Published in final edited form as:
Sci Transl Med. 2017 March 29; 9(383): . doi:10.1126/scitranslmed.aag1166.

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

Chris Finan[#1,3], **Anna Gaulton**[#2], **Felix. A Kruger**[1,3], **R. Thomas Lumbers**[1,3], **Tina Shah**[1,3], **Jorgen Engmann**[1,3], **Luana Galver**[5], **Ryan Kelley**[5], **Anneli Karlsson**[2], **Rita Santos**[2,6], **John P. Overington**[2,4,*], **Aroon D. Hingorani**[#1,3,*], and **Juan P. Casas**[#3,*]

[1]Institute of Cardiovascular Science, Faculty of Population Health, University College London, London WC1E 6BT

[2]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

[3]Farr Institute in London, Institute of Health Informatics, University College London, London WC1E 6BT

[4]BenevolentAI, 40 Churchway, London, UK

[5]Illumina Inc, 5200 Illumina Way, San Diego, CA 92122 USA

[#] These authors contributed equally to this work.

Abstract

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 (GWAS) 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.

[*]Corresponding Author - John Overington: jpo@benevolent.ai, Aroon Hingorani: a.hingorani@ucl.ac.uk, Juan Casas: jp.casas@ucl.ac.uk..

[6]GlaxoSmithKline, Medicines Research Centre, Gunnels Wood Road, Stevenage, Herts, SG1 2NY, United Kingdom. (current address)

Author contributions: C.F., A.G., F.K., J.O., A.H., and J.P.C. developed the idea for the project and approaches to accurately connect genetic associations to drug targets and compounds. A.G., F.K.and J.O. updated estimates of the druggable genome. L.G.and R.K.worked withA.G., C.F., T.S., and J.E. to develop SNP content for the Illumina DrugDev array. A.K. curated target information for clinical stage drugs, R.S. curated target information for FDA approved drugs in the ChEMBL database, and T.L. and A.H. compared indications and adverse effects of licensed drugs with disease associations from GWAS.

Competing interests: The authors declare that they have no competing interests.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Introduction

Only 4% of drug development programs yield licensed drugs (1, 2), largely because of two unresolved systemic flaws: (1) 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 (2) 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 over-interpretation 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, of short duration and measure a narrow range of surrogate outcomes, 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 - MR) (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 (15, 16, 11, 17, 13, 18, 19), including the striking 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 datasets 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 GWAS could be explicitly interpreted as an under-used source of randomised 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 identified 130 protein families and domains found in targets of drug-like small molecules known at the time, and over 3000 potentially druggable proteins containing these domains (28). A similar estimate was made by Russ and Lampel, using a later human genome build (29). Kumar *et al*. 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 (30). 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 PharmGKB (33), Therapeutic Target Database (TTD) (34), and 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 bio-therapeutics, which formed more than a quarter of the 45 new drugs approved by the 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 (NHGRI) and European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI) GWAS catalog (37) to this updated gene set, and also to encoded proteins and associated drugs or drug-like compounds curated in the ChEMBL (38) and First Databank (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. Additionally, 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

Re-defining the druggable genome

We estimated 4,479 (22%) of the 20,300 protein coding genes annotated in Ensembl v.73 to be drugged or druggable. This adds 2,282 genes to previous estimates made by Hopkins and Groom, Russ and Lampel, or Kumar, 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 pre-clinical 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 bio-therapeutics. 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 (40, 41) (Fig. 1). We stratified the druggable gene set into 3 tiers corresponding to position in the drug-development pipeline. Tier 1 (1,427 genes) included efficacy targets of approved small molecules and biotherapeutic drugs as well as clinical-phase drug candidates. Tier 2 was comprised 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 2,370 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 Tiers 1 or 2 (GPCRs, nuclear hormone receptors, ion channels, kinases, and phosphodiesterases). A full list of genes is provided in table S1. An overview of the 15 most frequently occurring protein domain types for each tier can be found in table S2, based on the Pfam-A database of protein families.

Connecting loci identified by GWAS to the druggable genome

We retrieved 21,406 associations from 2,155 GWAS, of which 9,178 surpassed the significance threshold of $p \geq 5 \times 10^{-8}$. The retrieved associations spanned 315 Medical Subject Heading (MeSH) disease terms, which can be stratified into twenty-four MeSH root disease areas and three MeSH Psychiatry and Psychology areas (Table 1). Variants associated with common diseases and biomarkers had median minor allele frequency 0.29 (interquartile range, IQR 0.21) based on a subset of 7,387 records with risk allele frequency data), reflecting the preponderance of common variants on widely used genotyping arrays. The median odds ratio (OR) for studies of disease end-points was 1.24 (IQR 0.31) (based on the 3,367 results with effect size data). We examined sequence ontology consequence types (42) of disease and biomarker-associated variants and found most to be non-coding, mainly intronic, presumably altering or marking variants that alter mRNA expression or availability, or marking variants that alter structure or activity of encoded proteins (fig. S1).

Of the 9,178 GWAS significant associations ($p \leq 5 \times 10^{-8}$), 8,879 mapped to 5,084 unique intervals defined as containing all SNPs in linkage disequilibrium (LD) (with an $r^2 \geq 0.5$) with the SNP exhibiting the most significant association, applying an upper physical bound of 1 Mbp on either side of this variant. The remaining 299 associations were either not in LD with any other variants, or not present in the 1000 genomes reference panel (phase 3 version). Such associations were assigned a nominal interval of 2.5 kbp on either side of the variant. The frequency distribution of genes and druggable genes in such LD intervals were right skewed (Fig. 2), and there was a correlation between LD interval size and the number of resident genes (fig. S2).

Of the 5,084 unique LD intervals, 1,533 (30.2%) contained a single gene. Of these, 532 contained a gene from the druggable set: 233 from Tier 1, 76 from Tier 2, and 223 from Tier 3. Of the remaining genomic intervals, 17.3% (880) mapped to intervals containing two genes, 10.1% (511) contained three genes, 6.7% (343) contained four genes, and 25.2% (1281) contained five or more genes. Additionally, 536 (10.5%) regions had no gene in the LD interval. For the 1624 LD intervals containing two or more genes, of which at least one was druggable, the median distance of the closest druggable gene to the reported GWAS

variant was 4.98 kbp (IQR 37.7 kbp), where the distance was set to 0 bp for GWAS variants lying within a gene, and a druggable gene was among the two most proximal genes in 67.1 % of these LD intervals (1089) (Fig. 3). We identified a total of 3,052 genes in the druggable set that were not represented in any of the LD intervals corresponding to a GWAS association; 62.7%, 69.2%, and 71.6% of Tier 1, 2, and 3 genes, respectively.

Linking GWAS associations to licensed drug targets

We found that 1,291 GWAS associations defined 1,072 LD intervals containing 532 druggable genes from Tier 1, which includes the targets of licensed drugs. 479 of the intervals contained a single drug target, and 593 contained two or more targets. For the set of LD intervals containing genes encoding the targets of licensed drugs, two clinically qualified curators blinded to the identity of the genes independently evaluated the correspondence between the disease association from the GWAS and the treatment indication(s) for drug(s) acting on the target(s) encoded by a druggable gene in the interval (Table 2). Our curators identified 56 unique associations (30 unique drug targets) where the treatment indication and genetic association were precisely concordant and 13 associations (9 targets) where the indication and association came from the same disease area (for example a GWAS in one form of epilepsy identifying a drug target for a different form of epilepsy). 97 associations (mapping to 37 licensed drug targets) corresponded to biomarkers known to be altered by treatment with the corresponding drug (for example, an LD interval containing the gene encoding the interleukin-6 receptor was identified in a GWAS of C-reactive protein, a biomarker altered by the action of the interleukin-6 receptor blocker, tocilizumab (43). A further 76 associations (27 licensed drug targets) were identified through a genetic association with a mechanism-based adverse effect, such as in a GWAS of heart rate, where the SNP rs3143709 defined an LD interval containing the gene *ACHE* (acetylcholinesterase), encoding the target of cholinesterase inhibitors used in the treatment of myasthenia gravis, which have the side effect of lowering heart rate (44). A further 32 genetic associations (corresponding to 8 targets) were with a quantitative trait that could be either a marker of therapeutic efficacy or a mechanism-based side effect, as in the case of QT interval in the context of anti-arrhythmic drug therapy. In all, GWAS ‘rediscovered’ 74 licensed drug targets through disease indications, mechanism of action, or mechanism-based adverse effects (the numbers for the categories above are non-additive because some targets overlap categories). Illustrative examples of the curation are shown in table S3.

Manual curation identified 1,523 discordant pairings of drug indications and disease associations, corresponding to 144 drug targets that were interpreted as plausible repurposing opportunities (Fig. 4). After manual curation, uncertainty remained for 108 associations (52 targets) as to whether discordance represented a repurposing opportunity, or an unrecognized mechanism-based side effect. The remaining targets of licensed drugs mapped to LD intervals corresponding to GWAS traits unlikely to be of therapeutic interest (for example, hair color) or to a genetic association with a new biomarker of uncertain biological function (such as a metabolite measured by a new metabolomics platform). Curators disagreed on the coding for GWAS associations corresponding to 4 licensed targets. For LD intervals corresponding to GWAS rediscoveries, the interval length was

Europe PMC Funders Author Manuscripts

smaller, contained fewer genes, and the druggable gene was closer to the lead SNP than for those LD intervals where the indication and genetic association were discordant (table S4).

Translational opportunities unveiled by the data linkage

Fig. 5 and fig. S3 and S4 illustrate the result of mapping disease associations in the GWAS catalog to the full set of druggable genes, the encoded proteins, and compounds exhibiting binding affinity to these targets, regardless of development phase. For example, 84 studies in the GWAS catalog reported findings pertaining to cardiovascular system diseases (39 disease subcategories), reporting 388 GWAS associations, mapping to 228 unique LD intervals containing 670 genes, of which 135 were in the druggable set. Of these, 29 genes were either the solitary occupant or one of only a pair of genes in the LD interval. We linked all 135 druggable genes identified in the cardiovascular category to 19,844 compounds with measured activities in ChEMBL, 512 of which had a United States Adopted Name (USAN) International Non-Proprietary Name (INN) or were in late phase development, and 168 of which were previously licensed drugs. Based on comparisons between GWAS phenotype terms and treatment indications in the cardiovascular category, 8 drug target indications and genetic associations were concordant (target ‘rediscovery’) and 19 were discordant. Fig. 6 illustrates the results of a similar mapping exercise for seven specific diseases (type 2 diabetes, hypertension, inflammatory bowel disease, asthma, coronary heart disease, schizophrenia, and Alzheimer’s disease).

The proportion of druggable genes in LD intervals defined by GWAS SNPs for digestive system diseases (0.20, 95% CI: 0.12-0.27), neoplasms (0.15, 95%CI: 0.10-0.20), nervous system diseases (0.17, 95%CI: 0.10-0.24), cardiovascular diseases (0.20, 95%CI: 0.12-0.29), respiratory diseases (0.19, 95%CI: 0.08-0.31), skin and connective tissue diseases (0.17, 95%CI: 0.10-0.24), immune system diseases (0.19, 95%CI: 0.12-0.26), and mental health (0.16, 95%CI: 0.08-0.24) was similar to the proportion of druggable genes in the genome overall (4479/20,300 = 0.22).

