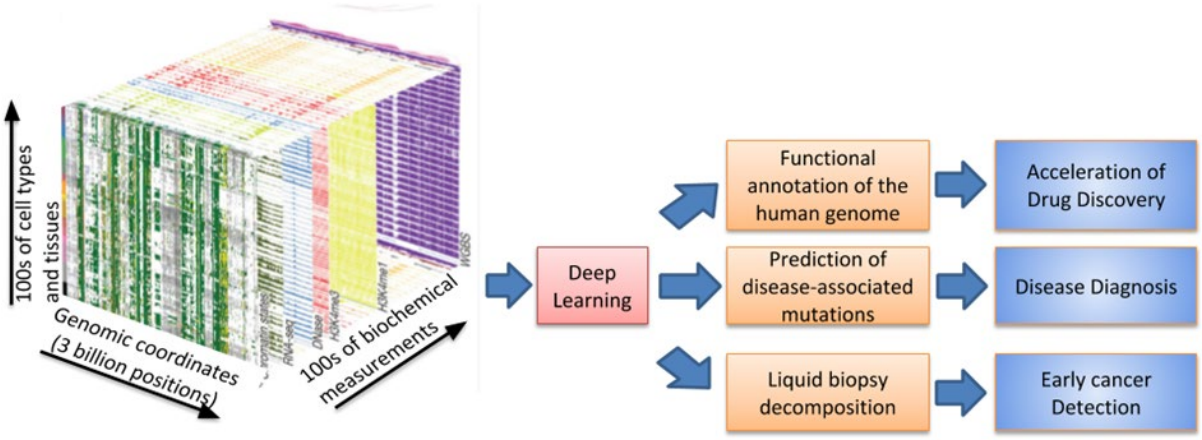

Figure 6. Opportunities for deep learning in genomics.

On the left side different types of data are shown of cell types and tissues of DNA and RNA-data and biochemical measurements. Deep learning on these data can assist in functional annotation of the human genome, prediction of disease-associated mutations and determine the composition of a liquid biopsy.

Faster and cheaper sequencing

Techniques in sequencing have evolved quickly and moved from sequencing short oligonucleotides to millions of bases, for an overview see Heather and Chain, 2016. Over the years, the technological capabilities of sequencing have increased, while costs have decreased. Current progress is also made in applying nanotechnology (Figure 7), in which it is used for portable analysis of DNA and other biological molecules.

Figure 7. SmidgION and Android basecaller (Nov 2017), which is under development for portable DNA sequencing.

Potential applications of 'genomics big data' in the regulatory process and the drug's life cycle

The value of genomics data in the regulatory process lies mainly in coupling of genomics data to sources of phenotypic and/or clinical outcome data. The coupling to phenotypic data (e.g. disease) could lead to discovery of new disease pathways, and subsequently allow discovery of new drug targets. Coupling of genomics data to clinical outcome data, on the other hand, could lead for example to improved (post-authorisation) pharmacovigilance, or identification of biomarkers for efficacy. Three examples are provided below in more detail:

1. Genomics-driven pharmacovigilance.
2. Genomic biomarkers predictive of efficacy.
3. Genomic biomarkers to monitor drug response.

1. Genomics-driven pharmacovigilance

It is known that severe/fatal ADRs are sometimes related to genetic predisposition and genetic association studies to determine links between genetic predisposition and ADRs have yielded very promising results. As such, genomics/genetic aberrations can in some cases be used to predict adverse events (consider e.g. *DPYD* genotypes in relation to 5-FU-associated toxicity; SmPC Xeloda®).

Most genetic association studies are currently performed in the research setting. It can be envisaged that in the future, whole-genome sequencing data will be used in pharmacovigilance activities to improve drug safety. Envisioning this, and if the regulatory network would decide to take an active role in the use of 'genomics big data' in pharmacovigilance, then it could be envisaged that MAHs are

requested to collect and submit genomics data linked to safety data from subjects treated within the clinical development programme. This would facilitate performing genetic association studies to determine whether there are genetic predictors for severe/fatal ADRs at the time of a marketing authorisation application. The information resulting from these analyses would potentially facilitate to improve the delineation of the target population, e.g. in case of the presence of very strong predictors of severe ADRs which result in a negative B/R, as well as improve information on genomics-safety associations in the product information.
Also in the *post-authorisation setting*, it could be envisioned that MAHs are requested to retrieve and submit genomics data from patients who experience severe/fatal ADRs. Comparing the genomics data from these patients with control patients who did not experience severe/fatal ADRs could for instance yield information on genomics-safety associations, including for rare severe/fatal ADRs.

2. Genomics biomarkers predictive of efficacy
Another obvious application of genomics data in the regulatory context is in the identification of biomarkers that are predictive of efficacy. Currently, submission of genomics data to analyse genomics-efficacy associations is voluntary and done infrequently. It could be envisioned that analyses similar to genomics-safety analyses could be used to determine in which patient populations a drug is likely to be more efficacious than in others. Such analyses could be performed both in the pre- and post-authorisation setting and could eventually be used to more specifically delineate the drug's indication, e.g. as for EGFR-targeted monoclonal antibodies which turned out to be effective only in non-KRAS-mutated colorectal cancer patients.

3. Genomic biomarkers to monitor drug response
A third example is the use of genomic biomarkers to monitor drug response, e.g. the use of circulating tumour DNA (liquid biopsy). This is a highly sensitive and non-invasive method, which can be used to determine the patient's response to treatment (Figure 8).

