ELSEVIER

Next-generation sequencing in drug development: target identification and genetically stratified clinical trials

Abolfazl Doostparast Torshizi[1,‡] **and Kai Wang**[1,2,3,‡]

[1] Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
[2] Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
[3] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

Next-generation sequencing (NGS) enabled high-throughput analysis of genotype–phenotype relationships on human populations, ushering in a new era of genetics-informed drug development. The year 2017 was remarkable, with the first FDA-approved gene therapy for cancer (Kymriah™) and for inherited diseases (LUXTURNA™), the first multiplex NGS panel for companion diagnostics (MSK-IMPACT™) and the first drug targeting a genetic signature rather than a disease (Keytruda®). We envision that population-scale NGS with paired electronic health records (EHRs) will become a routine measure in the drug development process for the identification of novel drug targets, and that genetically stratified clinical trials could be widely adopted to improve power in precision-medicine-guided drug development.

Introduction

A large number of human diseases are known to be caused or influenced by genetic factors, which can provide potential insights into disease pathogenesis, leading to the development of novel treatment strategies [1]. For over four decades, genetic-linkage-based studies have been successfully applied to the identification of causal genetic factors in Mendelian disorders [2]. Despite all of these breakthroughs, many common diseases do not exhibit Mendelian inheritance but represent complex multifactorial inheritance patterns, making genetic-linkage-based studies less successful at capturing allelic determinants of such disorders [3]. To circumvent such limitations, genome-wide association studies (GWAS) were then widely employed to characterize susceptibility loci associated with complex phenotypes. Ever since 2005 when the first GWAS was published on patients with age-related macular degeneration [4], GWAS have been employed in a variety of human diseases and traits [5,6]. Thus far, almost 10,000 strong associations [e.g., statistically significant associations with *P*-value threshold of 5×10^{-8} excluding single-nucleotide polymorphisms (SNPs) in linkage disequilibrium] have been reported between genetic variants and one or more complex traits [6]. Among these findings, there are several examples of disease-associated genes that have been identified as being effective drug targets, such as *HMGCR* being the target of statins which is associated with serum cholesterol levels [7]. GWAS can also enable the discovery of biological pathways that confer susceptibility to diseases [8,9]. Some early examples include the confirmation of interleukin (IL)-12/IL-23 pathways in inflammatory bowel diseases [10], and the discovery of the autophagy pathway in inflammatory bowel diseases through the *ATG16L1*/*IRGM*associations [11,12]; indeed, it has now become clear that the modulation of autophagy has strong therapeutic implications for drug development [13]. Pleiotropic SNPs for multiple related diseases have also been reported in GWAS [5,6], sometimes with opposing effects [14]. For example, SNPs discovered at the *IL23R* locus were associated with several autoimmune conditions such as ankylosing spondylitis, inflammatory bowel disease and psoriasis [15–17]. Such pleiotropic genes discovered by GWAS can be useful for drug repurposing or enable basket trials that recruit patients with diverse diseases. Monoclonal

Corresponding author: Wang, K. (wangk@email.chop.edu)
‡ Prior address: Department of Biomedical Informatics and Institute for Genomic Medicine, Columbia University Medical Center, New York, NY 10032, USA.

https://doi.org/10.1016/j.drudis.2018.05.015

antibodies targeting IL-23 and/or IL-12 are already in clinical trials for the treatment of various autoimmune diseases [18–20]. A recent study demonstrated that, among well-studied indications, the proportion of drug mechanisms being supported by direct genetic evidence across the drug development pipeline can increase from 2% at the preclinical stage to 8.2% for approved drugs [21]. Therefore, human genetic studies on well-phenotyped populations can guide the selection of the best targets and indications, with a measurable impact on the successful development of new drugs.

