Leveraging human genetics to guide cancer drug development

Ben Kinnersley[1*+], Amit Sud[1*], Elizabeth A Coker[2*], Joseph E Tym[2], Patrizio Di Micco[2], Bissan Al-Lazikani[2], Richard S Houlston[1]

1. Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK
2. Cancer Research UK Cancer Therapeutics Unit, The Institute of Cancer Research, London, UK

*These authors contributed equally to the work

+To whom correspondence should be addressed:

Ben Kinnersley (ben.kinnersley@icr.ac.uk), Tel: +44 (0)208 722 4424 Fax: +44 208 (0)722 4365

The Institute of Cancer Research, London, UK

Acknowledgement of research support for the study

This project was funded by a Team Science project grant from the Institute of Cancer Research. Additional support from Cancer Research UK is acknowledged. A.S. is supported by charitable funds from the Royal Marsden Hospital.

Running head: “Human genetics informing cancer drug discovery”

Disclaimers: The authors declare no competing interests

Keywords: Cancer, drug, target, discovery, human, genetics

Aspects of this study have been presented at the American Association for Cancer Research Annual Meeting 2018

ABSTRACT

PURPOSE

The high attrition rate of cancer drug development programs is a barrier to realising the promise of precision oncology. We have examined if the genetic insights from genome-wide association studies (GWAS) of cancer can guide drug development and repurposing in oncology.

MATERIALS AND METHODS

Across 37 cancers we identified 955 genetic risk variants from the NHGRI-EBI GWAS Catalog. We linked these variants to target genes using strategies based on information on linkage-disequilbrium, DNA 3D-structure and integration of predicted gene function and expression. Using the Informa Pharmaprojects database we identified genes that are targets of unique drugs and assessed the level of enrichment that would be afforded by incorporation of genetic information in pre-clinical and Phase II studies. For targets not under development we implemented machine learning approaches to assess druggability.

RESULTS

For all pre-clinical targets incorporation of genetic information a 2.00-fold enrichment of a drug being successfully approved could be achieved (95% confidence interval (CI): 1.14-3.48, *P*= 0.02). For Phase II targets a 2.75-fold enrichment was shown (95% CI: 1.42-5.35, *P*= 4.2×10^{-3}). Application of genetic information suggested potential repurposing of 15 approved non-oncology drugs.

CONCLUSION

Our findings serve to illustrate the value of using insights from the genetics of inherited cancer susceptibility discovery projects as part of a data-driven strategy to inform drug discovery. Supporting cancer germline genetic information for prospective targets is available from https://cansar.icr.ac.uk/.

INTRODUCTION

The high attrition rate of drug development programs represents a significant barrier to fully realising the vision of precision oncology[1]. The failure of preclinical model systems to adequately predict efficacy in humans is leading drug developers to seek additional sources of evidence to inform decisions about which targets to pursue[2,3].

Following completion of the Human Genome Project there has been rapid progress in identifying inherited genetic variants influencing cancer risk through genome-wide association studies (GWAS) and large-scale sequencing projects[4]. Genome-wide association studies have now have been performed for most common malignancies and many rare tumor types, and over 900 genetic variants have been robustly demonstrated to influence risk[4].

The insights from these GWAS potentially offer an additional mechanism for selecting drug targets and indications, both key requirements in drug discovery. Risk single-nucleotide polymorphisms (SNPs) in or near a gene that may associate with the activity or expression of the encoded protein therefore can be used as a tool to infer the effect of pharmacological action on the same protein in a trial. Specifically, by extension, disease-associated SNPs identified by GWAS can be explicitly interpreted as a source of randomized human evidence to aid drug target identification and validation.

Several examples serve to illustrate the application of human genetics to inform drug discovery by utilising knowledge of variation in genes associated with disease risk. These include the targeting of 3-hydroxy-3-methyl-glutaryl-coenzyme A reductase (HMGCR) by statins for treatment of coronary heart disease[5] and ustekinumab, a monoclonal inhibitor of interleukin-12 (IL-12) and IL-23 used to treat inflammatory bowel disease[6].

