Bioinformatics for Human Genetics: Promises and Challenges

Annika Elisabeth Lindblom, Peter N. Robinson

Human Mutation

Bioinformatics for Human Genetics: Promises and Challenges

Journal:	*Human Mutation*
Manuscript ID:	humu-2011-0043.R1
Wiley - Manuscript type:	Overview
Date Submitted by the Author:	28-Jan-2011
Complete List of Authors:	Lindblom, Annika; Karolinska Institutet, Department of Molecular Medicine & Surgery Robinson, Peter; Charite University Hospital, Humboldt University Berlin, Institute of Medical Genetics
Key Words:	Bioinformatics, HVP, genetics, ontology

Humu-2011-0043

OVERVIEW

Bioinformatics for Human Genetics: Promises and Challenges

Annika Lindblom[1] and Peter N. Robinson[2,3,4]

Deleted: ,

1Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm S17176

Deleted: ,

Deleted: Sweden Email: Annika.Lindblom@ki.se

2) Institute for Medical Genetics and Human Genetics, Charité - Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany

3) Berlin-Brandenburg Center for Regenerative Therapies (BCRT), Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany

4) Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany

Deleted: Email: peter.robinson@charite.de

Corresponding Authors:

Peter Robinson

E-mail: peter.robinson@charite.de

Annika Lindblom

Email: Annika.Lindblom@ki.se

Formatted: Font: Bold

Abstract

Recent developments, including next-generation sequencing (NGS), bio-ontologies and the Semantic Web, and the growing role of hospital information technology (IT) systems and electronic health records, amass ever-increasing amounts of data before human genetics scientists and clinicians. However, they have ever-improving tools to analyze those data for research and clinical care. Correspondingly, the field of bioinformatics is turning to research questions in the field of human genetics, and the field of human genetics is making greater use of bioinformatic algorithms and tools. The choice of “Bioinformatics and Human Genetics” as the topic of this special issue of *Human Mutation* reflects this new importance of bioinformatics and medical informatics in human genetics. Experts from among the attendees of the Paris 2010 Human Variome Project symposium provide a survey of some of the “hot” computational topics over the next decade. These experts identify the promise — what human geneticists who are not themselves bioinformaticians stand to gain — as well as the challenges and unmet needs that are likely to represent fruitful areas of research.

Key words: Bioinformatics, HVP, genetics, ontology

Deleted: Computers have been used for decades to get practical work done in the field of human genetics for the benefit of patients and the advancement of science. To name only two examples, computer algorithms and programs for linkage analysis have formed the foundation of most disease-gene discovery projects, and databases of clinical findings have been widely used to support diagnostic decisions in the field of dysmorphology and general human genetics. A number of trends and r

Deleted: ,

Formatted: Highlight

Deleted: ever increasin

Deleted: g

Deleted:

Deleted: are confronting scientists and clinicians with ever increasing amounts of

Deleted: but

Deleted: also with

Deleted:

Deleted: these

Deleted: increasingly

Deleted: increasingly

Deleted: H

Deleted: G

Deleted: , and we have been fortunate to be able to recruit some of the leaders in this field as authors. There are now so many areas in which Human Genetics stands to profit from improved computational support, that it would be impossible to include all relevant topics in a single issue. We have therefore put

Formatted: Highlight

Deleted: together a highly subjective

Deleted: the issues that we think will be

Deleted: and have tried to

Introduction

Formatted: Font: Bold

Formatted: Font: Bold

Computers have been used for decades to get practical work done in the field of human genetics for the benefit of patients and the advancement of science. Computer algorithms and programs for linkage analysis have formed the foundation of most disease-gene discovery projects, and databases of clinical findings have been widely used to support diagnostic decisions in the fields of dysmorphology and general human genetics. A number of trends and recent developments, including next-generation sequencing (NGS), bio-ontologies and the Semantic Web, and the ever-increasing role of hospital information technology (IT) systems and electronic health records, are confronting human genetics researchers with growing amounts of information that can be difficult to interpret. However, the rapid development of cutting-edge bioinformatic tools is allowing clinicians and scientists to analyze these data more efficiently and accurately, with the end result being improved clinical care and research.