However, GWAS have several intrinsic limitations, such as the inability to account for a large proportion of genetic variance in complex traits and the difficulty to identify casual genes or mutations from proxy markers [6,22]. These limitations can be partially attributed to the fact that most GWAS only target common SNPs, yet common and complex disorders such as schizophrenia are genetically heterogeneous with only a proportion of genetic risks caused by SNPs of small-to-moderate effect sizes [23]. Therefore, even with carefully designed exome arrays, GWAS are also unable to or inefficient at detecting loci harboring very rare variations with larger effect sizes. In fact, despite considerable and unquestionable successes of GWAS, it has become clear that many disease-associated genetic loci with therapeutic implications have not been discovered by GWAS [24]. An illustrative example is the association of hypertension drug target genes with blood pressure [25]. Despite the discovery of over 30 genes before the GWAS era, only a small fraction has been re-identified by GWAS [26]. The need for extremely large cohort sizes to capture disease-associated variants using GWAS is also a major drawback, leading to a diminishing return after a study of the first several thousand subjects has identified the strongest determinants among common variants [27]. For instance, for a long time, no large-scale GWAS were able to identify *DRD2*, a traditional target of most antipsychotic drugs in the market, to be associated with schizophrenia until the largest GWAS was kicked off involving 37,000 schizophrenia individuals and 113,000 controls [28]. The ultimate challenge of effective use of GWAS in drug development is well probed by an omnigenic model [29], where regulatory networks are sufficiently interconnected such that all genes expressed in disease-relevant cells are liable to affect the functions of core disease-related genes. Finding the set of 'core' genes with biologically interpretable roles in disease thus becomes especially important to guide drug development efforts.

The advent of next-generation sequencing (NGS) technologies and their applications on human populations have led to the identification of a large catalog of rare and common variants [30–32]. The use of NGS on large isolated population cohorts [33] and on specific disease cohorts [34–37] has led to the discovery of a number of new disease genes. Although NGS cannot completely address the limitations of GWAS, its avoidance of proxy markers allows direct biological interpretation of rare genetic mutations in the context of disease phenotypes. In particular, NGS can be leveraged to identify genetic variations that can inactivate drug target genes. Such mutations can mimic the action of therapeutic antagonism of these targets, thus providing a means to infer possible clinical effects of antagonist drugs on such gene products.

The pharmaceutical industry is no exception in opting for NGS technologies in diverse stages of drug development, because these technologies have brought about unique opportunities to ameliorate the discovery of novel therapeutic targets [38,39], facilitate the precise design of clinical trials on target populations more likely to benefit from the treatment [40,41], and characterize novel indications of existing drugs that are approved or under development [42–44]. While acknowledging the combined use of various types of NGS-based genomic techniques (including transcriptome and methylation profiling) in drug development, in this review we will only discuss the use of DNA sequencing in the pharmaceutical industry for target discovery and for genetically stratified clinical trials. Additionally, given the existence of several reviews on the oncology space [45,46], this article will place more emphasis on inherited diseases.

Target identification

It has been clear that the wealth of human genetic information can be leveraged to identify drug targets, validate therapeutic hypotheses and predict the potential safety of inhibitory compounds aimed at molecular targets [47,48]. NGS has the potential to uncover many mutations associated with genetic diseases and identify target genes for future drug development endeavors. Compared with GWAS which rely on proxy markers for unknown causal variants or genes, NGS is gaining momentum as a means of choice for drug target identification efforts. For such purposes, a particular trend in the field is to sequence well-phenotyped populations coupled with longitudinal electronic health records (EHRs) as a test bed to identify genes associated with a variety of phenotypic traits (Fig. 1a). A pioneering example is the DiscovEHR study as a collaboration between Regeneron and Geisinger Health System [49], in which whole-exome sequencing was performed on 50,726 subjects with paired EHRs. By leveraging rich phenotype information such as lipid levels extracted from EHRs, such studies can examine associations between loss-of-function (LoF) variants in candidate drug targets and selected phenotypes of interest. Some of the associations between predicted LoFs (pLoFs) in drug targets of low-density lipoprotein cholesterol (LDL-C) levels were confirmed. For example, pLoF mutations in *NPC1L1* (the drug target of ezetimibe [50]) and *PCSK9* (the drug target of alirocumab and evolocumab) were confirmed to be significantly associated with LDL-C levels [49]. Additionally, such LoF variants linked to EHR data can uncover novel associations, such as the association of LoF variants in *CSF2RB* with basophil and eosinophil counts [49]. In another study to investigate whether and how whole-genome sequencing (WGS) data can be used to implement genomic medicine, we examined WGS data and lifetime EHRs from 300 deceased patients [51]. Among them, five carried pathogenic or likely pathogenic variants in cancer-predisposing genes (*APC*, *BRCA1*, *BRCA2*, *NF1* and *TP53*). Based on EHRs, each of the five patients had one or more different types of cancers, fully consistent with their genetic profiles. Therefore, this is a clear demonstration that the discovery of pathogenic or likely pathogenic germline mutations from population-wide WGS correlates with clinical outcome on EHRs, thus the use of WGS could have clinical impacts to improve healthcare delivery.

While recognizing the importance of population-scale genome sequencing to find associations with common diseases or traits, NGS analysis on extreme phenotypes of therapeutic implications is also a growing opportunity to discover underlying drivers of

Please cite this article in press as: