Genome, and beyond
Findings from the Section on Bioinformatics

Y. L. Yip, Section Editor for the IMIA Yearbook Section on Bioinformatics
Merck Serono International S.A., Geneva, Switzerland

Summary
Objectives: To summarize current excellent research in the field of bioinformatics, with an emphasis on those that have direct application in the medical domain.
Method: Synopsis of the articles selected for the IMIA Yearbook 2011.
Results: The selection process for this yearbook's section on Bioinformatics results in six excellent articles highlighting the continuous progress towards a better understanding of human phenotype. Compared to the selection in Yearbook 2010, several key advancements can be noted. First, year 2010 marked the inaugural use of a complete human genome in a clinical context. This proof-of-principle study represents a large step towards personalized medicine. Second, there is a clear trend to understand diseases beyond the genome level, namely to include environmental and epigenetic information. Third, an innovative framework making use of the web to harness participant-driven genotype-phenotype information sets a new scene for conducting research in an era where social media plays an increasingly important role.
Conclusions: The current literature showed that all pieces are now present to enable a much more comprehensive understanding of human diseases and traits, beyond the highly focused genetic or genomic studies seen previously.

Keywords
Bioinformatics, epigenetics, environment, human genome, phenotype

Yearb Med Inform 2011: 156-9

Introduction

It is well acknowledged that many diseases are the result of a complex combination of genetic and environmental factors. While most studies today focus on the genetic aspects, one can witness a rapid expansion of the field of epigenetics or epigenomics in the past few years accompanied by numerous technological and informatics breakthroughs, such as the recent characterization of a human DNA methylome at single nucleotide resolution, and the unveiling of the genome-wide map of chromatin structure [1-3]. Epigenetics is the study of non-DNA sequence-related heredity, and involves mechanisms such as DNA methylation or histone modifications that serve to regulate gene expression without altering the gene sequence. It is certainly perceived as the main missing link between genetics, environment and diseases. To further complete the picture in terms of environmental factors, studies tackling the direct association between environmental exposure and health outcome are also starting to emerge [4].

On the genotype-phenotype association front, important advances continue. The direct exploitation of a complete human genome in a defined clinical context is first reported [5]. A tightening link between biobank materials providing genotype information and electronic patient records with phenotype information can also be observed [6, 7]. Methodology-wise, these advances are further fuelled by innovative conceptual framework of studies [8, 9], and also computational resources made widely available through cloud computing platforms [10].

Best Paper Selection

The best paper selection of articles for the section 'bioinformatics' in the IMIA Yearbook 2011 follows the tradition of previous yearbooks [11-13] in presenting examples of excellent research in bioinformatics that are most relevant to medical informatics. As a result of a comprehensive review process, six articles were selected from international peer-reviewed journals in the fields of medicine, medical informatics, and bioinformatics.

The first paper described the inaugural use of whole-genome information from an individual for clinical evaluation [5]. Ritchie *et al.*, on the other hand, reported the coupling of genetic information from biobanks with the phenotype and clinical information present in electronic patient records [6]. Taken together, these two papers illustrate important steps towards individualized therapy, the holy grail of clinical medicine. In parallel, two papers showed advances achieved in the study of epigenetics [14] and environmental influences [4]. Although not a methodology nor a research paper, the article from Fingerman *et al*. reported the new NCBI resource focusing on epigenomic information [14]. This provides the direct evidence that epigenetics is moving from a secondary role to the center stage of current biomedical research. At the same time, the first Environment-Wide Association Study (EWAS), prototyped by Patel *et al*, completes the scene where the interplay between genetics, epigenetics and environment should

allow a deeper understanding of phenotype and diseases [4]. Finally, two studies demonstrated how novel syndrome-pathogen, as well as genotype and phenotype associations can be gathered from literature-wide scan [9], and participant-driven web surveys [8], respectively. The latter method is particularly interesting as it represents one of the 'crowdsourcing' activities that are now being used to tackle research problems [15].

Table 1 presents the selected papers. A brief summary of the selected best papers can be found in the appendix of this report.

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2011 in the section 'Bioinformatics'. The articles are listed in alphabetical order of the first author's surname.