This special issue of *Human Mutation* reflects the views of a number of international leaders in bioinformatics as applied to human genetics. Many were recruited from among the esteemed attendees of the May 2010 Human Variome Project meeting at UNESCO in Paris (Kohonen-Corish et al., 2010). There are now so many areas in which human genetics stands to profit from improved computational support, that it would be impossible to include all relevant topics in a single special issue. The articles presented aim to identify the promise — what human geneticists who are not

Formatted: Font: Italic

Formatted: Font: Arial

themselves bioinformaticians stand to gain — as well as the challenges and unmet needs that are likely to represent fruitful areas of research in the not-too-distant future.

Standards

One of the major challenges in biomedical informatics is to overcome the lack of standards for the collection, manipulation and transmission of information in biomedical research. The need for standards is especially pressing in the field of human genetics with its focus on DNA sequences and descriptions of rare phenotypic abnormalities that tend not to be adequately represented in general-use biomedical vocabularies such as SNOMED.

Ontology is the philosophical discipline which studies the nature of existence and aims to understand what things exist in the world, how these things can be divided into categories according to distinguishing characteristics, and how these categories are related to one another. In computer science, the word ontology is used with a similar meaning, to refer to a structured, computable representation of knowledge within certain domains, such as science, government, industry, and healthcare. Computationally, an ontology provides a classification of the entities within a domain, each associated with a human-readable term and often additional synonyms and other information. Furthermore, an ontology specifies the semantic relationships between these entities. Thus, an ontology can be used to define a standard, controlled vocabulary for a scientific field as well as to enable reasoning and inference across this knowledge to drive new hypotheses and further research.

One of the most widely used ontologies is the Gene Ontology (GO), which provides structured, controlled vocabularies and classifications for several domains of molecular and cellular biology and is structured into three domains, molecular function, biological process and cellular component (Ashburner et al. 2000). GO has been extensively adopted by the molecular biology community as a kind of *lingua franca* for describing the biological function of gene products in humans and model organisms using a consistent and computable language. Although the GO was originally developed primarily to provide a means for integration, retrieval, and computation of data, it is now commonly used to help understand the results of high-throughput expression profiling experiments, network modeling, analysis of semantic similarity, and many other applications.

Clinical medicine and research in human genetics have not yet incorporated ontologies and information technology to the same extent. Consider for a minute how useful it would be if the genotype and phenotype data for all published studies on genetic disorders over the last several decades had been recorded in a standard fashion in a central database in the same way as molecular geneticists have made sequence data freely available. The development and widespread adoption of a number of standards will be required to achieve this in the next decades.

Victor McKusick provided a classic description of the challenges of the nosology of genetic disease in 1969, which in many ways boils down to a dispute between “lumpers” and “splitters” (McKusick, 1969). This debate is impossible to resolve because, among other things, of the numerous instances of “many from one” (multiple phenotypes from different mutations in the same gene) and “one from many” (the same

Deleted: V A

phenotype from mutations in two or more separate genes) (McKusick, 2007). This means that no one classification scheme is likely to satisfy all needs. There are currently two important schemes that should be widely adopted (and computational links will need to be made between them). The first naming scheme for genetic diseases which was developed over decades by Victor McKusick and colleagues at Online Mendelian Inheritance in Man (OMIM) has long been used in the research community. Ada Hamosh and colleagues provide an overview of the logic behind the OMIM nomenclature in this issue (Hamosh et al. 2011). Although the OMIM nomenclature has been widely adopted in the research community, it has not been used extensively in hospital IT systems.

Deleted: Victor A

Deleted: over decades

Orphanet, which is the most important portal especially for clinical aspects of rare disease (http://www.orpha.net), offers information for professionals and the general public on rare diseases, orphan drugs, expert centers, clinical trials, patients' organizations, and other related topics. The International Classification of Diseases, version 10 (ICD-10) is the international standard diagnostic classification of the World Health Organization (WHO) for epidemiological and health management purposes as well as for clinical use. Only about 240 hereditary diseases have a specific code in the ICD 10, and no systematic classification scheme was used. Therefore, Ségolène Aymé and colleagues at Orphanet are developing a clinically oriented nosology of rare diseases that will be adopted by the ICD-11, and thus is likely to have an enormous influence on human genetics simply by making the great majority of rare diseases visible for statistics and hospital IT systems that make use of ICD-11. OMIM provides separate identifiers for each of the distinct subtypes of heterogeneous diseases (for

Deleted: offering

Deleted: information

instance, there are identifiers for each of the forms of long QT syndrome, which are caused by mutations in distinct genes but characterized by highly similar phenotypes). On the other hand, the Orphanet nosology will generally provide a single identifier for such syndromes, since this is generally more useful in a clinical setting. Thus, it seems fair to regard Orphanet as the lumpers, and OMIM as the splitters within the field of nosology for human genetics.

