Opportunities for precision oncology revealed by whole genome sequencing 10,478 cancers - the UK 100,000 Genomes Project

Ben Kinnersley[*1,2], Amit Sud[*1,3], Alex J. Cornish[1], Daniel Chubb[1], Richard Culliford[1], Andrew Everall[1], Andreas Gruber[4], Adrian Lärkeryd[5], Costas Mitsopoulos[6], Genomics England Research Consortium[7], David Wedge[8], Richard Houlston[1]

1. Division of Genetics and Epidemiology, The Institute of Cancer Research, Sutton, Surrey, SM2 NG, UK.
2. University College London Cancer Institute, University College London, London, UK.
3. Haemato-oncology Unit, The Royal Marsden Hospital NHS Foundation Trust, Sutton, UK.
4. Systems Biology & Biomedical Data Science Laboratory, University of Konstanz, Germany.
5. Division of Molecular Pathology, The Institute of Cancer Research, Sutton, Surrey, SM2 NG, UK.
6. Division of Cancer Therapeutics, The Institute of Cancer Research, Sutton, Surrey, SM2 NG, UK.
7. Genomics England Research Consortium, Genomics Unit, NHS England and NHS Improvement, London, UK.
8. Manchester Cancer Research Centre, University of Manchester, Manchester, UK.

*contributed equally

Correspondence to: Richard Houlston; +44(0) 208 722 4175; e-mail: richard.houlston@icr.ac.uk

Running title: 100,000 Genomes Project to Inform Precision Oncology

Key words: Precision oncology, cancer driver, mutation, therapy

Word count: 2,271 (2,804 including methods)

ABSTRACT

Identifying cancer driver genes is key for delivering the vision of precision oncology. The falling cost makes whole genome sequencing (WGS) attractive as a single all-encompassing test to identify uncommon cancer drivers in a patient, which may not be captured by panel testing but are potentially targeted by small molecules. We analysed 10,478 patient tumour genomes spanning 35 cancer types recruited to the UK 100,000 Genomes Project. We identified 330 driver genes, including 74 which are novel to any cancer. Across all cancer types 16% of the patients are eligible for a currently approved therapy. Computational chemogenomic analysis of cancer mutations identified 96 additional targets of compounds that are potentially active and represent candidates for future clinical trials, expanding opportunities for improved patient care.

INTRODUCTION

The rationale for the one-size-fits-all medical treatment model is being challenged with the move towards individualised therapy[1]. This is epitomised in oncology where standard therapies are ineffective in around 75% of patients, representing one of the highest therapy failure rates in all diseases[2].

Precision oncology describes a set of strategies tailored to the unique biology of a patient's disease. The potential promise of this approach includes improved treatment efficacy, more favourable toxicity profiles and a reduction in the administration of ineffective treatments[3]. Underpinning precision oncology is the concept of somatic driver mutations as the foundation of cancer development[4]. There are already a number of approved therapies for tumours with specific "actionable" driver mutations, with additional ones in development[5]. Knowledge of the actionable driver mutational landscape in cancers has recently become central to delivering precision oncology.

Currently multiple standalone tests or a panel are typically used to capture a set of genomic features for a given tumour type. However, falling costs make whole genome sequencing (WGS) a potentially attractive proposition as a single all-encompassing test[6]. Moreover, it provides the opportunity to identify additional cancer drivers in a patient which may not be captured by panel testing but are potentially targeted by small molecule therapies. To examine this proposition we analysed WGS data from 10,478 cancer patients recruited to the Genomics England (GEL) 100,000 Genomes Project (100kGP) (**Fig 1A**)[7].

METHODS

The GEL cohort

We restricted our analysis to high-quality data derived from PCR-free, flash-frozen primary tumour samples (v8 data release) from 10,470 individuals (**Supplementary Tables 1 and 2**). In addition to using variant calls from 100kGP we removed alignment bias introduced by Isaac soft-clipping[8]. This yielded 285,233,025 somatic single base substitutions (SBS), 2,272,616 doublet base substitutions (DBS), 142,764,499 small insertions and deletions (ID). For this evaluation, we restricted our analysis to high-quality data derived from PCR-free, flash-frozen tumour samples from 10,478 tumours (34 bile duct, 305 bladder, 2,306 breast, 2,324 colorectal, 440 central nervous system, 91 esophageal, 201 head and neck, 1,045 renal cell, 24 liver, 1,466 lung, 35 mesothelioma , 607 soft-tissue, 454 ovarian, 94 pancreas, 366 prostate, 270 melanoma, 72 gastric, 51 testicular, 649 uterus) (**Supplementary Tables 1 and 2**). Complete details on sample curation, WGS, somatic variant calling, mutation annotation and power calculations are provided in the **Supplementary Methods**.