Here we have, using GWAS association data for 37 cancers, examined the potential for human genetics to guide cancer drug development and repurposing of current approved drugs.

MATERIALS AND METHODS

Compiling GWAS data

To curate cancer risk SNPs identified by GWAS we queried the National Human Genome Research Institute (NHGRI) GWAS catalogue[7] (https://www.ebi.ac.uk/gwas/; accessed July 2017). We imposed a number of quality control metrics, filtering by association *P*-value < 5 × 10^{-8} and including only SNPs associated with the cancer rather than another cancer-related phenotype such as progression. We additionally manually added SNPs from recent cancer GWAS that had not yet been added to the catalog (**Supplementary Table 1**). We considered GWAS associations irrespective of their ethnicity. Gene transcript information, including gene annotations and transcript start sites for human build 37 were obtained from Ensembl biomart Genes 89 dataset (http://grch37.ensembl.org/biomart/martview/).

Linking risk SNPs to target genes

To the extent that they have been deciphered, most GWAS risk SNPs map to non-coding regions of the genome and influence gene regulation. Since spatial proximity between specific genomic regions and chromatin looping interactions are central for the regulation of gene expression the 3D structure of DNA means that gene proximity to the risk SNP does always necessarily equate to target gene. It is however, the case that regulatory effects and hence target genes are generally confined within topologically associated domains (TADs) of the genome. To link risk SNPs to target genes we therefore adopted three strategies.

For linkage disequilibrium (LD) based annotation, an approach similar to that adopted by Finan *et al*., 2017[8] was undertaken. For each cancer risk SNP, correlated SNPs were obtained for European (CEU), east Asian (CHB) and African (YRI) populations from 1000 Genomes Project Phase 3 using the LDlink[9] web application (https://analysistools.nci.nih.gov/LDlink/). LD boundaries were designated by the smallest and largest genomic location of SNPs correlated (r^2 values 0.1 to 0.9) with the reported cancer risk SNP. For SNPs where LD information could not be obtained, the boundaries were taken as 2.5kb on either side of the SNP genomic position. Gene transcription start sites were then mapped to these LD boundaries.

Topologically associating domain boundaries encompassing each risk locus were based on H1 Human Embryonic Stem Cells were obtained from Schmitt *et al.*, 2016[10]. These data makes use of Hi-C data described in Dixon *et al.*, 2015[11]. TAD boundaries were identified using the insulation score approach proposed by Crane *et al.,* 2015[12] at 40kb resolution.

To further explore target gene prioritisation, we used DEPICT[13] (https://data.broadinstitute.org/mpg/depict/); an integrative tool, which based on predicted gene function, prioritizes the most likely target genes of risk SNPs uses gene expression data from multiple sources. SNP associations were pruned to a set of independent signals by r^2>0.05 in YRI, CEU and CHB populations additionally retaining SNPs for which LD metrics could not be obtained. We considered all target genes with a FDR *Q*<0.05 as well as the top gene per SNP.

Finally, as an adjunct to our GWAS-based analysis, we also considered the classical cancer susceptibility genes (CSGs) whose mutation in the germline is responsible for the various

Mendelian forms of cancer. These were obtained from the COSMIC Cancer Gene Census[14] (https://cancer.sanger.ac.uk/census; accessed February 2018).

