2010 Translational bioinformatics year in review

Russ B Altman,1 Katharine S Miller2

ABSTRACT
A review of 2010 research in translational bioinformatics provides much to marvel at. We have seen notable advances in personal genomics, pharmacogenetics, and sequencing. At the same time, the infrastructure for the field has burgeoned. While acknowledging that, according to researchers, the members of this field tend to be overly optimistic, the authors predict a bright future.

INTRODUCTION
Reviewing a year’s worth of papers in translational bioinformatics is extremely rewarding, but it is stressful to choose only a few to highlight. Progress is marvelous, but an annual snapshot might simply amuse future generations for its inaccuracy rather than identify what is really important.

The process is straightforward: solicit nominations from colleagues, review the journal table of contents for the year, think about which sources to trust, and attempt to spot research that either has had an early impact or has the potential to create valuable opportunities in the future. But in the end the choices are subjective, and there is much room for debate and disagreement. We can also apply a translational bioinformatics “filter,” asking if the paper is about informatics research that develops or applies methods that link basic molecular, genetic, and cellular data to clinical concepts such as drugs, diseases, symptoms, and patients—a very rough definition of the topic, and one with which others might disagree.

This year, an initial review of approximately 12–14 months of papers identified 62 semi-finalists, all of which are listed in Appendix A (apologies to the high quality papers missed). Twenty-five of these papers are highlighted here and organized into five topics: personal genomics; drugs and genes; infrastructure for translational bioinformatics; sequencing and science; and warnings and causes for hope. At the end, we look into a crystal ball to predict what might happen in 2011–2012. These are exciting times for this nascent field.

PERSONAL GENOMICS
Genetic knowledge has clinical value
One question raised in this era of personal genomics is whether the availability of genetic knowledge can already provide clinically useful information. That is, are we at the point where we know enough to actually help people who possess their genomic data?

In a test of that question, Ashley et al1 looked at the genome of a 40-year-old patient with a family history of vascular disease and early sudden death. The researchers assessed the patient’s rare genetic disease carrier status, common disease risk, environmental risks, and drug response to determine whether they could come up with clinically actionable advice. And they did: the analysis indicated the patient had a high risk for coronary artery disease, that he had rare variants of three genes associated with sudden cardiac death, that statins were likely to work, that he might require only a low dose of warfarin, and that he had no risk alleles indicating that the statins would cause negative side effects. In light of his family history, his genome, and the clinical evaluation, the researchers then recommended that he start taking statins.

The lesson: While whole genome data is now providing clinically useful information, there is a need for methods to comprehensively analyze such data in a defined clinical context. And while Ashley et al’s approach provides a proof of principle, it also points out a few challenges, including the fact that we have no way to interpret millions of genetic variants. That is where the future lies.

Empowerment through do-it-yourself genetic testing
In 2010, in an apparently distinct violation of Myriad Pharmaceuticals patents, researchers Salzberg and Pertea2 successfully empowered individuals to run their own tests for mutations in breast cancer genes. The researchers developed a computational assay that patients can use in the privacy of their own homes to test for 68 known BRCA variants.

The authors’ goal was to demonstrate the folly of gene patents. Myriad has made a good business out of testing for BRCA1 and BRCA2 mutations. But genetic testing, the authors say (and many agree) is no different from measuring one’s temperature or blood pressure. Requiring permission from a private company before doing such testing just does not make sense. The authors write: ‘It is hard to envision how the patent holders can enforce their claims in this scenario. Our contention is that these patents never should have been awarded and that no private entity should have rights to the naturally occurring gene sequences in every human individual.’ (This quote elicited applause at the annual AMIA Translational Bioinformatics summit). For this brave paper we owe a debt of thanks to Dr Salzberg and Dr Pertea.