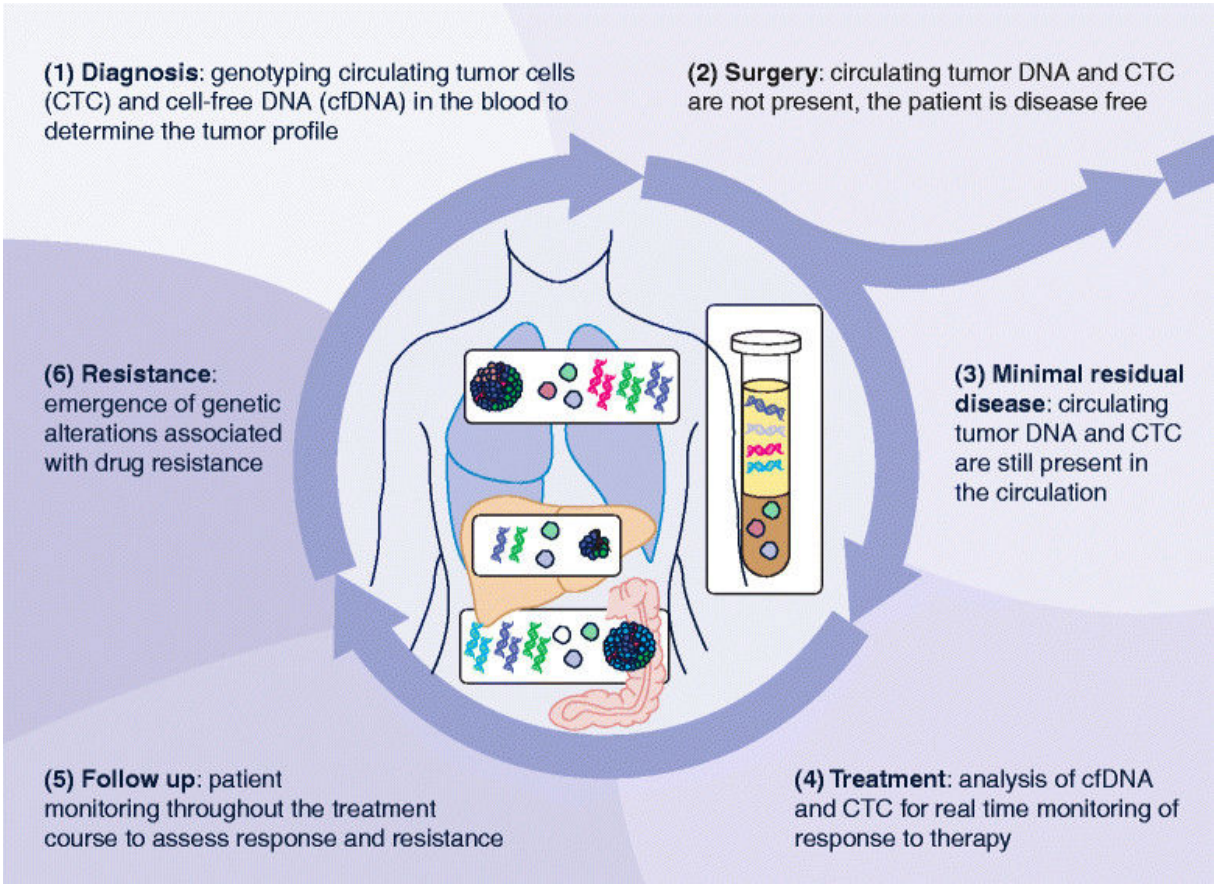

Figure 8. Circulating tumour DNA. Clinical applications of cell-free DNA analysis. cfDNA can be used in (1) diagnosis (2,3) to detect residual disease after surgery, (4) to monitor the response to therapy and (5) follow-up, and (6) to detect resistance. cfDNA, cell-free DNA; CTC, circulating tumor cells. (Siravegna and Bardelli, 2014).

Conclusions

Genomics is a fast-moving field with a lot of potential for personalising medicine by for instance reducing adverse events and/or optimising efficacy. A key recurring question is whether regulatory bodies will actively stimulate/require genomics data submission and/or be involved in 'genomics big data' analyses themselves. It is considered that as a start, it could be considered to actively stimulate applicants/MAHs to make use of available big data for defined/recommended purposes. In addition, it has to be considered whether applicants/MAHs should be stimulated to publicly share data to facilitate analyses by academia/third parties (see also section 6.2 Specific recommendations from the analysis).

5.9.2. Identify any uncertainties or unknowns which require further exploration

As the value of genomics data in the regulatory process lies mainly in coupling of genomics data to sources of phenotypic and/or clinical outcome data, optimisation of this linkage is essential. The following issues are therefore considered important (as mentioned in Brookes and Robinson, 2015): reach agreement on the minimum amount of data that should be made available (both for genomic and clinical data), promote responsible data sharing, maximize the ability to aggregate data by standardising informed consent related to data sharing, implement standards for analysing and reporting data quality, have globally accepted identifiers for patients, have standard APIs, and address sustainability of databases. Moreover, the extent of required phenotypic information is important. Sometimes, longitudinal data would be necessary, acquired over the life of a patient.

An important uncertainty is how feasible it is on a large scale, to link relevant sources of clinical outcome data to genomics data in practice. Uncertainties in this respect are e.g. technical aspects, ethical and privacy aspects, and security aspects. However, there are several examples/initiatives that are currently already doing this, and lessons can be learned from these initiatives.

In addition, what remains to be established, is the extent to which the regulatory system may indeed be improved by having these sources of data. A pilot study to demonstrate the value of genomics data submission/analysis could be useful, to demonstrate the added value of genomics data for the regulatory system e.g. for pharmacovigilance purposes in the post-marketing setting, or for the identification of biomarkers for efficacy in the post-marketing setting.

5.9.3. Possible gaps in current European guidance

- Current guidance on technical validation of advanced genomics (e.g. sequencing) methods is limited.
- There is limited guidance on standardisation of genomics analysis and data processing techniques, as well as for standardisation of data formats for genomics data and/or clinical outcome data linked to genomics data.
- There is no regulatory guidance related to data sharing practices.

5.9.4. Data Accessibility - consider privacy and governance challenges and the limitations to access as this will affect the value from a regulatory context

The genomic databases described are mostly publicly available, and the data can thus also be accessed by national competent authorities (NCAs). However, there is also data that cannot be accessed, because it is not publicly available, either from companies or obtained in a diagnostic setting. Thus, big data analyses will be limited to those sources of genomics data that are actively shared.
As described, especially the linkage of genomics data to other data sources (e.g. electronic health records) is considered relevant. To be able to link electronic health record data, there should be an incentive for e.g. hospitals to contribute to the process of linking these data.

Privacy issues will be an important hurdle to sharing of genomics and medical data. When linking data from different sources, a system would be required that ensures privacy of the patients, and informed consent would have to be adequately arranged (e.g. for the reuse of the material or data sharing). Sometimes patients have only provided informed consent to the use of their data for the trial they participated in, and not for data sharing purposes, and consequently those data cannot be shared.

Even though technical solutions are available that could facilitate anonymisation of the data, such as the data management system ProMISe (https://www.msbi.nl/promise/), with respect to genomics data there will be a risk that the patient is identifiable once their complete genomics data are publicly available, i.e., simply because every genome is unique. This is especially of concern for the rare diseases where sometimes there are only a few patients with a specific mutation worldwide.

5.9.5. Data analytics - discuss current and potential new approaches

Currently, the underlying genomics data are not submitted to the regulatory authorities. It could be considered to request companies, as is done by peer-reviewed journals, to make their genomics data available in public primary databases (i.e. GenBank, DNA DataBank of Japan, European Nucleotide Archive). In this way the data will be accessible for the research community for further analyses. To be able to optimally profit, it would be important to link the most important parameters related to phenotype and/or treatment outcome to the genomics dataset. Linking clinical and phenotype variables across data sets will both power precision medicine studies and introduce new privacy risks (Craig, 2016); e.g. Harmanci and Gerstein (2016) examined the increased privacy risk from linkage attacks when information about an individual is present in multiple high-dimensional genotypic and phenotypic data sets.

5.9.6. Regulatory challenges

Knowledge/expertise gaps within Agencies

Assessing genomics data requires specific expertise, i.e. expertise in bioinformatics that may not be available in all regulatory bodies. This knowledge/expertise gap needs to be addressed in order to be able to adequately assess big data analyses for regulatory purposes. Furthermore, keeping knowledge up to date could be challenging, as progress in the genomics field is rapid.

Feasibility of implementing genomic tests in EU member states

There are different hurdles that have to be taken into account when considering implementation of use of genomic data in routine clinical practice; financial, technical/logistical and physician/patient acceptance. Consequently, differences between EU Member States could possibly exist in the feasibility of conducting genetic tests in clinical practice. As an example, for *DPYD* genotyping, the PGWP asked

EU Member States on the knowledge or expectations regarding the feasibility of *DPYD* genotyping in their country (EMA/PRAC/22272/2017). Several countries (e.g. Bulgaria, Croatia, Slovenia, Serbia) indicated that the *DPYD* genotyping methodology had been or could be implemented in their country. At least in Croatia, Bulgaria, Norway, UK, and the Netherlands, DPYD genotyping was already conducted in some centres, either pre-emptively, or after development of serious drug reactions. Therefore, the PGWP did not anticipate problems with setting up the *DPYD* genotyping methodology in the EU. However, this example concerns genotyping of single variants. The use of e.g. whole-genome sequencing in clinical practice will be much more complicated and might not be feasible in all countries in the EU at this moment.

Practical feasibility of linking large numbers of data sources

Although technically possible, the practical feasibility of linking large numbers of data sources is a concern that needs to be addressed before truly big data analyses are possible. For example, in the past it has turned out that even linking different sources of electronic health records within one country is difficult in practice. Therefore, it remains to be established how feasible it is to link different sources of genomics data with different sources of electronic health record data, for example from different countries.

Feasibility of linking large numbers of data sources from a privacy/security perspective

Currently, no framework is in place that addresses the security and privacy issues associated with sharing of genomics and clinical outcome data within the regulatory context. It remains to be established whether these issues can be addressed in a broad European or global context.

