From Genome Sequencing to Bedside

Findings from the Section on Bioinformatics and Translational Informatics

T. Lecroq, L. F. Soualmia, Section Editors for the IMIA Yearbook Section on Bioinformatics and Translational Informatics

Normandie University, University of Rouen, LITIS EA 4108, Information Processing in Biology & Health, Mont-Saint-Aignan, France

Summary

Objectives: To summarize excellent current research in the field of Bioinformatics and Translational Informatics with application in the health domain and evidence-based medicine.

Method: We provide a synopsis of the articles selected for the IMIA Yearbook 2013, from which we attempt to derive a synthetic overview of current and future activities in the field. Three steps of selection were performed by querying PubMed and Web of Science. A first set of 5,549 articles was refined into a second set of 1,272 articles from which 15 articles were retained for peer-review.

Results: The selection and evaluation process of this Yearbook's section on Bioinformatics and Translational Informatics yielded four excellent articles regarding the Human Genome and Medicine. Exploiting genomic data depends on having the appropriate reference annotation available. In the first article, the goal of the GENCODE Consortium is to produce and publish The GENCODE human reference gene set. As a result it is composed by merged manual and automatic annotations, which are frequently updated from public experimental databases. The quality of genome sequencing is platform-dependant. In the second article, a generic database independent from the sequencing technologies, Huvariome, can help to identify errors and inconsistencies in sequencing. To understand complex diseases of patients it will be of great importance to detect rare gene variants. This is the aim of the third study. Finally, in the last article, the plasma's DNA of healthy individual and patients suffering from cancer is compared.

Conclusions: The current research activities attest to the continuous convergence of Bioinformatics and Medical Informatics for clinical practice. For instance, a direct use of high throughput sequencing technologies for patients could aid the diagnosis of complex diseases (such as cancer) without invasive surgery (such as biopsy) but only with blood analysis. However, ongoing genomic tests will generate massive amounts of data and will imply new trends in the near future: "Big Data" and smart health management.

Keywords

Translational Medical Research, Computational Biology, Gene expression, Genome, Medical Informatics

Yearb Med Inform 2013:175-7

Introduction

Pursuing the increasing research on "bedside to bench" as mentioned in the last year's Yearbook [1], main ongoing works on Bioinformatics and Translational Informatics are related to Genome Medicine. Indeed, high throughput sequencing technologies, also known as Next Generation Sequencing (NGS) technologies have changed Biology in the last years. On one hand, these technologies have reduced both the cost and the time needed to obtain genomic sequences and on the other hand they need much more bioinformatics resources than the traditional Sanger sequencing technology [2]. So far, NGS already has been successfully used to characterize various genomic elements in numerous species and especially in the Human Genome. The next step, after changing Biology, is to change Medicine. The availability of genomic data from NGS experiments allows the analysis of the disease-related biomolecular networks, which are expected to couple genotypes and disease phenotypes to determine the biological mechanisms of complex diseases.

Electronic Health Records (EHRs) are a valuable source of information for knowledge discovery. The "Integrating Biology and the Bedside" (i2b2) platform is a well-known framework that enables researchers to use clinical data for discovery research [3]. The integration of genomic data into EHRs as well as the development of genomic tests and their increasing clinical utility will change the medical decision process as highlighted by a survey on last year's articles (2012).

For instance, Roden et al. [4] review the use of EHRs for clinical pharmacology. EHRs are considered a knowledge resource for discovering new drug actions secondary to genomic influences on disease phenotypes and drug responses. In a recent study, Liu et al. [5] review the existing network biology efforts to study complex diseases, such as breast cancer, diabetes, and Alzheimer's disease, using high throughput data and computational tools.

