DRUG DEVELOPMENT

The druggable genome and support for target identification and validation in drug development

Chris Finan,[1,2]* Anna Gaulton,[3]* Felix A. Kruger,[1,4] R. Thomas Lumbers,[1,2] Tina Shah,[1,2] Jorgen Engmann,[1,2] Luana Galver,[5] Ryan Kelley,[5] Anneli Karlsson,[3] Rita Santos,[3†] John P. Overington,[3,4‡] Aroon D. Hingorani,[1,2*‡] Juan P. Casas[2*‡]

Target identification (determining the correct drug targets for a disease) and target validation (demonstrating an effect of target perturbation on disease biomarkers and disease end points) are important steps in drug development. Clinically relevant associations of variants in genes encoding drug targets model the effect of modifying the same targets pharmacologically. To delineate drug development (including repurposing) opportunities arising from this paradigm, we connected complex disease- and biomarker-associated loci from genome-wide association studies to an updated set of genes encoding druggable human proteins, to agents with bioactivity against these targets, and, where there were licensed drugs, to clinical indications. We used this set of genes to inform the design of a new genotyping array, which will enable association studies of druggable genes for drug target selection and validation in human disease.

INTRODUCTION

Only 4% of drug development programs yield licensed drugs (*1*, *2*), largely because of two unresolved systemic flaws: (i) Preclinical experiments in cells, tissues, and animal models and early-phase clinical testing to support drug target identification and validation are poorly predictive of eventual therapeutic efficacy and (ii) definitive evidence of the validity of a new drug target for a disease is not obtained until late-phase development [in phase 2 or 3 randomized controlled trials (RCTs)]. Reasons for poor reliability of preclinical studies include suboptimal experimental design with infrequent use of randomization and blinding (*3*), species differences, inaccuracy of animal models of human disease (*4*, *5*), and overinterpretation of nominally significant experimental results (*6*–*8*). Human observational studies can mislead for reasons of confounding and reverse causation. Evidence of target validity from phase 1 clinical studies can also be inadequate (because phase 1 studies primarily investigate pharmacokinetics and tolerability, are typically small in size, are of short duration and measure a narrow range of surrogate outcomes, and are often of uncertain relevance to perturbation of the target of interest) (*9*). Because the target hypothesis advanced by preclinical and early-phase clinical studies is all too frequently false, expensive late-stage failure in RCTs from lack of efficacy is a common problem affecting many therapeutic areas (*10*), posing a threat to the economic sustainability of the current model of drug development.

Genetic studies in human populations can imitate the design of an RCT without requiring a drug intervention (*11*–*13*). This is because genotype is determined by a random allocation at conception according to Mendel's second law (Mendelian randomization) (*12*, *14*). Single-nucleotide polymorphisms (SNPs) acting in cis (variants in or near a gene that associate with the activity or expression of the encoded protein) can therefore be used as a tool to deduce the effect of pharmacological action on the same protein in an RCT. Numerous proof-of-concept examples have now been reported (*11*, *13*, *15*–*19*), including the marked correlation between 80 circulating metabolites' association with a SNP in the *HMGCR* gene that encodes the target for statin drugs and the effect of statin treatment on the same set of metabolites (*20*). SNPs acting in cis are a general feature of the human genome (*21*), and population and patient data sets with stored DNA and genotypes linked to biological phenotypes and disease outcome measures are now widely available for this type of study.

By extension, disease-associated SNPs identified by genome-wide association studies (GWAS) could be explicitly interpreted as an underused source of randomized human evidence to aid drug target identification and validation. For illustration, loci for type 2 diabetes identified by GWAS include genes encoding targets for the glitazone and sulphonylurea drug classes already used to treat diabetes (*22*, *23*). Apparently, sporadic observations such as this suggest that numerous, currently unexploited disease-specific drug targets should exist among the thousands of other loci identified by GWAS and similar high-quality genetic association studies. Recent studies of advanced or completed drug development programs (mostly based on established approaches to target identification) have also indicated that those with incidental genomic support had a higher rate of developmental success (*24*–*27*).

Fulfilling the potential of GWAS (and studies using disease-focused genotyping arrays) for drug development requires mapping disease- or biomarker-associated SNPs to genes encoding druggable proteins and to their cognate drugs and drug-like compounds. The set of proteins with potential to be modulated by a drug-like small molecule has been predicted on the basis of sequence and structural similarity to the targets of existing drugs, the set of encoding genes being referred to as the druggable genome. Hopkins and Groom (*28*) identified 130 protein families and domains found in targets of drug-like small molecules known at the time and more than 3000 potentially druggable proteins containing these domains. A similar estimate was made by Russ and Lampel (*29*), using a later human genome build. Kumar *et al.* (*30*) used these protein families (plus other families of particular relevance to cancer) to manually curate lists of druggable proteins for inclusion in the dGene data set. More recently, the Drug Gene Interaction database (DGIdb) has been developed (*31*), which integrates data from each of the previous efforts together with a recently compiled list of drug candidates and targets in clinical development (*32*) as well as information from the

[1]Institute of Cardiovascular Science, Faculty of Population Health, University College London, London WC1E 6BT, U.K. [2]Farr Institute of Health Informatics, University College London, London WC1E 6BT, U.K. [3]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, U.K. [4]BenevolentAI, 40 Churchway, London, U.K. [5]Illumina Inc., 5200 Illumina Way, San Diego, CA 92122, USA.
*These authors contributed equally to this work.
†Present address: GlaxoSmithKline Medicines Research Centre, Gunnels Wood Road, Stevenage, Herts SG1 2NY, U.K.
‡Corresponding author. Email: jpo@md.catapult.org.uk (J.P.O.); a.hingorani@ucl.ac.uk (A.D.H.); jp.casas@ucl.ac.uk (J.P.C.)