Feasibility of linking large numbers of data sources from an ethical perspective

Currently, no framework is in place that addresses the ethical issues associated with sharing of genomics and clinical outcome data within the regulatory context. An example of an ethical challenge is that of the reporting of incidental findings. In a diagnostic setting, frequently filters are applied so that only the genes are investigated for variants that are possibly the cause of the phenotype. There is a risk that incidental findings will be found in a regulatory context when big data analyses are performed. It needs to be discussed how to address this and, in case there are incidental findings with consequences for individual patients, how to deal with these findings. Several guidelines are published on incidental findings by the American College of Medical Genetics and Genomics (ACMG), the European Society of Human Genetics (ESHG; Hehir-Kwa et al., 2015), and the Canadian College of Medical Geneticists.

Handling genomic data from third parties

Another challenge to consider is how signals from third parties (such as research groups), that may arise e.g. from big data genetic association studies based on publicly available genomics data, can be used in the (post-marketing) regulatory process.

6. Conclusions

6.1. Summarised key messages

The following four points summarise the key messages from the mapping exercise:

- There are many publicly available data sources on genomics. However, these data sources usually couple genomics data to information on disease, but do not couple genomics data to treatment outcomes. The latter would be useful in the regulatory context.
- Genomics analyses require highly specific skills and knowledge. Therefore, although it is anticipated that regulators will not do these highly specialised analyses themselves, knowledge should be available within the regulatory agencies to be able to assess big data analyses when part of a drug application. Collaborations with skilled academic groups, as well as clustering expertise (as it is done in the Pharmacogenomics Working Party) or educating assessors using the EU NTC platform could be considered.
- Openness of data (data sharing) should be strived for, as this would be beneficial for medicine development and research.
- Privacy and security are key issues in the context of sharing genomics data from patients.

6.2. Specific recommendations from the analysis

Based on the mapping analysis performed in this report, the genomics subgroup has drafted the following recommendations revolving around the following topics:

- Sharing of genomic data, including privacy and security.
- Data standardisation.
- Data linkage.
- Data quality requirements.
- Skills and knowledge within the network.
- Regulation of genomics diagnostic tests.
- Availability of clinically meaningful information regarding genomics data.
- Demonstration of value.
- Exploiting the value of genomics data in post-authorisation setting.
- Need for regulatory guidance.

For the specific recommendations on each of these topics reference is made in Table 12 below. Moreover, for each of these topics additional information is provided below.

Sharing of genomic data, including privacy and security

There are many data sources where genomics data can be freely accessed. These data sources usually couple genomics data to disease, but do not couple genomics data to individual treatment outcome. However, some of the data sources integrate peer-reviewed literature in their recommendations, such as the Pharmacogenomics Knowledgebase. The information on these websites could be consulted as a reference during assessment of a dossier, similar to peer-reviewed literature. In case data linkage of individual genomics data to treatment outcome is applied, these data will become more valuable from a regulator's perspective. The current data of genomics data linked to treatment outcome comes from the data that is submitted by the MAH. These data can also be very valuable for academia, and it could be considered to request the company to make the data publicly available. Important aspects in this are, however, privacy and security. Moreover, it is recommended to explore whether the EMA should

provide a central secure platform for sharing of clinical trial data, or whether EMA could provide a portal linking to industry owned data.

There is a risk that the privacy of patients becomes jeopardised when genomics data are shared. This could severely impact the individual patient's life, e.g. in getting a mortgage or life-insurance, when for instance information becomes public that an individual has a chronic severe condition. Thus, although openness of data should be strived for, as it would be beneficial for medicine development and research, privacy is a requirement.

Security of the system, which prevents that privacy of the patient could get impaired should be up to date. Like privacy, security is a requirement.

Data standardisation

Standardisation will be an important objective in relation to use of genomics data in the regulatory context. Standardisation applies to the ways samples are analysed, but also to how the resulting data are analysed (the analytical pipeline), and how genomic association studies based on genomics data in relation to clinical data are performed. As described earlier, Bertier et al. (2016) indicated that there is a need to share WGS/WES results and to develop more complete, less biased databases containing fewer false positive and false negative variant-phenotype associations. Further, for a good interpretation of variants it is a necessity that the phenotype is collected accurately and in a standardised fashion (Bowdin et al., 2016). The same applies for the clinical outcome measurements. These should be standardised to be able to couple them to each other.

Data linkage

The value of genomics data in the regulatory process lies mainly in coupling genomics data to sources of phenotypic and/or clinical outcome data, optimisation of this linkage is essential. Several recommendations are made in the table that would enhance this linkage.

Data quality requirements

A challenge with regard to the available public genomics data repositories is data quality. Exponential growth of the amount of data makes it difficult to maintain accuracy and accessibility across the public databases. Although researchers are able to update sequences they have submitted to e.g. GenBank and other repositories, a large portion of the stored data may be incorrect or incomplete due to the volume of the submitted information and the nature of research (e.g., researchers move on to other projects, mistakes in the original data go unnoticed, etc.). There are also issues of duplication with minor variations and redundancy. Quality control of genomics data is thus an important issue in the context of using big data genomics in the regulatory context.

Several guidances exist on genomics from the ICH, EMA and FDA. In many of these guidelines quality is discussed, however, the guidance that is provided is rather general. This indicates that the discussion on data quality and in particular data quality metrics has not fully crystallised. It is therefore recommended to establish a working group to determine data quality requirements, standards, etc., and to initiate international collaboration regarding setting the standards for data quality requirements.

Skills and knowledge within the network

It requires specialised knowledge and skills to know when and how to use which data resource. Combining different data sources such as expression data in different tissues can provide information on when certain genes are co-expressed, and provide you with a gene network (www.genenetwork.nl). These data can function as a starting point for further research for academia or for medicine development for new targets/pathways. It should be kept in mind though that these possible relations always need to be confirmed.

From a regulatory perspective it is not anticipated that regulators will do these highly specialised analyses themselves. However, knowledge should be available to be able to assess these analyses when part of a marketing authorisation application. Collaborations with skilled academic groups could be considered, as well as clustering expertise within a working party (similar to the Pharmacogenomics Working Party) and/or different agencies in the network or educating assessors using the EU NTC platform.

Regulation of genomics diagnostic tests (in vitro diagnostic medical devices)

As mentioned in the description of available regulatory guidance, the responsibility in Europe for regulation of *in vitro* (companion) diagnostic tests – including genomics tests that are used to make treatment decisions for approved medicinal products – lies with the notified bodies and is not within the direct remit of the national competent authorities or EMA. This is in contrast to the situation in the US, where the FDA centrally assesses and approves such diagnostic tests (refer also to sections 5.2 and 5.3, as well as Table 4). In the EU, when a medicinal product is administered based on a genomic diagnostic test (e.g. based on the presence of a mutation in tumour tissue, such as BRAF), the requirement for industry at the time of marketing authorisation of the medicinal product is that a CE marked diagnostic test, as certified by a notified body, is available on the market.

With the increasing complexity of genomics tests and the increasing importance of these tests in the adequate use of medicinal products (in particular in oncology, as reflected by a sharp increase in the number of oncology medicinal products, which are used specifically in patients with certain genomic characteristics), it needs to be considered whether the current system for assessment of in vitro (companion) diagnostic tests is still fit for purpose.

A new IVD regulation will come into action in 2022 (EU 2017/746; see also http://www.ema.europa.eu/docs/en_GB/document_library/Presentation/2018/06/WC500250068.pdf). According to this new regulation, for companion diagnostics the notified body shall consult a concerned competent authority or EMA. Furthermore, the notified body, before issuing an EU technical documentation assessment certificate for the companion diagnostic, will consult the competent authority regarding the suitability of the device in relation to the medicinal product concerned. This means there will be more interaction between notified bodies and competent authorities in the future in the assessment of *in vitro* (companion) diagnostic tests including genomics tests used with medicinal products.

However, the responsibility for assessment will still lie with a variety of notified bodies, potentially leading to variability in the assessment of diagnostics tests and hence variability in performance characteristics that may ultimately affect the benefit/risk balance of medicinal products that are used specifically based on the results of a diagnostic test. A system in which central assessment of *in vitro* (companion) diagnostic tests is performed could potentially be beneficial in this respect.