NGS technologies not only enable better characterization of protein-coding regions but they were also used for studying non-coding regions [6]. When studying genetic diseases, researches focus mainly on coding regions: It is then fundamental to distinguish true new variations from already known variations. In [7], Stubbs et al. aim at building a database of high quality annotated variations. To understand complex diseases it will be of great importance to detect rare variants (with a Minor Allele Frequency < 0.5%) as highlighted by Tenenssen et al. [8]. The main studied complex disease is cancer. Tian et al. [9] present a new paradigm in Medicine to complete Evidence-Based Medicine: Predictive, Preventive, Personalized, and Participatory Medicine (P4), the challenge of treating cancer is its complexity. This complexity is in part due to genomic data.

Akan et al. [10] present a study, in which whole genome and transcriptome data for three human cancer cell lines were analyzed in conjunction with protein data. The authors demonstrate the advantage for integrative analysis for identifying tumor-related genes. Among several results, another direct use of these high throughput technologies in patient cares could to diagnose cancer without biopsy [11].

Best Paper Selection

The best paper selection for the section Bioinformatics and Translational Informatics follows a generic method, commonly use in all the sections of the IMIA Yearbook 2013. The search is performed on MEDLINE via PubMed and completed on the Web of Science for articles not indexed in MEDLINE. The queries include MeSH headings related to the

domain of computational biology and medical genetics and with a restriction to international peer-reviewed journals. Only original research articles published in 2012 are considered; we exclude the publications types reviews, editorials, comments, letters to the editors …*etc.*

As computational biology and translational research is a very active publication field (over 20,000 articles in 2012), we limited the search on only the major MeSH headings leading 5,549 articles. To reduce this set, we add another feature: the 2011 Impact Factor of international peer-reviewed journals. The search was performed on the top 15 journals of the Bioinformatics and Translational Informatics section (such as Genome Research, Science Translational Medicine…) and top six journals in Medicine (such as Lancet, Nature …). The analysis of a set of 1,272 articles resulted in 15 articles for peer-review. Finally, four papers [6-9] are retained by the reviewers. As mentioned in the introduction, Genome Medicine and complex disease analysis characterize this year's researches. Harrow et al. [6] introduce release 7 of GENCODE the gene set that has been adopted by the ENCODE Project consortium [11] whose goal it is to find all functional elements in the human genome; Stubbs et al. [7] present a database called Huvariome that can help to distinguish true new genomic variations from know variations; Tennessen et al. [8] characterize rare variants in two populations: American individuals with European ancestry and American individuals with African ancestry. Leary et al. [11] show that some cancers could be diagnosed by sequencing blood DNA thus avoiding invasive surgery (such as biopsy).

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2013 in the sections 'Bioinformatics and Translational Informatics'. A brief content of each one can be found in the appendix of this synopsis. The articles are listed in alphabetical order of the first author's surname.

Section
Bioinformatics and Translational Informatics
▪ Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigó R, Hubbard TJ. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 2012 Sep;22(9):1760-74.
▪ Leary RJ, Sausen M, Kinde I, Papadopoulos N, Carpten JD, Craig D, O'Shaughnessy J, Kinzler KW, Parmigiani G, Vogelstein B, Diaz LA Jr, Velculescu VE. Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. Sci Transl Med 2012 Nov 28;4(162):162ra154
▪ Stubbs A, McClellan EA, Horsman S, Hiltemann SD, Palli I, Nouwens S, Koning AH, Hoogland F, Reumers J, Heijsman D, Swagemakers S, Kremer A, Meijerink J, Lambrechts D, van der Spek PJ. nHuvariome: a web server resource of whole genome next-generation sequencing allelic frequencies to aid in pathological candidate gene selection. J Clin Bioinforma 2012 Nov 19;2(1):19.
▪ Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, Bustamante CD, Bamshad MJ, Akey JM; Broad GO; Seattle GO; NHLBI Exome Sequencing Project. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 2012 Jul 6;337(6090):64-9