Section Bioinformatics
▪ Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, Dudley JT, Ormond KE, Pavlovic A, Hudgins L, Gong L, Hodges LM, Berlin DS, Thorn CF, Sangkuhl K, Hebert JM, Woon M, Sagreiya H, Whaley R, Morgan AA, Pushkarev D, Neff NF, Knowles JW, Chou M, Thakuria J, Rosenbaum A, Zaranek AW, Church G, Greely HT, Quake SR, Altman RB. Clinical assessment incorporating a personal genome. Lancet 2010;375(9725):1525-35.
▪ Eriksson N, J. Michael Macpherson JM, Tung JY, Hon LS, Naughton B, Saxonov S, Avey L, Wojcicki A, Pe'er I, Mountain J. Web-Based, Participant-Driven Studies Yield Novel Genetic Associations for Common Traits. PLoS Genetics 2010;6(6):e1000993
▪ Fingerman IM, McDaniel L, Zhang X, Ratzat W, Hassan T, Jiang Z, Cohen RF, Schuler GD. NCBI Epigenomics: a new public resource for exploring epigenomic data sets. Nucleic Acids Research 2011;39 (Database issue):D908-D912.
▪ Patel CJ, Bhattacharya J, Butte AJ. An Environment-Wide Association Study (EWAS) on Type 2 Diabetes Mellitus. PLoS One 2010;5(5):e10746.
▪ Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, Basford MA, Brown-Gentry K, Balser JR, Masys DR, Haines JL, Roden DM. Robust Replication of Genotype-Phenotype Associations across Multiple Diseases in an Electronic Medical Record. The American Journal of Human Genetics 2010;86:560–72.
▪ Sintchenko V, Anthony S, Phan XH, Lin F, Coiera EW. A PubMed-Wide Associational Study of Infectious Diseases. PLos One 2010;5(3):e9535

Conclusions and Outlook

An ever closer convergence of the two rapidly expanding technologies (i.e. biomedical informatics and high-throughput genotyping) highlights the unique opportunity to bring functional genomics to the bedside. As witnessed by this year's best paper selection, it is likely that entire genomic sequences will soon be linked to individual electronic patient records. At the same time, many large medical centers are constructing DNA or tissue biobanks. With the current momentum, it will be interesting to see how well the medical community is prepared to absorb the new tools, new technologies, and informatics solutions. Clearly, factors such as clinical usefulness, cost effectiveness, and ethical implications will also come into play to ultimately decide on the introduction of these developments into routine clinical practices.

Besides genomic, it can be anticipated that epigenomic and environmental information will become more and more available as the related data generation and collection technologies mature. The combination of these three types of information will open new ways of thinking about diseases. Taken together, it will further pave the way to a personalized, proactive healthcare delivery approach.

Acknowledgement

I would like to acknowledge the support of Martina Hutter and the reviewers in the selection process of the IMIA Yearbook.

References

1. Li Y, Zhu J, Tian G, Li N, Li Q, Ye M, et al. The DNA methylome of human peripheral blood mononuclear cells. PLoS Biol 2010; 8(11):e1000533.
2. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol 2010; 28(8):817-27.
3. Zhang Z, Pugh BF. High-resolution genome-wide mapping of the primary structure of chromatin. Cell 2011;144(2):175-86.
4. Patel CJ, Bhattacharya J, Butte AJ. An Environment-Wide Association Study (EWAS) on Type 2 Diabetes Mellitus. PLoS One 2010; 5(5):e10746.
5. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, et al. Clinical assessment incorporating a personal genome. Lancet 2010; 375(9725):1525-35.
6. Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, et al. Robust Replication of Genotype-Phenotype Associations across Multiple Diseases in an Electronic Medical Record. Am J Hum Genet 2010;86:560–72.
7. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics 2010;26(9):1205-10.
8. Eriksson N, Macpherson JM, Tung JY, Hon LS, Naughton B, Saxonov S, et al. Web-Based, Participant-Driven Studies Yield Novel Genetic Associations for Common Traits. PLoS Genet 2010; 6(6): e1000993
9. Sintchenko V, Anthony S, Phan XH, Lin F, Coiera EW. A PubMed-Wide Associational Study of Infectious Diseases. PLos One 2010; 5(3): e9535.
10. Dudley JT, Pouliot Y, Chen R, Morgan AA, Butte AJ. Translational bioinformatics in the cloud: an affordable alternative. Genome Med 2010;2:51.
11. Yip YL. Closing the genotype-phenotype gap. Findings from the Yearbook 2010 Section on Bioinformatics. Yearb Med Inform 2010;82-5.
12. Yip YL. Accelerating knowledge discovery through community data sharing and integration. Findings from the Yearbook 2009 Section on Bioinformatics. Yearb Med Inform 2009;117-20.
13. Yip YL. The promise of systems biology in clinical applications. Findings from the Yearbook 2008 Section on Bioinformatics. Yearb Med Inform 2008;102-4.
14. Fingerman IM, McDaniel L, Zhang X, Ratzat W, Hassan T, Jiang Z, et al. NCBI Epigenomics: a new public resource for exploring epigenomic data sets. Nucleic Acids Res 2011;39 (Database issue): D908-D912.
15. Sansom C. The power of many. Nature biotechnology 2011;29(3):201-3.