Identification and actionability of driver genes

Cancer driver genes were identified using IntOGen[9], which combines seven computational methods to detect signals of positive selection from the mutational patterns of genes in each cancer type (**Supplementary Methods**). Details of pre-processing of mutations, combining driver gene identification methodologies, post-processing and annotation of driver gene mutations are provided in the **Supplementary Methods**. Information on the clinical actionability of cancer drivers were retrieved from OncoKB[5] and COSMIC[10].

Mutation signature analysis

Tumours with microsatellite instability (MSI) were identified using MSINGS[11] and homologous recombination deficiency was assessed using HRDetect[12].

Chemogenomic annotation of cancer networks

To construct networks for each cancer type, we used protein products of the cancer driver genes to seed a search for all interacting proteins in the canSAR interactome[13], which is

based on information from 8 databases, including the IMeX consortium[14], Phosphosite[15], and key publications. The canSAR interactome features interactions where there are: (i) ≥2 publications with experimental evidence of binary interaction between the two proteins; (ii) 3D protein evidence of a complex; (iii) ≥2 reports that one protein is a substrate of the other; (iv) ≥2 publications reporting that one protein is the product of a gene under the direct regulatory control of the other. Each tumour-specific interactome was seeded using cancer driver proteins retrieving interacting proteins that had supporting experimental evidence. To ensure only additional proteins are likely to function primarily through interaction with proteins in the network we adopted the following strategy: Starting with the input list of proteins we obtained all possible first neighbours. We then computed, for each new protein, the proportion of its first neighbours in the original input list. To define proteins likely to function through the network, we calculated the probability of these occurring randomly, by permuting the interactome 10,000 times. We corrected empiric *P*-values for multiple testing retaining only proteins having a FDR < 0.05. For each cancer type we minimised the network by retaining only proteins connected to more than one cancer protein, or whose only connection was to a cancer-specific protein. We then annotated proteins with pharmacological and druggability data using canSAR's Cancer Protein Annotation Tool (CPAT). Essential and selective genes including lineage specificity were ascertained from the ShinyDepMap analysis[16] .

RESULTS

We analysed genomes from 10,478 cancers comprising 35 different cancer types (**Fig. 1B** and **Supplementary Table 2**) including colorectal cancer (n=2,324), invasive ductal carcinoma of the breast (n=2,001), clear cell renal cell carcinoma (n=872) and lung adenocarcinoma (n=677), as well as rare tumour types such as hepatocellular (n=24) and large-cell lung cancer (n=25). Mutation rates varied across the different cancer types with cutaneous melanoma possessing the highest median SNV mutation count and meningioma possessing the lowest median SNV mutation count (**Supplementary Fig. 1)**. Furthermore, 945 samples, notably colorectal and uterine cancers, were hypermutated, either as result of deficient mismatch repair (dMMR) or *POLE* mutation.

Invasive ductal carcinoma of the breast had the highest power for driver gene detection (>90% power to detect driver genes with a non-silent mutation rate 2% higher than the background) and LCLC had the lowest power to detect driver genes (<90% power to detect driver genes with a non-silent mutation rate 10% higher than background) (**Fig. 1C** and **Supplementary Table 3**). Compared with the recent PCAWG pan-cancer genome analysis[17], for 19 cancer types, there was an increase in power for non-silent mutation detection. Notably, there was a >10-fold increase in sample size in breast, colorectal, oesophageal and uterine cancer, lung adenocarcinoma and bladder transitional cell carcinoma.

Spectrum of cancer driver genes

Across cancer types we identified 770 unique tumour-driver gene pairs corresponding to 330 unique cancer driver genes; 246 high confidence genes and an additional 84 genes recovered after post-processing (**Fig. 2A** and **Supplementary Table 4**). When compared to the largest pan-cancer driver analysis, in 21 of 31 cancer types with matching tumour histologies we recovered 61% of all cancer drivers reported by COSMIC, IntOGen and the TCGA pan-cancer analysis reported in Bailey et al., 2018 (**Supplementary Table 4**)[18]. We were able to detect 80% of drivers reported for colorectal, breast, lung and ovarian cancers but <20% of drivers for hepatocellular and stomach cancers, which is likely a result of differing sample size. The number of identified cancer driver genes varied between cancer types, with colorectal and uterine cancers having the most (60 genes) and spindle cell