Coverage of the druggable genome by Illumina DrugDev and other widely used genotyping arrays

Capture of variation in druggable genes by the widely used genotyping arrays is illustrated in Fig. 7, with reference to the 1000 genome European super population ancestry panels (45). Disease-focused genotyping arrays and whole genome arrays with fewer than 600,000 SNPs used for many of the discoveries curated in the GWAS catalogue provided less comprehensive capture of variation in the druggable genome than the more recently developed arrays with several million SNPs (such as the Illumina Human Omni 2.5 Exome 8 and Illumina Omni 5). However, because no array to date has been designed specifically to ensure capture of variation in genes encoding druggable targets, we designed the content for an array (the Illumina DrugDev array) using the Illumina Infinium platform, which combines genome-wide tag SNP content of the Illumina Human Core array with 182,375 bespoke markers in 4479 druggable genes. The median number of variants captured per kbp of the druggable genome was very similar to that of the Illumina Human Omni 2.5 Exome 8 and Illumina Omni 5 (Fig. 7 and fig. S5 and S6) with an average of around 2.5 SNPs per kbp

of the druggable genome, at an average of nearly 50 variants per gene array wide, with even denser coverage of Tier 1 and 2 genes.

With the exception of Illumina Omni products, all available genotyping arrays captured druggable genome variation most efficiently among populations of European descent and most poorly among populations of African descent (Fig. 7 and fig. S5 and S6). Outside of the European populations, the high density Illumina Omni arrays gave superior coverage (for both directly genotyped variants and tagged variants) to all other genotyping arrays. The Affymetrix UK Biobank array displayed similar coverage to the Illumina DrugDev array in European populations but less complete coverage in non-European populations. A heat map summarizing the coverage for each druggable gene, stratified by tier and 1000 genomes population groups, is shown in Fig. 8. Results for directly typed and tagged variants in 1000 genomes sub-populations are shown in fig. S7 and fig. S8, respectively.

Discussion

By first re-estimating the boundaries of the druggable genome and then mapping biomarker and disease-associated loci from GWAS to genes encoding druggable targets, we have demonstrated the extent to which GWAS have already rediscovered target-disease indications or mechanism-based adverse effects of licensed drugs. These findings indicate the potential of genetic association studies to systematically and accurately identify disease-specific drug targets across the spectrum of human diseases, addressing one of the key productivity-limiting steps in drug development.

For example, we found substantial potential for repurposing of drugs with licensed indications from one disease area to another (Fig. 4), in keeping with previous analyses from the GWAS catalog that indicated that 17% of genes exhibit associations with more than one phenotype (46). We also mapped genetic associations to drug target and compound annotations in ChEMBL to evaluate the potential for progressing or repositioning compounds at earlier developmental stages (Fig. 5)

Estimating the expected number of licensed drug target rediscoveries by GWAS is not straightforward. It involves an estimate of the extent to which GWAS have already been done for diseases and biomarkers that have at least one licensed drug target available for rediscovery; enumerating the total number of licensed drug targets represented across these conditions, since some diseases have multiple licensed drug targets; and estimating the number of GWAS that have been completed for diseases and biomarkers that reflect the mechanism-based adverse effects of licensed drugs. It also requires an assumption about the average power of eligible GWAS to detect a true association at a gene encoding a licensed drug target in a relevant disease. This effort is hindered by inconsistent vocabularies of disease terms in GWAS and drug indications in licensing documents and product information leaflets. Separating the important mechanism-based (often rare) and idiosyncratic adverse effects listed in product information and other relevant sources is also challenging. Nevertheless, the rediscovery of 70 of the 600 or so known licensed targets (32, 47) by GWAS, suggests that this approach shows promise as a means to more systematically identify target-disease indication pairings in the future.

Despite the many therapeutic opportunities already arising from the mapping of existing genetic association findings to drug targets and compounds, there are strong reasons to suspect that the potential of this approach has yet to be maximized. Our analysis identified target-disease indication pairings (defined as a gene encoding a druggable target mapping to an LD interval containing a lead SNP from a GWAS) for 1,427 of the 4,479 druggable genes and 240 of the 652 genes encoding targets of licensed drugs. We might not have discovered associations for all genes in our druggable set because targets of drugs in development may truly play no role in any disease. However, alternative explanations are that only a fraction of diseases have been subjected to GWAS [451 out of 3022 conditions (the denominator is based on the number of bottom level MeSH disease areas)]; that for many of the diseases that have been investigated by GWAS the sample sizes have been too small to detect all the responsible genes; or that there may have been incomplete coverage of certain druggable genes by the arrays most widely deployed in GWAS.

Genome wide association analyses continue to be published in new disease areas and in new ethnic groups. Additional genetic discoveries are also being made with other types of arrays, such as the dense, locus-centric SNP arrays following up on GWAS findings that are currently not systematically captured by the GWAS catalog, including Cardiochip (48), CardioMetabochip (49), and Immunochip (50), and by increases in sample size. Exome-array analyses are also unveiling rare, disease-associated variants under-represented in whole-genome arrays. Therefore, we anticipate that the current gap between druggable genes and GWAS findings will be reduced over time, particularly if such studies are extended to electronic health record datasets, which form rich repositories of phenotypic traits and diagnostic codes. In the future, as cost falls and the pipelines for interpreting individual sequence variation are streamlined, whole genome sequencing may replace genotyping arrays as the major of source of information on genetic variation used for drug target identification and validation.

Genetic profiling of a promising target against a range of outcomes can help evaluate the efficacy and safety of a target for the primary indication as well as the identification of additional disease indications to help plan drug development priorities. To stimulate the wider use of genetic association studies in drug development and to ensure that such studies have comprehensive coverage of the druggable genome, we designed the content of a new array that combines focused coverage of the druggable genome with a whole genome scaffold. This array could be deployed to boost sample size and power in diseases already studied by GWAS to identify additional susceptibility loci and druggable targets. The Illumina list price for the array DrugDev ($56/sample) is lower than that of the Omni 2.5 Exome array ($177/sample) and Omni 5 array ($273/sample), thus allowing a 3-5 fold increase in sample size under a fixed budget. It could also help stimulate new druggable GWAS prioritized according to unmet therapeutic need. This would automatically result in an abundance of target profiling information encompassing both efficacy and safety outcomes. This will need to be captured systematically and curated consistently to help develop a repository of human drug targets linked to the predicted consequences of their pharmacological modification.

Some limitations of our analysis are noteworthy. The identification of repurposing opportunities in the current dataset relied on detecting discordance between a gene-disease association and the corresponding target-disease indication for a licensed drug, and excluding instances where this was likely to be due to a mechanism-based adverse effect. However, the lack of standardized vocabulary in licensing agency approval documents and the scientific literature currently hampers this effort. We therefore used a combination of EFO and MeSH terms to harmonize nomenclature. Two qualified physicians then compared the annotations using a pre-specified classification system developed in a pilot study involving one fifth of the dataset. Greater efforts to harmonize terms from the different ontologies [EFO, MeSH terms, the Disease Ontology (DO), and the Human Phenotype Ontology (HPO)] (51–53), as well as from vocabularies for drug indications from the Anatomical Therapeutic Chemical (ATC) classification, electronic BNF, and eMC+ terms would help generate standardized terminology to improve the efficiency of similar efforts in the future.

In general, antagonist or inhibitor drugs are easier to develop than agonists or activators. However, it was not straightforward to establish a single workflow that would allow recommendation of agonist or antagonist development in the light of a GWAS finding. This is because alleles reported in GWAS sometimes associate with increased, and sometimes with reduced disease risk. Moreover, alleles reported for their association with a disease biomarker could have an opposite (yet unreported) association with disease outcome if the biomarker and disease risk are inversely correlated. We recommend that this issue should be considered on a case-by-case basis whenever a specific drug development program is predicated on a genetic association at a locus encoding a druggable target

Where several genes occupy the same LD interval as a GWAS SNP, it may be difficult to determine which is responsible for the disease or biomarker association. We took a pragmatic approach to this problem by classifying LD intervals containing druggable genes according to the total number of genes in the interval and the number and proximity of any druggable gene to the associated SNP. Approximately 529 unique LD intervals containing a variant with a significant association from a GWAS ($p \leq 5\times10^{-8}$) contained a single druggable gene. Such genes are strong positional candidates for the association. For the remainder, the LD interval included 2-146 genes (median 4 genes; excluding the 536 regions containing 0 genes,Fig. 3), but a druggable gene was first or next most proximal gene to the association signal in 36.1% of these cases. The rediscovery of 183 target-indication or mechanism-based adverse pairings for licensed drugs using this information supports the validity of this approach. Previous Mendelian randomization studies also provide reassurance that associations of SNPs in proximity to genes encoding druggable targets recapitulate the effects of drugs modifying the encoded proteins pharmacologically (13, 43, 18). Nevertheless, we recognize that some misclassification is possible, for example a causal signal arising from a gene encoding a non-druggable protein that occupies the same LD interval as a gene encoding a druggable target (confounding by linkage disequilibrium). Integrating information from feature annotation databases such as ENCODE (54), NIH Roadmap (55), and the Single Amino Acid Polymorphism Database (SAAP) (56) could help reduce misclassification. Localization of causal genes could also be aided by evidence on the effect of genetic variants on RNA transcription or on the activity or concentration of proteins

and metabolites, combining new proteomic and metabolomics technologies that are scalable to large population studies (57, 58) with statistical approaches to assess whether association signals from the same region are consistent with the same causal variant (59). It should be noted, also, that even where GWAS identify a gene outside the druggable set, the findings also have the potential to inform drug development indirectly, by highlighting pathways and processes involved in disease pathogenesis that may contain other druggable targets.

The Mendelian randomization paradigm that underpins this strategy validates targets (within a defined disease context) and not compounds, although comparing the profile of effects of a genetic variant with those of a drug or developmental compound can help distinguish on- from off-target effects (13, 18). For this reason, RCTs will not be superseded by the approach we describe, because any new molecule developed for a target of interest could have off-target actions that cannot be modelled genetically. Additionally, the effect of altering the expression or function of a target may only be seen beyond some lower threshold, so that a weak genetic effect may not adequately model the effect of modifying the target pharmacologically (26). Genetic evidence of a causal mechanism also does not guarantee its reversibility through pharmacological modification. For example, immune system-related genetic variants associate with the risk of developing type I diabetes, but useful therapies arising from this knowledge may be difficult to realize, because by the time the disease is diagnosed, immune-mediated damage to the pancreatic beta cells may be too advanced (26). Despite these theoretical limitations, evidence is emerging that Mendelian randomisation studies have wide-ranging potential to improve the efficiency of drug development and reduce the risk of expensive late-stage failure.