PharmGKB (*33*), Therapeutic Target Database (*34*), DrugBank (*35*) databases, and others.

However, earlier estimates of the druggable genome predated contemporary genome builds and gene annotations and also did not explicitly include the targets of biotherapeutics, which formed more than a quarter of the 45 new drugs approved by the U.S. Food and Drug Administration's (FDA's) Center for Drug Evaluation and Research in 2015 (*36*), reflecting their increasing importance in pharmaceutical development. We therefore updated the set of genes comprising the druggable genome. We then linked GWAS findings curated by the National Human Genome Research Institute and European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI) GWAS catalog (*37*) to this updated gene set, as well as to encoded proteins and associated drugs or drug-like compounds curated in the ChEMBL (*38*) and First Databank (FDB) (*39*) databases. We used the connection to explore the potential for genetic associations with complex diseases and traits for informing drug target identification and validation, as well as to repurpose drugs from one indication for another. In addition, to better support future genetic studies for disease-specific drug target identification and validation, we assembled the marker content of a new genotyping array designed for high-density coverage of the druggable genome and compared this focused array with genotyping arrays previously used in GWAS.

RESULTS

Redefining the druggable genome

We estimated 4479 (22%) of the 20,300 protein-coding genes annotated in Ensembl version 73 to be drugged or druggable. This adds 2282 genes to previous estimates made by Hopkins and Groom, Russ and Lampel, or Kumar *et al.*, by inclusion of targets of first-in-class drugs licensed since 2005; the targets of drugs currently in late-phase clinical development; information on the growing number of preclinical phase small molecules with protein binding measurements reported in the ChEMBL database; as well as genes encoding secreted or plasma membrane proteins that form potential targets of monoclonal antibodies and other biotherapeutics. A set of 432 genes that was included in all other proposed druggable gene sets but not the DrugDev set consists mainly of olfactory receptors and phosphatases; both protein families have major limitations for future exploitation as drug targets (Fig. 1) (*40*, *41*). We stratified the druggable gene set into three tiers corresponding to position in the drug development pipeline. Tier 1 (1427 genes) included efficacy targets of approved small molecules and biotherapeutic drugs as well as clinical-phase drug candidates. Tier 2 was composed of 682 genes encoding targets with known bioactive drug-like small-molecule binding partners as well as those with ≥50% identity (over ≥75% of the sequence) with approved drug targets. Tier 3 contained 2370 genes encoding secreted or extracellular proteins, proteins with more distant similarity to approved drug targets, and members of key druggable gene families not already included in tier 1 or 2 [G protein (heterotrimeric guanine nucleotide–binding

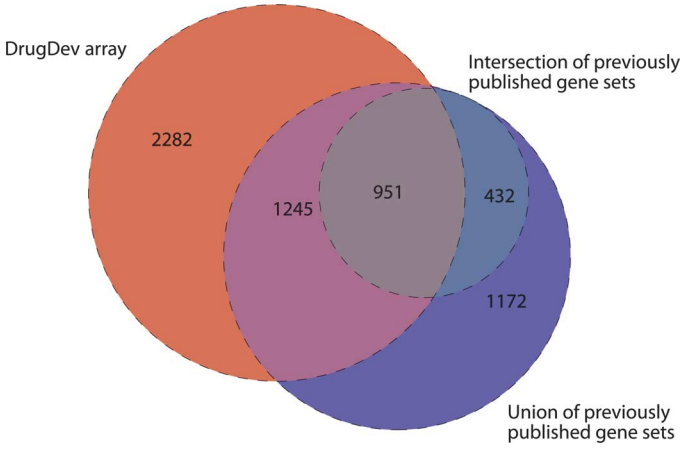

Fig. 1. Overlap between targets on the DrugDev array and three previously published sets. The Venn diagram shows the overlap of targets on the DrugDev array with the union (circle composed of blue, purple, gray, and turquoise segments), as well as the intersection (circle composed of gray and turquoise segments) of the druggable gene sets defined by Hopkins and Groom (*28*), Russ and Lampel (*29*), and Kumar *et al.* (*30*).

Table 1. Count of GWAS published per disease area.

MeSH term	Count
Neoplasms	187
Immune system diseases	130
Skin and connective tissue diseases	107
Digestive system diseases	106
Nervous system diseases	104
Mental disorders	85
Cardiovascular diseases	84
Nutritional and metabolic diseases	83
Endocrine diseases	77
Musculoskeletal diseases	57
Male urogenital disorders	52
Eye diseases	50
Respiratory diseases	47
Hematological diseases	43
Female urogenital diseases and pregnancy complications	41
Pathological signs and symptoms	34
Congenital disorders	29
Viral diseases	19
Oral diseases	17
Substance-related disorders	11
Diseases of the ear, nose, or throat	8
Parasitic diseases	4
Bacterial and fungal infections	2
Behavioral disorders	1
Wounds and injuries	1
Psychological phenomena and processes	1
Occupational diseases	1