Electronic medical records (EMRs) replicate genetic signals
Many of us have wondered whether EMRs can replicate the genetic signals found in the much more controlled research setting of a genome-wide association study (GWAS). Indeed, some have argued that clinical medical records are tangled messes. But last year, Ritchie et al3 demonstrated that they are in fact goldmines, as shown in figure 1. This is really

...
Figure 1  Genome-wide association study (GWAS) signals replicated in an electronic medical record (EMR) study at somewhat lower ORs. For each disease, the odds of identifying a particular single nucleotide polymorphism (SNP) involved (column 2) in a carefully controlled GWAS study (red squares) was compared to the researchers’ ability to identify it in their EMR study (red diamonds). And although the blue diamonds are a bit to the left of the red squares, most of the blues (except perhaps the three at the bottom) are above the 1.0 OR. AF, atrial fibrillation; CD, Crohn’s disease; MS, multiple sclerosis; RA, rheumatoid arthritis; T2D, type 2 diabetes. Reprinted from Ritchie et al.4 with permission from Elsevier.

a landmark outcome: the researchers tested 21 previously reported gene–disease associations (discovered with carefully controlled cases and controls) within an EMR and replicated most GWAS results, although typically with smaller ORs. With this finding, some researchers should perhaps contemplate conducting gene–disease association studies using a cost-effective EMR rather than a case–control trial. The National Institutes of Health (NIH) eMERGE network continues to explore the utility of the EMR for genetic discovery.

Novel associations found using web-based phenotypes in a GWAS study

In the same way that Ritchie et al.’s paper4 showed the value of EMRs for gene–disease associations, 23andMe (for which RBA serves as an advisor) showed the value of phenotype information gathered through a web-based questionnaire.5 23andMe provides personal genomic information to customers who have provided only a vial of saliva, an email address, and a credit card number (to pay for the service). But many customers also consent to answer an online questionnaire about their phenotypes.

In a paper published in PLoS Genetics,6 the company used these questionnaire responses and associated single nucleotide polymorphism (SNP) data to perform multiple GWAS studies. They were able to replicate several known associations and also identify many new genotype–phenotype associations, as illustrated in figure 2, in diverse areas such as freckling, photic sneeze reflex, hair curl, and asparagus anosmia. Thus, self-reported data collected over the web can be used to support research and discovery.

Phenome-wide association study (PheWAS): understanding the web of disease

Biologists know that certain SNPs that are associated with particular diseases might also be associated with other diseases. A greater understanding of which SNPs are involved in multiple diseases could provide insight into the web of diseases and how they are related at the molecular level.

To investigate that idea, researchers at Vanderbilt University used EMRs to identify cases and controls for 776 diseases.7 They looked at five disease-associated SNPs across that population. As shown in figure 3, the study replicated known associations with the index diseases and also found 19 new (statistically significant) associations between these SNPs and diseases that had not been previously associated with any kind of genetic association. Some of these provide intuitive connections between related diseases, and some are big surprises.

While further analysis is needed to determine the validity of the SNP–diseases associations and the statistical thresholds for clinical significance, the paper suggests that EMRs can connect SNP data with new diseases and yield interesting hypotheses that may merit experimental follow-up.

Environment-wide association study (EWAS): giving environmental affects their due

We know that environmental factors combine with genetics to change the risk of developing many diseases. Yet such factors are notoriously hard to study. Patel et al.8 took a novel approach to the problem by borrowing the GWAS methodology to search for environmental factors associated with disease on a broad scale. They call the approach an EWAS—environmental-wide association study.

For this study, the authors used the National Health and Nutrition Examination Survey (NHANES) dataset, a collection of longitudinal data that is very rich with lifestyle and exposure information, to look for associations between 266 environmental factors and glucose values in the range typical for type 2 diabetes. They found significant associations between diabetes diagnoses and exposure to the pesticide heptachlor as well as exposure to polychlorinated biphenyls (PCBs), as illustrated in figure 4. They also found a protective effect of β-carotenes.

This paper is an important early step in balancing the equation relating phenotype to a combination of genetics and environment. We can measure a million SNPs, but our ability to measure environmental variables is extremely limited. So kudos to Patel et al. for this work.

GWAS reality check: rare variants may create synthetic associations

The success of GWAS studies has been much debated over the last few years. Some people have concluded that GWAS have been a failure, while others argue that GWAS have led to insights
Figure 2 Using phenotype data from online questionnaires, 23andMe found novel associations between certain SNPs and (A) hair curl, (B) asparagus anosmia, (C) photic sneeze reflex, and (D) freckling. In these Manhattan plots, the non-hits are the low-rises and the high-rise peaks represent areas where there were strong, statistically significant hits, appropriately corrected for multiple hypothesis testing. Reprinted from Eriksson et al.5

into the mechanisms of disease. New work by Dickson et al. is both sobering and helpful in sorting this out.

