Biomarker and Drug Discovery for Gastroenterology Through Translational Bioinformatics

JOEL T. DUDLEY[*,‡,§] and ATUL J. BUTTE[‡,§]

[*]Program in Biomedical Informatics; [‡]Department of Pediatrics, Stanford University School of Medicine, Stanford; and [§]Lucile Packard Children's Hospital, Palo Alto, California

Mandates from federal funding agencies to refocus research toward translational medicine and heightened expectations of patients have put substantial pressure on clinicians to serve active and critical roles in the translational process. Although the clinician scientist is well-poised to interpret and synthesize knowledge discoveries in genomic, molecular, and clinical sciences into improved patient outcomes, the landscape of modern data-driven molecular medicine has come to be profoundly unaccommodating to those best poised to explore it.

A deluge of publicly available biomedical data and the commoditization of technologies enabling high-throughput molecular profiling of patients in the clinic have given rise to unique challenges and opportunities in translating fundamental molecular findings into clinical innovations. Genome-wide molecular measurements, such as those provided by genotyping arrays and gene expression microarrays, are now inexpensively and commonly obtained in the course of clinical investigation—generating millions to billions of data points in a single study. Navigation of published findings now entails searching large public data repositories such as the NCBI Gene Expression Omnibus (GEO; available: http://www.ncbi.nlm.nih.gov/geo/), which contains tens of thousands of published samples for nearly any disease condition. Because of this data abundance, computational approaches are becoming a critical component of translational clinical investigation.[1]

A number of translational bioinformatics methods have been developed for integrating and analyzing high-throughput molecular data toward evaluation of clinical hypotheses. Herein we present a synopsis of some recent work in translational bioinformatics that are enabling entirely new modalities for clinical investigation through systems medicine. We also discuss the implications of these methods and findings for the clinical investigation of gastrointestinal disease, and highlight potential future applications of translational bioinformatics in the search for novel biomarkers and therapeutics in gastroenterology.

Enabling Systems Medicine Through Translational Bioinformatics

The field of translational bioinformatics, defined as "the development of storage, analytic and interpretive methods to optimize the transformation of increasingly voluminous biomedical data into proactive, predictive, preventative, and participatory health" (available: https://www.amia.org/inside/stratplan), has emerged to address the informatics challenges in formulating and evaluating translational clinical hypotheses that draw from vast and diverse molecular and clinical vantages.[1] The approaches of translational bioinformatics are oriented toward the broad unmet needs of medicine, including the desire for diagnostic and prognostic biomarkers, improved therapies, identification of novel drug targets, and basic insights into the fundamental molecular basis of disease pathophysiology.

Translational bioinformatics can enable a systems view of medicine, by which many of the available molecular measurements, clinical data, and other characteristics of a disease or treatment can be integrated in a systematic fashion. This view offers novel modes for approaching translational clinical investigation. For example, translational bioinformatics might be applied to model and investigate the molecular basis of a specific disease condition through integration of patient laboratory biomarkers values with genome-wide gene expression profiles from affected tissues (Figure 1). Alternatively, one could attempt to infer novel characteristics about a disease through comparison with other related and presumably unrelated diseases, based on molecular or clinical characteristics. In either case, the expanded perspective can offer novel insights that only become apparent through consideration of a broader context of a clinical state (ie, finding a gene signature that is only up-regulated in 1 gastrointestinal disease versus all other diseases).