Correspondence to:
Dr. Yum Lina Yip
R&D Knowledge Management
Merck Serono S.A.
9 Chemin des Mines Geneva
Switzerland
Tel: +41 22 414 3937
Fax: +41 22 414 3059
E-mail: lina.yip.sonderegger@merckserono.net

Appendix: Content Summaries of Selected Best Papers for the IMIA Yearbook 2011, Section Bioinformatics*

Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, Dudley JT, Ormond KE, Pavlovic A, Hudgins L, Gong L, Hodges LM, Berlin DS, Thorn CF, Sangkuhl K, Hebert JM, Woon M, Sagreiya H, Whaley R, Morgan AA, Pushkarev D, Neff NF, Knowles JW, Chou M, Thakuria J, Rosenbaum A, Zaranek AW, Church G, Greely HT, Quake SR, Altman RB

Clinical assessment incorporating a personal genome

Lancet 2010;375(9725):1525-35

As the cost of whole genome sequencing falls rapidly, it is essential to derive analysis methods to allow more comprehensive genetic risk assessment and personalized therapies. In this study, the authors provided the first integrated approach to analyze a complete human genome in a defined clinical context. An individual with a family history of vascular disease and early sudden death was assessed clinically. Genetic analysis was then performed on whole genome sequence data, with a focus on predicting genetic risk of genes associated with known Mendelian diseases, response to medications, and the significance of novel variants. The authors also developed methods to integrate disease risk across multiple common polymorphisms and to account for gene-environment interactions. The analysis results showed increased genetic risk for myocardial infarction, type II diabetes and certain cancers. Rare variants in *LPA* were also found which are consistent with the family history of coronary artery disease. In addition, pharmacogenomic analysis suggested potential responses to a range of therapies. Overall, this is the first study providing a proof-of-principle that whole genome sequence data can have clinical utility at an individual level, thus making a large step towards personalized medicine. Significant challenges however remain, which include the lack of a comprehensive rare mutation database, a robust statistical framework, and the difficulties in quantifying gene-environment interactions.

Eriksson N, J. Michael Macpherson JM, Tung JY, Hon LS, Naughton B, Saxonov S, Avey L, Wojcicki A, Pe'er I, Mountain J

Web-Based, Participant-Driven Studies Yield Novel Genetic Associations for Common Traits

PLoS Genetics 2010;6(6):e1000993

A major challenge in the understanding of the genetic basis for variation in common human traits is the efficient, coordinated collection of genotype and phenotype data. In this study, the authors have developed an innovative research framework that facilitates the parallel study of 22 common human traits within a single, nearly 10,000 participants cohort. Their approach makes use of the customers of 23andMe (a personal genetics company) who submit samples of saliva for genotyping. The participants can later take part in a web-based surveys and self report information related to their own phenotypes. As incentive, the participants receive interpretations of their genetic data upon completion of surveys. The 22 traits studied were selected based on indications of heritability, ease of phenotype data collection via web-based surveys, and wide interest. The initial results from this study verified the association of a large number of previously identified genes with variation in hair color, eye color, and freckling, thus validating the web-based self-reporting approach. In addition, novel associations among several single-nucleotide polymorphisms and freckling, hair curl, asparagus anosmia, and photic sneeze reflex were identified. In conclusion, this novel research framework is demonstrated to be a viable alternative to traditional methods for the understanding of much of the unexplained human variation.

Fingerman IM, McDaniel L, Zhang X, Ratzat W, Hassan T, Jiang Z, Cohen RF, Schuler GD

NCBI Epigenomics: a new public resource for exploring epigenomic data sets

Nucleic Acids Research 2011;39 (Database issue):D908-D912

Epigenetics is the study of stable and heritable changes in gene expression that occur independently of the primary DNA sequence. In the past decade, this field has gained ever wider interest as it has been shown that misregulation of epigenetic processes such as regulation of gene expression and DNA repair can be associated with human disease. Epigenetic mechanisms include post-translational modifications of histones, DNA methylation, chromatin conformation and non-coding RNAs. This report presents the Epigenomics database at the National Center for Biotechnology Information (NCBI). The database has been created to serve as a comprehensive public resource for epigenetic and epigenomic data sets (www.ncbi.nlm.nih.gov/epigenomics). Data in the database derived primarily from data originally submitted to archival databases at the NCBI, namely the Gene Expression Omnibus (GEO) and the Sequence Read Archive (SRA) which contain data collected from large scale projects such as the NIH Roadmap Epigenomics project, ENCODE and modENCODE projects as well as from smaller single laboratory studies. The subset of epigenetics-specific data are subjected to further review, annotation and reorganization. The Epigenomics resource also provides the user with an intuitive interface to view, explore, analyze and manipulate these data.