For many kinds of research in human genetics, it is important to record the phenotypic abnormalities in addition to the specific disease entity or diagnosis. Human genetic diseases can be clustered on the basis of their phenotypic similarities, and it has been proposed that the clustering reflects true biological relationships of the genes involved (Oti et al., 2007). Several mutually incompatible nomenclatures are in use to describe phenotypic abnormalities (signs, symptoms, laboratory and imaging findings) in the field of human genetics, including those of the commercial programs London Dysmorphology Database and POSSUM. One of the guest editors of this issue has led the development of the Human Phenotype Ontology (Robinson et al., 2008), which currently provides over 10,000 terms to describe individual phenotypic abnormalities and over 50,000 annotations to hereditary diseases. The HPO owes an enormous debt to OMIM, from which much of the initial information in the HPO was derived. Continued development of the HPO, which is freely available to all, is being conducted with the help of a number of groups including OMIM, Orphanet, and other interested clinicians. The further involvement of experts from the community is highly welcome. The HPO has been adopted by the International Standard Cytogenomic Array (ISCA) Consortium, and the implementation of the HPO by several other groups is currently in the planning stage.

The third kind of standard that is absolutely essential for human genetics regards the nomenclature of mutations. The Human Genome Variation Society (HGVS) has developed a nomenclature for mutations that is now required in this and in many other journals (www.hgvs.org/mutnomen). In this issue, den Dunnen et al., (2011) describe an extension of this standard for several kinds of complex mutation. These standards deserve broad support in the community for reporting and publishing scientific findings. Finally, there is currently no widely accepted way of reporting all sequence variants found in an exome or genome. The recently presented Genome Variation Format (GVF), which uses terms from the Sequence Ontology to describe genome variation data (Reese et al., 2010), seems to have the potential to meet this need.

Deleted: and colleagues

Deleted: (den Dunnen et al. 2011)

Deleted: ¶

Biomedical ontologies and standards exist to serve integration of data, their conditions of success can usefully be compared to those of a telephone network. The utility of an ontology or of a telephone network is critically related to the number of users. The more datasets are annotated to a given set of interoperable ontologies, the higher the value not only of the ontologies but also of the annotated data, which can now be integrated and compared with ever more comprehensive datasets (Smith et al., 2010). This has been one of the many reasons for the continued success of the Gene Ontology, which in 2000 comprised about 2500 terms and was used to annotate only yeast, mouse, and fly genes, and has subsequently grown and matured such that it now contains over 33 thousand terms and is used to annotate all major model organisms.

Deleted: ¶
¶

Next-Generation Sequencing (NGS)

NGS is revolutionizing biomedical research in many ways, not least by the ability to resequence the great majority of coding exons in the human genome by hybridization-based enrichment techniques followed by NGS analysis, a procedure which has come to be called exome sequencing. Multiple novel disease genes have been discovered following the pioneering work of Ng et al. (2009), and it is widely expected that exome sequencing and related technologies will dramatically accelerate the pace in the discovery of the disease genes for the estimated 3,750 Mendelian disorders whose molecular basis is currently unknown. Presumably, our understanding of the genetic causes of oligogenic and polygenic disorders will also benefit in a similar way.

Deleted: and coworkers (Ng

In 1965, Gordon Moore described a trend in computing hardware, according to which the number of transistors on a chip roughly doubles every two years. A similar trend is likely to apply to DNA sequencing technologies, and the primary challenge in diagnostics in human genetics is likely to shift from the mere identification of sequence variants to the interpretation of the variants. Bioinformatics plays a key role at all levels of data analysis and interpretation. In addition to alignment of short reads to the reference genome and variant calling (Li and Homer, 2010), a number of integrative analysis procedures have been developed to deal with the primary challenge in the medical interpretation of exome data, which is simply making sense of the sheer number of variants found in each individual patient. It has been surprising to discover just how common sequence changes with presumably deleterious effects on protein function are in a typical human genome. For instance, it has been reported that on