0016-5085/$36.00
doi:10.1053/j.gastro.2010.07.024

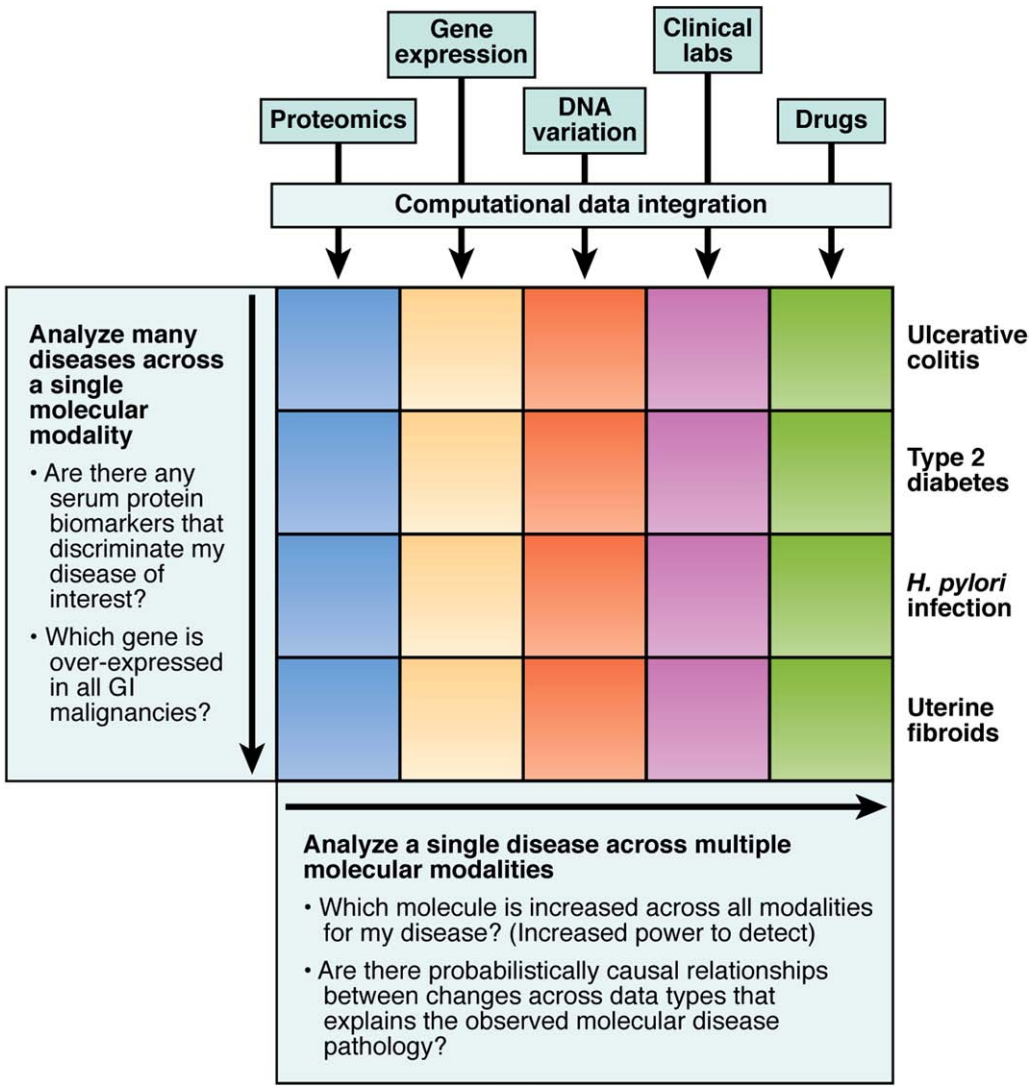

Figure 1. A conceptualization of the hypothesis space enabled by integrative systems medicine. Using publicly available data, it becomes possible to investigate a single disease simultaneously across many different molecular and clinical modalities (*rows*) and to simultaneously investigate many diseases from a single modality (*columns*).

We and others have developed and applied a number of translational bioinformatics methods aimed at taking advantage of the wealth of clinically relevant, genome-wide molecular measurements found abundantly in the public domain. This work includes the development of methods enabling systematic discovery of clinically relevant data from public repositories,[2] and establishment of the quality of public molecular data for downstream clinical investigation.[3] Enabled by the ability to systematically identify and obtain clinically relevant data from public repositories, we performed an integrative analysis of the genetic basis of obesity, in which data from 49 published obesity experiments comprising a range of molecular measurement modalities (eg, gene expression, proteomics, and others) were combined to form a predictive model for identifying obesity-associated genes.[4] The results of this study demonstrated that an integrative model that incorporated multiple experiments obtained from the public domain was able to outperform any single experiment in the task of predicting known obesity-associated genes.