* The complete papers can be accessed in the Yearbook's full electronic version, provided that permission has been granted by the copyright holder(s).

Patel CJ, Bhattacharya J, Butte AJ

An Environment-Wide Association Study (EWAS) on Type 2 Diabetes Mellitus

PLoS One 2010;5(5):e10746

While it is well acknowledged that most diseases are a result of a complex combination of genetic and environmental factors, nearly all studies today focus on the genetic components. In this work, the authors conducted a pilot Environmental-Wide Association Study (EWAS) to comprehensively associate specific environmental factors with disease. Type 2 Diabetes (T2D) was chosen as prototype and the study is made possible by the examination of multiple cohorts present in the nationally representative National Health and Nutrition Examination Survey (NHANES) dataset. The EWAS consists of two methodological steps that are analogous to a Genome Wide Association Study (GWAS). First, a panel of 266 unique environmental assays (or environmental "loci") measured across cases of diabetics and controls were considered. This has yielded several environmental factors with significantly high association with T2D. Second, the associations were validated with other cohorts in NHANES. With EWAS, the authors are able to rediscover factors such as carotenes and PCBs with previously known association with T2D. They also hypothesize about new associations with T2D. In conclusion, the EWAS is proved to be a promising way to search and consider potential environmental factors as associated with disease or other clinical phenotypes. These results also indicate that it may be high time to usher in "enviromics", the study of a wide array of environmental factors in relation to health and biology.

Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, Basford MA, Brown-Gentry K, Balser JR, Masys DR, Haines JL, Roden DM

Robust Replication of Genotype-Phenotype Associations across Multiple Diseases in an Electronic Medical Record

The American Journal of Human Genetics 2010;86:560–72

Electronic medical record (EMR) systems offer the potential to act as platforms for generating sets of cases and controls for clinical and translational research as they contain large populations with diverse diseases. An especially attractive vision is one in which large-scale DNA databanks are linked to EMRs, and thus enabling discovery and incorporation into practice new genotype-phenotype associations. In this study, the authors have genotyped, from the BioVU (the Vanderbilt DNA biobank), about 10,000 samples accrued (over 4 mo) for 21 SNP sites reproducibly associated with five diseases: atrial fibrillation, Crohn disease, multiple sclerosis, rheumatoid arthritis, or type 2 diabetes. For each phenotype, natural language processing techniques and billing-code queries were used to identify cases and controls from deidentified health records. It was found that at least one previously reported genotype-phenpotype association was replicated. These results demonstrate that common genetic variants associated with disease can be replicated with the use of samples from a DNA databank coupled to a deidentified EMR, thus validating the concept that biorepositories linked to EMR systems can represent robust tools for accelerating genome-driven diagnostics and therapeutics. In addition, the method can lead ultimately to the identification of common genotype variants that are useful in clinical medicine, as healthcare information accumulates and genotype-phenotype associations become increasingly well defined.

Sintchenko V, Anthony S, Phan XH, Lin F, Coiera EW

A PubMed-Wide Associational Study of Infectious Diseases.

PLos One 2010; 5(3): e9535

With the rapid accumulation of scientific information, it is increasingly challenging to synthesize a collective view of knowledge in different disciplines and enable the discovery of new findings. Text mining offers a powerful computational approach to uncover hidden associations. So far, in the field of human genetics and basic microbiology, several information retrieval systems relying on the mining of MEDLINE text have been proposed to uncover molecular level host-pathogen interactions. In this study, the authors have developed a Pubmed-wide association study to find and quantify existing but hidden relationships between clinical syndromes and individual pathogens in order to detect meaningful associations across multiple scales. The knowledge space is explored by mapping of the infectious disease literature. It involves different types of biomedical entities (e.g., genes, pathogens, drugs) and events (syndromes and diseases), which are expressed using standard terms or as relationships among such objects. A new and clinically focused reclassification of the microbial world is created by using syndromic signatures for different pathogens. This work illustrates how multilevel syndrome and pathogen network representations can provide insights on unexpected biological similarities in disease pathogenesis and epidemiology, and accelerate the translational research enterprise.