Genetic association enrichment for approved drugs

Data on the status of drug-target combinations along the various stages of drug development from pre-clinical through to regulatory approval were obtained by interrogation of the Informa Pharmaprojects database (https://pharmaintelligence.informa.com/; accessed January 2018). In addition to drugs assessed by Pharmaprojects, cancer drugs approved for use in cancer susceptibility gene carriers were also considered. Drugs with a specific indication for symptom control only, were excluded. Records were retained if target genes could be unambiguously mapped to HUGO Gene Nomenclature Committee at the European Bioinformatics Institute (HGNC; https://www.genenames.org) identifiers. We assessed whether drug targets with supporting genetic evidence were more likely to be approved in the drug development pipeline, by constructing a two by two table of genes and counts corresponding to whether a gene product has genetic support as a drug target at respective stages of development (*e.g.* comparing approved drugs with those only reaching preclinical stages). Test of association was Fisher's exact test, with the Wald test used to quantity effect size and 95% confidence intervals. A *P*-value of 0.05 (two-sided) was considered as statistically significant. All statistical calculations were performed using R version 3.2 software.

Druggability annotation of target genes

Targets of FDA-approved drugs were obtained from Santos *et al.*, 2017[15]. Genes were filtered for protein-coding genes and canSAR v4 Cancer Protein Annotation Tool (CPAT)[16] used to identify proteins with >95% sequence homology to existing drug targets. CPAT was also used to extract structure- and ligand-based druggability assessments from canSAR (https://cansar.icr.ac.uk/; accessed 2018). Network-based druggability scores for proteins were based on Mitsopoulos *et al*., 2015[17].

Finally, we assessed all 355,305 active compounds identified by canSAR against their targets using Probe Miner[18], which catalogues >1.8 million compounds for their suitability as chemical tools against 2,220 Uniprot-defined human targets (http://probeminer.icr.ac.uk/#/download).

RESULTS

Linking risk SNPs to target genes

Across 37 cancers we identified 955 risk loci. To link sentinel risk SNPs to respective target gene(s), we first considered genes within regions of LD to which risk SNPs mapped, imposing a range of r^2 thresholds. After which, we considered all genes localising within the risk SNP-defined TAD boundaries. Finally, we based linkage on the gene prioritisation approach implemented in DEPICT[13]. These three approaches yielded between 394 and 7,379 protein-coding target genes (**Fig. 1, Supplementary Tables 1-3**).

Genetic association enrichment for approved drugs

By interrogating the Informa Pharmaprojects database, we identified 1,706 unique genes that were the target of 3,435 unique therapeutic agents for cancer (**Supplementary Table 4**). These were grouped according to the furthest point reached across five stages of drug development pipeline: (1) Pre-clinical (*i.e. in vitro* and *in vivo* dosing and toxicity assessment), (2) Phase I (safety and dosage), (3) Phase II (efficacy and side effects), (4) Phase III and pre-registration (efficacy and monitoring of adverse reactions), (5) Approved.

We first considered all targets from the Pre-clinical stage and assessed the level of enrichment for being successfully approved conferred by genetic information. All of the methods linking SNPs to target genes provided evidence for enrichment. For the LD-based assessment enrichment was strongly correlated with r^2 values; imposing a r^2 value >0.9 resulted in 2.00-fold improvement in targeting of Pre-clinical drugs (95% CI: 1.14-3.48, *P*=0.02, **Fig. 2A, Table 1**). The comparative enrichment associated with COSMIC catalogued CSGs was 6.61-fold (95% CI: 3.17-13.78, *P* = 2.23 x 10^{-6}, **Fig. 2A, Table 1**).

We reasoned that a target's failure to progress along the Pre-clinical and Phase I stages is often for reasons unrelated to efficacy, and therefore next considered all targets from Phase II and above, and assessed the degree of enrichment for approval conferred by genetic information. As with the analysis of pre-clinical targets incorporating genetic association information led to enrichment for approval (**Fig. 2B, Table 2**). The strongest enrichment from the LD-based approach was attained after imposing an r^2 value >0.9 which was associated with a significant 2.75-fold difference (95% CI: 1.42-5.35, *P* = 4.2 x 10^{-3}, **Fig. 2B**). The comparative enrichment associated with COSMIC catalogued CSGs was 5.72-fold (95% CI: 2.35-13.89, *P* = 8.41 x 10^{-5}).