Translational Insights From Systems of Disease Relationships

The notion that understanding of how diseases relate to one another can be informative of disease pathophysiology is not novel; however, new types of genome-wide molecular measurements and integrative bioinformatics approaches have provided the means to explore the spectrum of human disease from a molecular basis. The traditional view of disease relationships, or nosology, is based largely on anatomy and symptoms, which reflects the organization of the delivery of medical care into its specialties concerned with particular organs or phys-

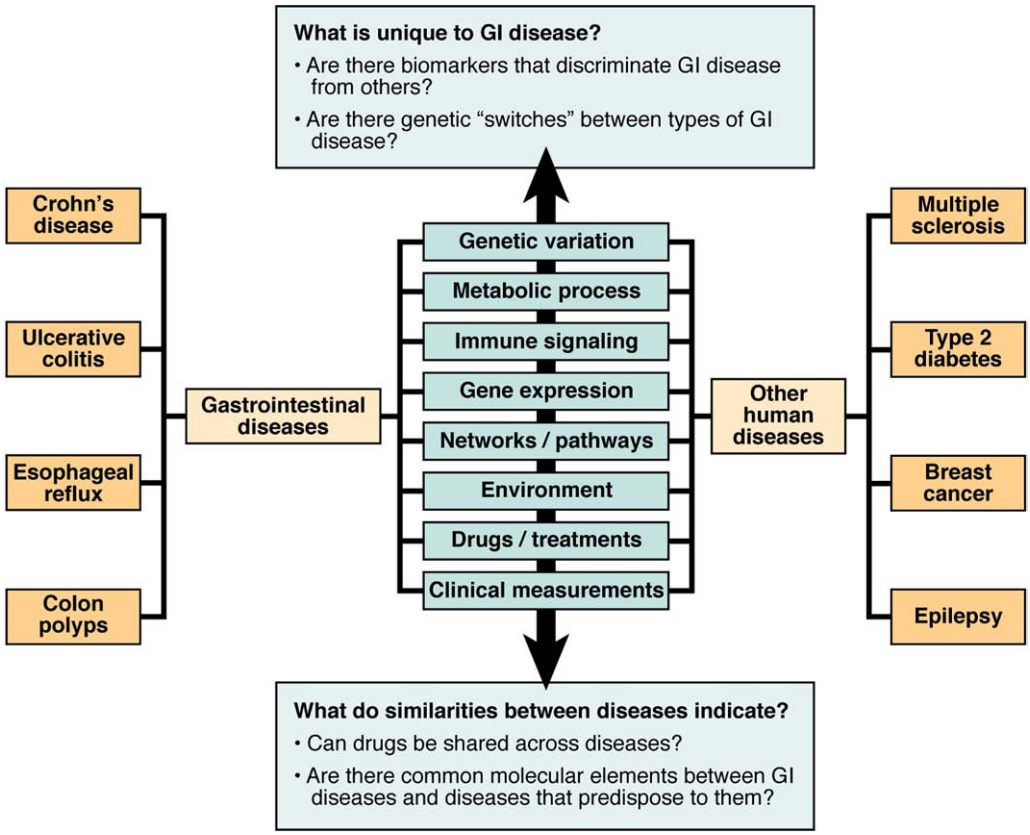

Figure 2. A schematic representation of a data-driven approach for exploring systems of disease relationships. Novel translational hypotheses are enabled through systematic evaluation of differences and similarities between diseases and disease groups.

iologic systems. Data-driven efforts to establish systems of disease relationships can enable new opportunities to identify molecular or clinical axes of commonalities or distinctions within or among diseases that may point toward new directions in biomarker discovery or therapeutic development (Figure 2).

The creation and exploration of a system of disease relationships based on genome-wide molecular measurements offers a powerful perspective for understanding the molecular architecture of disease. As an example, we recently compared the published genetic architecture of several autoimmune diseases, represented by the data available from genome-wide association studies on these conditions.[5] Looking across diseases from this vantage, we discovered a set of "toggle" variants that distinguished autoimmune diseases into exclusive classes. This work also revealed that diseases known to share drugs and clinical characteristics, such as Crohn's disease and rheumatoid arthritis, exhibited little similarity in DNA variation profiles, suggesting distinct molecular etiologies.

In other work, we integrated gene expression profiles on 54 diseases using protein–protein interaction data to build networks of disease relationships based on shared functional modules.[6] This work uncovered the molecular principles salient across a broad number of diseases, including expected pathways, such as inflammation and DNA repair, and unexpected pathways, such as insulin and somatostatin receptor signaling. More surprising was the discovery that drug targets found in these common, pan-disease modules were already known to be broad spectrum in activity, namely, drugs targeting genes in these modules are already known to have significantly more therapeutic indications on average.

