HHS Public Access

Author manuscript

Hum Genet. Author manuscript; available in PMC 2016 May 01.

Published in final edited form as:

Hum Genet. 2015 May ; 134(5): 455–457. doi:10.1007/s00439-015-1545-6.

Human Genetics Special Issue on Computational Molecular Medicine

Rachel Karchin, Ph.D[1,2] and **Melissa S. Cline, Ph.D**[3]

[1]Department of Biomedical Engineering and Institute for Computational Medicine Johns Hopkins University rachel.karchin@gmail.com

[2]Department of Oncology, Johns Hopkins Medical Institutions

[3]Center for Biomolecular Engineering, University of California, Santa Cruz cline@soe.ucsc.edu

Computational analysis of genomic big data has the potential to transform how clinical medicine is practiced, leading to increasingly personalized diagnosis, prognosis and therapeutic decision making. The past ten years have seen an explosion of large, publicly-funded sequencing and genotyping projects studying both healthy individuals, and those with complex or Mendelian disease phenotypes. As costs decrease, whole exome sequencing (WES) has become common, and many researchers are now shifting to whole-genome sequencing (WGS). In spite of the challenges involved in collection of sequencing data, many now consider the greatest challenge to be one of interpretation; as the joke goes, the $1,000 genome is now coupled with the $10,000 interpretation. In this special issue of *Human Genetics*, we present reviews on progress and some of the remaining challenges in several broad areas of modern and medically-relevant genomic and transcriptomic - and other "omics" - interpretation, which we define as computational molecular medicine.

Modern sequencing methods have revolutionized the study of Mendelian diseases. Bahlo et al. review approaches to identifying Mendelian causal variants in the era before high-throughput DNA sequencing. They describe how family information and older statistical and computational methods used to identify linkage and identity-by-descent can now be leveraged to improve WES/WGS data quality and identification of inheritance models. Some of these models have previously been hard to identify, e.g., rare *de novo* germline mutations and sporadic somatic variants.

The study of complex diseases has also been transformed by large-scale DNA and RNA sequencing projects. However, many questions remain unanswered. Sadee et al. review the problem of "missing heritability" in modern genomics studies of complex disease and explore possible solutions. Potential causes include the failure of additive models of heritability to account for epistatic effects, the confounding influences of positive and balancing selection on detecting causal variants and ascertainment bias in current WES studies. They are optimistic about the growing popularity of WGS and RNAseq, which will

Correspondence to: Rachel Karchin.

enable discovery of previously unknown causal variants that effect gene regulation or affect RNA function through changes in conformation, stability and binding interactions.

Large-scale DNA and RNA sequencing projects have generated an abundance of data that present researchers with the temptation to “just throw data at the modeling problem”. However, Geman et al. argue that most methods fail in either their reproducibility or their inability to generate new biological knowledge, because they do not represent biological mechanisms in the structure of the model. This has led to two problems: overfitting and abstraction. While overfitting might not seem like a lasting problem in the era of big data, they argue that it is here to stay, given factors such as greater patient stratification within personalized medicine. Abstraction further confounds the issue. When the structure of the model fails to mirror the structure seen in the underlying biology, the results of the model become difficult to interpret in anything other than a post-hoc analysis. They argue that both problems can be addressed by applying prior knowledge in defining the structure of the model, which can at the same time reduce the complexity of the modeling problem. They review examples from the modeling of metabolic processes, signaling networks and tumorigenesis. They end with encouragement that encoding mechanisms into predictive models offers a win-win situation: to the computationalist in reducing overfitting, and to the biologist by improving the ability of the models to offer new hypotheses on causal mechanisms.

Pharmacogenomics is at the forefront of application of genomics to medical practice. Mooney reviews currently available resources for computational analysis, recent advances and remaining challenges to bringing genomic analysis of personalized drug response into the clinic. Computational work in this area is supported by initiatives to systematically extract patient data from electronic health records (EHR) and also by well-curated databases such as PharmGKB. EHR data is central to the Phenome Wide Association Study (PheWAS) approach, in which genetically matched populations can be tested for association with a phenotype, i.e., lab test results indicative of drug efficacy and adverse events. Like Sadee et al., he is optimistic about the potential of WGS, since many pharmacogenomic variants lie outside the exome. However, computational, medical and regulatory challenges to progress in this area are significant. Computational methods to predict the impact of pharmacogenomic variants have thus far been less effective than methods to predict deleterious or disease-causing variants. Even the "poster child" of the early days of pharmacogenomics -- genotype-based dosing of the anti-coagulation drug warfarin -- has not significantly reduced major adverse events [1], in spite of well-studied associations between warfarin response and variants in the *CYP2C9* and *VKORC1* genes.