carcinoma having the fewest (4 genes). Across the 35 cancers, we found no correlation between average mutation burden and the number of driver genes in each cancer (Pearson R=0.19, *P*-value=0.27). The consensus list also includes 330 tumour-driver pairs that have not previously been reported by either the Cancer Gene Census (CGC), Intogen or the PanCancerSoftware analysis of TCGA[18] (**Supplementary Table 4**), and 74 that have not previously been associated with any specific tissue. Almost all of the novel drivers identified were uncommon with 65/74 (88%) possessing a mutated frequency <10% in each cancer type. We observed the highest number of new cancer driver genes for uterine (n=42), bladder (n=40) and colorectal (n=37) cancers. Predictions of known cancer driver genes in new cancer types include *SPTA1*, *CHD4* and *ASXL1* in colorectal cancer, *FOXO3*, *MUC16* and *ZFPM1* in breast cancer and *CNTNAP2*, *CTNND2* and *TRRAP* in lung adenocarcinoma. Entirely novel predictions include *MAP3K21* in colorectal cancer, *USP17L22* in breast ductal carcinoma, and *TPTE* in lung adenocarcinoma (**Supplementary Table 4**).

Considering the prevalence of driver genes across cancer types, some genes were seen to act as drivers across multiple cancer types, while others tended to be more specific. Eighty-five genes were identified as a driver in more than two tumour types, with 26 genes functioning as drivers in more than five tumour types (**Figure 2B**). As expected, *TP53* was identified as a driver gene in the largest number of tumour types, followed by *PIK3CA*, *ARID1A* and *PTEN*, acting as cancer driver genes in 29, 18, 16 and 14 different tumour types respectively. While many genes function as drivers in multiple cancer types, some drivers are mutated at high frequencies only in specific tumours, such as *VHL* in clear cell renal cell carcinoma and *FGFR3* in bladder cancer (**Figure 2B**). Hierarchical clustering of cancers based on the presence of identified driver mutations and their respective *Q*-value demonstrated clustering of cancer types by cell of origin (*e.g.* head and neck and lung squamous cell carcinoma) and by organ (*e.g.* breast ductal and lobular carcinomas) (**Supplementary Fig. 2**). The ratio of predicted activating to tumour suppressor driver genes varied across tumour types with meningioma and myxofibrosarcoma possessing the highest and lowest ratio respectively (**Figure 2C** and **Supplementary Table 4**).

Across the 35 different tumour types we identified 12,606 distinct oncogenic mutations in tumour-relevant cancer driver genes, in 9,070 unique samples. The median number of

oncogenic mutations in cancer driver genes per sample was 2 across all tumours. The highest median number of oncogenic mutations in cancer driver genes per sample was seen in uterine cancer (**Supplementary Fig. 3**). We observed significant differences in oncogenic mutation frequency in cancer driver genes across different tumour histologies arising from the same organ. Examples include *CDH1*, *TBX3* and *TP53* in breast cancers, *ATRX*, *CIC*, *IDH1*, *PTEN* and *TP53* in central nervous system tumours, *IDH1* and *TP53* in connective tissue tumours, *PBRM1* and *VHL* in renal cancers and *EGFR*, *KMT2D*, *KRAS*, *NFE2L2*, *PTEN*, *STK11* and *TP53* in lung cancer (**Figure 2D)**.

Clinical implications of genomic features

Systematic analyses of cancer genomes provide an opportunity of estimating the number of individuals eligible for a targeted therapy and identify potentially novel therapeutic interventions. We first used two different databases to evaluate the therapeutic implications of the genetic events across this cohort: Precision Oncology Knowledge Base (OncoKB) and the COSMIC Mutation Actionability in Precision Oncology Product[5,10]. Both databases catalogue approved marketed drugs having demonstrated efficacy in tumours with specified driver gene mutations, based on clinical trials and published clinical evidence. OncoKB also provides compelling biological evidence supporting the cancer driver gene as being predictive of a response to a given drug.

We observed that both the fraction of samples and proportion of alteration types varied across tissue types. Data from COSMIC indicates that 85% of all samples (8,874/10,478) possess at least one putatively actionable alteration being targeted in a clinical setting (**Fig. 3A** and **Supplementary Table 5**), while 55% of samples (5,805/10,478) had at least one putatively actionable or biologically relevant alteration from OncoKB (**Fig. 3B** and **Supplementary Table 6**). Across all cancer types, 16% (1,633/10,470) of the patients would be eligible for a currently approved therapy as defined by OncoKB.

