Findings from the Section on Bioinformatics and Translational Informatics

H. Dauchel, T. Lecroq, Section Editors for the IMIA Yearbook Section on Bioinformatics and Translational Informatics
LITIS EA 4108, UNIROUEN, Normandy University, Rouen, France

Summary

Objectives: To summarize excellent current research and propose a selection of best papers published in 2016 in the field of Bioinformatics and Translational Informatics with applications in the health domain and clinical care.

Method: We provide a synopsis of the articles selected for the IMIA Yearbook 2017, from which we attempt to derive a synthetic overview of current and future activities in the field. As in 2016, a first step of selection was performed by querying MEDLINE with a list of MeSH descriptors completed by a list of terms adapted to the section coverage. Each section editor evaluated separately the set of 951 articles returned and evaluation results were merged for retaining 15 candidate best papers for peer-review.

Results: The selection and evaluation process of papers published in the Bioinformatics and Translational Informatics field yielded four excellent articles focusing this year on the secondary use and massive integration of multi-omics data for cancer genomics and non-cancer complex diseases. Papers present methods to study the functional impact of genetic variations, either at the level of the transcription or at the levels of pathway and network.

Conclusions: Current research activities in Bioinformatics and Translational Informatics with applications in the health domain continue to explore new algorithms and statistical models to manage, integrate, and interpret large-scale genomic datasets. As addressed by some of the selected papers, future trends would include the question of the international collaborative sharing of clinical and omics data, and the implementation of intelligent systems to enhance routine medical genomics.

Keywords

Translational medical research; computational biology; gene genome expression; genome; medical informatics

Yearb Med Inform 2017:188-92
Published online August 18, 2017

Introduction

As mentioned in [1], main ongoing works on Bioinformatics and Translational Informatics (BTI) are related to clinical genomics i.e., the identification from data of genes and mutations underlying human diseases, the promotion of research on "bedside to bench", the management of "Big Data", and personalized medicine. Currently, the availability of large-scale genomic data from Next Generation Sequencing (NGS) experiments allows the analysis and prediction of disease-related genes and biochemical networks, which are expected to associate genotypes and phenotypes of complex diseases. In 2017, the selection of best papers in the BTI section of the IMIA Yearbook intends to highlight the current progress of secondary use and massive integration of multi-omics data from patients or populations in order to provide better comprehensive causative cellular and molecular mechanisms leading to the identification of novel diagnoses and predictive biomarkers, and therefore to opportunities for new drug discovery and development. Well known resources for secondary use such as The Cancer Genome Atlas repository (TCGA, http://cancergenome.nih.gov/), the Genome Wide-Association Study repository (GWAS, https://grasp.nhlbi.nih.gov/) for non-cancer complex diseases, or the 1000 Genomes Projects repository [2] continue to be exploited, while the Clinical Proteomic Tumor Analysis Consortium (CPTAC, https://proteomics.cancer.gov/data-portal) [3] appears to be a great contributor to emerging proteogenomic approaches. This year, the selection of best papers in the BTI section also highlights initiatives for sharing and standardizing resources.

Best Paper Selection Method

The selection of best papers for the section Bioinformatics and Translational Informatics follows a generic method, commonly used in all the sections of the IMIA Yearbook. As in previous years, the search was performed on MEDLINE by querying PubMed. The Boolean query includes MeSH descriptors related to the domain of computational biology and medical genetics with a restriction to international peer-reviewed journals. Only original research articles published in 2016 (from 01/01/2016 to 12/31/2016) were considered; we excluded reviews, editorials, comments, letters to the editors …etc. We limited the search to major MeSH descriptors in order to avoid returning a large set of articles, and we completed it with non-MeSH terms searched on titles and abstracts of articles. In 2017, the PubMed query yielded a set of 951 articles (vs. 1,566 last year) that were evaluated separately by each section editor (HD & TL) using the BibReview tool and the generic method described by Lamy et al. in [4]. BibReview takes as an entry a PubMed file (in XML format) and shows all metadata. Thanks to the interface, the user can tag the returned articles as "Accepted", "To Revise", "Conflict" or "Reject". Opinions of different reviewers can be merged and the results can be filtered. This year, only four articles were tagged "Accepted" by both section editors. Each section editor proposed his/her list of interesting papers to reach a set of 15 candidate best papers. Ten external reviewers, specialized in the BTI field, reviewed these 15 articles, which allowed ranking them according to criteria such as: topic significance, coverage of literature, quality of research, results, and presentation [5]. Finally, the

editorial team of the IMIA Yearbook selected four papers [6-9] as the best papers published in 2016 in the BTI field.