Potential for re-purposing non cancer drugs

To explore the application of genetics to inform drug re-purposing we first identified approved drugs used in the treatment of non-oncological disease. We then examined discordant pairing of drug indications and cancer associations. We identified 15 genes for which an approved drug is currently available with genetic support (**Table 3**). Notable examples included: (1) TGFB1 at 19q13.2, where a targeted drug is used in the treatment of rheumatoid arthritis and is the site of an association with colorectal cancer risk[19]; (2) VDR at 12q13.11, which is targeted by drugs treating osteoporosis and is a risk locus for prostate cancer[20]; (3) At 11q14.3 TYR is the target of an approved drug used in the treatment of skin disorders, which is also the site of a risk locus for melanoma[21], squamous cell carcinoma[22] and basal cell carcinoma[23]; (4) PTGIR at 19q13.32 which is targeted by a drug used in the treatment of transplant rejection and peripheral vascular disease, and is the site of a chronic lymphocytic leukaemia risk locus[24].

Availability of cancer germline genetic information

Supporting cancer germline genetic information for prospective targets is available from https://cansar.icr.ac.uk/ (**Figure 3**). For each uniprot identifier, a report has been generated detailing whether the given gene has been annotated as containing cancer-causing germline mutations by the COSMIC germline cancer gene census[14], as well as whether any variants from cancer genome-wide association studies map to the gene (**Figure 3**).

DISCUSSION

Our findings support the potential of human genetics to guide the identification of drug targets, addressing a productivity-limiting step in drug development and a bottleneck to realising the vision of precision oncology. Specifically, we have demonstrated that knowledge of cancer susceptibility genes identified by GWAS can be used to maximise discovery of likely Pre-clinical and Phase II targets, thereby empowering drug development programs. Our analysis benefits from the larger of risk loci for cancer that have been identified over recent years thereby providing greater power than earlier studies[1].

Significant enrichment of pre-clinical and phase II targets was also shown by incorporating information on the classical CSGs. Given that many of the CSGs are somatically mutated these targets may have already directly influenced recent drug development programs. Indeed, we observed a highly significant enrichment for CSGs being selected for pre-clinical validation *per se* (OR=11.37; CI=7.44-17.37; P=5.19 × 10^{-20}), which is greater than that afforded to genes simply implicated by GWAS (r^2>0.9 targets (OR=1.44; CI=1.14-1.81; P=0.003).

We employed a number of methods to map target genes to cancer risk SNPs, incorporating LD blocks, TAD regions and gene expression. We found that genes implicated by LD r^2>0.9 method showed the greatest enrichment for drug approval. While compatible with the functional basis of many GWAS associations being due to the most proximal gene(s), this does not preclude the possibility of longer-range tissue-specific mechanisms that are less amenable to detection by our approach. Therefore future endeavours of this kind will likely benefit from more detailed experimental investigation of the biological mechanism underpinning cancer risk loci. While the TAD-based strategy is likely to be always beset by the issue of capturing too many genes, strategies based on integration of GWAS and multi-omics as per DEPICT[13] are likely to improve making them attractive sources of genetic information. To investigate regulatory interactions across all cancer risk loci we made use of publicly available Hi-C data from human embryonic stem-cells, noting the observation of Dixon *et al.,* 2012[25] that TAD boundaries are relatively stable across cell types. However, the increasing availability of tissue- and cancer-specific Hi-C data is likely to improve efforts to identify target genes of specific cancer risk regions.

In concert with our primary analysis we identified a number of possible opportunities for drug re-purposing, informed by cancer germline genetics. These extend the potential of pre-existing therapies and highlight that pathways subverted by cancers may also be altered in other diseases.