In summary, we have shown an approach to focus and catalyze the use of genomic information to support drug target validation, accurately match targets to disease indications, and identify rational repurposing opportunities for licensed drugs. The approach aligns well with proposals to ‘re-engineer’ translational science (60). It could help address the efficiency and innovation problem and serve as a basis for reinvigorating drug development.

Materials and Methods

Study design

Work in this paper extended the concept of Mendelian randomization studies for drug development from individual targets to the whole genome. The study (1) defined a set of genes that encode actual (or potential) drug targets and are likely to be responsible for genetic associations with complex diseases from earlier genome wide association studies (GWAS); (2) allowed us to design a genotyping array with enriched SNP coverage of these genes; and (3) linked the proteins encoded by this gene set to licensed drugs or to compounds with bioactivities against these targets. A variety of bioinformatics resources and other in silico tools were used to achieve these aims. The integrity of the analysis was evaluated through a comparison of the consistency between licensed drug indications and GWAS associations through manual curation and blinded clinical expert review. This analysis showed that GWAS have already ‘rediscovered’ around 70 or so of the approximately 600 targets of licensed drugs through associations with disease indications, disease-related biomarkers, or mechanism-based adverse effects. The dataset was then used

to draw inferences about the potential for drug repositioning and the more systematic application of genomics for drug target-disease indication mapping in the future.

Assembly of a druggable gene set

The reference set of genes used to redefine the druggable genome was comprised of gene annotations from Ensembl v.73 with a biotype of 'protein coding'. To this, we added T cell receptor and immunoglobulin genes, polymorphic pseudogenes, and a number of additional genes that were annotated in Ensembl v.73 as non-protein coding but which were nevertheless believed to encode important proteins (for example *SRD5A2*, *CYP4F8*). Data were extracted via Biomart (http://www.ensembl.org/biomartt). The content was assembled in three tiers:

Tier 1—This tier incorporated the targets of approved drugs and drugs in clinical development. Proteins that are targets of approved small molecule and biotherapeutic drugs were identified using manually curated efficacy target information from release 17 of the ChEMBL database (61). An efficacy target was defined as the target for the intended drug indication as opposed to any other potential targets for which the drug shows high affinity binding. Where binding site information was available in ChEMBL, a non-drug-binding subunit of a protein complex was assigned to Tier 3, whereas the drug-binding subunit was included in Tier 1. Drugs in clinical development were identified from a number of sources: investor pipeline information from a number of large pharmaceutical companies [including Pfizer, Roche, GlaxoSmithKline, Novartis (oncology only), AstraZeneca, Sanofi, Lilly, Merck, Bayer, and Johnson & Johnson – accessed June-August 2013] monoclonal antibody candidates and USAN applications from the ChEMBL database (release 17), and drugs in active clinical trials from clinicaltrials.gov (62). Targets for these drug candidates were assigned from company pipeline information and scientific literature, where available. Where no reported target information could be found, a potential target was assigned through analysis of bioactivity data in ChEMBL, with the target having the highest dose-response measurement ≤ 100 nM for the compound being assigned. All other human targets having an IC50/EC50/GI50/XC50/AC50/Kd/Ki/potency ≤100 nM for an approved drug or USAN compound were also included in Tier 1. Genes involved in ADME/drug disposition (phase I and II metabolic enzymes, transporters, and modifiers) were identified from the PharmaADME.org extended set (63).

Tier 2—This tier incorporated proteins closely related to drug targets or with associated drug-like compounds. Proteins closely related to targets of approved drugs were identified through a BLAST search (blastp) of Ensembl peptide sequences against the set of approved drug efficacy targets identified from ChEMBL previously (38). Any genes where one or more Ensembl peptide sequences shared ≥50% identity (over ≥75% of the sequence) with an approved drug target were included. Putative targets with drug-like (Lipinski rule-of-five compliant) compounds having an IC50/EC50/GI50/XC50/AC50/Kd/Ki/potency ≤1 μM were identified from ChEMBL and were also included in Tier 2.

Tier 3—This tier incorporated extracellular proteins and members of key drug-target families. Proteins distantly related to drug targets were identified through a BLAST search

against the set of approved drug targets (as above), with any proteins sharing ≥25% identity over ≥75% of the sequence and with E-value ≤0.001 being included in the set. Members of five major 'druggable' protein families (GPCRs, kinases, ion channels, nuclear hormone receptors, and phosphodiesterases) were extracted from KinaseSarfari (64), GPCRSarfari (65), and IUPHARdb (66) and included in the Tier 3. Extracellular proteins were identified using annotation in UniProt (67) and Gene Ontology (GO) (68). Because the potential size of the secreted/extracellular portion of the proteome (potential targets for monoclonal antibodies) is large, and the number of markers available for inclusion on the array was limited, this dataset was restricted to those proteins for which higher confidence annotations of extracellular localization were available (not solely prediction of a signal peptide). Proteins annotated in UniProt as having a 'secreted' subcellular location, those containing a signal peptide, or those annotated as 'Extracellular' (where these annotations were supported by the following evidence types: experimental, probable, by_similarity) were included in Tier 3. Proteins annotated in GO with Cellular Component terms: GO:0005576 : extracellular region, GO:0005615 : extracellular space, GO:0005578 : proteinaceous extracellular matrix, GO:0031233 : intrinsic to external side of plasma membrane, GO: 0031232 : extrinsic to external side of plasma membrane, GO:0071575 : integral to external side of plasma membrane, GO:0031362 : anchored to external side of plasma membrane, GO:0009897 : external side of plasma membrane, GO:0044214 : fully spanning plasma membrane, and supported by strong evidence (EXP, IDA, TAS), were also included in the tier. Finally, proteins known to be cluster of differentiation antigens (CD antigens) according to UniProt were also added to Tier 3. Because the final set of genes included in Tier 3 was large (2370 genes), this Tier was further subdivided to prioritize those genes that were in proximity (+/- 50 kbp) to a GWAS SNP and had an extracellular location (Tier 3A). The remainder of the genes was assigned to Tier 3B.

Pfam-A domain content

To evaluate the Pfam-A domain content for druggable genes, gene identifiers were converted to UniProt accession keys using the UniProt web services (67). Only UniProt accessions matching the regular expression pattern '[OPQ][0-9][A-Z0-9]{3}[0-9]' were retained for further analysis. Pfam-A domains were extracted using the Xfam API (69). For genes mapping to multiple UniProt accessions, we retained domain annotations for the UniProt accession mapping to the highest number of unique Pfam-A domains.

Comparison of druggable gene sets

For comparison with genes covered on the Illumina DrugDev array, sets of druggable genes defined by Hopkins and Groom in 2002 (28), Russ and Lampel in 2005 (29), and Kumar (30) were obtained from DGIdb (31). Gene names were converted to Ensembl gene identifiers using the Ensembl REST API (70). The overlap between the three sets was determined and visualized using the Python module matplotlib_venn.

Compilation of GWAS results

The GWAS catalog was downloaded from (http://www.ebi.ac.uk/gwas/api/search/downloads/alternative) on 21/07/2015. Several quality control and further post processing steps were then taken. The identifiers of associated variants were validated against Ensembl

(version 79, build 37) using the perl API. This step returned the latest identifier and the human genome build 37 chromosome coordinates; 707 associated variants could not be validated and were excluded. The GWAS catalog provides numerical effect estimates but does not specify the type of effect, such as odds ratio (OR) or beta co-efficient. We attempted to resolve the effect type by using data in other fields (such as the presence of case or control in the discovery population fields) to classify the effect type as OR, beta, or unknown. The discovery population field was also processed using a set of regular expressions to determine the sample size and populations used. The populations were then mapped to an appropriate 1000 genomes super population. Where no population name could be identified, EUR was used as a default because the majority of studies in the GWAS catalog were performed on Europeans. The pubmed identifier field was used to search Pubmed using the Biopython API. MeSH terms for the publications were mapped to the association to provide structured phenotype descriptions. However, these study level descriptions may not apply to every association reported by the study, therefore the MeSH terms were manually curated for each association. These supplemented the experimental factor ontology terms (EFO) that are already present in the GWAS catalog. Finally, the associations were filtered for those that are $\leq 5\times10^{-8}$ so all data used in this study exceeded genome-wide significance.

Assignment of LD intervals

The complete 1000 genomes phase 3 data (release 5) was downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502. BCFTools (v1.2 using HTSlib 1.2.1) and used to subset the vcf files into sub- and super- population files (71). For each population group, Plink v1.90b3d (72) was used to perform pairwise LD (r^2) calculations between all variants in the processed GWAS catalog and bi-allelic 1000 genomes variants within a 1 Mbp flank on either side of the GWAS variant having a maf ≥ 0.005. To reduce file size, only r^2 values ≥ 0.2 were output. The boundaries of the LD region surrounding each GWAS SNP were defined by the positions of the variants furthest upstream and downstream of this SNP with an r^2 value ≥ 0.5. Associated variants that were not present in the 1000 genomes panel that were not in LD with any other variants were given a nominal flank of 2.5 kbp on either size of the association.

Linking GWAS and drug target data

Gene annotations were extracted from Ensembl version 79. After filtering out pseudogenes, 38,352 genes remained. The set of genes was further reduced to those that overlapped an LD region surrounding an association. Within each associated LD region, the absolute base pair distance of the closest point of a gene from the associated variant was calculated. Variants located within a gene were given a distance of 0 bp. Genes were given a distance rank value according to their base pair distance. In the event of a distance rank tie, the gene with the oldest annotation date was assigned the lower rank.

Drug targets in ChEMBL 20 are annotated with UniProt accessions. The accessions were converted to Ensembl gene identifiers using the UniProt ID mapper (http://www.uniprot.org/uploadlists/). Drug target Ensembl gene IDs were then intersected with the IDs of genes within LD regions to give a set of drug targets in the proximity of associated variants.

Evaluation of consistency between licensed drug indications and GWAS disease/ biomarker traits

We evaluated the concordance between drug indication and disease association for those LD intervals defined by a GWAS SNP containing one or more genes encoding the target or targets of licensed drugs (fig. S9). Two experienced clinicians used a pre-specified classification system developed in a pilot study of one-fifth of the total data set. Each physician was blinded to the identity of the gene encoding the druggable target within each LD interval. The outputs from the two physician-curators were then compared, any coding errors corrected, and inconsistencies between curators resolved by consensus where agreement could be reached. Category 0 referred to a situation where coding could not be completed because of missing data; 1 to a precise drug indication-target gene-disease association match; 2 to a drug indication-target gene-disease area association match; and 3 to a drug indication-target gene-mechanism-of-action association match. Categories 1 to 3 were defined as 'concordant'. Category 4 referred to a drug mechanism based adverse effect-target gene-disease-association match; 5 to a drug indication-target gene-disease association mismatch with prior biological plausibility, and 6 without prior biological plausibility; 7 to a trait unlikely to be of therapeutic interest (such as hair color); and 8 to a genetic association with a new biomarker of uncertain biological function (such as a metabolite measured by a metabolomics platform). For certain drug targets/genes, a 34 code was used to indicate that the genetic association finding could reflect both a mechanism of action and mechanism-based adverse effect rediscovery. For example, the modification of certain electrocardiographic parameters by variants in the targets of certain antiarrhythmic drugs could reflect both their mechanism of action and the mechanism by which such drugs produce their adverse effects. A 54 code was used when there was uncertainty about the direction of effect. A 9 code was assigned to the four cases where consensus could not be reached between the two curators. Categories 4, 5, 54, and 6 were referred to as discordant. Categories 1-4 and 34 were referred to collectively as 'GWAS rediscoveries' of known drug effects.