The most common putatively actionable alterations across all of the 35 cancer types were mutations in *PIK3CA*, *KRAS* and *PTEN*. Specific oncogenic missense mutations in *PIK3CA* are present in 40% of lobular breast cancers and 30% of ductal breast cancers and their presence are an indication for the use of PI3Kα- inhibitor Alpelisib[19]. These mutations are

present in a number of cancers including colorectal (20%) and uterine cancers (47%) and are subject to early clinical studies with an allosteric inhibitor of PI3Kα. We found high fractions of patients with pancreatic, colorectal cancers and lung adenocarcinoma with actionable *KRAS* mutations (between 34% and 69% of all cases). The G12C mutation was present in 17% of lung adenocarcinoma cases and is targeted by mutation specific selective covalent inhibition with Adagrasib or Sotorasib[20,21]. PI3Kβ inhibition is of significant biological interest in patients with oncogenic *PTEN* mutations as PI3Kβ is thought to drive cellular proliferation in these tumours. These PTEN mutations were prevalent in melanoma (10%), hepatocellular carcinoma (13%), squamous cell carcinoma of the lung (15%), glioblastoma multiforme (29%) and uterine carcinoma (66%) and their presence would result in eligibility for early studies of PI3Kβ inhibition[22].

319 tumours possessed a HRD mutational signature, providing support for the use of a PARP inhibitor in these individuals. Furthermore, 1,309 tumours possessed a high coding tumour mutational burden (≥10 mutations/megabase[5,23]) and 144 cancers had evidence of defective mismatch repair. Considering these collectively would suggest that 1,312 patients may be eligible for checkpoint inhibition.

Expanding the druggable cancer genome

An opportunity emerging from the systematic analysis of cancer genomes is the identification of novel therapeutic intervention strategies. Of the 330 cancer driver genes identified in this study, 261/330 (79%) are not currently identified as targets in either COSMIC or OncoKB databases. As a means of triaging these genes as candidates for therapeutic intervention, we assessed the essentiality and selectivity of driver genes and their druggability using RNAi/CRISPR DepMap data and the integrative cancer-focused knowledgebase, canSAR respectively. We found 96/261 (37%) of these genes are predicted to be commonly essential and of these 12/96 (13%) have a chemical probe available and 35/96 (36%) have a ligandable 3-dimensional structure (**Supplementary Table 7**).

Motivated by the observation that targeting proteins that interact with cancer driver genes can result in successful precision oncology strategies, we sought to expand the network of druggable targets in cancer. To this end, we used canSAR to map and pharmacologically

annotate networks of the cancer genes identified for each tumour type. Specifically, we seeded networks with driver genes identified in each tumour group and used transcriptional and curated protein-protein interactions to recover a refined cancer specific-network of proteins, each protein being annotated based on multiple assessments of 'druggability', *i.e*. the likelihood of the protein being amenable to small molecule drug intervention. After seeding each cancer specific network with their respective drivers, we yielded a total of 631 distinct proteins across all cancers (**Supplementary Table 8**). The median number of unique proteins in each network across all cohorts was 57 with colorectal cancer possessing the largest network (n=231) and spindle cell carcinoma (n=10) possessing the smallest network. As expected there was a correlation between network size and number of identified drivers for each cancer type (Pearson R = 0.9, P=1.23×10^{-9}).

Of these 631 proteins, 58% (n=369) were retrieved solely through network analysis of which the majority (n=323) were novel to any of the cancer types (hereon referred to as cancer-network proteins). Notable examples include *HDAC1*, *CDK2* and *CDK1* which were present in 31, 29 and 28 cohorts respectively. We observed 70% (n=225) of these cancer-network proteins as being targetable by existing approved or investigational therapies with notable examples including *BCL2* and *BTK*. Of the remaining 97 genes, 34 are commonly essential, 11 possess concordant lineage specificity, 48 are ligandable by 3D structure and 11 have an existing high quality probe available (**Supplementary Table 9**). Collectively these data provide potential future opportunities for therapy for a number of cancers.