Deleted: &

average, each genome carries 165 homozygous protein-truncating or similar variants in genes representing a diverse set of pathways (Pelak et al., 2010). This means that the mere finding of a sequence variant that appears to be a pathogenic mutation cannot be taken as proof that the change is causally related to the disease being investigated. Groups involved in exome sequencing have developed a number of strategies to deal with this. The "intersection" approach searches for rare, potentially deleterious variants affecting the same gene in multiple unrelated individuals with the same disease (Ng et al., 2010) Family-based approaches use linkage analysis in addition to NGS analysis in order to narrow down the candidate region in families (Volpi et al., 2010). The group of one of the guest editors of this special issue has developed procedures to perform linkage analysis directly in the exome sequence data of individuals affected by autosomal recessive or X-linked diseases by using a Hidden Markov Model approach to identify regions identical by descent in all affected individuals (Krawitz et al., 2010). Recently, it was shown that exome sequencing can be used to identify *de novo* mutations in sporadic cases of mental retardation (Vissers et al., 2010).

It is to be expected that new algorithms and analysis procedures will continue to be developed for making sense of exome data, but it is still unclear how to bring exome sequencing into routine clinical care outside of big research centers for human genetics in the near future because, among other things, of the complexity of the computational analysis required for this type of data. It seems more likely that targeted resequencing of panels of genes known to be involved in genetically heterogeneous disorders such as hereditary deafness or cardiomyopathy will be the first widespread application of NGS technologies in routine molecular genetic diagnostics. One key point for any technology in which variants in tens, hundreds, or thousands of genes are interrogated is the use of databases to help in interpretation and reporting. In addition to centralized registries for mutations and phenotypes, which will be covered below, specialized software for analyzing and reporting variants will be needed to keep up with the increased demands of NGS data.

Deleted: ¶
¶

Human Genetics and Personalized Medicine

One of the articles in this issue brings up the urgent need of placing human genetics in the perspective of personalized medicine (Ullman-Culliaire et al., 2011). In fact, it is the rapid and immense progress in the field of human genetics in combination with an equally rapid development in bioinformatics which calls for the establishment of electronic health records, a functioning health care IT and communication with genetic/genomic information on an individual level to be able to offer optimal health care.

In contemporary medical practice, a patient's genetic/genomic data is becoming ever more important for clinical decision making, in disease risk assessment, diagnosis, and drug therapy. The electronic health record will typically need to include a segment of key elements for the individual, including phenotypic and other clinical data, previous and current diagnoses, previous and current medication, accessibility to genetic data such as genetic test results from individuals, encoded using standard formats discussed in the articles (Taschner et al., 2011).

For some fields including especially oncology, there is a need to distinguish between tissues from which samples have been taken. This is because not only constitutional genetic information, but also genetic information related to malignant tissues are of extreme importance for clinical decision making. The amount of information necessary today for optimal decision making requires computer-based management of data rather the traditional paper based system not only for security but also because of the complexity of medicine today. The electronic management will all the use of tools to improve patient safety such a triggering for guidelines to choose the right drug and avoid adverse drug effects such as patient allergy or drug interaction, or by receiving pharmacogenetic-pharmacogenomic information important for drug dosing, efficacy or toxicity. The rapid development of techniques in the area of genome analysis has facilitated identification of new pharmacogenomic biomarkers that can be used as predictive tools for improved drug response and fewer adverse drug reactions. Such biomarkers mainly originate from genes encoding drug-metabolizing enzymes, drug transporters, drug targets and human leukocyte antigens (Ingelman-Sundberg et al., 2011).

Electronic Patient Records and Hospital Information Technology

An electronic health record will also facilitate control of clinical work-flows and interaction with genetic/genomic, pharmacogenetic/genomic databases, and databases to be used for research purposes. BioGrid Australia, described in this issue (Merriel et al., 2011), is an example of this type of warehouse database to be used for research. It is a federated data linkage and integration structure and integration infrastructure that uses the internet to enable patient specific information to be utilized for research from multiple databases of various types, from a range of diseases and across more than 20 health services, universities and medical institutes. There is an open access for any researcher to view available data. Data is transported every night from the local databases into BioGrid and researchers only have access to de-identified data. Research can also be done using electronic health records directly, in particular when these are in communication with databases with genetic/genomic information. There is much to be gained by having as few databases as possible and if scientific studies used clinical information from the primary health records it would be safer, cheaper and give more accurate data. It will, however, require that data is requested in a systematic and ethically controlled fashion and that data is delivered to researchers in a de-identified format. Medical records in general, electronic or not, harbor most information relevant for one individual. However, not all information is relevant, it is not structured, and typically it is divided in many different caretakers' registers. A nationwide system of electronic health records with access from all registered health care services within the country (and in the future perhaps also outside the country) would serve the patient best. This database should then be able to communicate with all necessary databases, such as those with genetic-genomic information for individuals. If for practical reasons this is not possible, all existing health record electronic systems should be able to communicate with each other and all other necessary databases.

