APPLICATIONS OF GENETICS, GENOMICS AND BIOINFORMATICS IN DRUG DISCOVERY

RICHARD BOURGON
Genentech Inc.
South San Francisco, CA 94080
Email: bourgon.richard@gene.com

FREDERICK E. DEWEY
Regeneron Genetics Center
Tarrytown, NY 10591
Email: frederick.dewey@regeneron.com

ZHENGYAN KAN
Pfizer Inc.
San Diego, CA 92121
Email: Zhengyan.Kan@pfizer.com

SHUYU D. LI
Sema4, a Mount Sinai venture
Stamford, CT 06902
Icahn School of Medicine at Mount Sinai
New York, NY 10029
Email: shuyu.li@sema4genomics.com

As the impact of genetics, genomics, and bioinformatics on drug discovery has been increasingly recognized, this session of the 2018 Pacific Symposium on Biocomputing (PSB) aims to facilitate scientific discussions between academia and pharmaceutical industry on how to best apply genetics, genomics and bioinformatics to enable drug discovery. The selected papers focus on developing and applying computational approaches to understand drug mechanisms of action and develop drug combination strategies, to enable *in silico* drug screening, and to further delineate disease pathways for target identification and validation.

1. Introduction

Drug discovery and development continues to face the challenges of rising cost and declining productivity. While the estimated average cost to bring a new molecular entity to market has exceeded US$ 1.5 billion, R&D return on investment fell considerably from 10.1% in 2010 to 3.7% in 2016 [1]. Recent advances in genetic and genomic research has not only accelerated our studies of disease mechanisms, but also enabled drug discovery in many areas. For example, the power of human genetics in therapeutic target validation has been underscored by a retrospective analysis that selecting targets with supportive human genetics evidence doubled the success rate in clinical development [2]. A recent report on the clinical impact of loss-of-function (LoF) genetic variants in 50,726 exomes confirmed previously known associations between genes such as PCSK9 and cardiovascular disease-related phenotypic traits, and identified novel associations with therapeutic implications [3]. Genomics and genetics also play an increasingly important role in other areas in drug discovery such as biomarker identification for drug efficacy [4] and safety [5], understanding drug mechanisms of action [6], and selecting disease relevant experimental models [7]. To facilitate the application of genomics in drug discovery, data quality and reproducibility have been systematically assessed [8] to increase our confidence on findings from pharmacogenomic studies. Furthermore, new methods and tools have been developed for integrative genomic data analysis [9].

Although the impact of genetics, genomics and bioinformatics in drug discovery has been recognized by both academia and pharmaceutical industry, the coverage of the topic in scientific conferences is very limited. The main objective of this session "Applications of genetics, genomics and bioinformatics in drug discovery" in the 2018 PSB is to cover recent advances in developing and applying computational approaches to enable drug discovery in the above-described areas. Furthermore, the session is also intended to promote more interactions and collaborations between academic and industry experts. We believe such a session dedicated to bioinformatics in the context of drug discovery could significantly benefit academic preclinical drug discovery activities, as a large number of academic drug discovery centers have been established in recent years [10]. We define the following topics and problems are within the scope of this session.

- Target identification and validation: integrative analysis of molecular data at scale, coupling genetic, epigenetic, gene expression profiling, proteomic, metabolomic, phenotypic trait measurements to disease diagnosis and clinical outcome data to generate hypotheses on molecular etiology of diseases in service of identification or validation of novel therapeutic targets.
- Biomarker discovery: utilizing genetic and genomic data derived from cell lines, animal models, human disease tissues and PBMC to develop preclinical or clinical biomarkers for target engagement, pharmacodynamics, drug response, prognosis, and patient stratification; applying genomic profiling in clinical trials to identify early response markers to predict clinical end points.
- Pharmacogenomics: identify associations between germline SNPs, somatic mutations, gene expression and other molecular alterations and drug responses.
- Toxicogenomics: integrative analysis of genomic, histopathology, and clinical chemistry data to develop predictive toxicology biomarkers in preclinical 4-day, 14-day and 30-day studies and clinical studies.
- Understanding drug mechanisms of action (MoA): applying genomic profiling to de-convolute targets and delineate MoA of non-selective drugs or drugs from phenotypic screening.
- Characterization of mechanisms of acquired resistance: analysis of genetic and genomic data derived from preclinical isogenic models or clinical patient samples to study the mechanisms of acquired resistance.