DISCUSSION

Delivering precision oncology to all patients is partly constrained because routine patient tumours are generally only testing for a restricted set of common actionable mutations. Herein we have analysed WGS data on 10,470 patients recruited to the 100kGP study to explore the value of WGS to inform patient care. One of the main aims of the 100,000 Genomes Project is to improve cancer care for NHS patients through personalised medicine by implementing WGS as part of routine care. Although we primarily focussed on point mutations and small indels we identified 330 cancer driver genes, 74 of which are novel to a cancer type. Importantly, many recurring mutations are highly frequent across different tumour types. If clinically translated, these observations suggest approximately 55-85% of patients harbour an actionable mutation, either in terms of predicting sensitivity to certain treatments or clinical trial eligibility. This contrasts to around 22-36% achievable if based on many routine small variant testing panels in widespread usage[24]. This does however assume that the use of approved drugs as a proxy for effective cancer therapies. However, a recent comprehensive analysis of cancer drug approvals by the U.S. Food and Drug Administration has concluded that new cancer drug approvals reduce the risk of death and tumour progression[25].

To inform potential future therapeutic opportunities, we applied established chemogenomic technologies to map and pharmacologically annotate the cellular network of cancer genes identified by WGS. Based on broader criteria, we estimate that 55-85% of the cases harbour at least one potentially clinically actionable target thus supporting the assertion that increased numbers of patients will benefit from wider and more intensive genomic sequencing in line with the aspirations of the 100kGP[7].

ACKNOWLEDGEMENTS

Funding was provided by the Wellcome Trust (214388), Cancer Research UK (C1298/A8362) and the Medical Research Council. A.S. is in receipt of a National Institute for Health Research (NIHR) Academic Clinical Lectureship, funding from the Royal Marsden Biomedical Research Centre, a starter grant for clinical lecturers from the Academy of Medical Sciences, and is recipient of the Whitney-Wood Scholarship from the Royal College of Physicians. This is a summary of independent research supported by the NIHR Biomedical Research Centre at the Royal Marsden NHS Foundation Trust and the Institute of Cancer Research.

This research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support.

GENOMICS ENGLAND RESEARCH CONSORTIUM

Ambrose, J. C.[1]; Arumugam, P.[1]; Bevers, R.[1]; Bleda, M.[1]; Boardman-Pretty, F. [1,2]; Boustred, C. R.[1]; Brittain, H.[1] ; Brown, M.A.[1]; Caulfield, M. J.[1,2]; Chan, G. C.[1]; Giess A.[1]; Griffin, J. N.[1]; Hamblin, A.[1]; Henderson, S.[1,2]; Hubbard, T. J. P. [1]; Jackson, R.[1] ; Jones, L. J.[1,2]; Kasperaviciute, D.[1,2]; Kayikci, M.[1]; Kousathanas, A.[1]; Lahnstein, L.[1]; Lakey, A.[1]; Leigh, S. E. A.[1]; Leong, I. U. S.[1]; Lopez, F. J.[1]; Maleady-Crowe, F.[1]; McEntagart, M.[1]; Minneci F.[1]; Mitchell, J.[1]; Moutsianas, L.[1,2]; Mueller, M.[1,2]; Murugaesu, N.[1]; Need, A. C.[1,2]; O'Donovan P.[1]; Odhams, C. A.[1]; Patch, C.[1,2]; Perez-Gil, D.[1]; Pereira, M. B.[1]; Pullinger, J.[1]; Rahim, T.[1]; Rendon, A.[1]; Rogers, T.[1]; Savage, K.[1]; Sawant, K.[1]; Scott, R. H.[1]; Siddiq, A.[1]; Sieghart, A.[1]; Smith, S. C.[1]; Sosinsky, A.[1,2]; Stuckey, A.[1]; Tanguy M.[1]; Taylor Tavares, A. L.[1]; Thomas, E. R. A.[1,2]; Thompson, S. R.[1]; Tucci, A.[1,2]; Welland, M. J.[1]; Williams, E.[1]; Witkowska, K.[1,2]; Wood, S. M.[1,2]; Zarowiecki, M.[1].

1. Genomics England, London, UK.

2. William Harvey Research Institute, Queen Mary University of London, London, UK.

AUTHOR CONTRIBUTIONS

B.K, A.S, and R.S.H designed the study. B.K., A.S., A.C, D.C performed sample curation, B.K., A.S., A.C, D.C, R.C, A.E, A. G, A.L, K. M, D. W, performed bioinformatic and statistical analysis. B.K. A.S., and R.S.H. drafted the manuscript; all authors reviewed, read, and approved the final manuscript.

CONFLICTS OF INTERESTS

The authors declare no competing financial interests.

DATA AVAILABILITY

The whole genome sequencing data in this study are available within the National Genomics Research Library, a secure cloud workspace. To access genomic and clinical data within this Research Environment, researchers apply to become a member of either the Genomics England Clinical Interpretation Partnership (https://www.genomicsengland.co.uk/research/academic) or the Discovery Forum (industry partners) https://www.genomicsengland.co.uk/research/research-environment. The COSMIC and OncoKB clinical actionability data are available from https://cancer.sanger.ac.uk/actionability and https://www.oncokb.org/actionableGenes#sections=Tx, respectively. The canSAR chemogenomics data are available from https://cansar.ai/.

CODE AVAILABILITY

Details and code for using the Intogen framework are available here (https://intogen.readthedocs.io/en/latest/index.html).

FIGURE LEGENDS

Figure 1. (A) Study design; (B) Number of samples per tumour type; (C) Power estimates for driver gene identification per tumour type. The number of samples needed to achieve 90% power for 90% of genes (*y* axis). Grey vertical lines indicate exome-wide background mutation rates (*x* axis). Black dots indicate sample sizes in the current study. For most tumour types, the current sample size is inadequate to reliably detect genes mutated at 5% or less above background. BileDuct-AdenoCA, cholangiocarcinoma; Bladder-TCC, bladder transitional cell carcinoma; Breast-DuctalCA, breast ductal carcinoma; Breast-LobularCA, breast lobular carcinoma; CNS-Astro, astrocytoma; CNS-GBM-IDHmut, IDH mutated glioblastoma; CNS-GBM-IDHwt, IDH wild type glioblastoma; CNS-Menin, meningioma; CNS-Oligo, oligodendroglioma; ColoRect-AdenoCA, colorectal adenocarcinoma; Connective-Chondro, chondrosarcoma; Connective-Leiomyo, leiomyosarcoma; Connective-Liposarc, liposarcoma; Connective-Myxofibro, myxofibrosarcoma; Connective-Osteosarc, osteosarcoma; Connective-SCS spindle cell sarcoma; Connective-SS, synovial sarcoma; Eso-AdenoCA, esophageal adenocarcinoma; HeadNeck-SCC, squamous cell carcinoma of the head and neck; Kidney-CCRCC, clear cell renal cell carcinoma; Kidney-ChRCC, chromophobe renal cell carcinoma; Kidney-PRCC, papillary renal cell carcinoma; Liver-HCC, hepatocellular carcinoma;Lung-AdenoCA, lung adenocarcinoma; Lung-LargeCell, large-cell lung cancer; Lung-SCC, squamous cell carcinoma of the lung; Lung-SmallCell, small cell carcinoma of the lung; Mes-Mesothelioma, mesothelioma; Ovary-AdenoCA, ovarian adenocarcinoma; Panc-AdenoCA, pancreatic adenocarcinoma; Prost-AdenoCA, prostate adenocarcinom; Skin-Melanoma, melanoma of the skin; Stomach-AdenoCA, gastric adenocarcinoma; Testis-GCT, testicular germ cell tumour; Uterus-AdenoCA, uterine adenocarcinoma.

Figure 2. (A) Circos heatmap of cancer driver genes identified. Heatmap intensity proportional to Stouffer *P*-value. **(B) Distribution of driver genes across different types of cancer.** y-axis, maximal mutational prevalence in a tumour type, x-axis, number of tumour types in which the driver gene is identified. **(C) Distribution of cancer driver gene function associated with each cancer type.** Y-axis, tumour group, x-axis, percentage of tumour specific driver genes. **(D) Comparison of driver gene somatic mutation rates between tumour histologies.** Expected mutation rate of each driver in the cohort based on the

number of samples in which the driver gene is mutated for the given tumour histology. Binomial *P*-values are shown.

FIgure 3. Clinical actionability ascribable to each cancer driver gene according to COSMIC and OncoKB by cancer type. Tumours were annotated by the highest scoring gene mutation - indication pairing, with "None" indicating no actionable mutations were detected in the tumour. (A) Catalogued by COSMIC: 1, Approved marketed drug with demonstrated efficacy at the mutation; 2, Phase 2/3 clinical results meet primary outcome measures; 3 Drug in ongoing clinical trials. **(B) Catalogued by OncoKB:** 1, FDA approved drug in the cancer type; 2, standard of care in the cancer type; 3, clinical evidence in the cancer type or standard of care in a different cancer type; 4, supported by compelling biological evidence.