For pragmatic purposes we considered all cancers assuming generic effects exist at least across some cancers in order to maximise study power. We do however acknowledge that this is in essence crude since certain cancer subtypes can show specific associations with risk SNPs, reflective of differences in their biology. For example, ER-positive and negative breast cancers[26,27] as well as combinations of 1p/19q co-deletion, *TERT* promoter and IDH mutation in glioma[28,29]. The future availability of larger datasets which will afford the identification of additional risk SNPs will open up the possibility of fine-tuned analyses. In addition we make the assumption that cancer risk variants act directly to influence cancer initiation or progression. However, this does not preclude the existence of a limited subset which may have indirect mechanisms, such as at 15q25.1 where the association with lung cancer is likely due to smoking[30].

One caveat to using all forms of germline genetics as a mechanism for prioritisation of drug development is the assumption that susceptibility *per se* is also reflective of progression, which may not always be the case. As with other studies, we have used drug approval as a surrogate for drug efficacy. This assumption will only however serve to make our estimates conservative. We additionally acknowledge our lack of inclusion of generic drugs, however as the vast majority of these have a broad range of targets we do not regard this as significantly impacting our findings. Considering the extent to which cancer genes implicated by GWAS that are not currently in the drug development pipeline might represent good candidates we performed multi-faceted druggability analyses incorporating assessments of the 3D structures of the target protein and any associated protein complexes, chemical properties of known ligands of the target, and the target's position and role within the human interactome. Ranking target-indication pairings by criteria including novelty relative to existing targets and predicted attrition risk (**Supplementary Tables 5 and 6**). Of 1,292 genes annotated to GWAS SNPs by r^2>0.9; 977, 486 and 1,287 had druggability assessments by network, structure and ligand-based prediction respectively. Of note is the observation that 29 of these can be targeted by existing high-quality probes and thus represent good candidates for being prioritised in for future studies.

In conclusion, we have demonstrated enrichment for targets implicated by cancer risk variants being more successful in the drug development pipeline, providing a rationale for germline genetics empowering cancer drug discovery. Mapping approved drug targets back to cancer GWAS signals enables identification of both novel drug targets and patient populations. To benefit the wider community the cancer germline information used in this study is available at https://cansar.icr.ac.uk. Collectively our findings show the value of incorporating information from germline cancer genetics as part of interdisciplinary, data-driven approaches to inform drug discovery in oncology.

ACKNOWLEDGEMENTS AND FUNDING

This project was funded by a Team Science project grant from the Institute of Cancer Research. Additional support from Cancer Research UK is acknowledged. A.S. is supported by charitable funds from the Royal Marsden Hospital. We are grateful to John Whittaker of GlaxoSmithKline for helpful comments.

Conflict of interest statement

The authors have no conflicts of interest to disclose.

Author contributions

BK, AS, RSH and EC conceived the study; BK, AS, and EC carried out statistical and bioinformatics analyses; all authors contributed to the final manuscript.