- Selection of disease-relevant experimental models: comparative analysis of genetic and genomic data to assess and select cell line and animal models in drug discovery that best represent the disease indications.
- Developing drug combination strategies: analysis of genetic and genomic data to identify synthetic lethality genes as drug combination targets; computational analysis to understand gene regulatory networks to develop combination strategies that target parallel pathways or reverse drug resistance.
- Drug repurposing: applying *in silico* approaches to identify new disease indications for existing drugs.
- Novel methods and tools for multi-omics data integration, analyses, and visualization.

2. Session Contributions

A total of eight papers were selected from the submissions. We categorized the eight papers into the following three groups. The papers in the first group focus on drug mechanisms of action and drug combinations. The second group includes studies that can be applied to enable computational drug screening. The papers in the third group apply various computational approaches on genetic and genomic data to further understand diseases.

2.1. *Drug mechanisms of action and drug combinations*

Understanding drug mechanisms of action is critical in clinical development and precision medicine, particularly in identifying early response markers as surrogates for clinical end-point as well as biomarkers for patient stratification. In addition, a better knowledge of MoA may allow us to reposition the existing drugs for new indications. In the study by *Luo et al.* [11], the authors developed a novel method, referred as Mania, for scalable data integration incorporating chemical structure, drug sensitivity and gene expression changes in response to drug treatment. Drug similarity networks were first constructed based on each of these data sources, followed by integration through Mania into a low-dimensional vector representation of each drug. It was shown that integration of various data sources improves quantification of drug-drug similarities, and achieves more accurate prediction of drug targets and MoA. Functionally enriched "drug communities", as referred by the study, was also identified using the low-dimensional vector representation matrix. Finally, the authors illustrated potential utilities of their new method by analyzing the most significantly mutated genes across 21 tumor types in the cancer genome atlas (TCGA) and presented examples of drugs that are predicted to target some of the significantly mutated cancer genes.

Gene expression profiling in cell lines in response to drug perturbation has provided a valuable tool to study drug MoA. Although a large number of drugs have been profiled in many cancer cell lines of various tissue origins, there are still substantial missing drug-cell line combinations in these data sources. *Hodos et al.* [12] attempted to fill the gaps by predicting cell specific drug perturbation expression profiles. The authors developed a computational framework to first arrange existing gene expression profiles into a three-dimensional array (or tensor) indexed by drugs, genes, and cell types, and then use either local (nearest-neighbors) or global (tensor completion) information to predict unmeasured profiles. The prediction accuracy was thoroughly evaluated and it was found that the two methods (local vs. global) have complementary performance, each superior in different regions in the drug-cell space. Finally, the authors demonstrated that the predicted profiles add value for downstream prediction of drug targets and therapeutic classes. For example, it was shown the classifiers trained on the complete dataset are of higher quality than those trained only on the measured dataset, with particularly significant impact on those cell types with fewer measured profiles available.

Drug treatment may induce alternative splicing as a key response event with functional consequences. However, limitation of short-read sequencing poses a barrier to accurately detect different splicing isoforms. *Chen et al.* [13] described characterization of the transcriptional splicing landscape in a prostate cancer cell line treated with a previously identified synergistic drug combination, by using a combination of third generation long-read RNA sequencing technology and short-read RNA-seq to create a high-fidelity map of expressed isoforms and fusions to quantify splicing events triggered by treatment. The authors found strong evidence for drug-induced, coherent splicing changes that disrupt the function of oncogenic proteins, and detected novel transcripts arising from previously unreported fusion events. The study demonstrated the benefit of long-read technology in identifying highly homologous isoforms routinely and with high fidelity.