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2017 in the section 'Bioinformatics and Translational Informatics'. The articles are listed in alphabetical order of the first author's surname.

Section
Bioinformatics and Translational Informatics
▪ Chen J, Rozowsky J, Galeev TR, Harmanci A, Kitchen R, Bedford J, Abyzov A, Kong Y, Regan L, Gerstein M. A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. Nat Commun 2016 Apr 18;7:11101.
▪ Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z, Bergmann S. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. Nat Methods 2016 Apr;13(4):366-70.
▪ Zhang D, Chen P, Zheng CH, Xia J. Identification of ovarian cancer subtype-specific network modules and candidate drivers through an integrative genomics approach. Oncotarget 2016 Jan 26;7(4):4298-309.
▪ Zhang J, White NM, Schmidt HK, Fulton RS, Tomlinson C, Warren WC, Wilson RK, Maher CA. INTEGRATE: gene fusion discovery using whole genome and transcriptome data. Genome Res 2016;26(1):108-18.

Description of Candidate Best Papers and Selected Best Papers

The 15 candidate best papers illustrate different relevant trends in bioinformatics integration of large-scale multi-omics data in clinical genomics. The first axis illustrated by three of these articles [10-12] is the secondary use of data collections from the TCGA repository to gain insight and deeper understanding of the molecular mechanisms of tumorigenesis at genomic and transcriptomic levels and in order to propose translational methods and tools for data integration and patients care. The TCGA Research Network has generated and made publicly available comprehensive multiplatforms for genomic data of more than 11,000 tumor samples representing 33 cancer types. It is an enabling resource for cancer research. The study published in [10] contributes to the better understanding of the roles of miRNAs and their regulatory targets in cancer biology. Aiming at helping clinicians to interpret microRNA data from their own cancer cohorts, the authors described a computational pipeline that allows to compare them to the largest worldwide microRNA data of TCGA and to characterize microRNA expressions across large patient cohorts. In [11], the authors investigate the population genetics theory of the dynamics of neutral evolutionary processes in the context of the subclone dynamics in human cancers. They provided a new statistical model describing the expected pattern of subclonal mutations within a tumor. By applying the model to large pre-existing cancer genomics datasets from TCGA, they were able to reconstruct the mutational timeline of each tumor. While the common idea is that cancer evolution is dominated by strong selection dynamics, the authors were able to identify malignancies that evolve neutrally. In these neutral cancers, all tumor-driven alterations responsible of cancer expansion were present in the first malignant cell and subsequent tumor evolution was effectively neutral. The model provides a new analytical method to study cancer evolution and gain clinically relevant insight from commonly available genomic data from patients. In [12], the authors present a new method of multi-modal data analysis designed for heterogeneous data at the levels of transcription and of the regulation of transcription. While supervised approaches already exist, here the authors used a non-negative matrix factorization for data reduction and exploration. They provided an algorithm for jointly decomposing the data matrices involved that also included a sparsity option for high-dimensional settings. They evaluated the performance of the proposed method on synthetic data and on real DNA methylation, gene expression, and miRNA expression data from ovarian cancer samples extracted from TCGA.

The second axis illustrated by four of the 15 pre-selected articles [13-16] is the rapid growth of proteomics data resource analysis and integration with genomics data enabling the emerging of proteogenomic approaches. The Clinical Proteomic Tumor Analysis Consortium (CPTAC) [3] has produced large proteomics data sets from the mass spectrometric interrogation of tumor samples previously analyzed by TCGA program. In [13], the authors describe a tool named Common Data Analysis Platform (CDAP) produced by the CPTAC. Thanks to a dedicated pipeline, the goal of the CDAP is to provide standard and uniform reports for all CPTAC data (peptide-spectrum-match reports and gene-level reports), hence reducing the variability generated by disparate data analysis platforms and enabling comparisons between different samples and cancer types as well as across the major omics fields. In [14], the authors study the effects of somatic alterations on the proteomic landscape in breast cancer. They used a proteogenomic approach integrating quantitative mass-spectrometry-based proteomic and phosphoproteomic analyses of 105 genomically annotated breast cancers from TCGA. They were able to connect genome and transcriptomic variations to the peptide level, describing the functional impact of single amino-acid variants (SAAVs) and highlighting activated key regulators of signaling pathways in breast cancer. In [15], the authors use pan-cancer reverse phase protein array (RPPA) data, and compared the performance of 13 inference methods on protein-protein interactions (PPI) networks for retrieving curated Pathway Commons interactions. From a set of six high performers, they were able to build consensus networks on data for 11 cancer types from TCGA. In [16], the authors analyze the prioritization of somatic mutations for druggable targets that can be further studied for precision cancer medicine. They presented SGDriver which is a structural genomics-based method (supported by large-scale protein structural data (X-ray)) incorporating somatic missense mutations into protein–ligand binding site residues using a Bayes inference statistical framework. The authors applied their method on data from 16 cancer types from TCGA and identified potential druggable significantly mutated proteins (SMPs).

The third axis illustrated by one article [17] is the study of non-cancer complex diseases and the secondary use in the light of new omics data, new annotation tools, and the databases of the rich collection of publicly available Genome Wide Association Studies (GWASs) accumulated in the last decade. In [17], the authors describe a ranking method for signal prioritization loci called GenoWAP that integrates a genomic functional annotation of variations and a statistical model. They showed the effectiveness of their tool through its applications to Crohn's disease and schizophrenia from previous data.

The fourth axis illustrated by three of the pre-selected articles [18-20] describes the urgent need of international shared and standardized resources for omics and clinical data to translate genomic and phenotypic information into medical practice. In [18], the authors present the first public database and its web interface for storing and exploring large-scale data and profile analysis at the level of transcription involved in various cancer types (coding and non-coding RNA-seq), without the restriction of data origin (TCGA exclusive resource) or the restriction of a raw international data archive (Gene Expression Omnibus, Sequence Read Archive). In [19], the authors describe a network called IGNITE (Implementing GeNomics In pracTicE) funded by the National Institute of Health for collaborative initiatives. It was established to support the development, investigation, and dissemination of genomic medicine practice models that seamlessly integrate genomic data into electronic health records. The article presents clinical decision support strategies, methods for returning genomic test results, and educational initiatives for patients and providers. In [20], the Global Alliance for Genomics and Health (GA4GH) - federating together hundreds of individuals and organizations - describes how sharing the data can help researchers and clinicians. It presents a framework that can help institutions with different architectures to share data without requiring compatible data sets or compromising patient identity. Furthermore, this federated data ecosystem offers a standardized Application Programming Interface (API) to allow disparate technology services of institutions around the globe to communicate with one another and exchange genotypic and phenotypic information.

Among the 15 pre-selected papers, four papers were finally selected as best papers (Table 1). All fall in the theme of the secondary use and massive integration of multi-omics data. They explore disease data (cancer data from TCGA [8, 9], GWAS data for non-cancer complex diseases [7]), or population data from the 1000 Genomes Project [6]). They all have a common interest in the functional impact of genetic variations, either at the level of the transcription (variability of transcripts by gene fusion [9], annotation of variations with allelic functional imbalance [6]), or at the pathway and network levels [7, 9]. A content summary of the selected best papers can be found in the appendix of this synopsis.

Conclusions and Outlook

Despite extraordinary efforts to profile cancer genomes and non-cancer genetic disease genomes, interpreting and integrating the vast amount of multi-omics data with clinical data remains challenging. For bioinformaticians and biostatisticians, the challenge is not only to develop new data mining methods and models to discover molecular mechanisms underlying diseases, but it is also to conduct efficient collaborative works with clinicians to implement the conditions of the use of new knowledge in the routine clinical management. In this aim, in the next few years, bioinformatics and translational informatics should have to work together more closely to implement intelligent systems and tools integrating the power of bioinformatics methods, machine-learning methods, web semantics and ontology methods, as well as new efficient storage and computational technologies. Initiatives for the international collaborative sharing of raw, analyzed, and standardized omics data and clinical data should also contribute to the expansion of the genomic medicine.

Acknowledgements

We would like to acknowledge the valuable support of Martina Hutter, Adrien Ugon, and all the reviewers in the evaluation process of the best papers of the Bioinformatics and Translational Informatics section of the IMIA Yearbook. We also would like to greatly thank the IMIA Yearbook editors and managing editors Brigitte Séroussi, Lina Soualmia, and John Holmes.

References

1. Dauchel H, Lecroq T. Findings from the Section on Bioinformatics and Translational Informatics. Yearb Med Inform 2016:207-10.
2. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature 2015;526(7571):68-74.
3. Edwards NJ, Oberti M, Thangudu RR, Cai S, McGarvey PB, Jacob S, et al. The CPTAC Data Portal: A Resource for Cancer Proteomics Research. J Proteome Res 2015;14(6):2707-13.
4. Lamy J-B, Séroussi B, Griffon N, Kerdelhué G, Jaulent M-C, Bouaud J. Toward a formalization of the process to select IMIA Yearbook best papers. Methods Inf Med 54(2) (2015) 135-44.
5. Ammenwerth E, Wolff AC, Knaup P, Ulmer H, Skonetzki S, van Bemmel JH, et al. Developing and evaluating criteria to help reviewers of biomedical informatics manuscripts. J Am Med Inform Assoc 2003;10(5):512-4.
6. Chen J, Rozowsky J, Galeev TR, Harmanci A, Kitchen R, Bedford J, et al. A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. Nat Commun 2016;7:11101.
7. Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z, Bergmann S. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. Nat Methods 2016;13(4):366-70.
8. Zhang D, Chen P, Zheng CH, Xia J. Identification of ovarian cancer subtype-specific network modules and candidate drivers through an integrative genomics approach. Oncotarget 2016;7(4):4298-309.
9. Zhang J, White NM, Schmidt HK, Fulton RS, Tomlinson C, Warren WC, et al. INTEGRATE: gene fusion discovery using whole genome and transcriptome data. Genome Res 2016;26(1):108-18.
10. Chu A, Robertson G, Brooks D, Mungall AJ, Birol I, Coope R, et al. Large-scale profiling of microRNAs for The Cancer Genome Atlas. Nucleic Acids Res 2016:44(1):e3.
11. Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. Identification of neutral tumor evolution across cancer types. Nat Genet 2016;48(3):238-44.
12. Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. Bioinformatics 2016;32(1):1-8.
13. Rudnick PA, Markey SP, Roth J, Mirokhin Y, Yan X, Tchekhovskoi DV, et al. A Description of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) Common Data Analysis Pipeline. J Proteome Res 2016;15(3):1023-32.
14. Mertins P, Mani DR, Ruggles KV, Gillette MA,

Clauser KR, Wang P, et al. Proteogenomics connects somatic mutations to signaling in breast cancer. Nature 2016;534(7605):55-62.
15. Şenbabaoğlu Y, Sümer SO, Sánchez-Vega F, Bemis D, Ciriello G, N. Schultz N, et al. A Multi-Method Approach for Proteomic Network Inference in 11 Human Cancers. PLoS Comput Biol 2016;12(2):e1004765.
16. Zhao J, Cheng F, Wang Y, Arteaga CL, Zhao Z. Systematic Prioritization of Druggable Mutations in ~5000 Genomes Across 16 Cancer Types Using a Structural Genomics-based Approach. Mol Cell Proteomics 2016;15(2):642-56.
17. Lu Q, Yao X, Hu Y, Zhao H. GenoWAP: GWAS signal prioritization through integrated analysis of genomic functional annotation. Bioinformatics 2016;32(4):542-8.
18. Li JR, Sun CH, Li W, Chao RF, Huang CC, Zhou XJ, et al. Cancer RNA-Seq Nexus: a database of phenotype-specific transcriptome profiling in cancer cells, Nucleic Acids Res 2016;44(D1):D944-D951.
19. Weitzel KW, Alexander M, Bernhardt BA, Calman N, Carey DJ, Cavallari LH, et al. The IGNITE network: a model for genomic medicine implementation and research. BMC Med Genomics 2016;9:1.
20. Global Alliance for Genomics and Health. GENOMICS. A federated ecosystem for sharing genomic, clinical data. Science 2016;352(6291):1278-80.

Correspondance to:
Dr Hélène Dauchel
Normandy University, UNIROUEN, LITIS EA 4108
76821 Mont-Saint-Aignan Cedex, France
Tel : +33 235 146 389
E-mail: Helene.Dauchel@univ-rouen.fr

Appendix: Content Summaries of Selected Best Papers for the IMIA Yearbook 2017, Section Bioinformatics and Translational Informatics

Chen J, Rozowsky J, Galeev TR, Harmanci A, Kitchen R, Bedford J, Abyzov A, Kong Y, Regan L, Gerstein M