Design of the Illumina DrugDev Array and comparative analysis of coverage of variation in the druggable genome

Selection of custom SNP content—The design was based on three tiers, corresponding to the level of evidence for druggability of the encoded proteins, with highest priority given to genes in Tiers 1 and 2. Tag SNPs were selected from the 1000 genomes European ancestry populations (CEU/GBR/FIN/TSI). Associations (tagging) between SNPs were identified based on linkage disequilibrium (r2 >0.8). SNPs already covered, or tagged by the Human Core base content were not duplicated. Only SNPs with a minor allele frequency ≥1.5% were considered for inclusion. The tagging threshold was defined as the number of variants a SNP tags (including itself) and was varied according to the tier. For Tiers 1 and 2, a tagging threshold of 1 was applied, meaning that all SNPs were considered for inclusion, even if they only tag themselves. For Tier 3A, we used a tagging threshold of 3, and for Tier 3B, a threshold of 4. SNPs were selected only if they were positioned within +/-2.5 kbp of the druggable genes selected in the three tiers (defined as a region of 2.5 kbp upstream of the Ensembl gene start position to 2.5 kbp downstream of the Ensembl gene end position). SNPs

from the Illumina Exome array were also included in the custom content where these were found within genes in Tiers 1, 2, and 3A. Again, any redundancy with the Human Core and selected tag SNP content was eliminated. A collection of mitochondrial tag SNPs from the Broad Institute, designed to capture common variation within the mitochondrial genome, was also included in the custom content ((http://www.broadinstitute.org/mpg/tagger/mito.html). This set is comprised of 64 SNPs, but only 56 of these loci were designable and included in the array. Finally, remaining space was filled with lead SNPs for any disease or trait association from the GWAS catalog, prioritizing SNPs located within 50 kbp of a druggable gene, or within the boundaries of any protein-coding gene.

For Tier 1 genes, 99,102 custom markers were selected, including tag SNPs and HumanExome content. A further 17,944 of the HumanCore markers also fell within Tier 1 gene regions, giving 117,046 markers in total. Tier 2 included 40,943 custom markers, and an additional 6,270 markers from the HumanCore fell within Tier 2 gene regions, resulting in a total of 47,213 markers. Genes in Tier 3 were represented by 38,858 custom markers. A further 21,626 HumanCore markers fell within Tier 3 gene regions, yielding 60,484 markers in total. In addition to coverage of genes encoding druggable targets, 6,400 SNPs associated with complex diseases or traits identified from the GWAS catalog and from selected gene-centric studies were also incorporated in the array content. Of these SNPs, 2,996 were already covered in the Human Core or previously included in the custom content, leaving 3,410 variants to be added (of which 1,395 were within Tier 1-3 gene regions). Finally, 53 mitochondrial genome tag SNPs were also included, along with 9 mitochondrial genome exome SNPs. Considering all content, 226,138 markers were located in, or within +/-2.5 kbp of, genes in the selected drugged, druggable, and ADME sets. For the array as a whole, 78,175 markers were exonic, 286,577 intronic, and 27,393 located in 5'-, and 41,171 in 3'-untranslated regions.

We used variants in the 1000 genomes phase 3 reference panel populations to compare coverage of the druggable genome by the new array and other commonly used genotyping arrays (see previous section). For this analysis, the variants in each array were first mapped to the 1000 genomes phase 3 reference panel, and coverage was then compared using two metrics: variant density (per kbp of the druggable gene) and the proportion of the variants in the druggable genome that were captured. We defined complete coverage of druggable genome as capture of all the bi-alleilic variants in a 1000 genomes phase 3 reference panel population with a minor allele frequency ≥ 0.005 (representing low-frequency to common variants). Because of differences in variant content reported in successive genome builds, not all the content of the genotyping arrays could be mapped back to the 1000 genomes phase 3 reference set. However, the proportion of variants captured by each array that could be mapped to the 1000 genomes reference panel was very similar (fig. S10).

Evaluating genotyping array coverage of the DrugDev array—The build 37 genotyping array content for the Illumina arrays was downloaded from Will Rayner's array strand website (http://www.well.ox.ac.uk/~wrayner/strand). Where multiple versions of an array exist, the latest version number was downloaded. The Affymetrix array annotations were downloaded as SQLite databases from the Affymetrix website. 1000 genomes data were processed as described in the method for creating LD regions. Variants present on the

genotyping arrays were mapped to 1000 genomes phase 3 using the following sequence: variants with rs identifiers were searched against the 1000 genomes sites file, and if no match was obtained, then synonyms of the rs identifier (obtained from Ensembl version 79 build 37) were searched. Variants not mapping by rs identifier were then mapped by chromosome, position, and alleles (flipping the strand of the alleles where appropriate). Allele frequencies and variant tagging for each sub-population group were calculated using Plink(v1.90b3d (73)). Tagging was restricted to bi-allelic low-frequency and common variants (maf ≥ 0.005) within 1 Mbp of the source SNP. Baseline 1000 genomes coverage of the druggable genome in the different sub-populations was ascertained using Bedtools (v2.22.1) to intersect 1000 genomes variants with a maf ≥ 0.005 against the druggable gene list (including 2.5 kbp up/down stream). Proportional coverage of the druggable genome by the different genotyping arrays was then ascertained by intersecting the baseline coverage with the 1000 genomes mapped array content.

Indications and adverse effects of licensed therapies—Drug indication data were obtained from several sources. The primary source was the First Databank database (FDB, http://www.fdbhealth.co.uk/). This is a commercial database used by University College London Hospitals (UCLH), and a one off single release was kindly provided for research purposes by First Databank Europe Ltd. Because FDB is used clinically, this was regarded as the “gold standard” indication set used for the manual categorization of concordant/discordant drug/GWAS links (see above). FDB drug indications are tagged with Universal Medical Language System concept identifiers (CUIs) and could be mapped into MeSH and other ontologies within the UMLS meta-thesaurus (51, 74). Drug indication data were obtained from ChEMBL 21 by manual curation and mapping of data from FDA drug labels (https://dailymed.nlm.nih.gov/dailymed/), WHO ATC classification (http://www.whocc.no/atc_ddd_index/), and ClinicalTrials.gov (https://clinicaltrials.gov). This was used to supplement the FDB data and fill in indication data for drugs that were not present in the FDB release.

Side effect data were obtained from the Side Effect Resource (SIDER) database (75). The drug identifiers used in SIDER were mapped back to ChEMBL identifiers using a mapping file provided by SIDER. The side effects are provided as MedRA terms and UMLS CUIs and were mapped to MeSH terms using the UMLS.

Statistical analysis

The proportion of druggable genes in LD intervals specified by GWAS associations in each MeSH disease or MeSH psychiatry category was calculated by dividing the number of druggable genes by the number of all genes with 95% confidence intervals calculated assuming a binomial distribution, on the assumption that each study was independent.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr Cora Vacher and colleagues for helping to facilitate the design of the Illumina DrugDev Array. We would like to thank Dr. Reecha Sofat and Anita Jena-Smol for facilitating access to the First Databank Database and advice on designing database queries, and First Databank Europe Ltd for providing a single copy of the database for research purposes.

Funding: Work in this paper was supported by awards from University College London Hospitals National Institute of Health Research (NIHR) Biomedical Research Centre, British Heart Foundation (BHF Project Grant PG12/71/29684), a Strategic Award from the Wellcome Trust (WT086151/Z/08/Z) and Member States of the European Molecular Biology Laboratory (EMBL), and the the Rosetrees Trust. A.H.is an NIHR Senior Investigator. R.T.L is supported by an NIHR Clinical Lectureship. For J.O.and A.K., the work in this publication was entirely performed while employees of EMBL-EBI, funded under grants from the Wellcome Trust WT086151/Z/08/Z and WT104104/Z/14/Z. Since that time, they have both been employed by BenevolentAI. For F.K., the work in this publication was entirely performed while an employee of UCL under British Heart Foundation and Rosetrees Trust Grants. Since that time, he has been employed by BenevolentAI.

References

1. Munos B. Lessons from 60 years of pharmaceutical innovation. Nat Rev Drug Discov. 2009; 8:959–968. [PubMed: 19949401]
2. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov. 2010; 9:203–214. [PubMed: 20168317]
3. Macleod MR, Lawson McLean A, Kyriakopoulou A, Serghiou S, de Wilde A, Sherratt N, Hirst T, Hemblade R, Bahor Z, Nunes-Fonseca C, Potluru A, et al. Sena, Risk of Bias in Reports of In Vivo Research: A Focus for Improvement. PLoS Biol. 2015; 13:e1002273. [PubMed: 26460723]
4. Perel P, Roberts I, Sena E, Wheble P, Briscoe C, Sandercock P, Macleod M, Mignini LE, Jayaram P, Khan KS. Comparison of treatment effects between animal experiments and clinical trials: systematic review. BMJ. 2007; 334:197. [PubMed: 17175568]
5. van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, Macleod MR. Can Animal Models of Disease Reliably Inform Human Studies? PLoS Med. 2010; 7:e1000245. [PubMed: 20361020]
6. Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Med. 2005; 2:e124. [PubMed: 16060722]
7. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. R Soc Open Sci. 2014; 1:140216. [PubMed: 26064558]
8. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nat Methods. 2015; 12:179–185. [PubMed: 25719825]
9. Naci H, Ioannidis JPA. How Good Is "Evidence" from Clinical Studies of Drug Effects and Why Might Such Evidence Fail in the Prediction of the Clinical Utility of Drugs? Annu Rev Pharmacol Toxicol. 2015; 55:169–189. [PubMed: 25149917]
10. Arrowsmith J, Miller P. Trial Watch: Phase II and Phase III attrition rates 2011-2012. Nat Rev Drug Discov. 2013; 12:569–569. [PubMed: 23903212]
11. Hingorani A, Casas JP. Interleukin-6 receptor pathways in coronary heart disease: a collaborative meta-analysis of 82 studies. The Lancet. 2012; 379:1205–1213.
12. Hingorani A, Humphries S. Nature's randomised trials. The Lancet. 2005; 366:1906–1908.
13. Sofat R, Hingorani AD, Smeeth L, Humphries SE, Talmud PJ, Cooper J, Shah T, Sandhu MS, Ricketts SL, Boekholdt SM, Wareham N, et al. Separating the Mechanism-Based and Off-Target Actions of Cholesteryl Ester Transfer Protein Inhibitors With CETP Gene Polymorphisms. Circulation. 2010; 121:52–62. [PubMed: 20026784]
14. Davey Smith G. Capitalizing on Mendelian randomization to assess the effects of treatments. J R Soc Med. 2007; 100:432–5. [PubMed: 17766918]
15. Casas JP, Ninio E, Panayiotou A, Palmen J, Cooper JA, Ricketts SL, Sofat R, Nicolaides AN, Corsetti JP, Fowkes FGR, Tzoulaki I, et al. PLA2G7 Genotype, Lipoprotein-Associated