REFERENCES

1. Nelson MR, Tipney H, Painter JL, et al: The support of human genetic evidence for approved drug indications. Nat Genet 47:856-60, 2015
2. Shih HP, Zhang X, Aronov AM: Drug discovery effectiveness from the standpoint of therapeutic mechanisms and indications. Nat Rev Drug Discov 17:78, 2018
3. Workman P, Draetta GF, Schellens JHM, et al: How Much Longer Will We Put Up With $100,000 Cancer Drugs? Cell 168:579-583, 2017
4. Sud A, Kinnersley B, Houlston RS: Genome-wide association studies of cancer: current insights and future perspectives. Nat Rev Cancer 17:692-704, 2017
5. Taylor F, Huffman MD, Macedo AF, et al: Statins for the primary prevention of cardiovascular disease. Cochrane Database Syst Rev:CD004816, 2013
6. Feagan BG, Sandborn WJ, Gasink C, et al: Ustekinumab as Induction and Maintenance Therapy for Crohn's Disease. N Engl J Med 375:1946-1960, 2016
7. Welter D, MacArthur J, Morales J, et al: The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42:D1001-6, 2014
8. Finan C, Gaulton A, Kruger FA, et al: The druggable genome and support for target identification and validation in drug development. Sci Transl Med 9, 2017
9. Machiela MJ, Chanock SJ: LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics 31:3555-7, 2015
10. Schmitt AD, Hu M, Jung I, et al: A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Rep 17:2042-2059, 2016
11. Dixon JR, Jung I, Selvaraj S, et al: Chromatin architecture reorganization during stem cell differentiation. Nature 518:331-6, 2015
12. Crane E, Bian Q, McCord RP, et al: Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature 523:240-4, 2015
13. Pers TH, Karjalainen JM, Chan Y, et al: Biological interpretation of genome-wide association studies using predicted gene functions. Nat Commun 6:5890, 2015
14. Futreal PA, Coin L, Marshall M, et al: A census of human cancer genes. Nat Rev Cancer 4:177-83, 2004
15. Santos R, Ursu O, Gaulton A, et al: A comprehensive map of molecular drug targets. Nat Rev Drug Discov 16:19-34, 2017
16. Tym JE, Mitsopoulos C, Coker EA, et al: canSAR: an updated cancer research and drug discovery knowledgebase. Nucleic Acids Res 44:D938-43, 2016
17. Mitsopoulos C, Schierz AC, Workman P, et al: Distinctive Behaviors of Druggable Proteins in Cellular Networks. PLoS Comput Biol 11:e1004597, 2015
18. Albert A. Antolin JET, Angeliki Komianou, Ian Collins, Paul Workman, Bissan Al-Lazikani: Objective, Quantitative, Data-Driven Assessment of Chemical Probes. biorxiv, 2017
19. Zhang B, Jia WH, Matsuda K, et al: Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. Nat Genet 46:533-42, 2014
20. Al Olama AA, Kote-Jarai Z, Berndt SI, et al: A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. Nat Genet 46:1103-9, 2014
21. Bishop DT, Demenais F, Iles MM, et al: Genome-wide association study identifies three loci associated with melanoma risk. Nat Genet 41:920-5, 2009

22. Asgari MM, Wang W, Ioannidis NM, et al: Identification of Susceptibility Loci for Cutaneous Squamous Cell Carcinoma. J Invest Dermatol 136:930-7, 2016

23. Chahal HS, Wu W, Ransohoff KJ, et al: Genome-wide association study identifies 14 novel risk alleles associated with basal cell carcinoma. Nat Commun 7:12510, 2016

24. Di Bernardo MC, Crowther-Swanepoel D, Broderick P, et al: A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. Nat Genet 40:1204-10, 2008

25. Dixon JR, Selvaraj S, Yue F, et al: Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485:376-80, 2012

26. Michailidou K, Lindstrom S, Dennis J, et al: Association analysis identifies 65 new breast cancer risk loci. Nature 551:92-94, 2017

27. Milne RL, Kuchenbaecker KB, Michailidou K, et al: Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nat Genet 49:1767-1778, 2017

28. Labreche K, Kinnersley B, Berzero G, et al: Diffuse gliomas classified by 1p/19q co-deletion, TERT promoter and IDH mutation status are associated with specific genetic risk loci. Acta Neuropathol, 2018

29. Eckel-Passow JE, Lachance DH, Molinaro AM, et al: Glioma Groups Based on 1p/19q, IDH, and TERT Promoter Mutations in Tumors. N Engl J Med 372:2499-508, 2015

30. Amos CI, Wu X, Broderick P, et al: Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat Genet 40:616-22, 2008

FIGURE AND SUPPLEMENTARY TABLE LEGENDS

Figure 1: Summary of analytical strategy.

Figure 2: Enrichment of approved cancer drug targets incorporating genetic evidence relative to pre-clinical (A) and phase II (B) targets. TAD, topologically associating domain; N, number; DEPICT, Data-driven expressed prioritised integration for complex traits. Data based on Tables 1-2.