Most patients with advanced cancers ultimately develop drug resistance to chemotherapy or targeted therapy due to reactivation of the same pathway or compensatory pathways. Combination therapy targets multiple pathways, therefore may improve efficacy and also overcome drug resistance in some cases. *Xu et al.* [14] presented a novel computational approach to predict combinations through assessing the potential impact of inhibiting a drug target on disease signaling network. Using melanoma as an example to apply the approach, the authors first constructed a disease network by integrating gene expression profiling and protein-protein interaction data. A drug-disease “impact matrix” was computed using network diffusion distance from drug targets to signaling network elements. The drugs were then clustered into “communities” that are supposed to share similar mechanisms of action. Finally, drug combinations maximally impacting signaling sub-networks are ranked and proposed as potential combination strategies for melanoma.

2.2. *Drug metabolism and in silico drug screening*

Human gut bacteria have the ability to activate, deactivate, and reactivate drugs with huge implications in drug efficacy and toxicity at individual patient level. Understanding the complete space of drug metabolism by the human gut microbiome is critical for predicting bacteria-drug relationships and their effects on drug response. To address the challenge that there are limited computational tools for predicting drug metabolism by the gut microbiome, *Mallory et al.* [15] developed a pipeline for comparing and characterizing chemical transformations using continuous vector representations of molecular structure based on unsupervised learning, and characterized the utility of vector representations for chemical reaction transformations. After clustering molecular and reaction vectors, enriched enzyme names, Gene Ontology terms, and Enzyme

Consortium (EC) classes were detected within the reaction clusters. Finally the authors queried reactions against drug-metabolite transformations known to be metabolized by the human gut microbiome, and showed the top results for these known drug transformations contained similar substructure modifications to the original drug pair. The method described in this study could be potentially applied in high throughput screening of drugs and their resulting metabolites against chemical reactions common to gut bacteria.

The study by *Greenside et al.* [16] addresses a critical component in drug discovery, identification of small molecule ligands that bind to the target proteins as a first step in drug screening. While the currently available computational tools for predicting protein-ligand binding largely rely on 3D protein structure, this study described an interpretable confidence-rated boosting algorithm to predict protein-ligand interactions with high accuracy from ligand chemical substructures and protein 1D sequence motifs, without relying on 3D protein structures. The authors showed that their models can be generalized to unseen proteins and ligands, demonstrating the possibility to predict protein-ligand interactions using only motif-based features and that interpretation of these features can reveal new insights into the molecular mechanics underlying each interaction.

2.3. *Disease genes and pathways*

Novel computational approaches have been continuously developed and applied to analyze genetic and genomic data. Recently, deep learning has emerged as a novel class of machine leaning methods. While deep learning has been applied in many domains such as speech recognition, image recognition, natural language processing, its application in analyzing genomic data is very limited. *Way et al.* [17] applied variational autoencoders (VAEs), an unsupervised deep neural network approach to analyze TCGA gene expression profiling data. Specifically, the extent to which a VAE can be trained to model cancer gene expression, and whether or not such a VAE would capture biologically relevant features were evaluated. The paper introduced a VAE trained on TCGA pan-cancer RNA-seq data, identified specific patterns in the VAE encoded features, and discussed potential merits of the approach. To illustrate the utility of VAEs in further delineating cancers, the authors described examples from their analyses on significant pathways separating primary and metastatic melanoma, and on pathways over-represented in different subtypes of high-grade serous ovarian cancer.

Human genetic data based on genome-wide sequencing or genotyping, coupled with hospital electronic medical records (EMRs) have provided a powerful tool to study the genetic basis of human diseases. *Smith et al.* [18] described integrative analysis of genetic data derived from DNA samples in a biobank and the accompanying clinical diagnosis information in EMRs to identify several neuroplasticity genes associated with neurodevelopmental diseases. The authors first developed a neuroplasticity gene signature from two independent gene expression profiling datasets. Subsequently, carriers of loss-of-function (LoF) genetic variants in the neuroplasticity genes were identified in the biobank cohort. The authors then performed an association analysis to discover significant associations between LoF in neuroplasticity genes and neurodevelopmental diseases. Finally, a thorough literature review was described to demonstrate the validity of the results.