Phospholipase A2 Activity, and Coronary Heart Disease Risk in 10 494 Cases and 15 624 Controls of European Ancestry. Circulation. 2010; 121:2284–2293. [PubMed: 20479152]

16. CRP Coronary Heart Disease Genetics Collaboration. Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. BMJ. 2011; 342:d548. [PubMed: 21325005]
17. Holmes MV, Simon T, Exeter HJ, Folkersen L, Asselbergs FW, Guardiola M, Cooper Ja, Palmen J, Hubacek Ja, Carruthers KF, Horne BD, et al. Secretory phospholipase A(2)-IIA and cardiovascular disease: a mendelian randomization study. J Am Coll Cardiol. 2013; 62:1966–76. [PubMed: 23916927]
18. Swerdlow DI, Preiss D, Kuchenbaecker KB, Holmes MV, Engmann JEL, Shah T, Sofat R, Stender S, Johnson PCD, Scott RA, Leusink M, et al. HMG-coenzyme A reductase inhibition, type 2 diabetes, and bodyweight: evidence from genetic analysis and randomised trials. The Lancet. 2014; 385:351–361.
19. Scott RA, Freitag DF, Li L, Chu AY, Surendran P, Young R, Grarup N, Stancáková A, Chen Y, Varga TV, et al. A genomic approach to therapeutic target validation identifies a glucose-lowering GLP1R variant protective for coronary heart disease. Sci Transl Med. 2016; 8
20. Würtz P, Wang Q, Soininen P, Kangas AJ, Fatemifar G, Tynkkynen T, Tiainen M, Perola M, Tillin T, Hughes AD, Mäntyselkä P, et al. Metabolomic Profiling of Statin Use and Genetic Inhibition of HMG-CoA Reductase. J Am Coll Cardiol. 2016; 67:1200–1210. [PubMed: 26965542]
21. Melzer D, Perry JRB, Hernandez D, Corsi A-M, Stevens K, Rafferty I, Lauretani F, Murray A, Gibbs JR, Paolisso G, Rafiq S, et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). PLoS Genet. 2008; 4:e1000072. [PubMed: 18464913]
22. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PIW, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, et al. Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels. Science. 2007; 316:1331–1336. [PubMed: 17463246]
23. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, et al. A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants. Science. 2007; 316:1341–1345. [PubMed: 17463248]
24. Cook D, Brown D, Alexander R, March R, Morgan P, Satterthwaite G, Pangalos MN. Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. Nat Rev Drug Discov. 2014; 13:419–431. [PubMed: 24833294]
25. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, Floratos A, Sham PC, Li MJ, Wang J, Cardon LR, et al. The support of human genetic evidence for approved drug indications. Nat Genet. 2015; doi: 10.1038/ng.3314
26. Plenge RM, Scolnick EM, Altshuler D. Validating therapeutic targets through human genetics. Nat Rev Drug Discov. 2013; 12:581–594. [PubMed: 23868113]
27. Sanseau P, Agarwal P, Barnes MR, Pastinen T, Richards JB, Cardon LR, Mooser V. Use of genome-wide association studies for drug repositioning. Nat Biotechnol. 2012; 30:317–320. [PubMed: 22491277]
28. Hopkins AL, Groom CR. The druggable genome. Nat Rev Drug Discov. 2002; 1:727–30. [PubMed: 12209152]
29. Russ AP, Lampel S. The druggable genome: an update. Drug Discov Today. 2005; 10:1607–10. [PubMed: 16376820]
30. Kumar RD, Chang L-W, Ellis MJ, Bose R. Prioritizing Potentially Druggable Mutations with dGene: An Annotation Tool for Cancer Genome Sequencing Data. PLoS ONE. 2013; 8:e67980. [PubMed: 23826350]
31. Wagner AH, Coffman AC, Ainscough BJ, Spies NC, Skidmore ZL, Campbell KM, Krysiak K, Pan D, McMichael JF, Eldred JM, Walker JR, et al. DGIdb 2.0: mining clinically relevant drug–gene interactions. Nucleic Acids Res. 2016; 44:D1036–D1044. [PubMed: 26531824]
32. Rask-Andersen M, Masuram S, Schiöth HB. The Druggable Genome: Evaluation of Drug Targets in Clinical Trials Suggests Major Shifts in Molecular Class and Indication. Annu Rev Pharmacol Toxicol. 2014; 54:9–26. [PubMed: 24016212]

33. Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, Altman RB, Klein TE. Pharmacogenomics Knowledge for Personalized Medicine. Clin Pharmacol Ther. 2012; 92:414–417. [PubMed: 22992668]
34. Zhu F, Shi Z, Qin C, Tao L, Liu X, Xu F, Zhang L, Song Y, Liu X, Zhang J, Han B. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. Nucleic Acids Res. 2012; 40:D1128–1136. [PubMed: 21948793]
35. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. 2014; 42:D1091–1097. [PubMed: 24203711]
36. Mullard A. 2015 FDA drug approvals. Nat Rev Drug Discov. 2016; 15:73–76. [PubMed: 26837582]
37. Hindorff, La; Sethupathy, P; Junkins, Ha; Ramos, EM; Mehta, JP; Collins, FS; Manolio, Ta. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009; 106:9362–7. [PubMed: 19474294]
38. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, Krüger Fa, Light Y, Mak L, McGlinchey S, Nowotka M, et al. The ChEMBL bioactivity database: an update. Nucleic Acids Res. 2014; 42:D1083–90. [PubMed: 24214965]
39. First Databank Europe Ltd. Drug Data | FDB (First Databank). (available at http://www.fdbhealth.co.uk/)
40. Barr AJ. Protein tyrosine phosphatases as drug targets: strategies and challenges of inhibitor development. Future Med Chem. 2010; 2:1563–1576. [PubMed: 21426149]
41. Knapp S. Emerging Target Families: Intractable Targets. Handb Exp Pharmacol. 2015; doi: 10.1007/164_2015_28
42. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol. 2005; 6:R44. [PubMed: 15892872]
43. The Interleukin-6 Mendelian Randomisation Analysis Consortium. The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. The Lancet. 2012; 379:1214–1224.
44. Eijgelsheim M, Newton-Cheh C, Sotoodehnia N, de Bakker PIW, Müller M, Morrison AC, Smith AV, Isaacs A, Sanna S, Dörr M, Navarro P, et al. Genome-wide association analysis identifies multiple loci related to resting heart rate. Hum Mol Genet. 2010; 19:3885–3894. [PubMed: 20639392]
45. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015; 526:68–74. [PubMed: 26432245]
46. Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, Manolio T, Rudan I, McKeigue P, Wilson JF, Campbell H. Abundant Pleiotropy in Human Complex Diseases and Traits. Am J Hum Genet. 2011; 89:607–618. [PubMed: 22077970]
47. Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, Karlsson A, Al-Lazikani B, Hersey A, Oprea TI, Overington JP. A comprehensive map of molecular drug targets. Nat Rev Drug Discov. 2016; doi: 10.1038/nrd.2016.230
48. Keating BJ, Tischfield S, Murray SS, Bhangale T, Price TS, Glessner JT, Galver L, Barrett JC, Grant SFA, Farlow DN, Chandrupatla HR, et al. Concept, Design and Implementation of a Cardiovascular Gene-Centric 50 K SNP Array for Large-Scale Genomic Association Studies. PLoS ONE. 2008; 3:e3583. [PubMed: 18974833]
49. Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, Burtt NP, Fuchsberger C, Li Y, Erdmann J, Frayling TM, et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet. 2012; 8:e1002793. [PubMed: 22876189]
50. Cortes A, Brown MA. Promise and pitfalls of the Immunochip. Arthritis Res Ther. 2011; 13:101. [PubMed: 21345260]
51. Rogers FB. Communications to the Editor. Bull Med Libr Assoc. 1963; 51:114–116. [PubMed: 13982385]

52. Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, Felix V, Feng G, Kibbe WA. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. 2012; 40:D940–D946. [PubMed: 22080554]
53. Robinson P, Mundlos S. The Human Phenotype Ontology. Clin Genet. 2010; 77:525–534. [PubMed: 20412080]
54. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis Ca, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]
55. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, Farnham PJ, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol. 2010; 28:1045–1048. [PubMed: 20944595]
56. Al-Numair NS, Martin AC. The SAAP pipeline and database: tools to analyze the impact and predict the pathogenicity of mutations. BMC Genomics. 2013; 14:S4.
57. Lourdusamy A, Newhouse S, Lunnon K, Proitsi P, Powell J, Hodges A, Nelson SK, Stewart A, Williams S, Kloszewska I, Mecocci P, et al. Identification of cis-regulatory variation influencing protein abundance levels in human plasma. Hum Mol Genet. 2012; 21:3719–3726. [PubMed: 22595970]
58. Suhre K, Shin S-Y, Petersen A-K, Mohney RP, Meredith D, Wägele B, Altmaier E, Deloukas P, Erdmann J, Grundberg E, Hammond CJ, et al. Human metabolic individuality in biomedical and pharmaceutical research. Nature. 2011; :477.doi: 10.1038/nature10354
59. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, Plagnol V, Williams SM. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genet. 2014; 10:e1004383. [PubMed: 24830394]
60. Collins FS. Reengineering Translational Science: The Time Is Right. Sci Transl Med. 2011; 3
61. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012; 40:D1100–7. [PubMed: 21948594]
62. National Institute of Health. ClinicalTrials.gov*ClinicalTrials.gov*. (available at https://clinicaltrials.gov/)
63. Montreal Heart Institute Pharmacogenomics Center. www.pharmaadme.org - Home. (available at http://pharmaadme.org/joomla/index.php?option=com_frontpage&Itemid=1)
64. EMBL-EBI. Kinase SARfari. (available at https://www.ebi.ac.uk/chembl/sarfari/kinasesarfari/)
65. EMBL-EBI. GPCR SARfari. (available at https://www.ebi.ac.uk/chembl/sarfari/gpcrsarfari)
66. Pawson AJ, Sharman JL, Benson HE, Faccenda E, Alexander SPH, Buneman OP, Davenport AP, McGrath JC, Peters JA, Southan C, Spedding M, et al. Nc-Iuphar, The IUPHAR/BPS Guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands. Nucleic Acids Res. 2014; 42:D1098–D1106. [PubMed: 24234439]
67. The Uniprot Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2015; 43:D204–D212. [PubMed: 25348405]
68. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25:25–9. [PubMed: 10802651]
69. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, et al. Pfam: the protein families database. Nucleic Acids Res. 2014; 42:D222–D230. [PubMed: 24288371]
70. Yates A, Beal K, Keenan S, McLaren W, Pignatelli M, Ritchie GRS, Ruffier M, Taylor K, Vullo A, Flicek P. The Ensembl REST API: Ensembl Data for Any Language. Bioinformatics. 2014
71. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27:2156–2158. [PubMed: 21653522]
72. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MaR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559–75. [PubMed: 17701901]

73. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015; 4doi: 10.1186/s13742-015-0047-8

74. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004; 32:D267–D270. [PubMed: 14681409]

75. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res. 2016; 44:D1075–1079. [PubMed: 26481350]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

One Sentence Summary

The druggable genome and genome-wide association study data reveal new drug development and repurposing opportunities.