With the introduction of increased genetic testing in clinical practice (Presley et al., 2018; Bunn and Aisner, 2018), clinical education and decision support for personalised therapy will become more important. For instance, physicians were surveyed by Obeng et al. (2018) about their perceptions as to the clinical utility of genetic data as well as their preparedness to integrate it into practice. The majority believed that genetic testing was clinically useful; however, only a third believed that they had obtained adequate training to care for genetically "high-risk" patients. Consequently, the authors recommended exploring the use of simplified genetics-guided recommendations. Additional training, such as genetic e-learning resources, could be effective in improving genetic knowledge, skills and attitudes (Jackson et al., 2018).

Availability of clinically meaningful information regarding genomics data

Currently, pharmacogenomics data are commonly mentioned in the SmPC, describing e.g. the effect of certain polymorphisms on safety or efficacy of a drug. What is often lacking, however, is practical guidelines on how to use pharmacogenomics information to individualise treatment of patients. It could be envisioned that with increasing availability of genomics data, at some point a switch can be made from population-oriented product information, to more individual-oriented product information. For example, while now the benefit/risk is often determined in the whole population of patients eligible for treatment with a certain drug and irrespective of genomic differences between patients, it could be envisioned that with increasing availability of genomics data, a more personalised benefit/risk assessment can be made. In order to facilitate the use of genomics-guided treatment in clinical practice, it would be advisable to make clinically meaningful information on pharmacogenomics-guided treatment more readily available in the SmPC. Further, it could be considered to make this information online available in a separate database, which could be searched by pharmacists, geneticists and physicians, and which is linked to the most updated version of the SmPC. However, it would be important to have an adequate system in place to curate the presented information and be clear about the level of evidence available for clinical utility of the described genomics-outcome associations.

Demonstration of value

What remains to be established is the extent to which the regulatory process may indeed be improved by having sources of genomic big data coupled to clinical data for regulatory purposes. A retrospective analysis could be useful in this respect, to demonstrate the value of systematically gathered genomics data coupled to clinical data (efficacy/safety) from the pivotal trials, e.g. by performing a pilot study in oncology.

Exploiting the value of genomics data in post-authorisation setting

Likewise, to demonstrate the value of genomics data in the post-authorisation setting, it could be considered to initiate a pilot study to investigate the added value of genomics data for pharmacovigilance purposes in the post-marketing setting. In addition, an EMA-pilot study could be considered to demonstrate the added value of genomics for the identification of biomarkers for efficacy in the post-marketing setting (e.g. for determining in which patient populations a drug is likely to be more efficacious than in others).

Need for regulatory guidance

Providing guidance is critical in order to communicate regulatory expectations and hence improve quality of 'big data' elements with regulatory applications. For instance, some guidance exists on data standardisation for genomics data. However, the guidance that exists is rather general. To optimise data sharing it would be beneficial to agree on standardised data formats of genomics data and clinical outcome data to allow better linkage and ability to share data.

Possible gaps in relation to future regulatory assessment of genomics big data submissions and in the current EMA guidance documents have been identified, and recommendations are proposed in the table below (Table 12).

Table 12. Regulatory recommendations of the genomics subgroup

Topic	Core Recommendation	Reinforcing Actions	Strategic Goal
Sharing of genomic data, including privacy and security	Stimulate public sharing of genomics and clinical trial data by applicants/MAHs to facilitate analyses	- Promote the public sharing of genomics data from pivotal clinical trials submitted as part of a marketing authorisation application (MAA). In line with Policy 0070, data	Create openness (sharing) of genomic data linked to clinical data to advance

(under a general recommendation of promoting a data sharing culture)	by regulators/ academia/third parties.	should be shared irrespective of whether the MAA is successful, unsuccessful or subsequently withdrawn. Promote disclosure of historical clinical trial data. - As sharing of genomics data may be particularly challenging in terms of data anonymization/privacy protection/data security, a number of actions in this respect will be needed to deliver meaningful data sharing: (i) Establish an expert working group to establish conditions, which would enable the sharing of genomic data including data anonymization, data sharing mechanisms (access and security) and informed consent. (ii) Discuss in the expert working group ethical issues unique to genomics e.g. familial issues, secondary incidental findings. (iii) Consider international challenges and opportunities in enabling global data sharing. - Explore whether the EMA should provide a central secure platform for sharing of clinical trial data, or whether EMA could provide a portal linking to industry owned data. - Publish an EMA guideline on data sharing to facilitate the submission of industry genomics/clinical data to EMA. - Make society, industry, researchers, and funders of research aware of the value of data sharing. - Emphasise the need to link funding of clinical research to obligations regarding sharing of the generated data (data sharing obligations after research projects have been finalized; e.g. with 1-year data exclusivity).	personalised, genomics-guided medicine.
Data standardisation	Stimulate standardisation of genomics and clinical trial data in a structured way.	- Promote the sharing in a structured way of RCT data by generating incentives for researchers. - Consider the need to provide further guidance on standardisation of genomics analysis and data processing techniques, as well as for standardisation of data formats for genomics data and/or clinical outcome data linked to genomics data; because currently there is limited guidance available on these topics.	Create openness (sharing) of genomic data linked to clinical data to advance personalised, genomics-guided medicine.
Data linkage	Optimise linkage of the most important	- Promote linkage of the most important parameters (e.g. adverse	Linkage of data to the key parameters

	phenotypic and/or treatment parameters to genomics datasets.	events, primary efficacy outcomes) to the genomics dataset upon MAA - Determine any potential challenges when converting genomics data from one platform to another. - Study how this data linkage can be achieved across different trials, by for instance, unique identifiers. - Initiate an EMA pilot study to link genomics data to clinical outcome data from different studies (efficacy/safety) in the event that investigators or MAHs are willing to cooperate, but do not have the resources themselves. This could be relevant for both oncology and rare diseases. - Investigate the possibilities of using machine learning for linkage. - Consider the need to facilitate EMA sponsored research into tools for data linkage.	could reveal new correlations between genomics data and efficacy/safety.
Data quality requirements	Establish requirements regarding data quality.	- Establish a working group to determine data quality requirements, standards, etc. - Initiate international collaboration regarding setting the standards for data quality requirements. - Stimulate sharing of the following data by applicants/MAHs to assess data quality: 1. A minimal data standard to be considered. 2. Raw data in addition to processed data. 3. Meta-data to be attached to the data (i.e. descriptive information about the overall study, individual samples, all protocols, and references to processed and raw data file names).	Setting data quality standards/requirements is necessary to ensure reliability of the analyses performed on big data sources.
Skills and knowledge within the network	Address the knowledge/expertise gap across the European regulatory network to ensure big data applications can be reliability assessed.	- Document the current level of expertise within the European regulatory network. - Identify the gaps in knowledge/ expertise within the regulatory network (EMA and national competent authorities). - Determine mechanisms to address these gaps in knowledge. Consider collaborations with skilled academic groups, clustering of expertise within a Working party (such as the	To ensure genomics data in big data analyses are optimally assessed by the regulatory agencies.