Essentially, GWAS look for common gene variants that cause disease. But Dickson et al. note that common diseases might not be associated with common variants at all. They hypothesized that hits from GWAS might originate from rare variants that combine to create a signal—a ‘synthetic association.’ Because of the diversity of variation in the human genome, there could well be many ways to get a disease from rare, even personal, mutations across the genome.

Using simulations to explore that possibility, the researchers found that rare, causal mutations for hearing loss and sickle cell anemia create genome-wide significant associations over very large genomic regions even though GWAS identified a common variant that had nothing to do with the cause of the disease.

The conclusion, then, is that GWAS hits may be pointing to the right areas of the genome but the wrong individual variations. When researchers get a GWAS hit, they probably should re-sequence the surrounding region and look for rare variants that might be combining to cause the disease. This is a very important conclusion for the post-hoc reinterpretation of accumulated GWAS studies.
**DRUGS AND GENES**

**Data mining for drug side effects**

Kuhn et al.\(^7\) mined Food and Drug Administration (FDA) labels to create a publicly available database (SIDER) connecting 888 drugs to 1450 side effects. Many research groups have welcomed this resource, but there remains a caveat: since the data were mined from the FDA, they cover only the common side effects identified in phase III clinical trials. It is likely that this information will need to be augmented with off-label side effects. Nevertheless, it is a very valuable resource and we are all grateful for it.

**Data mining for multiple drug adverse events**

Last year, Harpaz et al.\(^8\) created a method to associate multiple drugs with adverse events using the FDA adverse event reporting system (AERS). Many consider the AERS nothing but a noisy, bias-filled data resource, but this paper proves that AERS can yield useful signals. The authors used association rule mining to discover multi-item adverse drug events in the AERS database. Roughly one third of the associations they found are novel.

This work is important because synergistic or antagonistic interactions between drugs will not be found during the FDA approval process, which focuses on single drugs in controlled settings. These interactions will not be noticed until patients, doctors, nurses, and pharmacists start reporting them.

Although this is very early work, it suggests that population-based analyses of drug effects can not only discover interactions between drugs but could also help discover the molecular basis of those effects. By identifying the molecular targets of interacting drugs, one could then hook into the pathways that create the interactions.

**Structure informs off-target drug effects**

It is always satisfying to see two distantly related informatics areas combined with good effect, as in the paper by Chang et al.\(^1^1\)

These researchers brought together structural and systems biology to explore how a drug creates unforeseen off-target effects.

First, they created a systems model of kidney function that included several hundred genes involved in more than 1500 reactions. Independently, they selected a small molecule for study: the cholesteryl ester transfer protein inhibitor torcetrapib, which was withdrawn from phase III trials for causing fatal hypertension in some patients. They screened torcetrapib for ligand binding across the entire human structural proteome.

This identified 41 off-target metabolic proteins that they then used to perturb the systems model of the kidney to assess potential effects. In fact, the perturbation successfully explained how torcetrapib causes hypertension in patients with certain genetic deficiencies.

These results show that it is possible to combine a math model with 3D structural interaction modeling to obtain actionable hypotheses about how a drug causes a side effect. It is an approach that has the potential to be generalized as an aid to drug discovery. Of course, we will also need to have models of the drug response for every drug on the market as well as the 3D structures of every human protein! We are not quite there yet, but we should all remember this paper as these models begin to accrue.

**Finding new drug targets for malaria**

Using the genome of the malarial parasite *Plasmodium falciparium* (the species that causes the most death in Africa and elsewhere), Plata et al.\(^1^2\) built a full metabolic model of all its capabilities. The model includes the metabolic enzymes involved and the transformations they effect. They then used an approach called flux balance analysis (FBA) to find places in the network where enzymes could be disrupted to hamper the metabolic process.

This analysis identified 90% of known effective knockouts as well as 40 new potential choke points—targets at key points in the network that appear to be uniquely critical. Some of these were shown in the laboratory to inhibit the growth of the parasite. This paper is exciting because it represents a very rational analysis of drug action and drug effects and seeks to find drugs to stop a very deadly parasite.