Deleted: ,

In contrast to general medicine, clinical genetics today mostly uses electronic health care records. Data included is normally limited to what is relevant in clinical genetic testing related to inherited disease and there is not often a systematic description of phenotypic and clinical data. For instance, if a woman with a family history of breast cancer is being referred for BRCA testing, there is no need for any other information and thus, the only data will be the resulting BRCA status. The extreme development in genetics and bioinformatics forced clinical genetics to deal with ever increasing amounts of data and to use bioinformatics and databases such as OMIM, ORPHANET and LSDBs over the internet on a routine basis in clinical care. In contrast to general medicine which treats individuals, clinical genetics deals with whole families to diagnose, treat and prevent disease and thus for diagnosis or interpretation of test results one cannot always rely on phenotypic/clinical data from a single patient but also from his or her relatives. This has created a need for various bioinformatic tools to facilitate the diagnosis of genetic disease such as CGEN and SISA (Möller et al., 2011a; Möller et al., 2011b) and also the Danish for familial colorectal cancer used for diagnosis and clinical care/prevention platform (Bernstein et al., 2011).

Deleted: -

Deleted: -

Deleted: ¶
¶

The Semantic Web for Human Genetics

The World-Wide Web (WWW) grew out of proposals by Tim Berners-Lee in 1989 and 1990 to create a web of hypertext documents using hyperlinks to allow users to browse between documents located on different web servers. Current Web browsers can easily parse Web pages in order to create a nice visual layout and links to other Web pages, but they have no reliable way of recognizing the meaning of the Web pages they present, much less of roaming the net to harvest the information needed to answer a question. The Semantic Web was proposed by Berners-Lee in 1999 as a framework in which computers are capable of analyzing the meaning -- the semantic content -- of the data on the Web in order to act as „intelligent agents.“ The WWW Consortium (W3C) has created languages for the Semantic Web that are specifically designed as a sort of semantic markup language for data: RDF, RDFS, and OWL. Especially OWL has also come to play an important role as a language for bio-ontologies.

It seems certain that the role of the Semantic Web in biomedical research and in human genetics will increase in the next decade. To harness the full power of the Semantic Web in a similar way for human genetics, formal representations of biological, genetic, and medical knowledge in the form of ontologies, annotations, and knowledge bases will need further development, and tools and applications for nonspecialists will need to be implemented. One interesting example of what is possible with Semantic Technologies for biomedical research is the Cell Cycle Ontology, for which tools have been developed for browsing, visualizing, advanced querying, and computational reasoning about cell cycle-related molecular network components (Antezana et al., 2009).

Deleted: .

Prioritization of Sequence Variants

As mentioned above, a major challenge in the interpretation of exome and genome data lies in the identification of the actual disease-causing mutation (or mutations) among very numerous candidate variants. Numerous tools have been developed over the last decade or so to predict the pathogenicity of nonsynonymous substitutions in protein coding sequences on the basis of attributes such as evolutionary conservation, physicochemical characteristics of the wild-type and mutant amino acids, and protein structure, including PolyPhen (Ramensky et al., 2002) , ProPhylER (Binkley et al., 2010) , and many others. Other groups have developed tools that integrate and visualize the output of multiple prediction routines (Schwarz et al., 2010). While these tools are highly useful for the prioritization of sequence variants, they are currently not reliable enough to make a definitive diagnosis in the setting of exome and genome analysis. One of the most important areas of research for bioinformatics in human genetics will be the development of analysis procedures that will combine the output of prioritization tools with phenotypic and pedigree data as well as data from databases of neutral variants and of disease-causing mutations in order to improve and facilitate the interpretation of exomic and genomic data in a clinical setting. Another interesting area of research will be the development of computational tools that use phenotype data from model organisms to help understand human disease.