		Pharmacogenomics Working Party) and specific training initiatives to fill these gaps.	
Medical devices regulation	Ensure effective regulation of genomic diagnostic tests which are associated with the use of medicinal products.	- Explore the legal basis for centralising medical device regulation. - Investigate whether the current CE system is fit for purpose. - Develop strong and systematic ties between device regulators in order that the different regulatory frameworks can operate in a complementary way. - Introduction of increased genomic testing in clinical practice must be complemented with clinician education and decision support to understand the importance of personalised therapy.	To deliver a device regulatory system with centralised competence and oversight to ensure data quality and reliability are fit for purpose.
Availability of clinically meaningful information regarding genomics data	Investigate how to optimize availability of clinically meaningful information regarding the impact of genomics on the benefits and risks of medicines to health-care providers and patients.	- In order to advance genomics-guided treatment in clinical practice, meaningful genomics data should be made more readily available to health-care providers, e.g. via the SmPC including the most up-to-date information. In light of this, reconsider the current process of keeping the SmPC up to date (i.e. applicant-driven nature). However, it will be important to have an adequate system in place to curate the presented information and be clear about the level of evidence available for clinical utility of the described genomics-outcome associations. - Explore other ways for publishing curated, clinically meaningful genomics data, e.g. by exploring the possibility of making the information online available in a separate database/app which can be searched by pharmacists/geneticists/physicians (such as pharmgkb.org), and which could be linked to the SmPC.	Accessibility of up-to-date and clinically meaningful genomics information will stimulate the use of genomics-guided personalised treatment in clinical practice.
Demonstration of value	Demonstrate the value of genomics/clinical big data analyses.	- Explore possibilities to demonstrate the value by a retrospective analysis of systematically gathered genomics data coupled to clinical data (efficacy/safety) from the pivotal trials, e.g. by performing a pilot study in oncology.	
Exploiting the value of genomics	Improve the regulatory process by employing	- Initiate a pilot study to demonstrate the added value of genomics data for pharmacovigilance purposes in the	Demonstration of added value is critical for

data in post-authorisation setting	genomic big data in the post-authorisation setting.	post-marketing setting. Such as by investigating the feasibility and the additional value of requesting MAHs to retrieve and submit genomics data from patients who experience severe/fatal ADRs. - Consider an EMA-pilot study to demonstrate the added value of genomics for the identification of biomarkers for efficacy in the post-marketing setting (e.g. for determining in which patient populations a drug is likely to be more efficacious than in others).	stimulating data sharing, to convince companies to submit data. Knowledge about the extent of the added value can result in further recommendations on how to optimally use/assess genomics data in the regulatory setting.
Need for regulatory guidance	Provide sufficient guidance for industry/academia regarding the use of big data in regulatory processes.	- Publish an EMA guideline on data sharing to facilitate the submission of industry genomics/clinical data to EMA. - Consider the need to provide guidance on technical validation of advanced genomics (e.g. sequencing) methods because current guidance is limited. - Consider the need to provide further guidance on standardisation of genomics analysis and data processing techniques, as well as for standardisation of data formats for genomics data and/or clinical outcome data linked to genomics data; because currently there is limited guidance available on these topics.	Providing guidance is critical in order to communicate regulatory expectations and hence improve quality of 'big data' elements with regulatory applications

7. References

Bao et al., 2014:13(Suppl 2):67–82, *Cancer Inform*. Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing.

Bertier et al., 2016:9:52, *BMC Med Genomics*. Unsolved challenges of clinical whole-exome sequencing: a systematic literature review of end-users' views.

Blute et al., 2015:25(1):83-88, *Curr Opin Urol.* The epigenetics of prostate cancer diagnosis and prognosis: update on clinical applications.

Bodenreiser, 2009:45-49, *AMIA Annu Symp Proc*. Using SNOMED CT in combination with MedDRA for reporting signal detection and adverse drug reactions reporting.

Bowdin et al., 2016, *Genetics in Medicine*. Recommendations for the integration of genomics into clinical practice.

Brookes and Robinson, 2015:16:702-715, *Nature Reviews Genetics*. Human genotype–phenotype databases: aims, challenges and opportunities.

Bunn and Aisner, 2018:320:445-446, *JAMA*. Broad-Based Molecular Testing for Lung Cancer: Precisely the Time for Precision.

Byun et al., 2009:18:4808-4817, *Hum Mol Genet*. Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns.

Cardoso et al., 2016:375:717-729, *N Engl J Med*. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer.

Cascorbi et al., 2016:99:468-470, *Clin Pharmacol Ther*. Epigenetics in Drug Response.

Craig, 2016:13:211-212, *Nature Methods*. Understanding the links between privacy and public data sharing.

Dawson et al., 2013:368:1199-1209. *N Engl J Med*. Analysis of Circulating Tumor DNA to Monitor Metastatic Breast Cancer.

Dean, 2016, *Medical Genetics Summaries*. Fluorouracil Therapy and *DPYD* Genotype. https://www.ncbi.nlm.nih.gov/books/NBK395610/

Dean, 2015, *Medical Genetics Summaries*. Clopidogrel Therapy and CYP2C19 Genotype. https://www.ncbi.nlm.nih.gov/books/NBK84114/

Fanini and Fabbri, 2016:99:485-93. *Clin Pharmacol Ther*. MicroRNAs and cancer resistance: A new molecular plot.

Fisel et al., 2016:99:512-27, *Clin Pharmacol Ther.* DNA Methylation of ADME Genes

Fox et al., 2014, *Next Gener Seq Appl*. Accuracy of Next Generation Sequencing Platforms.

Glavey et al., 2016:169:35-49. *Cancer Treat Res.* Epigenetics in Multiple Myeloma.

Gorzelak et al., 2015:10(8):e0134802, *PLoS One*. Methods for Improving Human Gut Microbiome Data by Reducing Variability through Sample Processing and Storage of Stool.

Haenssle et al., 2018, *Ann Oncol*. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists.

Han et al., 2017:15:59-72, *Genomics, Proteomics & Bioinformatics*. Circulating Tumor DNA as Biomarkers for Cancer Detection.

Harmanci and Gerstein, 2016:13:251-256, *Nature Methods*. Quantification of private information leakage from phenotype-genotype data: linking attacks.

Hehir-Kwa et al., 2015:23:1601-1606, *Eur J Hum Genet*. Towards a European consensus for reporting incidental findings during clinical NGS testing.

Herceg et al., 2018:142:874-882, *Int J Cancer*. Roadmap for investigating epigenome deregulation and environmental origins of cancer.

Jackson et al., 2018, epub ahead of print, *Genet Med*. The Gen-Equip Project: evaluation and impact of genetics e-learning resources for primary care in six European languages.

Kendziorski et al., 2006:62:19-27, *Biometrics.* Statistical methods for Expression Quantitative Trait Loci (eQTL) mapping.

Krop et al., 2017:35:2838-2847, *J Clin Oncol*. Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline Focused Update.

Lao et al., 2011:8:686-700. *Nat Rev Gastroenterol Hepatol.* Epigenetics and colorectal cancer.

Lathe et al., 2008:1:2, *Nature Education*. Genomic Data Resources: Challenges and Promises.

Lippman and Osborne, 2013:368:1249-1250, *N Engl J Med*. Circulating Tumor DNA — Ready for Prime Time?

Lynch and Pedersen, 2016:375:2369-2379, *N Engl J Med*. The Human Intestinal Microbiome in Health and Disease.

Mao et al., 2018:46:D92-99, *Nucleic Acids Res*. EpiDenovo: a platform for linking regulatory de novo mutations to developmental epigenetics and diseases.

Mitchell et al., 2008:105:10513-10518, *Proc Natl Acad Sci U S A*. Circulating microRNAs as stable blood-based markers for cancer detection.

Nature editorial staff, 2010:464:670-671, *Nature*. Human genome at ten: The sequence explosion.

Obeng et al., 2018:8(3), *J Pers Med*. Physician-Reported Benefits and Barriers to Clinical Implementation of Genomic Medicine: A Multi-Site IGNITE-Network Survey.

Presley et al., 2018:320:469-477, *JAMA*. Association of Broad-Based Genomic Sequencing With Survival Among Patients With Advanced Non-Small Cell Lung Cancer in the Community Oncology Setting.

Rakyan et al., 2011:12:529-41, *Nat Rev Genet.* Epigenome-Wide Association Studies for common human diseases.

Schiano et al., 2015:36:226-35, *Trends in Pharmacological Sciences*. Epigenetic-related therapeutic challenges in cardiovascular disease.

Seystahl et al., 2016:99:389-408, *Crit Rev Oncol Hematol*. Therapeutic options in recurrent glioblastoma--An update.

Sierra et al., 2015:16:435-440, *Curr Genomics*. Epigenetics of Aging.

Siravegna and Bardelli, 2014:15:449, *Genome Biol*. Genotyping cell-free tumor DNA in the blood to detect residual disease and drug resistance.

Swaminathan et al., 2015:14:8-15, *Comput Struct Biotechnol J*. A Review on Genomics APIs.

Telenti et al., 2018:27:R63-71, *Hum Mol Genet*. Deep learning of genomic variation and regulatory network data.