FIgure 4. Example druggability networks for colorectal cancer. Nodes acting as cancer-specific drivers are shaded purple. Edge visual properties are as follows: OncoKB interactions, red contiguous arrow; Signor interactions, green contiguous arrow; Signor inhibitors, black vertical slash; complex, black zigzag; direct interaction, red solid line; direct X-ray interaction, green solid line; direct non-protein data bank interaction, blue solid line; reaction, blue contiguous arrow; transcriptional interaction, black sinewave.

REFERENCES

1. Topol, E. J. Individualized medicine from prewomb to tomb. *Cell* **157**, 241–253 (2014).
2. Spear, B. B., Heath-Chiozzi, M. & Huff, J. Clinical application of pharmacogenetics. *Trends in Molecular Medicine* vol. 7 201–204 Preprint at https://doi.org/10.1016/s1471-4914(01)01986-4 (2001).
3. Schwartzberg, L., Kim, E. S., Liu, D. & Schrag, D. Precision Oncology: Who, How, What, When, and When Not? *American Society of Clinical Oncology Educational Book* 160–169 Preprint at https://doi.org/10.1200/edbk_174176 (2017).
4. Stratton, M. R., Campbell, P. J. & Andrew Futreal, P. The cancer genome. *Nature* vol. 458 719–724 Preprint at https://doi.org/10.1038/nature07943 (2009).
5. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* **2017**, (2017).
6. Freedman, A. N. *et al.* Use of Next-Generation Sequencing Tests to Guide Cancer Treatment: Results From a Nationally Representative Survey of Oncologists in the United States. *JCO Precis Oncol* **2**, (2018).
7. Turnbull, C. *et al.* The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ* **361**, k1687 (2018).
8. Cornish, A. J. *et al.* Reference bias in the Illumina Isaac aligner. *Bioinformatics* vol. 36 4671–4672 (2020).
9. Martínez-Jiménez, F. *et al.* A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**, 555–572 (2020).
10. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
11. Salipante, S. J., Scroggins, S. M., Hampel, H. L., Turner, E. H. & Pritchard, C. C. Microsatellite Instability Detection by Next Generation Sequencing. *Clinical Chemistry* vol. 60 1192–1199 Preprint at https://doi.org/10.1373/clinchem.2014.223677 (2014).

12. Davies, H. *et al.* HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* **23**, 517–525 (2017).

13. Mitsopoulos, C. *et al.* canSAR: update to the cancer translational research and drug discovery knowledgebase. *Nucleic Acids Res.* **49**, D1074–D1082 (2020).

14. Orchard, S. *et al.* The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358–63 (2014).

15. Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43**, D512–20 (2015).

16. Shimada, K., Bachman, J. A., Muhlich, J. L. & Mitchison, T. J. shinyDepMap, a tool to identify targetable cancer genes and their functional connections from Cancer Dependency Map data. *Elife* **10**, (2021).

17. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

18. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371–385.e18 (2018).

19. André, F. *et al.* Alpelisib for PIK3CA-Mutated, Hormone Receptor-Positive Advanced Breast Cancer. *N. Engl. J. Med.* **380**, 1929–1940 (2019).

20. Fell, J. B. *et al.* Identification of the Clinical Development Candidate , a Covalent KRAS Inhibitor for the Treatment of Cancer. *J. Med. Chem.* **63**, 6679–6693 (2020).

21. Canon, J. *et al.* The clinical KRAS(G12C) inhibitor AMG 510 drives anti-tumour immunity. *Nature* **575**, 217–223 (2019).

22. Mateo, J. *et al.* A First-Time-in-Human Study of GSK2636771, a Phosphoinositide 3 Kinase Beta-Selective Inhibitor, in Patients with Advanced Solid Tumors. *Clin. Cancer Res.* **23**, 5981–5992 (2017).

23. Marabelle, A. *et al.* Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: prospective biomarker analysis of the

multicohort, open-label, phase 2 KEYNOTE-158 study. *Lancet Oncol.* **21**, 1353–1365 (2020).

24. Website. 'National Genomic Test Directory.' n.d. https://www.england.nhs.uk/publication/national-genomic-test-directories/.

25. Michaeli, D. T. & Michaeli, T. Overall Survival, Progression-Free Survival, and Tumor Response Benefit Supporting Initial US Food and Drug Administration Approval and Indication Extension of New Cancer Drugs, 2003-2021. *J. Clin. Oncol.* **40**, 4095–4106 (2022).