Registries for mutations

it would be extremely useful to have a centralized, open-access database of genotypes and phenotypes that could be used to assess the significance of sequence variants for individual patients in a diagnostic setting. DECIPHER, the Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources, enables clinical scientists worldwide to maintain records of phenotype and chromosome rearrangement for their patients, which enables international data sharing and integration and has helped to delineate a number of new syndromes (Firth et al., 2009). At present, there is no such database for single-gene mutations, and much less for exome and genome data. While the ethical and privacy issues surrounding such databases are substantial and will need to be worked out in the near future, it seems clear that the full potential of exome sequencing for patient care will not be reached without a significant investment in public databases for medically relevant exome data that contains standardized descriptions of the genotype and the phenotype.

The Human Variome Project (HVP; http://www.humanvariomeproject.org/) is the natural successor to the Human Genome Project. It aims to improve human health by the establishment and maintenance of standards, systems, and infrastructure for the worldwide collection and sharing of all genetic variations affecting human disease. The goal is to make internationally available all existing genetic variants and tools to interpret and use these to diagnose, and treat individuals with the existing state-of-the-art knowledge regarding genetic-genomic information. It will do so by data collection into all existing and new locus-, gene- or disease-specific databases where data is reviewed by numerous experts in the field. The collection of data to these LSDBs could sometimes be facilitated through a network of HVP Country Nodes, created in partnership with Human Genetics Societies in individual countries across the globe.

Deleted:)

Deleted: (http://www.humanvariomeproject.org/).

Deleted:

Those nodes are data repositories that are run, managed and funded within each country and which could correspond with local databases and international LSDBs. In this issue is described how to initiate a HVP Country Node (Al Aama et al., 2011). Nodes exist already in Belgium, China, Egypt, Malaysia and Australia.

The GeneInsight Suit, similar to a country node, is an application and networking infrastructure enabling organizations to use a platform of software applications and consists of a laboratory/knowledge management application (GI Lab), a provider communicating with organizations (research or clinical) using GI Clinic (Aronson et al., 2011). This platform could be used to obtain high-quality state of the art reports from genetic testing regarding gene variants in every gene and their relation to disease, drug efficacy and drug response – and also manage to automatically deliver up-dates to these reports later when state of the art changes.

The GEN2PHEN project (http://www.gen2phen.org/), described in this issue by Webb et al. (2011), is another project that aims to assist in organizing the flood of existing genomic information. Within the GEN2PHEN project an evaluation of all existing locus-specific databases showed that there are more than 1,000 LSDBs covering almost 1,000 genes with some genes being represented in more than one LSDB (Webb et al., 2011). It is of value not to have overlapping LSDBs and concentrating the efforts in generating a high quality LSDB with professional curators with expertise in a certain topic will make the best use of existing data. The international society for inherited gastro-intestinal tumors (InSiGHT; http://www.insight-group.org/) is one example when an organization in line with the HVP made an effort to centralize and improve several

Deleted: (Webb et al. 2011)

Deleted: (http://www.gen2phen.org/)

Deleted: ,

existing LSDBs for the genes involved in familial gastrointestinal tumors, such as Lynch syndrome and familial adenomatous polyposis.

Often it is not clear what is relevant to submit to each LSDB. Typically, a genetic variant in a gene is reported related to a certain disease. The patient was screened for the gene because there was a clinical suspicion of the disease and the basis for this was symptoms or a diagnosis in the patient and his relatives. Normally, when the interpretation is a pathogenic mutation, the same variant will not be reported when it is found again in affected or healthy relatives – or even in unrelated subjects. There is an ongoing debate whether to report all genetic variants in all individuals or not. The reason to report all subjects is mainly for scientific purposes. For clinical purposes it is most important to report so called unclassified variants – i.e., those that are difficult to interpret and cannot be unambiguously classified as pathogenic or neutral at the point of diagnosis. To solve the issue of an unclassified variant needs research. Those studies could constitute *in silico* or *in vitro* functional tests, clinical information from many tested subjects, as well as for all relatives or (for not too rare variants), association studies in large cohorts of cases and controls. For scientific purposes it could be of value to obtain data of all individuals with a certain variant – however, it is likely more accurate for a scientist to approach the local databases to obtain full information since the individual health record is probably the source closest to the patient and the most relevant and updated. For clinicians or labs to submit data purely for unknown research interests will also likely result in low-quality or no data, low chance that data will be updated and low chance that data will be associated with relevant clinical parameters - because of low motivation for submitters.