Europe PMC Funders Author Manuscripts

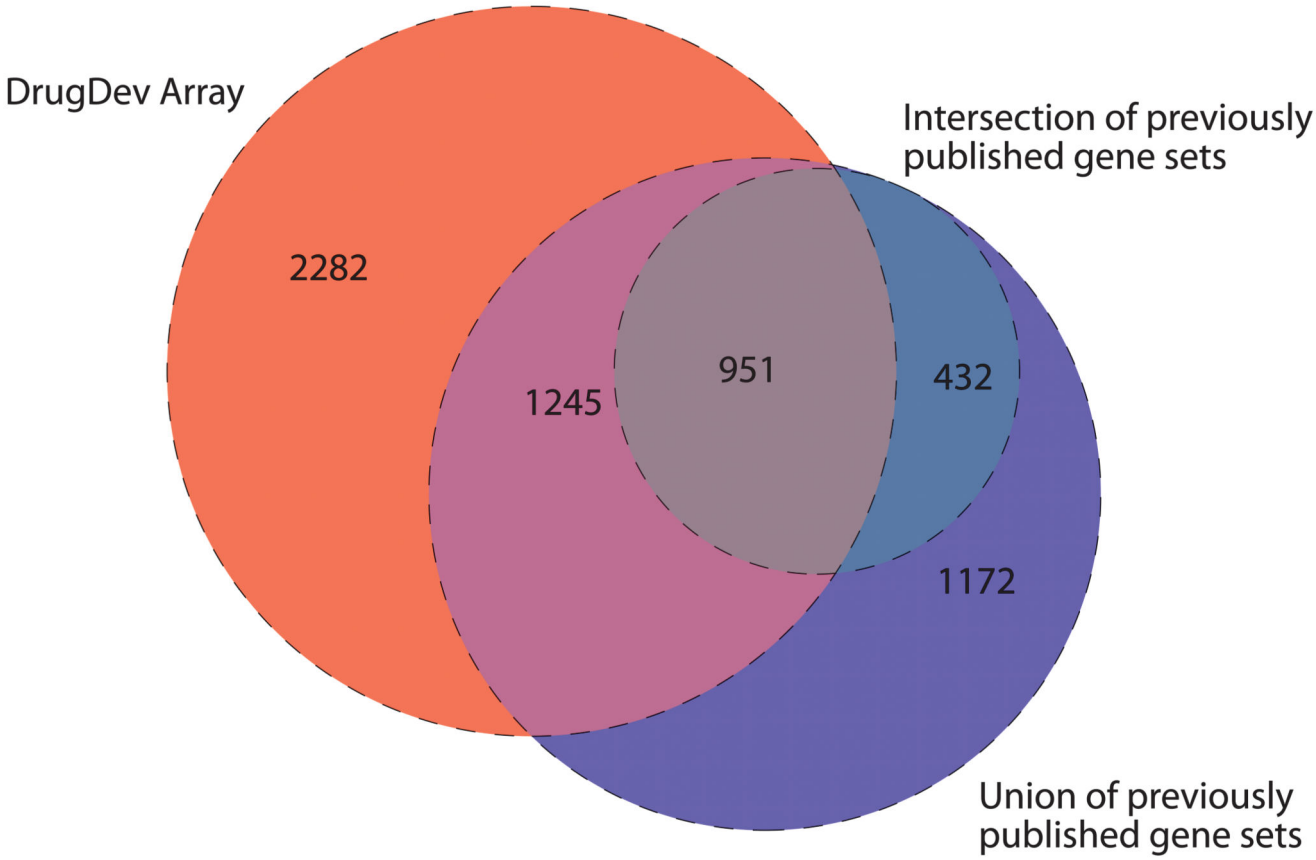

Fig. 1.

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 (30).

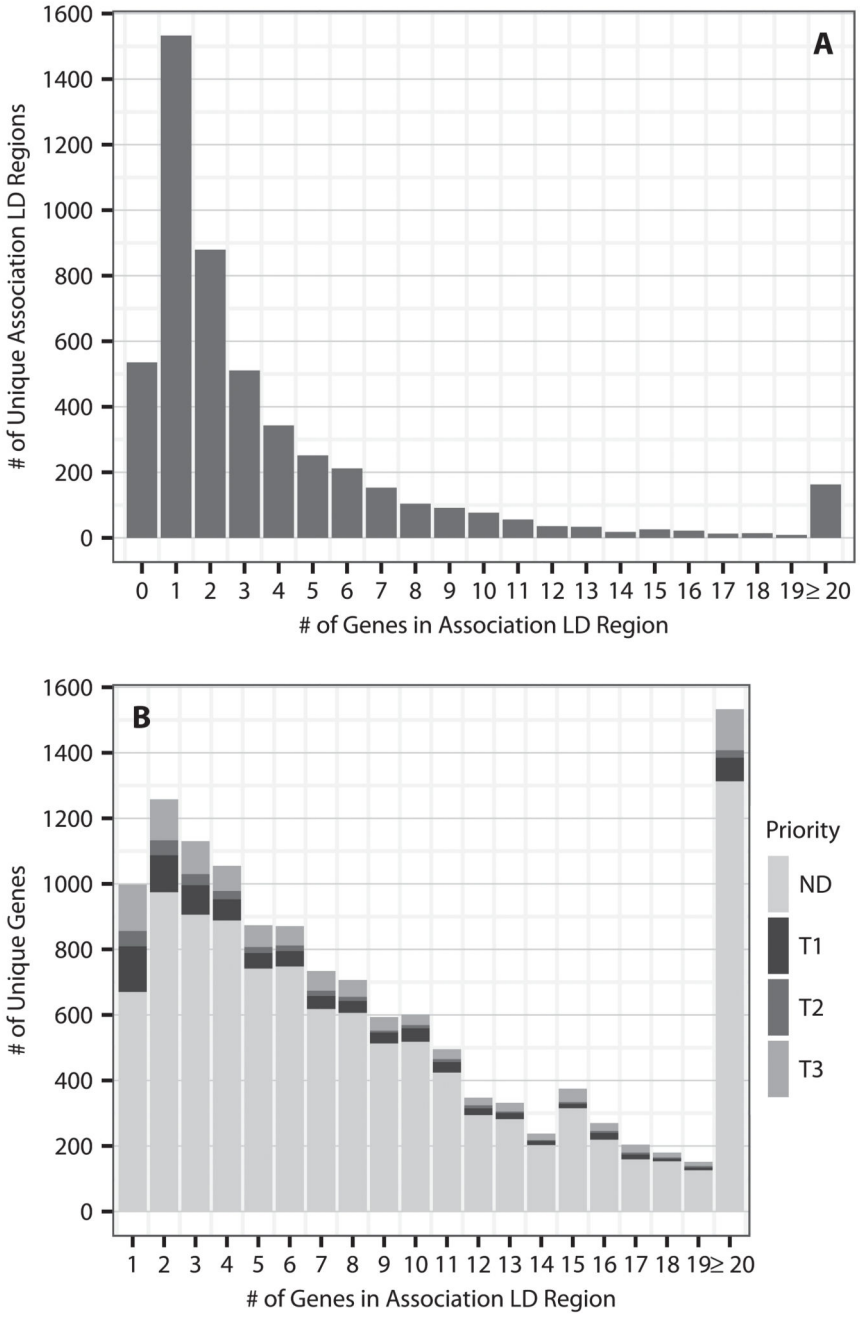

Fig. 2.

LD region summary. A shows the numbers of unique GWAS significant associations ($p \leq 5 \times 10^{-8}$) in the GWAS catalog that have 0 or more genes in their LD regions. Note that there are 299 associations that had no LD region or were not present in the 1000 genomes, which are not shown in this figure. B shows the number of unique genes that occupy LD regions with at least 1 gene. The counts are partitioned into genes that are not predicted (ND) to be druggable or the various druggable tiers (T1: Tier 1, T2: Tier2, T3: Tier 3A and Tier 3B combined)

Europe PMC Funders Author Manuscripts

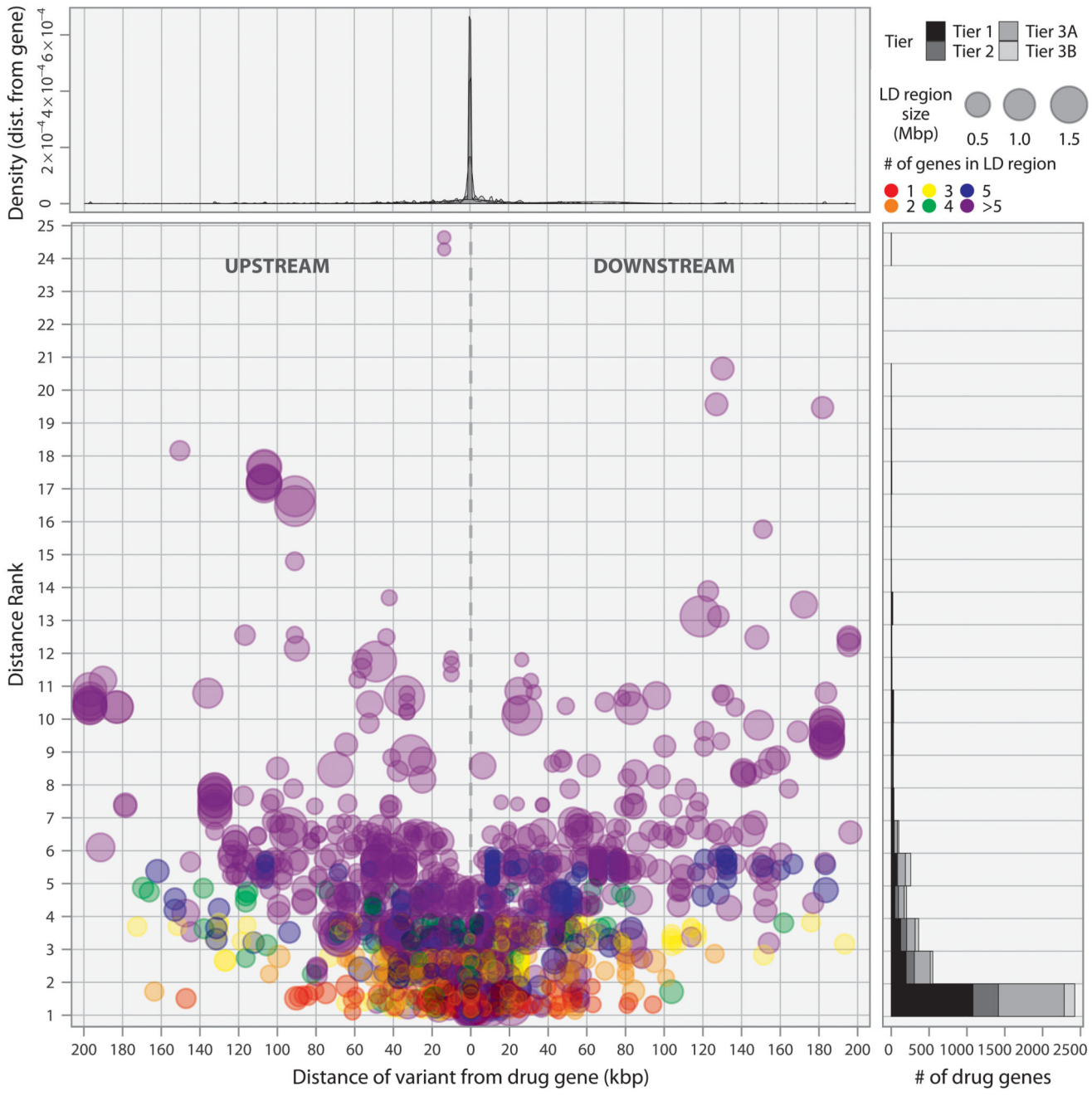

Fig. 3.