Reinhold et al. review publicly available database resources available to study the associations between genomic data and response to targeted cancer drugs. These include pre-clinical, cell-line models of drug activity for over 50 cancer types and 40,000+ drugs. The NCI-60 cell line collection includes extensive omics data, including WES, RNAseq, gene microarray expression, micro-RNA expression, proteomic analysis and metabolite profiling. The Cancer Cell Line Encyclopedia (CCLE) and Genomics of Drug Sensitivity in Cancer (GDSC) cell line collections contain more extensive cell lines but fewer drug activity profiles. They suggest that hypotheses of molecular associations with drug response derived

from these databases can be further explored for correlation with the large genomic data sets from clinical samples available from the Cancer Genome Atlas. For example, an association between increased expression of the connexin 43 (*GJA1*) gene and resistance to temozolimide has been supported by increased expression of connexin 43 in TCGA glioblastoma tumor samples [2]. A review of algorithms used to infer associations between drug response phenotypes and datasets of omics measurements is also provided.

Few human genes have been as extensively studied for their impact on cancer as the tumor suppressor *TP53*. Deleterious mutations in *TP53* have been associated with inherited cancer predisposition and have been used as biomarkers of cancer risk and prognosis (reviewed in [3]). Masica et al. present the first double-blinded and systematic study of the prognostic value of bioinformatics methods designed to predict the impact of specific mutations in *TP53*. *TP53* was sequenced in a cohort of 420 head and neck cancer (HNSCC) patients and somatic mutations were identified. Fourteen diverse bioinformatics classification methods designed to predict deleterious mutations and results of an *in vitro* yeast assay of mutant *TP53* function -- measured by *WAF1* transactivation -- were used to assign a status of "disruptive" or "non-disruptive" for each *TP53* mutation observed in the cohort. The classifications from each of the 15 methods were blindly assessed by the Eastern Cooperative Oncology Group (ECOG) to assess the prognostic power of each method, based on patient overall survival (OS) and progression-free survival (PFS). The bioinformatics methods performed well "as advertised" in that they were well correlated with the results of the functional assay, but they generally did not do well at predicting survival. One of the methods, based on a simple set of structural rules, yielded predictions that were significantly associated with survival. This study highlights challenges to clinical applications of bioinformatics classifiers of "deleterious mutations". A computationally classified deleterious mutation may be accurate if the end point is to predict protein inactivation, but if the end point is to contribute to predictions of patient prognosis, such classifications may not be the best choice. Methods specifically designed to learn the impact of mutation patterns in *TP53* and other clinically relevant genes on patient survival could provide more relevant predictions. WES and WGS data may contribute to development of such methods by enabling statistical analysis of survival and the combined impact of multiple mutations in critical pathways.

Finally, a major challenge to computational molecular medicine concerns the changing environment in which biological research is performed. As Woods et al. observe, biology is becoming more computationally intensive, but academic biology curricula often fail to provide their graduates with the computational and statistical training that they will need to analyze large datasets. While collaborating with a computational scientist may be an attractive option, a better option is for the biologist to develop some expertise to become an active partner in the computational analyses. They share the benefit of their ten years of experience in transitioning their work to include more computational methodologies, and outline sensible strategies on topics including avoiding batch effects, selecting computational analysis packages and managing data provenance for better reproducibility.

In summary, the advent of modern "omics" measurement technologies has led to computational innovations, which include expanded resources that would have been

unimaginable only a few years ago and new modeling strategies. However, these advances are only beginning to impact clinical practice. We encourage the readers of this special issue to join the efforts to bridge the gap between computational progress and translation.

References

1. Stergiopoulos K, Brown DL. Genotype-guided vs clinical dosing of warfarin and its analogues: meta-analysis of randomized clinical trials. JAMA Intern Med. 2014; 174(8):1330–1338. [PubMed: 24935087]
2. Munoz JL, et al. Temozolomide resistance in glioblastoma cells occurs partly through epidermal growth factor receptor-mediated induction of connexin 43. Cell Death Dis. 2014; 5:e1145. [PubMed: 24675463]
3. Robles AI, Harris CC. Clinical outcomes and correlates of TP53 mutations and cancer. Cold Spring Harb Perspect Biol. 2010; 2(3):a001016. [PubMed: 20300207]