**Optimizing treatments to prevent multi-drug resistance**

Multi-drug treatments have been very effective for tuberculosis, HIV, and other infectious diseases but open up the possibility that a pathogen may become resistant to multiple drugs at the same time. This spells trouble for both the patient and public health. Torella et al.\(^1^3\) created a mathematical model to explore ways of preventing this outcome.

Pairs of drugs can function either antagonistically or synergistically. Synergism tends to clear infections faster because the drugs kill the infection in two different ways. But synergy also creates a greater risk of multi-drug resistance because any bug with a mutation that escapes both drugs will have a strong selective advantage. It turns out that the optimal relationship between the drugs (on a scale from pure antagonism to pure synergism) depends on resource availability. If the bug is struggling for limited resources, it is better to treat with more antagonistic drugs that clear the infection more slowly but do not confer such a selective advantage on a particular mutant. If, on the other hand, there is less resource competition, more synergism can be tolerated without risking multi-drug resistance, as illustrated in figure 5.

This work shows again that math modeling can provide new intuition contrary to simple common sense. As we move toward increasing poly-pharmacy in the treatment of HIV, tuberculosis, cancer, schizophrenia, depression, and diabetes, we need to build our capability to rationally dissect the modes of drug interaction. This paper is a great start.

---

**Figure 5** Here, drug interactions (x-axis) range from antagonistic (they actually reduce each other’s effectiveness) to neutral (zero) to synergistic (the two drugs enhance each other’s ability to kill off an infection). One might expect that synergy would be ideal; let’s kill off as much as we can as fast as we can. But the treatment efficacy (blue line) flattens out above a certain level of synergy, while the prevention of multi-drug resistance (black line) continues to go down. Optimal results are found somewhere left of the shaded area. Reprinted from Torella et al.\(^1^3\).
INFRASTRUCURE FOR TRANSLATIONAL BIOINFORMATICS

Ontology recommender

We are often reminded, ‘Do not invent an ontology if it already exists.’ And then we think, ‘How can we learn if it exists?’ To answer that question, Jonquet et al.14 of the National Center for Biomedical Ontology (NCBO) created a tool to help researchers find the ontologies they need.

Here is how the NCBO web application works: after the researcher submits a textual corpus, the service scores the possible ontologies that might fit the corpus and presents them to the researcher. The score uses concepts of coverage, connectivity, and size based on keywords from the corpus and does a semantic analysis to determine a match between the corpus and the available domains covered by ontologies in the NCBO collection. This very useful tool should encourage ontology reuse, and discourage the creation of unnecessary competing ontologies.

Global identifiers for clinical data

Researchers studying autism found that data from multiple sites often came from the same patients, some of whom received care in multiple locations. Double-counting these patients threatens to confound key statistical analyses. To prevent double-counting, the researchers generated global IDs for participants in their large population-based studies.15 They implemented a one-way encrypted hashing algorithm based on commonly available demographics including mandatory fields such as first name, last name, and date of birth, and optional ones such as maternal and paternal information. They then assessed the ability of these hashes to identify patients who appear in more than one dataset. The algorithm performed well on samples of 1 million simulated patients and 8000 real people. The identifiers allowed resolution of similar cases in 96% of children and 77% of parents. In sum, generic unique identifiers provide a useful way to link datasets of patients. We may see increased use of these methods as we tap more and more into EMRs as a research resource.

Tools for conducting genomic-era research with EMR data

The Center for Informatics for Integrating Biology and the Bedside (i2b2) is one of the National Centers for Biomedical Computing that has been supported by the NIH for the last 6 years. As part of the i2b2 effort to create open-source software for translational research, Murphy et al.16 created software to identify research cohorts from EMR data. Initially implemented at Partners Healthcare at Harvard University, it is now used nationally at several sites. After these tools are linked to the local EMR, they enable clinical investigators to use i2b2 infrastructure for cohort finding, clinical analysis, and data mining. Thus, this tool creates a de facto standard for clinical research infrastructure.