Van El and Cornel, 2011:19:377–381, *Eur J Hum Genet*. Genetic testing and common disorders in a public health framework; Recommendations of the European Society of Human Genetics.

Van El et al., 2013:21:580–584, *Eur J Hum Genet*. Whole Genome Sequencing in health care. Recommendations of the European Society of Human Genetics.

Volinia et al., 2006:103:2257-2261, *Proc Natl Acad Sci U S A*. A microRNA expression signature of human solid tumors defines cancer gene targets.

Wilkinson et al., 2016:3:160018, *Sci Data*. The FAIR Guiding Principles for scientific data management and stewardship.

Yue and Wang, 2018. Invited chapter for Springer Book: Handbook of Deep Learning Applications. Deep Learning for Genomics: A Concise Overview

Appendix 1: Definitions

	Definition
Epigenetics	The study of modification of gene expression rather than alteration of the genetic code itself.
Genetics	The study of variations in DNA sequence and their function.
Genomic biomarker	A measurable DNA and/or RNA characteristic that is an indicator of normal biologic processes, pathogenic processes, and/or response to therapeutic or other interventions. A genomic biomarker could, for example, be a measurement of: the expression of a gene, the function of a gene, the regulation of a gene. A genomic biomarker can consist of one or more DNA and/or RNA characteristics. DNA characteristics include, but are not limited to: • Single nucleotide polymorphisms (SNPs). • Variability of short sequence repeats. • Haplotypes. • DNA modifications, e.g. methylation. • Deletions or insertions of (a) single nucleotide(s). • Copy number variations. • Cytogenetic rearrangements, e.g. translocations, duplications, deletions or inversions. RNA characteristics include, but are not limited to: • RNA sequences. • RNA expression levels. • RNA processing, e.g. splicing and editing. • microRNA levels. The definition of a genomic biomarker is not limited to human samples but includes samples from viruses and infectious agents as well as animal samples, i.e. for the application of genomic biomarkers to non-clinical and/or toxicological studies. The definition of a genomic biomarker does not include the measurement and characterisation of proteins or low molecular weight metabolites.
Genomics	The study of genes, including variations of DNA and RNA characteristics, and their function.
Microbioata	The microorganisms of a particular site, habitat, or geological period.
Microbiome	The combined genetic material of the microorganisms in a particular environment.
Microbiomics	The study of the microbiome.
Transcriptomics	The study of all RNA molecules in one cell or a population of cells, and their functions.

References:

- ICH Topic E15 Definitions for genomic biomarkers, pharmacogenomics, pharmacogenetics, genomic data and sample coding categories.
- Cambridge dictionary.

Appendix 2: Genomics initiatives

Name	Type	Additional Details	Cohort Size	Cohort Description	Type of Data	Project Timeline	Disease Area	Website
100k Wellness Project	Research Project	Non-profit research organization (academic & industry ties)	100000	Unaffected individuals		2014 - ongoing	Neurological Disease, Complex Diseases, Other	https://www.systemsbiology.org/research/100k-wellness-project/
AACR Project Genomics, Evidence, Neoplasia, Information, Exchange (GENIE)	Data-Sharing Initiative	International genomic and clinical data-sharing project	17000	International cancer patients		2015 - ongoing	Cancer	http://www.aacr.org/Research/Research/Pages/aacr-project-genie.aspx#.V3vTcPkrJaR
23andMe	Organisation/Company	Private company	1000000	Customers (>80% consented to research)	Variants	2008 - ongoing	Cancer, Rare Disease, Complex Diseases, Other	https://blog.23andme.com/
African Partnership for Chronic Disease Research (APCDR)	Consortium Research Project	Multi-centre collaboration of 18 institutions					Cancer, Neurological Disease, Complex Diseases	http://www.apcdr.org/
Ancestry.com	Organisation/Company	Private company	1400000	Customer DNA samples	Variants		Other	http://www.ancestry.com/
Asian Cancer Research Group (ACRG)	Organisation/Company	Not-for-profit company	176	HCC tumours and paired normal tissues	Whole-genome/ exome sequence	2010 - ongoing	Cancer	http://consortiapedia.fastercures.org/consortia/acrg/
AstraZeneca	Research Project	Pharmaceutical company	2000000	Individuals (includes 500,000 participants from AstraZeneca clinical trials)	Whole-genome/ exome sequence	2016 - 2026	Rare Disease, Complex Diseases, Other	https://www.astrazeneca.com/media-centre/press-releases/2016/AstraZeneca-launches-integrated-genomics-approach-to-transform-drug-discovery-and-development-22042016.html
Australian Genomics Health Alliance (AGHA)	Consortium	National consortium of clinical and academic centres (research institutes, hospitals, academic institutions)	1800	Cases (approximately 900 cancer, 900 rare disease)	Whole-genome/ exome sequence, Gene panel, mtDNA, RNAseq	2016 - 2020	Cancer, Rare Disease	http://www.australiangenomics.org.au

Beacon Project	Data-Sharing Initiative	GA4GH Demonstration Project	100000	Individuals	Variants	2014 - ongoing	Cancer, Rare Disease	Cancer Rare Disease
BioBank Japan	Repository Research Project	Biobank	260000	Patients only	Whole-genome/ exome sequence, Variants	2003 - ongoing	Cancer, Other	https://biobankjp.org/english/index.html
Biobanking and Biomolecular resources Research Infrastructure (BBMRI)	Consortium	International not-for-profit consortium of biobanks & biomolecular resources				2008 - ongoing		http://bbmri-eric.eu/
BRCA Challenge	Data-Sharing Initiative	GA4GH Demonstration Project	13500	BRCA1 and BRCA2 variants, assertions of disease risk, and evidence for pathogenicity classification	Variants	2015 - ongoing	Cancer	https://genomicsandhealth.org/work-products-demonstration-projects/brca-challenge-0
Broad Genomics Data Donation Platform	Repository Research Project	Cloud-based clinical and genomic data platform			Whole-genome/ exome sequence	2016 - ongoing	N/A	https://www.broadinstitute.org/scientific-community/science/platforms/genomics/genomics-platform
Broad-Novartis Cancer Cell Line Encyclopedia (CCLE)	Research Project	Collaboration between research/academic groups	1074	Cancer cell lines			Cancer	http://www.broadinstitute.org/ccle/about
Cancer Core Europe	Consortium	International consortium of research institutes and universities	60000	Newly diagnosed patients per year		2014 - ongoing	Cancer	http://www.cancercoreeurope.eu
Cancer MoonShot 2020	Consortium	National public-private consortium (government, industry, academia)	20000	Cancer patients (representing 20 tumour types)		2015 - 2020	Cancer	http://www.cancermoonshot2020.org/
Centre for Proteomic & Genomic Research (CPGR)	Organisation/Company	Non-profit company			Whole-exome sequence Variants		Cancer, Other	http://www.cpgr.org.za/
Children's Hospital of Philadelphia Biorepository	Repository	Biobank	8600000	Pediatric biological samples			Cancer, Rare Disease	http://www.research.chop.edu/cores/biorepository/