Figure 3: Integration of cancer germline genetics information into canSAR. Available germline genetic evidence can be searched for by target on cansar.icr.ac.uk or directly at https://cansar.icr.ac.uk/cansar/molecular-targets/P23458/germline_genetics/ where P23458 is the uniprot identifier for the target of interest.

Supplementary Table 1: Cancer risk SNPs.

Supplementary Table 2: Mapping cancer risk SNPs to gene transcripts by LD- and TAD-based approaches.

Supplementary Table 3: DEPICT gene prioritization of cancer risk SNPs

Supplementary Table 4: Number of unique genes targeted by cancer therapies at the last recorded stage of development.

Supplementary Table 5: Summary of canSAR druggability assessments of target genes implicated by cancer germline genetics. Druggability assessments were obtained from canSAR (https://cansar.icr.ac.uk/) and high-quality probe annotations obtained from ProbeMiner (http://probeminer.icr.ac.uk/).

Supplementary Table 6: CanSAR druggability assessments for target genes implicated by cancer germline genetics.

	Pre-clinical drug				**Enrichment (pre-clinical vs approved)**	
Method	**Approved with genetic support (N)**	**Not approved with genetic support (N)**	**Approved with no genetic support (N)**	**Not approved with no genetic support (N)**	***P*-value**	**OR (95% CI)**
TAD	49	264	66	457	0.22	1.28 (0.86-1.92)
r^2 > 0.1	36	186	79	535	0.21	1.31 (0.85-2.01)
r^2 > 0.2	28	141	87	580	0.26	1.32 (0.83-2.11)
r^2 > 0.3	26	125	89	596	0.19	1.39 (0.86-2.25)
r^2 > 0.4	24	109	91	612	0.13	1.48 (0.90-2.42)
r^2 > 0.5	22	92	93	629	0.08	1.62 (0.97-2.70)
r^2 > 0.6	21	89	94	632	0.10	1.59 (0.94-2.67)
r^2 > 0.7	19	77	96	644	0.08	1.65 (0.96-2.86)
r^2 > 0.8	19	69	96	652	0.03	1.87 (1.08-3.25)
r^2 > 0.9	19	65	96	656	0.02	2.00 (1.14-3.48)
COSMIC Germline	15	16	100	705	2.23×10^{-6}	6.61 (3.17-13.78)
DEPICT	8	33	107	688	0.25	1.56 (0.70-3.46)

Table 1: Enrichment of approved cancer drug targets supported by genetic evidence relative to pre-clinical targets. OR, odds ratio; CI, confidence interval; TAD, topologically associating domain; N, number; DEPICT, Data-driven expressed prioritised integration for complex traits. Enrichment was calculated by constructing a two by two table of genes and counts corresponding to whether a gene product has genetic support at the respective stages of drug development (*i.e.* approved compared with pre-clinical). Test of association was Fisher's exact test, with the Wald test used to quantity effect size and 95% confidence intervals. A *P*-value of 0.05 (two-sided) was considered as statistically significant.

	Phase II drug				Enrichment (phase II vs approved)	
Method	**Approved with genetic support (N)**	**Not approved with genetic support (N)**	**Approved with no genetic support (N)**	**Not approved with no genetic support (N)**	***P*-value**	**OR (95% CI)**
TAD	49	100	66	213	0.04	1.58 (1.02-2.45)
r^2 > 0.1	36	72	79	241	0.10	1.53 (0.95-2.45)
r^2 > 0.2	28	50	87	263	0.07	1.69 (1.00-2.85)
r^2 > 0.3	24	46	89	267	0.06	1.70 (0.99-2.90)
r^2 > 0.4	22	40	91	273	0.05	1.80 (1.03-3.15)
r^2 > 0.5	21	32	93	281	0.02	2.08 (1.15-3.75)
r^2 > 0.6	19	30	94	283	0.02	2.11 (1.15-3.86)
r^2 > 0.7	19	26	96	287	0.02	2.18 (1.16-4.12)
r^2 > 0.8	19	23	96	290	9.2×10^{-3}	2.50 (1.30-4.78)
r^2 > 0.9	19	21	96	292	4.2×10^{-3}	2.75 (1.42-5.35)
COSMIC Germline	15	8	100	305	8.41×10^{-5}	5.72 (2.35-13.89)
DEPICT	8	9	107	304	0.09	2.52 (0.95-6.71)