3. Acknowledgments

We thank our respective organizations for supporting our involvement in organizing the session. We also thank the following reviewers for providing expert reviews of the submitted manuscripts: Vinayagam Arunachalam, Kristin Ayer, Ronghua Chen, Keith Ching, Ying Ding, Di Feng, Julio Fernandez, Marc Fink, Rajarshi Guha, Yangyang Hao, Jon Hill, Kipp Johnson, Robert Kueffner, Samir Lal, Hai Lin, Meng Ma, Gianni Panagiotou, Chetanya Pandya, Kiran Patil, Jeff Sutherland, Alex Tropsha, Song Wu, Tao Xie, Kun Yu, Yong Yue, Baohong Zhang, Chi Zhang, Yan Zhang, Xiaotong Zhu, Daniel Ziemek.

References

1. Mullard, A. R&D returns continue to fall. *Nature reviews. Drug discovery* **16**, 9 (2016).
2. Nelson, M.R. et al. The support of human genetic evidence for approved drug indications. *Nature genetics* **47**, 856-860 (2015).
3. Dewey, F.E. et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science (New York, N.Y.)* **354** (2016).
4. Kelloff, G.J. & Sigman, C.C. Cancer biomarkers: selecting the right drug for the right patient. *Nature reviews. Drug discovery* **11**, 201-214 (2012).
5. Khan, S.R., Baghdasarian, A., Fahlman, R.P., Michail, K. & Siraki, A.G. Current status and future prospects of toxicogenomics in drug discovery. *Drug discovery today* **19**, 562-578 (2014).
6. Nijman, S.M. Functional genomics to uncover drug mechanism of action. *Nature chemical biology* **11**, 942-948 (2015).
7. Horvath, P. et al. Screening out irrelevant cell-based models of disease. *Nature reviews. Drug discovery* **15**, 751-769 (2016).
8. Haverty, P.M. et al. Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature* **533**, 333-337 (2016).
9. Fernandez-Banet, J. et al. OASIS: web-based platform for exploring cancer multi-omics data. *Nature methods* **13**, 9-10 (2016).
10. Dahlin, J.L., Inglese, J. & Walters, M.A. Mitigating risk in academic preclinical drug discovery. *Nature reviews. Drug discovery* **14**, 279-294 (2015).
11. Luo, Y., Wang, S., Xiao, J. & Peng, J. Large-Scale Integration of Heterogeneous Pharmacogenomic Data for Identifying Drug Mechanism of Action. *Pacific Symposium on Biocomputing* **23** (2017).
12. Hodos, R. et al. Cell-specific prediction and application of drug-induced gene expression profiles. *Pacific Symposium on Biocomputing* **23** (2017).
13. Chen, X. et al. Characterization of drug-induced splicing complexity in prostate cancer cell line using long read technology. *Pacific Symposium on Biocomputing* **23** (2017).
14. Xu, J. et al. Diffusion Mapping of Drug Targets on Disease Signaling Network Elements Reveals Drug Combination Strategies. *Pacific Symposium on Biocomputing* **23** (2017).

15. Mallory, E.K., Acharya, A., Rensi, S.E., Bright, R.A. & Altman, R.B. Chemical reaction vector embeddings: towards predicting drug metabolism in the human gut microbiome. *Pacific Symposium on Biocomputing* **23** (2017).
16. Greenside, P., Hillenmeyer, M. & Kundaje, A. Prediction of protein-ligand interactions from paired protein sequence motifs and ligand substructures. *Pacific Symposium on Biocomputing* **23** (2017).
17. Way, G.P. & Greene, C.S. Extracting a Biologically Relevant Latent Space from Cancer Transcriptomes with Variational Autoencoders. *Pacific Symposium on Biocomputing* **23** (2017).
18. Smith, M.R. et al. Loss-of-function of Neuroplasticity-related genes confers risk for human neurodevelopmental disorders. *Pacific Symposium on Biocomputing* **23** (2017).