Conclusions and Outlook

The current research activities attest the continuous convergence of Bioinformatics and Medical Informatics to improve clinical practice. There is still a strong effort to characterize and correctly annotate all the functional elements in the human genomes with the help of efficient sequencing technologies requiring elaborate bioinformatics software. In parallel with this identification of the common genomic elements, a major research emphasis focuses on distinguishing the pathogenic variations among all the natural polymorphisms. This effort is pursued by cataloguing and annotating known variation, and also by tracking rare variants responsible for complex diseases. All these studies will eventually contribute to personalize Medicine. For instance, a direct use of high throughput sequencing technologies in patients could be the diagnosis of complex diseases (such as cancer) without invasive surgery (such as biopsy) but only with blood analysis. One of the main challenges is to choose which genomic information, in addition to phenotypic information, should be stored in EHRs. However, ongoing genomic tests will keep on generating massive amounts of data and will suggest new trends in the near future: "Big Data" and smart health management.

Acknowledgements

We would like to acknowledge the valuable support of Martina Hutter and the reviewers in the evaluation process of the section Bioinformatics and Translational Informatics of the IMIA Yearbook.

References

1. Yip YL. Unlocking the potential of electronic health records for translational research. Findings from the section on the bioinformatics and translational informatics. Yearb Med Inform 2012;7(1):135-8.
2. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 1975;94(3):441–8.
3. Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. J Am Med Inform Assoc 2012 Mar-Apr;19(2):181-5.
4. Roden DM, Xu H, Denny JC, Wilke RA. Electronic medical records as a tool in clinical pharmacology: opportunities and challenges. Clin Pharmacol Ther 2012 Jun;91(6):1083-6.
5. Liu ZP, Wang Y, Zhang XS, Chen L. Network-based analysis of complex diseases. IET Syst Biol 2012 Feb;6(1):22-33.
6. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res.2012 Sep;22(9):1760-74.
7. Stubbs A, McClellan EA, Horsman S, Hiltemann SD, Palli I, Nouwens S, et al. Huvariome: a web server resource of whole genome next-generation sequencing allelic frequencies to aid in pathological candidate gene selection. J Clin Bioinforma 2012 Nov 19;2(1):19.
8. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, e al; NHLBI Exome Sequencing Project. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 2012 Jul 6;337(6090):64-9.
9. Tian Q, Price ND, Hood L. Systems cancer medicine: towards realization of predictive, preventive, personalized and participatory (P4) medicine. J Intern Med 2012 Feb;271(2):111-21.
10. Akan P, Alexeyenko A, Costea PI, Hedberg L, Solnestam BW, Lundin S, et al. Comprehensive analysis of the genome transcriptome and proteome landscapes of three tumor cell lines. Genome Med 2012 Nov 18;4(11):86.
11. Leary RJ, Sausen M, Kinde I, Papadopoulos N, Carpten JD, Craig D, et al. Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. Sci Transl Med 2012 Nov 28;4(162):162ra154.
12. The ENCODE Project Consortium. A user's guide to the ENCyclopedia Of DNA Elements (ENCODE). PLoS Biol 2011 9(4): e1001046.

Correspondence to:
Pr Thierry Lecroq
Normandie Univ., University of Rouen
LITIS EA 4108, Information Processing in Biology & Health
76821 Mont-Saint-Aignan Cedex, France
Tel : +33 235 146 581
E-mail: Thierry.lecroq@univ-rouen.fr

Appendix: Content Summaries of Selected Best Papers for the IMIA Yearbook 2013, Section Bioinformatics and Translational Informatics[1]

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigó R, Hubbard TJ

GENCODE: the reference human genome annotation for The ENCODE Project

Genome Res 2012 Sep;22(9):1760-74

The authors present the GENCODE gene set that has been adopted by the ENCODE Consortium, The 1000 Genomes Project Consortium, and The International Cancer Genome Consortium as their reference gene annotation. The GENCODE releases are a mix of manual and automatic annotation. Loci are classified into three main biotypes: protein-coding gene, long noncoding RNA gene, and pseudogene. From release 3c to release 7 the number of protein-coding genes has decreased (which is mainly due to the removal of poorly supported annotation models) while the number of coding locus transcripts has increased. The GENCODE 7 release consists of 20,687 protein-coding genes and 9,640 long noncoding RNA genes. Most annotated long noncoding RNA consists of two exons. A pseudogene ontology was created. An experimental validation was set up for part of these findings. GENCODE 7 is accessible from gencodegenes.org and via the Ensembl and UCSC Genome Browsers.