China Kadoorie Biobank	Repository	Biobank	512000	Unaffected individuals from China (genotyping data are available for ~100,000)		2008 - ongoing	Cancer, Complex Diseases	http://www.ckbiobank.org/site/
Chinese Newborn Sequencing Project	Repository/research project	Database	100000	Newborn babies	Whole-genome/ exome sequence	2016 - 2021	Rare Disease	http://news.xinhuanet.com/english/2016-08/07/c_135572902.htm
Clinical Sequencing Exploratory Research (CSER)	Consortium	National consortium of researchers, HCPs, labs, ELSI groups	6000	Pediatric and adult patients with various phenotypes, healthy adults, and physicians.	Whole-genome/ exome sequence	2009 - 2017	Cancer, Rare Disease, Neurological Disease, Complex Diseases, Other	https://cser-consortium.org/
ClinVar	Data-Sharing Initiative Repository	Database			Variants	2012 - ongoing	Cancer, Rare Disease, Infectious Disease, Neurological Disease, Complex Diseases, Other	http://www.ncbi.nlm.nih.gov/clinvar/
Critical Assessment of Genome Interpretation (CAGI)	Research Project	International research project / collaboration			Whole-genome/ exome sequence, Single gene, Variants	2012 - ongoing	Cancer, Rare Disease, Other	https://genomeinterpretation.org
DECIPHER	Data-Sharing Initiative Repository	Genomics interpretation and data-sharing platform and database	21475	International patients	Variants	2005 - ongoing	Rare Disease, Neurological Disease, Other	https://decipher.sanger.ac.uk/
deCODE Genetics	Organisation/Company	Private company	500000	International participants	Whole-genome/ exome sequence Variants	1996 - ongoing	Cancer, Complex Diseases	http://www.decode.com/
East London Genes & Health	Research Project	Not-for-profit research project	100000	Unaffected individuals from East London, of Pakistani or Bangladeshi heritage	Whole-exome sequence		Complex Diseases	http://www.genesandhealth.org/

Electronic Medical Records and Genomics (eMERGE)	Consortium Repository Research Project	National network of biorepositories and clinical sites	55028	Patients	Variants	2007 - ongoing	Other	https://emerge.mc.vanderbilt.edu/
ELIXIR	Organisation/Company	Inter-governmental organisation				2013 - ongoing	Rare Disease, Other	https://www.elixir-europe.org/
ENIGMA Consortium	Consortium	International consortium of researchers		Patients and relatives	Variants RNAseq	2009 - ongoing	Cancer	http://enigmaconsortium.org/
European Network for Genetic and Genomic Epidemiology (ENGAGE)	Research Project	International consortium of researchers and pharmaceutical companies	600000	Individuals	Variants	2008 - 2013	Complex Diseases, Other	http://www.euengage.org/
Exome Aggregation Consortium (ExAC)	Consortium	International consortium of researchers	60706	Unrelated individuals sequenced through various disease-specific and population genetic studies (not including severe paediatric disease patients)	Whole-exome sequence	2014 - ongoing	N/A	http://exac.broadinstitute.org/
FINDbase	Data-Sharing Initiative Repository		100000	Individuals from 92 populations worldwide, providing information on 3,800 disease-causing mutations across 26 genes	Variants	2006 - ongoing	Rare Disease, Other	http://findbase.org/
France Genomic Medicine 2025	Research Project	National precision medicine initiative			Whole-genome/ exome sequence	2016 - 2025	Cancer, Rare Disease, Complex Diseases	http://www.gouvernement.fr/sites/default/files/document/document/2016/06/22.06.2016_remise_du_rapport_dyves_levy_-_france_medecine_genomique_2025.pdf

Genome Asia 100K	Consortium	Non-profit consortium	100000	Individuals from across Asia	Whole-genome/ exome sequence	2016 - ongoing	Cancer, Rare Disease, Neurological Disease, Complex Diseases	http://genomeasia100k.com/
Genomic Data Commons (GDC)	Repository	Database				2014 - ongoing	Cancer	https://gdc.nci.nih.gov/index.html
Genomics England	Organisation/Company	Nationally-owned company	100000	Genomes derived from 70,000 rare disease and cancer patients and their relatives	Whole-genome/ exome sequence	2012 - 2017	Cancer, Rare Disease	http://www.genomicsengland.co.uk/
Genomics Research and Innovation Network (GRIN)	Consortium	National paediatric genomic research collaboration	100000	Targeted paediatric populations and familial studies.	Whole-genome/ exome sequence, Gene panel, Variants	2015 - ongoing	Rare Disease	http://grinnetwork.org/
Global Genomic Medicine Collaborative (G2MC)	Non/profit organisation						N/A	https://g2mc.github.io/
GoT2D	Consortium	International consortium of researchers		Multiple case-control cohorts	Whole-genome/ exome sequence Variants	2015 - ongoing	Complex Diseases	http://www.type2diabetesgenetics.org/projects/got2d
GTEx	Research Project	Collaboration among academic, government, and private sector scientists	8555	Samples, derived from 544 donors	Whole-genome/ exome sequence Variants		Other	http://www.gtexportal.org/home/
H3Africa	Consortium, Data-Sharing Initiative, Research Project	International consortium of researchers	60000		Whole-genome/ exome sequence Variants	2012 - ongoing	Cancer, Rare Disease, Infectious Disease, Neurological Disease, Other	http://h3africa.org/

Human Genome Variation Society (HGVS)	Consortium	International consortium of researchers					N/A	http://www.hgvs.org/
Human Longevity, Inc. (HLI)	Organisation/Company				Whole-genome/ exome sequence	2014 - ongoing	Cancer, Neurological Disease, Complex Diseases, Other	http://www.humanlongevity.com/
Implementing Genomics in Practice (IGNITE)	Consortium	National research network/program	73000		Gene panel Variants	2013 - 2018	Cancer, Rare Disease, Other	https://www.ignite-genomics.org/
International Cancer Genome Consortium (ICGC)	Consortium	International research consortium				2008 - ongoing	Cancer	https://icgc.org/
International Genomics of Alzheimer's Project (IGAP)	Consortium	International research consortium	40000	Patients with Alzheimer's disease		2011 - ongoing	Neurological Disease	http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php
International Inflammatory Bowel Disease Genetics Consortium (IIBDGC)	Consortium	International research consortium					Complex Diseases	http://www.ibdgenetics.org/
International Multiple Sclerosis Genetics (IMSG) Consortium	Consortium	International research consortium	50000	Patients with multiple sclerosis	Variants	2003 - ongoing	Complex Diseases	http://imsgenetics.org/
International Parkinson's Disease Genomics Consortium (IPDCG)	Consortium	International research consortium					Neurological Disease	http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000918.v1.p1
International Rare Diseases Research Consortium (IRDiRC)	Consortium	International public-private consortium (government, academia, industry, patient organisations)				2012 - ongoing	Rare disease	http://www.irdirc.org/
International Stroke Genetics Consortium (ISGC) Portal	Consortium	International consortium of researchers			Whole-genome/ exome sequence, Single gene, Variants	2007 - ongoing	Neurological Disease	http://www.strokegenetics.org/

Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH)	Repository Research Project	Database / regional research project	500000	Health plan members at Kaiser Permanente		2007 - ongoing	Cancer, Neurological Disease, Complex Diseases	https://www.dor.kaiser.org/external/DORExternal/rpgeh/index.aspx
Kaviar	Repository	Database of known human variants			Variants	2010 - ongoing	N/A	http://db.systemsbiology.net/kaviar/
Leiden Open Variation Database (LOVD)	Consortium Repository	Database	290000		Variants	1995 - ongoing	Cancer, Rare Disease, Infectious Disease, Neurological Disease, Complex Diseases, Other	http://databases.lovd.nl/shared/
Lung Genomics Research Consortium (LGRC)	Consortium	National research network/program				2009 - ongoing	Complex diseases	http://www.lung-genomics.org/
Matchmaker Exchange	Data-Sharing Initiative				Variants	2013 - ongoing	Rare disease	http://www.matchmakerexchange.org/
Medulloblastoma Advanced Genomics International Consortium (MAGIC)	Consortium	International research network/program	300	High-risk paediatric medulloblastoma cases			Cancer	http://www.bcgsc.ca/project/magic
Million Veteran Program	Research Project	National research network/program	1000000	Veterans from USA	Variants	2011 - ongoing	Cancer, Neurological Disease, Complex Diseases, Other	http://www.research.va.gov/mvp/
MSSNG	Data-Sharing Initiative Research Project	International research and data-sharing project	10000	Families affected with autism	Whole-genome/ exome sequence		Neurological Disease	https://www.mss.ng/
Multiple Myeloma Genomics Initiative	Research Project	Regional research project and data portal	17123	Multiple myeloma samples			Cancer	https://www.broadinstitute.org/mmgp/about