A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals

Nat Commun 2016 Apr 18;7:11101

In this article, existing large and diverse collections of individual genomes from the 1000 genomes Project, RNA-seq and ChIP-seq data sets were unified to build a comprehensive data corpus used to detect and functionally annotate allele-specific single nucleotide variants (SNVs) with allelic functional imbalance. The authors considered 1,263 functional genomics data sets from eight different studies to annotate variants associated with allele-specific binding and expression in 382 individuals consisting of 993 RNA-seq and 287 ChIP-seq data for coding and non-coding regions respectively. For each individual, the authors first built a diploid personal genome using the variants from the 1000 Genomes Project. Then expression data was mapped onto each of the haplotypes of the diploid genome, instead of the human reference genome. Results were then filtered to correct overdispersion and mapping bias, and finally enrichment analyses were performed thanks to a beta-binomial test to identify genomic regions that were enriched or depleted in allelic activity. Inheritance of allele-specific behavior was detected in autosomal protein-coding genes, untranslated regions (UTRs), introns and enhancers, and transcription factor (TF)-binding regions. Furthermore, considering the enrichment of rare variants, the authors examined selective constraints in allele-specific SNVs in coding DNA sections regions and TF motifs. The final data and results were organized into a distributed resource called AlleleDB that can be directly visualized as a UCSC (University of California, Santa Cruz) track in the UCSC Genome browser.

Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z, Bergmann S

Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases

Nat Methods 2016 Apr;13(4):366-70

Re-using previous 37 genome-wide association studies (GWASs) data for complex traits and diseases in the light of information supplied by diverse sequence data (CAGE-defined enhancers and promoters, ChIP-seq, and RNA-seq data) and expression quantitative trait loci (eQTL), the authors inferred and validated transcriptional regulatory circuits and the connectivity between trait-associated genes for 394 cell types or tissue-specific regulatory networks for human. Their integrative pipeline and network connectivity enrichment revealed that GWASs variants associated with specific diseases have impact on regulatory modules that are specific to disease-relevant cell types or tissues. All networks are freely available and they allow the systematic analysis of regulatory programs across hundreds of human cell types and tissues.

Zhang D, Chen P, Zheng CH, Xia J

Identification of ovarian cancer subtype-specific network modules and candidate drivers through an integrative genomics approach

Oncotarget 2016 Jan 26;7(4):4298-309

The identification of cancer subtypes is required to understand cancer heterogeneity and to propose the personalized therapy treatment appropriate to the different subtypes. In this study, the authors re-used large-scale ovarian cancer genomic data, including micro-array data (mRNA and microRNA expressions), SNP-array (copy number variations) and protein-protein interactions data in order to build a novel integrative procedure for defining ovarian cancer subtypes, identifying core pathways and candidate driver genes for each subtype. By applying a similarity network fusion approach to a patient cohort with 379 ovarian cancers from The Cancer Genome Atlas (TCGA) cancer samples, the authors were able to discover subnetworks enriched with genetic alterations. They identified two clinically relevant ovarian cancer subtypes with distinct molecular and clinical phenotypes and different survival profiles. Enrichment analysis of pathways associated with the two ovarian cancer subtype-specific networks revealed distinct molecular mechanisms of the tumorigenesis that could explain the different clinical outcomes.

Zhang J, White NM, Schmidt HK, Fulton RS, Tomlinson C, Warren WC, Wilson RK, Maher CA

INTEGRATE: gene fusion discovery using whole genome and transcriptome data

Genome Res 2016;26(1):108-18

Among somatic aberrations in cancer genome, gene fusions are the most prevalent chromosomal rearrangements. Especially in solid tumors, their detection can served as specific diagnostic markers, prognostic indicators, and therapeutic targets. Mono-modal data tools (structural variations with whole genome sequencing (WGS) or RNA-seq expression data) suffer from variability between fusion callers and from a poor sensitivity and specificity of fusion detection. In this article, the authors developed a new gene fusion discovery method that integrates both whole genome and transcriptome sequencing data from the same patient to reconstruct gene fusion junctions and genomic breakpoints by split-read mapping. INTEGRATE first utilizes mapped and unmapped RNA-seq reads, then analyzes WGS reads from tumors, and if available, from normal samples. INTEGRATE uses discordant RNA-seq reads to construct a gene fusion graph connecting genes involved in a putative fusion event. It finally proposes a prioritization of gene fusion candidates. INTEGRATE was evaluated by comparison to eight other gene fusion discovery tools by reusing data from a previously studied breast cancer cell line and peripheral blood lymphocytes derived from the same patient leading. INTEGRATE was also applied to a cohort of 62 breast cancer patients from The Cancer Genome Atlas (TCGA) and enabled the identification of novel gene fusions, a subset of which were recurrent. All together, by combining WGS and RNA-seq NGS data from a same patient, the authors demonstrated both high sensitivity and accuracy of INTEGRATE to detect novel causative mutations. Furthermore, unlike many gene fusion prediction tools that ignore read-through or trans-splicing events, INTEGRATE was able to provide valuable insight into RNA chimeras. The tool is freely available for academic use.