Wiki-based collaborative knowledge base

In relatively specialized fields that need a shared knowledge base, hiring a curator to create and maintain information resources can be prohibitively expensive. Some researchers have found a cost-effective solution: wikification. Barriot et al.17 provide a model for using wiki technology for shared community annotation: they created a knowledge base focusing on the gene-to-phenotype network for human congenital heart defects (CHD). It links thousands of concepts together and is edited by credentialed members of CHDWiki.

The lesson: In a field with a relatively small and close-knit community, serious scientists can collaboratively enter and integrate information in a structured format to create a useful shared resource.

Opportunities for cloud computing

Too often, we write papers knowing that our methods and results are reproducible only if a very motivated colleague hires someone to rewrite/decipher our code and then rerun the experiments on the difficult-to-find initial datasets. It is as if we are trying to make our work as difficult as possible to reproduce. So imagine a publishing system that includes a repository where researchers submit an image of the computational operating system on which they did the experiments described in a published paper. Imagine further that anyone can take that image and recreate the exact computational environment used by the authors, thus allowing anyone to rerun the scripts and re-analyze the data. It would generate an exact copy of the work reported in the paper. This is the proposal Dudley and Butte outline in their paper on in silico research in the era of cloud computing.18 They describe a ‘whole system snapshot exchange’ (WSSE) that is stored within a cloud computing infrastructure. This type of system could lead to a level of reproducibility that would minimize redundancy in research and likely accelerate the progress of science.

An opt-out approach to human consent for genomics

Genomics research requires human subjects to provide their consent for analysis of DNA from their blood and/or tissues. Pulley et al.19 describe Vanderbilt University’s success with the implementation of an opt-out consent procedure. When a patient is giving consent, the research is described and it is emphasized that the patient’s samples will be de-identified for that research. The patient then checks (or leaves blank) a box that says ‘Do not use my leftover blood for the DNA Databank’ and signs the form. Thus patients who leave the box blank are included in the biobank unless they opt out. The Vanderbilt procedure, accompanied by a substantial education and outreach effort, has led to a very large number of accrued participants in their genetic studies.

Many institutions have looked at and rejected the possibility of using such an opt-out rather than an opt-in (where the patient must affirmatively check a box to participate) approach. Some bioethicists and lawyers have said that opt-out is not ethical. At Vanderbilt, the consent form is part of the ‘agreement to pay’ and is likely one of many forms signed during admission or before a clinic visit. The fear is that the patient has not made a reasoned, fully informed choice.

Understanding this fear, it appears that Vanderbilt has made extensive efforts with billboards, posters, focus groups, and other activities to educate the local community about the value of participating. Their high accrual rate from the opt-out approach is impressive and gives them a large cohort with which to pursue scientific questions.

SEQUENCING AND SCIENCE

Evolutionary pressure in tumors

Many studies of cancer focus primarily on a limited set of genes known to be important. Lee et al.20 attempted to obtain a more complete picture of the mutational spectrum for cancer using high throughput DNA sequencing. They sequenced the entire genome from a lung cancer tumor and compared it to the genome from normal tissue just centimeters away. They found 50,000 high quality SNPs and many large scale rearrangements. They identified different types of apparent evolutionary pressure within the tumor, with selection against genes that are highly expressed in normal tissue, and a high mutation rate in the class of proteins called kinases. This is the first evidence for distinctive evolutionary pressures on tumors as they grow. I expect we will
see many similar papers in the future, and these may provide opportunities for new drug development programs.

Gut microbes and human health
The bacteria in our gut help us digest food, metabolize drugs, and secrete small molecules that enter the blood stream. To better understand the importance of gut microbes in human health, Qin et al. sequenced the DNA in the feces of 124 Europeans. The summary statistics are fascinating and amazing. For example, the researchers assembled and characterized 5.5 million microbial genes and were able to infer the presence of about 1100 different species of bacteria in the 124 people tested. On average, each person has 160 species in the colon. It appears that the flora in our digestive systems are not identical, and are quite diverse. This paper suggests a method for potentially associating gut microbes (a critical environmental factor) with health and disease.

Stem cell informatics
Stem cell biology is likely to be a key driver of informatics innovation in the next decade and Xu et al. provided a good review of the resources necessary to get started in stem cell informatics. They describe databases, algorithms, software, and data available for embryonic stem cell research. Existing tools are promising, but holes remain in our ability to broadly integrate modalities, so this is an area with both great promise and great need.