MyCode Community Health Initiative	Repository Research Project	Regional research project/biobank	250000	Patients from Geisinger	Whole-genome/ exome sequence Variants	2014 - 2019	N/A	https://www.geisinger.edu/en/research/departments-and-centers/genomic-medicine-institute/mycode-health-initiative
MyGene2	Data-Sharing Initiative Organization/ Company Repository Research Project	Not-for-profit	500	MyGene2 profiles are created by families with a rare disease or condition, clinicians/genetic counselors on behalf of such families, or researchers studying a rare condition.	Whole-genome/ exome sequence Gene panel Variants RNAseq mtDNA	2016 - ongoing	Rare Disease	https://mygene2.org/
Newborn Sequencing in Genomic Medicine and Public Health (NSIGHT)	Research Project	National research network/program made up of 4 separate NIH cooperative agreement awards.				2010 - ongoing	N/A	https://www.genome.gov/27558493/
openSNP	Repository	Database / Not-for-profit project	2500	2500	Variants	2012 - ongoing	N/A	https://opensnp.org/
PersonalGenomes.org	Repository	Database / charitable organisation				2005 - ongoing	N/A	http://www.personalgenomes.org/
Pharmacogenomics Research Network (PGRN)	Consortium Research Project	NIH funded network of research projects	200		Variants	2000 - ongoing	Other	http://www.pgrn.org/
Precision Link	Data-Sharing Initiative Repository Research Project Other		200000	Targeted pediatric populations.	Whole-genome/ exome sequence, Gene panel, Single gene, Variants	2015 - ongoing	Cancer, Rare Disease, Infectious Disease, Neurological Disease, Complex Diseases, Other	http://www.childrenshospital.org/research-and-innovation/innovation/initiatives/precision-link
Precision Medicine Initiative	Research Project	National research network/program	1000000	Participants from USA		2015 - ongoing	Complex Diseases Other	https://www.nih.gov/precision-medicine-initiative-cohort-program

Psychiatric Genomics Consortium (PGC)	Consortium	International consortium of researchers	170000	Psychiatric patients	Variants	2007 - ongoing	Neurological Disease	http://www.med.unc.edu/pgc
Public Health Genomics Knowledge Base (PHGKB)	Repository	Database					Cancer, Rare Disease, Neurological Disease, Complex Diseases, Other	https://phgkb.cdc.gov/GAPPKB/phgHome.do?action=home
Public Population Project in Genomics and Society (P3G)	Consortium Organization/ Company	Not-for-profit corporation					Cancer, Rare Disease, Infectious Disease, Neurological Disease, Complex Diseases, Other	http://www.p3g.org/
Qatar Genome Project	Research Project	National genomics and precision medicine initiative	1161		Whole-genome sequence, Whole-exome sequence	2013 - ongoing	Rare Disease	http://www.qatarbiobank.org.qa/qatar-genome/about-qatar-genome-programme
RD-Connect	Data-Sharing Initiative	Infrastructure hosting genomics data, providing online interface for gene discovery, and linking data from multiple sources.	2500		Whole-genome/ exome sequence, Gene panel, RNAseq	2012 - ongoing	Rare Disease	http://rd-connect.eu/
Reference Variant Store (RVS)	Repository	Database			Variants		Cancer, Rare Disease, Complex Diseases, Other	https://rvs.u.hpc.mssm.edu/
Repositive	Organisation/Company	One portal to search the world's human genomic data			Whole-genome/ exome sequence, Gene panel, Single gene, Variants, RNAseq, mtDNA	2014 - ongoing	Cancer, Rare Disease, Infectious Disease, Neurological Disease, Complex Diseases, Other	http://discover.repositive.io

Resilience Project	Research Project	International research project	589306		Whole-genome sequence Whole-exome sequence Variants	2016 - ongoing	Other	http://resilienceproject.com/
Saudi Human Genome Program	Research Project	National genomics and precision medicine initiative	100000	Patients and unaffected individuals from Saudi Arabia	Whole-genome/ exome sequence	2013 - 2018	Rare Disease, Complex Diseases	http://shgp.kacst.edu.sa/site/
Scottish Genomes Partnership (SGP)	Consortium Data-Sharing Initiative Research Project	National research network/program	3000	Individuals from Scotland	Whole-genome/ exome sequence	2016 - ongoing	Cancer, Rare Disease, Neurological Disease, Other	http://www.scottishgenomespartnership.org/
Sequence Bio 100K Genome Project	Research Project		100000	Individuals from Newfoundland & Labrador, Canada	Whole-genome/ exome sequence	2016 - ongoing	Rare Disease	http://www.sequencebio.co/#homepage
Stanley Center for Psychiatric Research	Research Project	National research network/program			Variants	2007 - ongoing	Neurological Disease	http://www.broadinstitute.org/scientific-community/science/programs/psychiatric-disease/stanley-center-psychiatric-research/stanle
T2D-GENES	Consortium	International consortium of researchers	10600		Whole-genome sequence Whole-exome sequence	2014 - ongoing	Complex Diseases	http://www.type2diabetesgenetics.org/projects/t2dGenes
TBResist	Consortium	International consortium of researchers	2600				Infectious Disease	http://projects.iq.harvard.edu/tbresist/home
The Clinical Genome Resource (ClinGen)	Data-Sharing Initiative	National research network			Single gene Variants	2013 - ongoing	Cancer, Rare Disease, Neurological Disease, Complex Diseases, Other	https://www.clinicalgenome.org/

The Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA)	Consortium	International consortium of researchers	46000	BRCA1 (n=>28,500) and BRCA2 (n=17,500) mutation carriers	Variants	2007 - ongoing	Cancer	http://apps.ccge.medschl.cam.ac.uk/consortia/cimba/about/about.html
Tohoku Medical Megabank Project	Repository Research Project	Repository; Biobank and Database; Research Project; National Project	150000	A population-based adult cohort study and a birth and three-generation cohort study.	Whole-genome/ exome sequence Gene panel Variants	2012 - 2022	Rare Disease, Infectious Disease, Neurological Disease, Complex Diseases, Other	http://www.megabank.tohoku.ac.jp/english/
Trans-Omics for Precision Medicine (TOPMed) Whole-Genome Sequencing project	Consortium Research Project	National research consortium	20000	Individuals with heart, lung, blood, and sleep disorders from across 26 NHLBI studies		2014 - ongoing	2014 - ongoing	http://www.nhlbi.nih.gov/research/resources/nhlbi-precision-medicine-initiative/topmed/wgs
Transforming Genetic Medicine Initiative (TGMI)	Consortium	International research consortium				2016 - ongoing	N/A	http://www.thetgmi.org/
Treehouse Childhood Cancer Initiative	Consortium Data-Sharing Initiative Research Project				Whole-genome/ exome sequence, Gene panel, Variants, RNAseq		Cancer	https://treehousegenomics.soe.ucsc.edu/
Type 2 Diabetes Knowledge Portal	Data-Sharing Initiative Repository	Open-access database, and tools for custom analysis.			Whole-genome/ exome sequence, Variants	2015 - ongoing	Complex Diseases	http://www.type2diabetesgenetics.org/
Ubiquitous Pharmacogenomics (U-PGx)	Consortium	International research project					Other	http://upgx.eu/
UK Biobank	Consortium Repository Research Project	Non-profit	500000	Individuals, aged 40-69 years, from across the UK	Variants	2006 - ongoing	Cancer, Neurological Disease, Complex Diseases	http://www.ukbiobank.ac.uk/

UK10K	Research Project	National research network/program	10000	Whole-genome/ exome sequence	2010 - ongoing	Rare Disease, Complex Diseases	http://www.uk10k.org/
Undiagnosed Diseases Network (UDN)	Research Project	National research network/program	8000	Whole-genome/ exome sequence	2015 - ongoing	Rare Disease	https://undiagnosed.hms.harvard.edu/
Universal Mutation Database (UMD) and BRCA Share	Consortium Repository	Consortium Repository		Variants	1992 - ongoing	Cancer, Rare Disease	http://www.umd.be/
Vanderbilt's BioVU	Repository	Biobank	215000		2007 - ongoing	Other	https://victr.vanderbilt.edu/pub/biovu/