Conclusions

We have mentioned just a few of the more important areas in which the field of bioinformatics is providing an indispensable contribution to research and clinical care in human genetics. The influence of bioinformatics is likely to continue to grow hand in hand with advances in next-generation sequencing and other high-throughput technologies in biomedicine. The full promise of these advances for patient care will only be attainable if geneticists and bioinformaticians are able to cope with the challenges of this kind of data analysis. Training programs in bioinformatics should include human genetics and genomics, and correspondingly, training programs in medicine and human genetics will need to increase coverage of statistics, genomics, and even bioinformatics.

Finally, it will be of utmost importance to create a health care system with electronic health records with systematic coding of clinical/phenotypic as well as genetic/genomic data. This will require an ontology for common clinical parameters and lab data, including genetic/genomic data for the individual and related tissues. It will also be necessary to complement the health care systems with software and tools for clinical decision making and messaging frameworks and to educate the health care system on how to use available systems on the internet.

References

[NOTE TO COPYEDITOR: DO NOT QUERY AUTHORS FOR PAGE NUMBERS IN THE INCOMPLETE REFERENCES, PAGES WILL BE SUPPLIED BY PRODUCTION.]

Al Aama J., Smith T.D., Lo A., Howard H., Kline A.A.,Lange M., Kaput J., Cotton R.G.H. 2011. Initiating a Human Variome Project Country Node. Hum Mutat 32: [**PLEASE COMPLETE REFERENCES**]

Formatted: Highlight

Deleted: **ADD PAGE NUMBERS**

Antezana E, Egaña M, Blondé W, Illarramendi A, Bilbao I, De Baets B, Stevens R, Mironov V, Kuiper M. 2009. The Cell Cycle Ontology: an application ontology for the representation and integrated analysis of the cell cycle process. *Genome Biology* 10:R58.

Aronson S et al 2011. The GeneInsight Suite: A Platform to Support Laboratory and Provider Use of DNA based Genetic Testing. Hum Mutat 32: [**PLEASE COMPLETE REFERENCES**]

Deleted: **ADD PAGE NUMBERS**

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25-29.

Bernstein I et al. 2011. Biomedical Informatics as support to individual healthcare in Hereditary Colon Cancer. The Danish HNPCC-system. Hum Mutat 32: [**PLEASE COMPLETE REFERENCES**]

Deleted: **ADD PAGE NUMBERS**

Binkley, J., Karra, K., Kirby, A., Hosobuchi, M., Stone, E. A., and Sidow, A. 2010. ProPhylER: a curated online resource for protein function and structure based on evolutionary constraint analyses. Genome Research 20:142-154.

den Dunnen J, Fokemma I, Taschner P. 2011. LOVD v.2.0: Facilitating simple creation of gene sequence variation databases. Hum Mutat 32: [**PLEASE COMPLETE REFERENCES**]

Deleted: **ADD PAGE NUMBERS**

Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, Van Vooren S, Moreau Y, Pettett RM, Carter NP 2009. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet 84:524-533.

Hamosh et al 2011. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM(R)). Hum Mutat 32: [**PLEASE COMPLETE REFERENCES**].

Deleted: **ADD PAGE NUMBERS**

Ingelman-Sundberg M et al. Databases in the area of Pharmacogenetics. Hum Mutat 32: [**PLEASE COMPLETE REFERENCES**].

Deleted: **ADD PAGE NUMBERS**

Maija R.J. Kohonen-Corish,1 Jumana Y. Al-Aama,2 Arleen D. Auerbach,3 Myles Axton,4 Carol Isaacson Barash,5 Inge Bernstein,6 Christophe Beroud,7 John Burn,8 Fiona Cunningham,9 Garry R. Cutting,10 Johan T. den Dunnen,11 Marc S. Greenblatt,12 Jim Kaput,13 Michael Katz,14 Annika Lindblom,15 Finlay Macrae,16 Donna Maglott,17 Gabriela Moslein,18 Sue Povey,19 Raj Ramesar,20 Sue Richards,21 Daniela Seminara,22 Marıa-Jesus Sobrido,23 Sean Tavtigian,24 Graham Taylor,25 Mauno Vihinen,26 Ingrid Winship,27 and Richard G.H. Cotton. 2010. How to Catch All Those Mutations-The Report of the Third Human Variome Project Meeting, UNESCO Paris, May 2010. Hum Mutat 31: 1374–1381. [CHRISTINE PLEASE CLEAN UP THIS REF PER SPEC]

Krawitz PM, Schweiger MR, Rödelsperger C, Marcelis C, Kölsch U, Meisel C, Stephani F, Kinoshita T, Murakami Y, Bauer S, Isau M, Fischer A, Dahl A, Kerick M, Hecht J, Köhler S, Jäger M, Grünhagen J, de Condor BJ, Doelken S, Brunner HG, Meinecke P, Passarge E, Thompson MD, Cole DE, Horn D, Roscioli T, Mundlos S, Robinson PN. 2010. Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. Nat Genet 42:827-829.