Moving from model organisms to translation
Funding agencies support research on model organisms with the understanding that because these organisms share common biology with humans, we can learn things that also apply to human health and disease. Too often, however, we do not adequately connect the dots between the model systems and the relevant human health problems.

McGary et al. report a novel approach for identifying non-obvious equivalences between model organisms and humans. Their idea is that a phenotype X in one organism may result from a set of genes whose homologues in a different organism conspire to create an apparently different phenotype Y. So phenotype X may be a phenotype-equivalent of phenotype Y in another organism because they use the same genetic network. They called these phenotype homologues or ‘phenologs.’

As an example, genes involved in human angiogenesis have yeast homologues. Yeast, of course, does not have blood vessels (!), but perhaps these genes in yeast can provide insights about the molecular details of angiogenesis. The researchers also identified a potential worm model of breast cancer, a mouse model for autism, and a plant model of neural crest development. This paper demonstrates a smart way to deliver on the promise of model organisms by tightening up the molecular criteria for what defines a model organism.

WARNINGS AND CAUSES FOR HOPE
Excessive optimism
According to Boulesteix, bioinformaticians are too optimistic. She analyzed the impact of our tendency to fish for statistical significance and publish only positive results. Her findings: there are fundamental flaws in current approaches. The recommendation: publish negative results; publish documented code and data more transparently; and perhaps develop approaches like the WSSE proposed by Dudley and Butte in the section on ‘Opportunities for cloud computing’ above. Based on this research by Boulesteix, some of what we report in this paper is probably wrong and overly optimistic. So that is a useful note of caution.

People sometimes do better than machines—so use them
It is also humbling (for fans of computers) and perhaps even frustrating to note that crowd-sourcing, where a bunch of people collaborate to perform a task (typically on the web), can perform as well or better than our best algorithms. Cooper et al. built an online game for protein folding and compared the results with those produced by Rosetta, the automatic algorithm that has been one of the best performers at the international protein folding competition for many years. The authors found that the top humans excel at this task; they work collaboratively and explore complex search strategies better than the computer program. The lesson is that, at least for now, when solving hard problems human expertise should probably be included.

Hope lies in great research groups
One paper this year suggests that a great research group needs to foster competence, autonomy, and social connectedness. Good projects lie at the intersection of an individual’s talents, objectives, and passions.

CRYSTAL BALL FOR THE NEXT YEAR
2009’s Crystal ball predicted some of 2010’s successes but missed the mark on others.

2010 Hits
► The linking of clinical records to genomics data has resulted in discoveries.
► There has been more emphasis on drugs and ancestry in direct-to-consumer (DTC) genetics companies.
► Whole genome sequencing has been carried out for a cohort with a common disease.
► Semantics has now been used in literature mining for knowledge discovery.

2010 Misses (but we still predict they will happen in 2011)
► We have not really seen consumer sequencing yet (as opposed to genotyping).
► We have not quite seen cloud computing contribute to biomedical discovery.

2011 Crystal ball
► Informatics applications for stem cell science will increase.
► Important biological discoveries will result from text mining.
► Population-based data mining will yield important biomedical insights.
► Systems modeling will suggest useful poly-pharmacy for common diseases.
► Immune system genomics will emerge as a very powerful source of data.

Next year I am hopeful that I can report a perfect batting average.

Acknowledgments To assist us in preparing this review, Robert Yao transcribed the talk RBA gave at the Third Annual AMIA Translational Bioinformatics Summit held in March 2011. We also thank those who nominated papers for consideration.

Funding The author is supported by National Institutes of Health grants LM05652, GM61374, and GM072970, and grants from Microsoft Research and Lightspeed Ventures.
Competing interests RBA is an advisor to https://www.23andme.com, mentioned in Erikson et al.18

Provenance and peer review Commissioned; externally peer reviewed.

REFERENCES
3. Altman RB, Kohnan IS. Clinical data repositories: less than meets the eye or more valuable than you’d expect? Biomedical Computation Review 2009;9:2–3.

APPENDIX A

Seminal papers reviewed

Perspective


365

Downloaded from https://academic.oup.com/jamia/article-abstract/18/4/358/732300 by guest on 23 April 2019