In each of these studies, traditional distinctions between diseases were often contrary to what the integrative models derived empirically from the data. For example, based on gene functional modules, we find molecular similarity between Crohn's disease and the infectious disease malaria, and also between actinic keratosis and colon cancer, the latter of which are known to share a therapeutic response to fluorouracil (Figure 3). Consideration of molecular and other links between diseases may offer novel opportunities for development of therapeutic interventions and fundamental understanding of disease pathophysiology. It is also possible to build and explore systems of disease relationships from many other possible clinical or molecular vantages, such as known etiologic causes.[7]

Future work in this area could provide novel insights into gastrointestinal disease through incorporation of

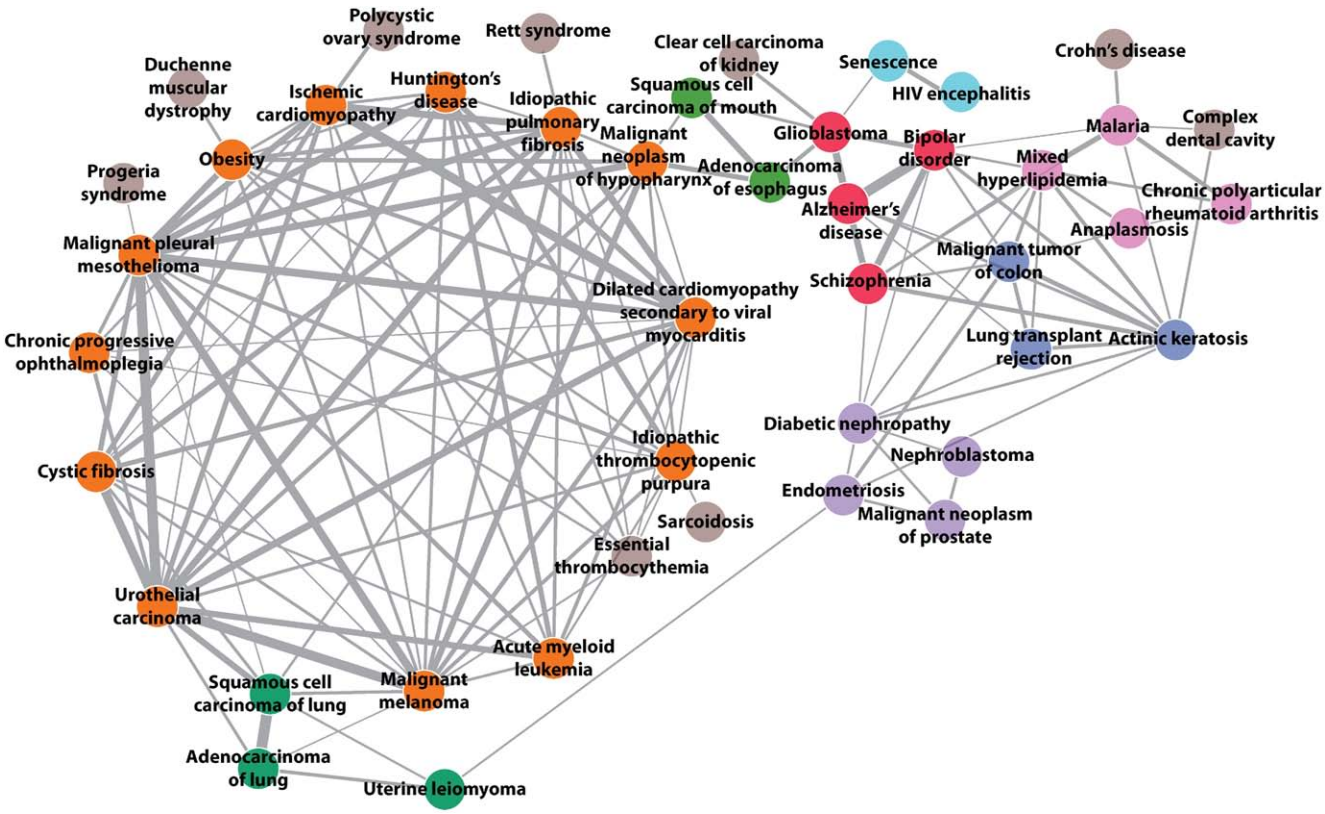

Figure 3. A network-based genomic nosology of molecular disease relationships defined by shared functional gene modules. The molecular relationships reveal that Crohn's disease is most similar to the infectious disease Malaria. Reprinted with permission from Suthram et al.[6]