Li H, Homer N. 2010. A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11:473-483.

Formatted: Font: Arial, 12 pt

Formatted: Line spacing: Double

Formatted: Font: Arial, 12 pt

Formatted: Font: Arial

Formatted: Font: Arial, Highlight

Formatted: Font: Arial

Formatted: Font: Arial, Highlight

Formatted: Font: Arial

McKusick VA. 1969. On lumpers and splitters, or the nosology of genetic disease. Perspectives in Biology and Medicine 12:298-312.

McKusick VA. 2007. Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet 80:588-604.

Merriel R et al. 2011. BioGrid Australia facilitates collaborative medical and bioinformatics research across hospitals and medical research institutes by linking data from diverse disease and data types. Hum Mutat 32: [**PLEASE** COMPLETE REFERENCES].

Deleted: **ADD PAGE NUMBERS**

Möller P et al. 2011a. CGEN – A Clinical GENetics software application. Hum Mutat 32: [**PLEASE** COMPLETE REFERENCES].

Deleted: **ADD PAGE NUMBERS**

Möller P et al. 2011b A Simplified method for Segregation Analysis (SISA) to determine penetrance and expression of a genetic variant in a family. Hum Mutat 32: [**PLEASE** COMPLETE REFERENCES].

Deleted: **ADD PAGE NUMBERS**

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ. 2010. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42:30-35.

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461:272-276.

Oti M, Brunner HG. 2007.. The modular nature of genetic diseases. Clin Genet 71:1-11.

Pelak K, Shianna KV, Ge D, Maia JM, Zhu M, Smith JP, Cirulli ET, Fellay J, Dickson SP, Gumbs CE, Heinzen EL, Need AC, Ruzzo EK, Singh A, Campbell CR, Hong LK, Lornsen KA, McKenzie AM, Sobreira NL, Hoover-Fong JE, Milner JD, Ottman R, Haynes BF, Goedert JJ, Goldstein DB. 2010. The Characterization of Twenty Sequenced Human Genomes. PLoS Genet 6:e1001111.

Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30:3894-3900.

Reese MG, Moore B, Batchelor C, Salas F, Cunningham F, Marth GT, Stein L, Flicek P, Yandell M, Eilbeck K. 2010. A standard variation file format for human genome sequences. Genome Biol 11:R88.

Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. 2008. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet 83:610-615.

Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. 2010. MutationTaster evaluates disease-causing potential of sequence alterations. Nature Methods 7:575-576.

Smith B, Brochhausen M. 2010. Putting biomedical ontologies to work. Methods of Information in Medicine 49:135-140.

Taschner P et al. 2011. Describing complex sequence variants by extending HGVS sequence variation nomenclature. Hum Mutat 32: [**PLEASE** COMPLETE REFERENCES].

Deleted: ADD PAGE NUMBERS

Ullman-Culliaire M et al. 2011. Bioinformatics in implementation of human variants into clinical health care. Hum Mutat 32: [**PLEASE** COMPLETE REFERENCES].

Deleted: ADD PAGE NUMBERS

Vissers LE, de Ligt J, Gilissen C, Janssen I, Steehouwer M, de Vries P, van Lier B, Arts P, Wieskamp N, del Rosario M, van Bon BW, Hoischen A, de Vries BB, Brunner HG, Veltman JA. 2010. A de novo paradigm for mental retardation. Nat Genet 42:1109-1112.

Volpi L, Roversi G, Colombo EA, Leijsten N, Concolino D, Calabria A, Mencarelli MA, Fimiani M, Macciardi F, Pfundt R, Schoenmakers EF, Larizza L. 2010. Targeted next-generation sequencing appoints c16orf57 as clericuzio-type poikiloderma with neutropenia gene. Am J Hum Genet 86:72-76.

Webb et al. 2011. An Informatics Project and Online ‘Knowledge Centre’ Supporting Modern Genotype-to-Phenotype Research. Hum Mutat 32: [**PLEASE** COMPLETE REFERENCES].

Deleted: ADD PAGE NUMBERS