Proximity and distance rank of druggable genes to GWAS SNPs. Each point in the scatterplot corresponds to a GWAS signal located in an interval containing a druggable gene. The position on the x-axis indicates the distance of the SNP from the druggable gene. Position in the y-axis indicates the number of genes in the same interval that are closer to the signal than the druggable gene. The top panel indicates the signal density for all such SNPs, and the side panel provides the counts of signals by the distance rank of the druggable gene divided by Tier.

Europe PMC Funders Author Manuscripts

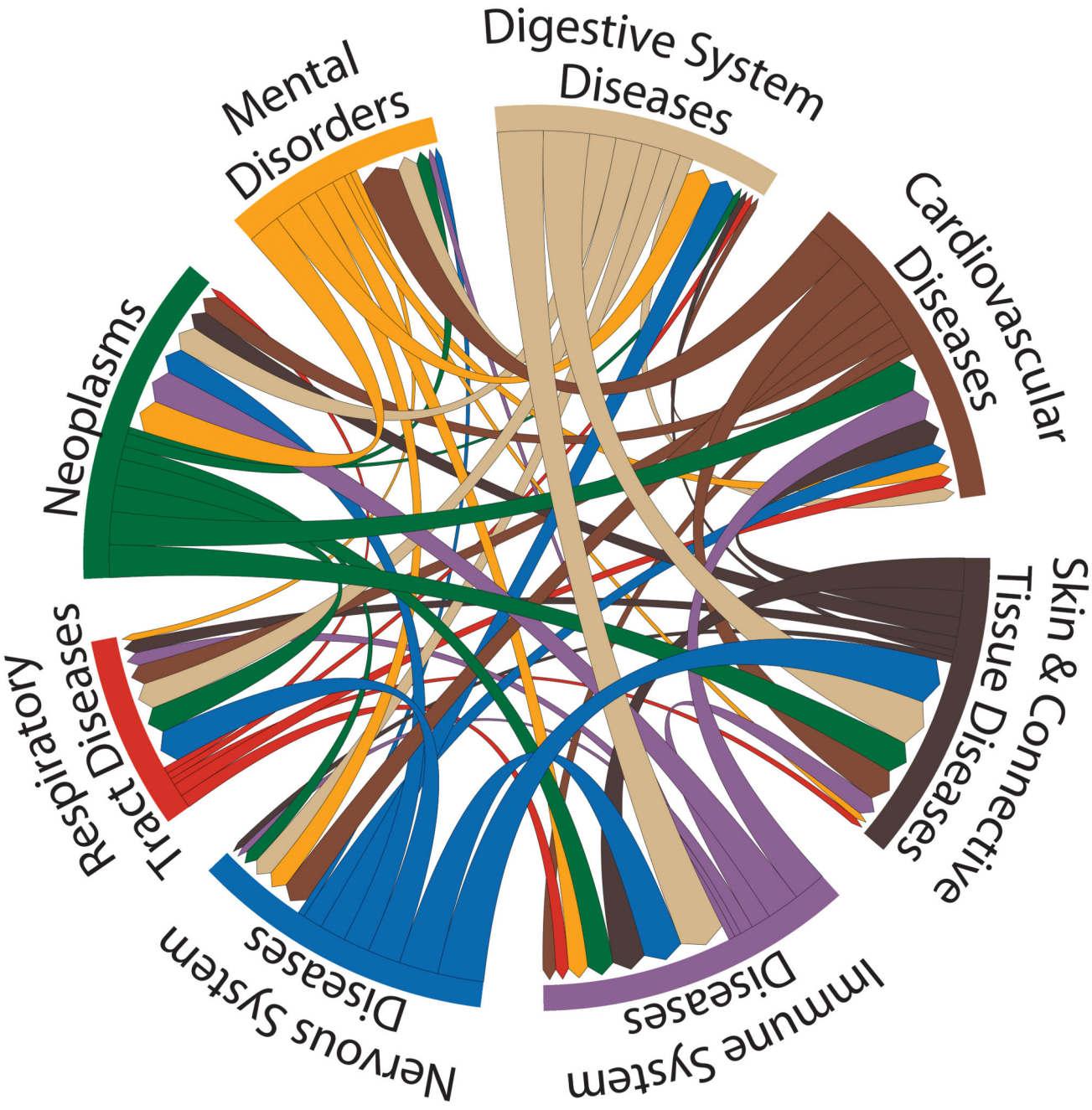

Fig. 4.

Potential repurposing opportunities from the discordant GWAS phenotype/drug indication matches. The disease categories on the circumference are MeSH root disease terms. The directional chords represent a connection from an indication class of drug to a GWAS phenotype. This connection is determined by a drug target gene occurring within 50 kbp of a GWAS association (a fixed distance was used to reduce the possibility of discordance due to confounding by linkage disequilibrium). The width of the chords is proportional to the number of genes connecting two therapeutic classes.

Europe PMC Funders Author Manuscripts

Diseases	Digestive System Diseases				Neoplasms			
Studies	106 (Est. N=234938)				187 (Est. N=478188)			
Assocs	705				783			
LD Regions	417				476			
Genes	1306				1466			
Drug Genes	256				219			
Drug Gene Priority	Tier 1 105		Tier >1 151		Tier 1 79		Tier >1 140	
Dist Rank	<= 2 57	>= 3 65	<= 2 76	>= 3 96	<= 2 46	>= 3 45	<= 2 72	>= 3 77
Compounds	16747	9303	4157	2168	32763	85817	7228	1592
USAN Compounds	351	204	17	19	519	729	71	24
Drugs	87	55	3	3	154	28	18	4
Drug I/ Disease P Comparison	C 10	D 45	C 2	D 42	C 0	D 2	C 0	D 2	C 24	D 69	C 1	D 17	C 2	D 8	C 0	D 2
Targets	3	12	2	9	0	2	0	1	6	8	1	6	1	3	0	3

Diseases	Nervous System Diseases				Cardiovascular Diseases			
Studies	104 (Est. N=323729)				84 (Est. N=426777)			
Assocs	425				388			
LD Regions	286				228			
Genes	997				670			
Drug Genes	170				135			
Drug Gene Priority	Tier 1 67		Tier >1 103		Tier 1 48		Tier >1 87	
Dist Rank	<= 2 40	>= 3 35	<= 2 48	>= 3 64	<= 2 29	>= 3 22	<= 2 49	>= 3 42
Compounds	83599	98334	867	1258	10802	5404	1995	2551
USAN Compounds	808	1057	30	18	359	216	31	20
Drugs	113	74	2	0	133	27	2	6
Drug I/ Disease P Comparison	C 12	D 35	C 9	D 42	C 0	D 1	C 0	D 0	C 4	D 60	C 10	D 11	C 0	D 1	C 0	D 3
Targets	2	12	3	8	0	1	0	0	4	12	4	4	0	2	0	2

Fig. 5.

Translational potential for the top 4 most studied MeSH root disease areas. For each disease area, the figure illustrates the estimated number of GWAS (Studies Row), the number of associations ($p \leq 5 \times 10^{-8}$) (Assocs), the number of LD regions corresponding to those associations (LD Regions), the number of genes in those regions (Genes), and the number of those genes that encode druggable targets (Drug Genes). Subsequent rows quantify the number of druggable genes by priority tier (Drug Gene Priority) and by distance rank of the druggable gene from the GWAS SNP (Dist Rank). The total numbers of compounds

(Compounds), compounds with an USAN/INN (USAN Compounds), and drugs corresponding to the drugged targets are also listed (Drugs). In the penultimate row, the numbers of drugs with an indication that is concordant (C) or discordant (D) with the GWAS phenotype are displayed (Drug I/Disease P Comparison). In the final row, the numbers of cognate targets for the concordant or discordant drugs are shown (Targets). Note that for the purposes of the figure, a drug target is a single gene even if it is part of a complex that is targeted by the drug. Within each cell, the values represent the number of unique entities, for example the cells in the Assocs row represent the number of unique associations (rsids). However, some values can be replicated across the figure because a GWAS study may have researched several of the disease areas. Additionally, there is some non-additivity between consecutive rows, namely Druggable Gene Priority (Drug Gene Priority) - Distance Rank (Dist Rank) and Drugs - Drug indication/Disease Phenotypes Comparison (Drug I/Disease P Comparison). In the case of the former, this is due to the same gene being further away from the associated variant in different studies, such that it falls into a different partition. For the latter, this is due to missing indications for some of the drugs, such that concordance or discordance could not be assigned. The estimated number of samples (Est. N) is the sum of all the cases involved in the respective studies.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