relevant types of clinical or molecular data into the calculation of disease similarity. For example, common elements seen across diseases in imaging data could be incorporated as a basis for relating diseases by noninvasive tissue morphology (eg, vascular growth patterns). Such data might elucidate structural changes in alimentary tissues that distinguish clinically informative relationships between chronic gastrointestinal diseases and formation of gastrointestinal neoplasms. Similarly, incorporation of metabolic profiling data could illuminate pathways or signals connecting diseases known to share similar gastrointestinal complications.

Integrative Approaches for Biomarker Detection

There is an imminent and pressing need to identify molecular biomarkers that can inform patient diagnosis or prognosis toward ideal outcomes. However, the task of using high-throughput molecular data to identify clinically relevant biomarkers has proven to be a considerable challenge. To be effective as a clinical tool, molecular biomarkers need to exhibit appreciable degrees of sensitivity, specificity, and reproducibility. Ideally, clinical biomarkers are obtained through noninvasive means, therefore placing a priority on measurements that can be drawn from peripheral tissues or biofluids such as blood or urine.

Although many studies have demonstrated the diagnostic or prognostic characteristics of gene expression (ie, RNA) measurements sampled from primary tissues, the inherent variability of the technology and invasiveness of the approach have limited their clinical utility. Most biomarkers in clinical use today are sampled from peripheral tissues, such as blood or urine, measuring compounds or molecules that serve as surrogates of some underlying clinical state. Thus, a major challenge in biomarker discovery is to connect the molecular basis of disease with those changes in molecular markers more easily measured in peripheral tissues, but which might be physiologically far removed from the primary affected site. Integrative approaches using translational bioinformatics can be used here too, to create models and networks of relationships, which enable the ability to connect underlying molecular phenomena with broader physiologic effects.[8]

Proteomics technologies have been applied to identify large catalogs of proteins that are detectable in peripheral tissues and fluids such as blood and urine, which serve as biomarker candidates. However, the ability to sort through these candidates to identify proteins uniquely

sensitive to a particular clinical state stands as a significant challenge. We proposed a bioinformatics method to identify candidate blood-based protein biomarkers by integrating gene expression profiles of disease with established catalogs of proteins detectable in plasma and urine.[9] We constructed a network in which proteins and diseases were connected to each other if the gene encoding that protein is found to be significantly up-regulated in the disease as measured by the RNA expression. Through this integrative approach, we determined that approximately 80% of proteins measurable in blood plasma could be connected to multiple diseases, putatively limiting their potential as a discriminating biomarker of a specific disease. Although it is a naïve assumption that changes in RNA expression will necessarily lead to changes in plasma levels of a protein, this method could reduce the candidate space of protein biomarkers to the 20% of the plasma proteome likely harboring the most specific candidate biomarkers.

Building on the method of connecting RNA expression to peripheral protein biomarkers, we identified and validated a novel serum protein biomarker for cross-organ transplant rejection.[10] In this study, we integrated 3 biopsy-based microarray studies of acute rejection (AR) from pediatric renal, adult renal, and adult cardiac transplantation and identified 45 genes up-regulated in all 3 experiments. We intersected this set of genes with genes whose products are found in the plasma proteome, and chose 10 proteins for serum enzyme-linked immunosorbent assays in 39 renal transplant patients. The results validated 3 proteins that were significantly higher in AR, which were also significantly higher during AR in the 63 cardiac transplant recipients studied. Our best marker, serum PECAM1, identified renal AR with 89% sensitivity and 75% specificity, and also showed increased expression in AR by immunohistochemistry in renal, hepatic and cardiac transplant biopsies.

Another example of improving biomarker specificity is in the integration of serologic "antibodyome" measurements with publicly available gene expression profiles of kidney compartment and other normal tissues to identify compartment specific, non-HLA antibody responses putatively associated with kidney transplant rejection.[11] Serum profiles of ~5000 antibodies were obtained from postoperative renal transplant patients using ProtoArrays. Antibodies found to be significantly elevated in postoperative patients were evaluated for enrichment in the genes specifically up-regulated in specific kidney compartment tissues or other nonkidney tissues. Although the antibody profiles and gene expression profiles were measured from unmatched individuals, several of the compartmental responses predicted by the integrative analysis were experimentally validated. These studies demonstrate that integrative analyses can be used to augment the primary clinical data with publicly available molecular measurements to yield novel translational findings.