Stubbs A, McClellan EA, Horsman S, Hiltemann SD, Palli I, Nouwens S, Koning AH, Hoogland F, Reumers J, Heijsman D, Swagemakers S, Kremer A, Meijerink J, Lambrechts D, van der Spek PJ

Huvariome: a web server resource of whole genome next-generation sequencing allelic frequencies to aid in pathological candidate gene selection

J Clin Bioinforma 2012 Nov 19;2(1):19

In this article the authors describe a database, called Huvariome, for storing genetic variations obtained by high throughput whole genome sequencing of human individuals. The sequencing was performed by Complete Genomics and the database is compliant with the Complete Genomics Analysis pipeline. However to be generic, the database was designed to be independent from the sequencing technologies. Variations occurring in genes are annotated. This database can be queried, via a web interface, by gene or by position. Results are then displayed sorted by variation frequencies. It has actually been populated with the genes of 165 individuals. It includes a set, called Huvariome Core, of 31 healthy individuals from the Benelux region. The aim of this database is to enable users to rule out common variants (with more than 5% of minor allele frequency in the database) when studying a given pathology. The authors show that Huvariome can help to disambiguate sequencing inconsistencies occurring in re-sequencing projects by providing accurate reference allele frequencies. Furthermore it can identify platform-dependent errors associated with specific regions in the human genome.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, Bustamante CD, Bamshad MJ, Akey JM; Broad GO; Seattle GO; NHLBI Exome Sequencing Project

Evolution and functional impact of rare coding variation from deep sequencing of human exomes

Science 2012 Jul 6;337(6090):64-9

The authors analyzed 15,585 human exomes with a median depth of 111 times of 1,351 American individuals with European ancestry and 1,088 American individuals with African ancestry. They identified 500,000 single nucleotides variants (SNVs), 86% with a minor allele frequency of less than 0.5%, 82% were previously unknown and 82% were population-specific. On average, each individual genome has 13,595 SNVs, 2.3% of which were predicted to have functional impact on the protein of approximately 313 genes and 95.7% of them are rare. The authors show that abundance of rare variation can be explained by human demographic history. The authors conclude by stating that large samples will be needed to associate rare variants with complex diseases.

Leary RJ, Sausen M, Kinde I, Papadopoulos N, Carpten JD, Craig D, O'Shaughnessy J, Kinzler KW, Parmigiani G, Vogelstein B, Diaz LA Jr, Velculescu VE

Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing

Sci Transl Med 2012 Nov 28;4(162):162ra154

The authors analyze the DNA from the plasma of 10 colorectal and breast cancer patients and 10 healthy individuals with high throughput sequencing. They identified, in all affected patients, structural alterations that were not present in the DNA from the plasma of healthy subjects. Among those detected alterations were chromosomal copy number changes and rearrangements, including amplification of pilot cancer driver genes such as ERBB2 and CDK6. Chromosomal copy number changes were estimated with a log-scale plasma aneuploidy score while rearrangements were detected with a technique named personalized analysis of rearranged ends. The level of circulating tumor DNA in cancer patients ranged from 1.4 to 47.9%. The sensitivity and specificity of this approach depends on the amount of sequence data obtained and is based on the fact that, in most cancers, multiple chromosomal alterations are observed that are unlikely to be present in normal cells. Since chromosomal alterations are present in almost all human cancers, this approach represents a useful method for the noninvasive detection of human tumors that do not rely on the availability of tumor biopsies.

[1] The complete papers can be accessed in the Yearbook's full electronic version, provided that the article is freely accessible or that your institution has access to the respective journal.