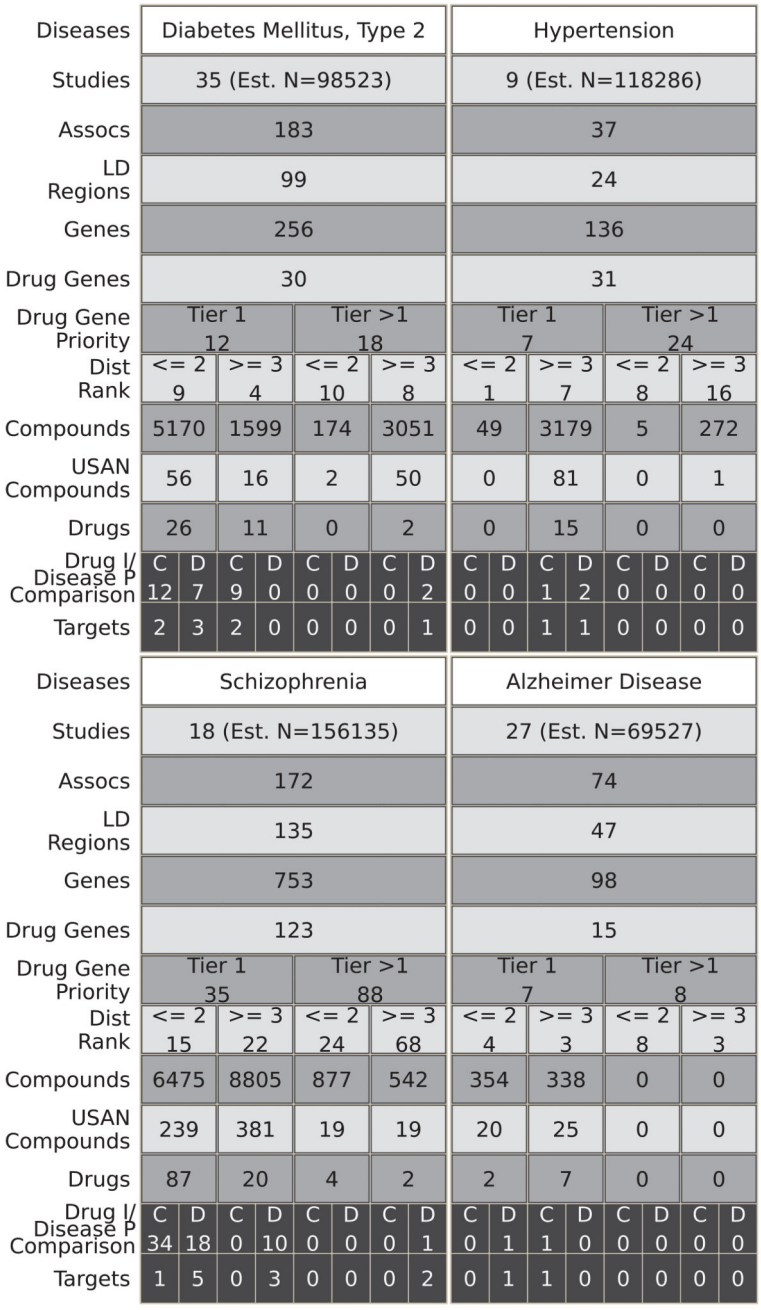

Fig. 6.

Translational potential for 4 specific diseases. Refer to Fig. 5 legend for detailed explanation.

Europe PMC Funders Author Manuscripts

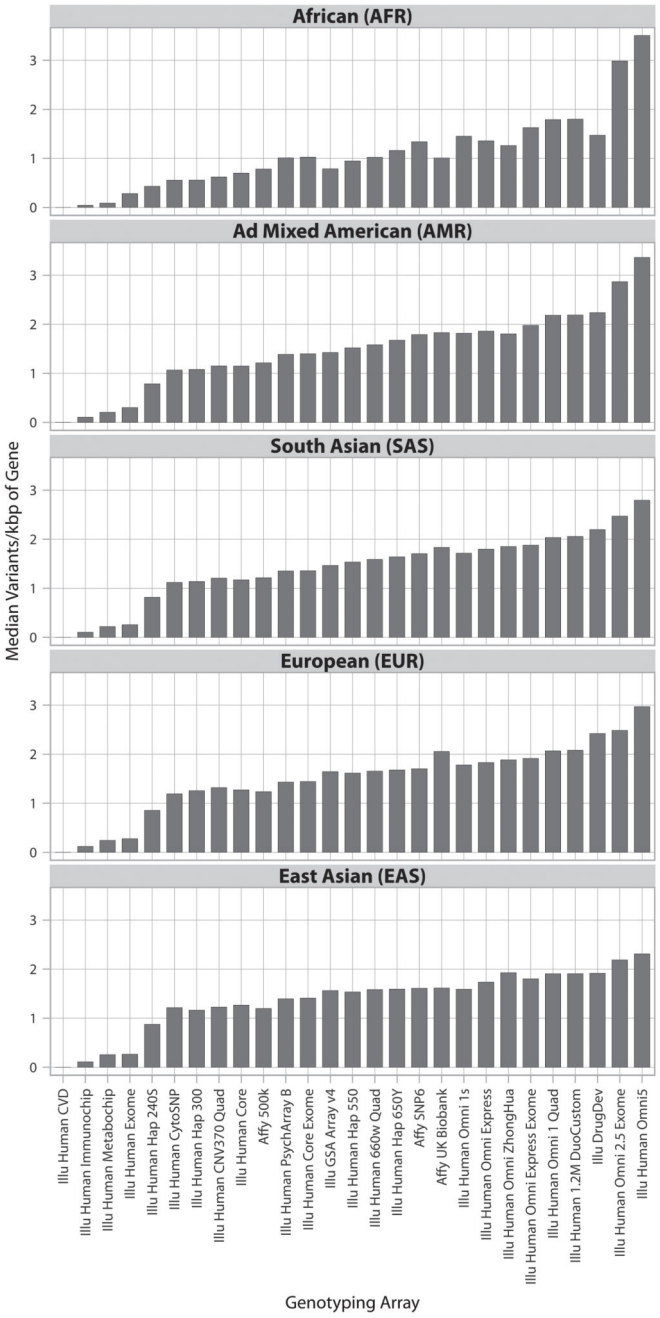

Fig. 7.

Tagged coverage of druggable genes in the 1000 genomes super populations. Coverage of the druggable gene set is represented as the median number of directly typed variants and variants in LD of $r^2 \geq 0.8$ (tagged) per kbp of druggable gene sequence.

Europe PMC Funders Author Manuscripts

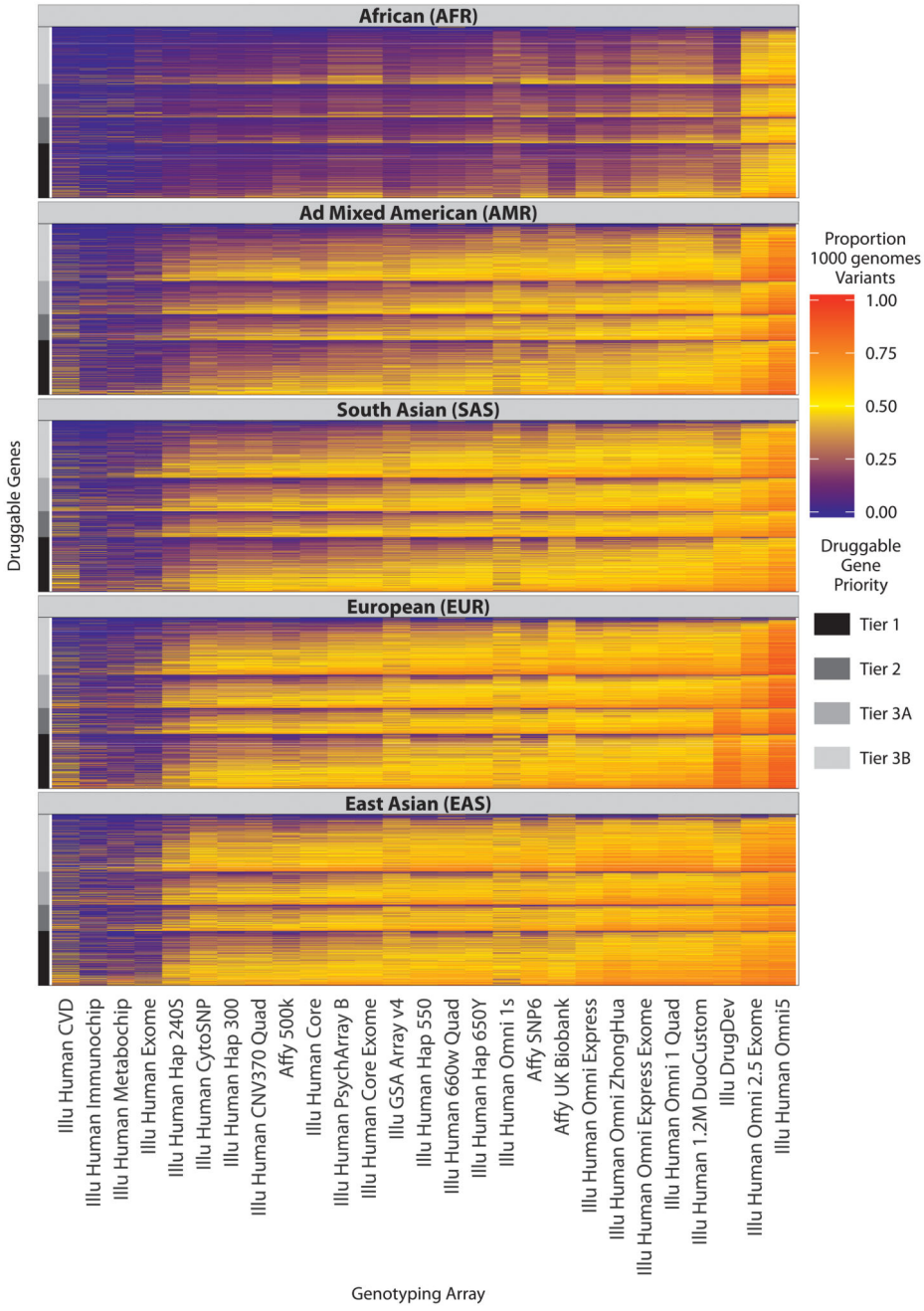

Fig. 8.

Tagged coverage of druggable genes in the 1000 genomes super populations. Coverage of the druggable gene set is represented as a proportion of 1000 genomes phase 3 variants (bi-allelic with maf ≥ 0.005) that are either directly typed or in LD with $r^2 \geq 0.8$ (tagged). Each column represents a genotyping array and each row a druggable gene. The druggable genes are grouped according to their druggability tier, which is indicated by the bar at the left side of each plot. To aid visualization, the druggable genes are further sorted within each tier on their median coverage across all the arrays, and the genotyping arrays are sorted based on

their median coverage of the druggable genome across all the 1000 genomes super populations. Note that all of the arrays contained some content that could not be mapped to the 1000 genomes phase 3 (see fig. S10). Note also that when deployed in real datasets, additional variation could be captured by all arrays through imputation.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

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

Europe PMC Funders Author Manuscripts

Table 2

Number of unique GWAS associations mapping to drug targets for licensed drugs curated according to the correspondence between the GWAS association and drug indication.

Category	# Associations	# drug targets
Disease association and treatment indication precisely concordant *	56	30
Disease association and treatment indication concordant within the same disease area *	13	9
Disease association concordant with a biomarker of therapeutic efficacy	97	37
Disease association corresponding to a mechanism-based adverse effect *	76	27
Disease association with a known biomarker of therapeutic efficacy that can also be responsible for a mechanism-based side effect *	32	8
Discordant disease association and target indication considered to imply a potential repurposing opportunity	1523	144
Discordant disease association and target indication considered to imply either a repurposing opportunity or mechanism-based side effect depending on the direction	108	52
Curators unable to agree		4

* Refers to a target effect rediscovery (see text)