Integrative Approaches to Connect Therapeutics and Disease

Integrative approaches to drug discovery are providing novel opportunities for therapeutic development that extend beyond the 1-drug, 1-target paradigm that typifies traditional drug discovery. The number of pharmaceutical compounds gaining US Food and Drug Administration approval is in decline, and therefore there is a pressing need to explore new approaches to connect drug compounds with therapeutic indications. Translational bioinformatics can be employed to develop systems-oriented approaches to exploring broader therapeutic relationships between drug therapeutics and disease indications. The advantage of using a systems-oriented view over the traditional approach is that many drug compounds can be computationally evaluated simultaneously for many disease indications. Furthermore, integrative approaches can incorporate diverse types of data into a unified integrated system, including molecular and chemical characteristics of drugs and various forms of genomic or clinical data.[12]

The network paradigm serves as a powerful basis for integrative methods that enable systems-oriented approaches to therapeutic discovery. Under the network paradigm, drugs, diseases, and other types of relevant data are represented as nodes, and edges (connections) are drawn between these nodes based on some established or computed associative metric (eg, drugs connected to diseases they are approved to treat). Once a network model is established, it can be systematically explored to discover novel therapeutic relationships revealed by the complex network of relationships inherent in the network. For example, novel associations may be drawn between drugs and disease if they share the same neighborhood or clique in the network model graph. Network models are flexible enough to allow integration of a broad assortment of chemical, genomic, or clinical data types that might be informative of therapeutic potential.

Such network approaches can be used on primary disease samples or even cell lines representative of diseases like cancer, where genes and chemotherapeutics are both represented as nodes, and connections are drawn between nodes based on strong correlations of chemotherapeutic efficacy and gene expression measurements.[13]

We leveraged knowledge of approved and practiced drug treatment indications to enable a "guilt-by-association" network approach for systematic analysis of drug–disease relationships.[14] In this study, diseases were repre-

sented as nodes, and connections were drawn between diseases if they were found to have similar therapeutic profiles. A therapeutic profile for a disease was composed of the set of drugs indicated as approved or used off-label in practice to treat a disease, and the connectivity metric considered the set-based similarity of therapeutic profiles between a pair of diseases. This network was explored for novel drug indications using a "guilt-by-association" approach, in which a drug was identified as a putatively novel indication for a disease if it was found in the therapeutic profile of closely connected diseases in the network. The results of this analysis suggested a number of putatively novel drug–disease associations that were found to be under active clinical investigation, as well as completely novel associations. These included a suggested therapeutic connection between atorvastatin and Crohn's disease, which was shown to have positive effect in a clinical trial (ClinicalTrials.gov NCT00454545), as well as a suggested therapeutic connection between rituximab and gastric ulcer, which has not yet received formal clinical investigation.

In other work, Campillos et al[15] used side effect information obtained from drug package inserts to construct a network that was used to infer novel drug target indications using side effect similarities. A number of novel target indications predicted by this method were experimentally validated, including novel target associations between rebeprazole, which is used to treat ulcers, gastroesophageal reflux disease, and Zollinger-Ellison syndrome, and the targets DRD3 and HTR1D. The network analysis placed rebeprazole in a subnetwork enriched for drugs with neurological indications, rather than other antiulcer agents.

Future Directions in Systems Gastroenterology

We have highlighted a number of integrative approaches that make use translational bioinformatics to enable systems-oriented perspectives for studying complex clinical phenomena. We have also demonstrated the translational utility of publicly available data, as well as integrative approaches that can be employed to incorporate public data into clinical investigation to augment primary data, or enable novel translational hypotheses. The approaches and methods underlying the highlighted studies are generalizable and extensible, suggesting an immediate opportunity to repurpose these integrative approaches toward long-standing problems in gastrointestinal disease.

The application of integrative and systems-oriented approaches have already uncovered a number of novel translational hypotheses relevant to gastroenterology. However, it is imperative that future work in this area gains full engagement of the community of clinical experts in gastroenterology, who, after consideration for the expansive opportunities and capabilities offered by public data and translational bioinformatics, can work to formulate and drive novel translational hypotheses in gastroenterology from integrative and systems oriented perspectives.