Table 2: Enrichment of approved cancer drug targets supported by genetic evidence relative to phase II targets. OR, odds ratio; CI, confidence interval; TAD, topologically associating domain; N, number; DEPICT, Data-driven expressed prioritised integration for complex traits. Enrichment was calculated by constructing a two by two table of genes and counts corresponding to whether a gene product has genetic support at the respective stages of drug development (*i.e.* approved compared with phase II). Test of association was Fisher's exact test, with the Wald test used to quantity effect size and 95% confidence intervals. A *P*-value of 0.05 (two-sided) was considered as statistically significant.

Gene	Entrez	Additional genes targeted by drug	Disease/s	Locus	Cancer type
ALOX5	240		Asthma, Chronic obstructive pulmonary disease	10q11.21	Prostate Cancer
CFTR	1080		Cystic fibrosis, Diarrhoea, short-bowel syndrome, Irritable bowel syndrome, diarrhoea-predominant infection, GI tract infection, HSV infection, HIV/AIDS	7q31.2	Barrett's esophagus/esophageal adenocarcinoma
CLCN2	1181		Chronic constipation, Irritable bowel syndrome, GI motility dysfunction,	3q27.1	Esophageal adenocarcinoma
CRHR1	1394	CRHR2 (1395)	Anxiety, unspecified insomnia	17q21.31	Ovarian cancer in BRCA1 carriers
DDC	1644		Parkinson's disease	7p12.1	Childhood ALL
GABBR1	2550		Spasticity, Multiple sclerosis, Alcohol addiction, Cerebral palsy, Spinal cord injury, Dystonia	6p22.1	Barrett's esophagus/esophageal adenocarcinoma
GBA	2629		Gaucher's disease	1q22	Gastric adenocarcinoma
INSR	3643		Diabetes Type 1, Diabetes Type 2	19p13.2	Renal Cell Carcinoma/Differentiated Thyroid Cancer
PDE4D	5144	PDE4A (5141)/PDE4B (5142)/PDE4C (5143)	COPD, Asthma, Non-alcoholic steatohepatitis, Eczma, Alzheimer's disease, Schizophrenia, Rhinitis, Psoriasis	5q12.1	Esophageal cancer/Breast cancer
PLG	5340		Venous thrombosis, Myocardial infarction, Pulmonary thrombosis	6q26	Prostate Cancer
PTGIR	5739		Pulmonary hypertension, Transplant rejection, Peripheral vascular disease, Limb ischaemia	19q13.32	CLL
SLC6A3	6531		Depression, CNS diagnosis, ADHD	5p15.33	Pancreatic cancer
TGFB1	7040		Wound healing, conjunctivitis, Asthma, Eczema, Rhinitis, Rheumatoid arthritis, Hyperuricaemia, Multiple Sclerosis, Restenosis	19q13.2	Colorectal cancer
TYR	7299		Skin disorder	11q14.3	Melanoma/Squamous cell carcinoma/Basal cell carcinoma
VDR	7421		Osteoperosis, Keratosis, Secondary hyperparathyroidism, Psoriasis, Osteodystrophy, Hypophosphataemia, Palmoplantar pustulosis, Ichthyosis	12q13.11	Prostate Cancer

Table 3: Opportunities for drug re-purposing informed by germline cancer genetics. Targets annotated to cancer risk SNPs by r^2>0.8 were assessed for overlap with approved non-oncology drugs.