Another important factor enabling advances in gastroenterology is the continued contribution of relevant molecular and clinical data into the public domain. The repurposing of these data through translational bioinformatics extends the value of these data well beyond their role in the originating experiment, and the accumulation of these data into public repositories enables novel opportunities in translational research that make use of power of aggregate data to enable completely new perspectives on long standing problems in clinical science.

Perhaps the most difficult challenge to address is the training of quantitatively minded individuals to conduct research in gastroenterology using the methodologies described herein. Driving trainees into this type of research will first take a community of mentors in gastroenterology who see the value of computational- or data-driven research. It will then take a second generation of mentors in gastroenterology who actually conduct research using these methods. Perhaps eventually, the community will see that a fellow who wants to learn and use computational methods for his or her research career is no more unusual than the fellow who wishes to learn wet-bench techniques. Those future fellows who are dual trained in both gastroenterology and computational methods will hold the most promise, and we look forward to many clinically useful data-driven tools to be developed by these individuals.

Supplementary Material

The first 5 references associated with this article are available below in print. The remaining references accompanying this article are available online only with the electronic version of the article. To access the remaining references, visit the online version of GASTROENTEROLOGY at www.gastrojournal.org, and at doi:10.1053/j.gastro.2010.07.024.

References

1. Butte AJ. Translational bioinformatics: coming of age. J Am Med Inform Assoc 2008;15:709–714.
2. Butte AJ, Chen R. Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics. AMIA Annu Symp Proc 2006:106–110.
3. Dudley JT, Tibshirani R, Deshpande T, et al. Disease signatures are robust across tissues and experiments. Mol Syst Biol 2009;5:307.
4. English SB, Butte AJ. Evaluation and integration of 49 genome-wide experiments and the prediction of previously unknown obesity-related genes. Bioinformatics (Oxford, England) 2007;23:2910.

5. Sirota M, Schaub MA, Batzoglou S, et al. Autoimmune disease classification by inverse association with SNP alleles. PLoS Genet 2009;5:e1000792.

Reprint requests

Address requests for reprints to: Atul J. Butte, Department of Pediatrics, Stanford University School of Medicine, 251 Campus Drive X-163, MS-5415, Stanford, CA, 94305-5415; e-mail: abutte@stanford.edu.

Conflicts of interest

The authors disclose the following: AJB is or has served as a scientific advisor and/or consultant to NuMedii, Genstruct, Prevendia, Tercica, Eli Lilly and Company, and Johnson and Johnson. JTD has served as a consultant to NuMedii.

Funding

Grant support from the National Library of Medicine (T15 LM007033 and R01 LM009719).

References (online only)

6. Suthram S, Dudley JT, Chiang AP, et al. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. PLoS Comput Biol 2010;6:e1000662.
7. Liu YI, Wise PH, Butte AJ. The "etiome": identification and clustering of human disease etiological factors. BMC Bioinformatics 2009;10(suppl 2):S14.
8. Yang X, Deignan JL, Qi H, et al. Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. Nat Genet 2009;41:415–423.
9. Dudley JT, Butte AJ. Identification of discriminating biomarkers for human disease using integrative network biology. Pac Symp Biocomput 2009:27–38.
10. Chen R, Sigdel TK, Li L, et al. Differentially expressed RNA from public microarray data identifies serum biomarkers for cross-organ transplant rejection and other conditions. PLoS Comp Bio. Accepted.
11. Li L, Wadia P, Chen R, et al. Identifying compartment-specific non-HLA targets after renal transplantation by integrating transcriptome and "antibodyome" measures. Proc Natl Acad Sci U S A 2009;106:4148–4153.
12. Perakslis ED, Van Dam J, Szalma S. How informatics can potentiate precompetitive open-source collaboration to jump-start drug discovery and development. Clin Pharmacol Ther 2010;87:614–616.
13. Butte AJ, Tamayo P, Slonim D, et al. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci U S A 2000;97:12182.
14. Chiang AP, Butte AJ. Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. Clin Pharmacol Ther 2009;86:507–510.
15. Campillos M, Kuhn M, Gavin AC, et al. Drug target identification using side-effect similarity. Science 2008;321:263.