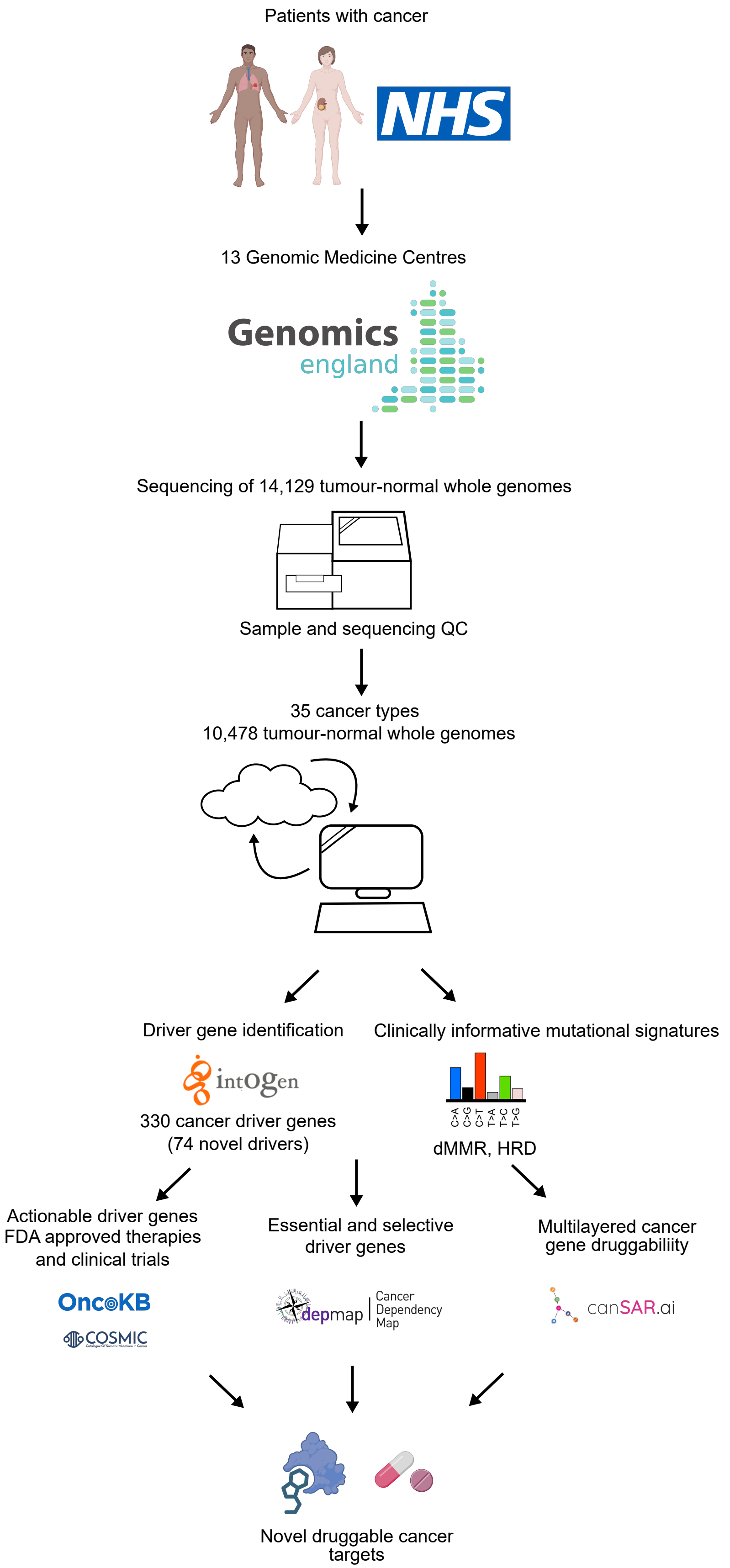
Patients with cancer
NHS
13 Genomic Medicine Centres
Genomics
england
Sequencing of 14,129 tumour-normal whole genomes
Sample and sequencing QC
35 cancer types
10,478 tumour-normal whole genomes
Driver gene identification
Clinically informative mutational signatures
intOgen
330 cancer driver genes
(74 novel drivers)
C>A
C>G
C>T
T>A
T>C
T>G
dMMR, HRD
Actionable driver genes
FDA approved therapies
and clinical trials
OncoKB
COSMIC
Essential and selective
driver genes
depmap
Cancer
Dependency
Map
Multilayered cancer
gene druggabiliity
canSAR.ai
Novel druggable cancer
targets

Figure 1A

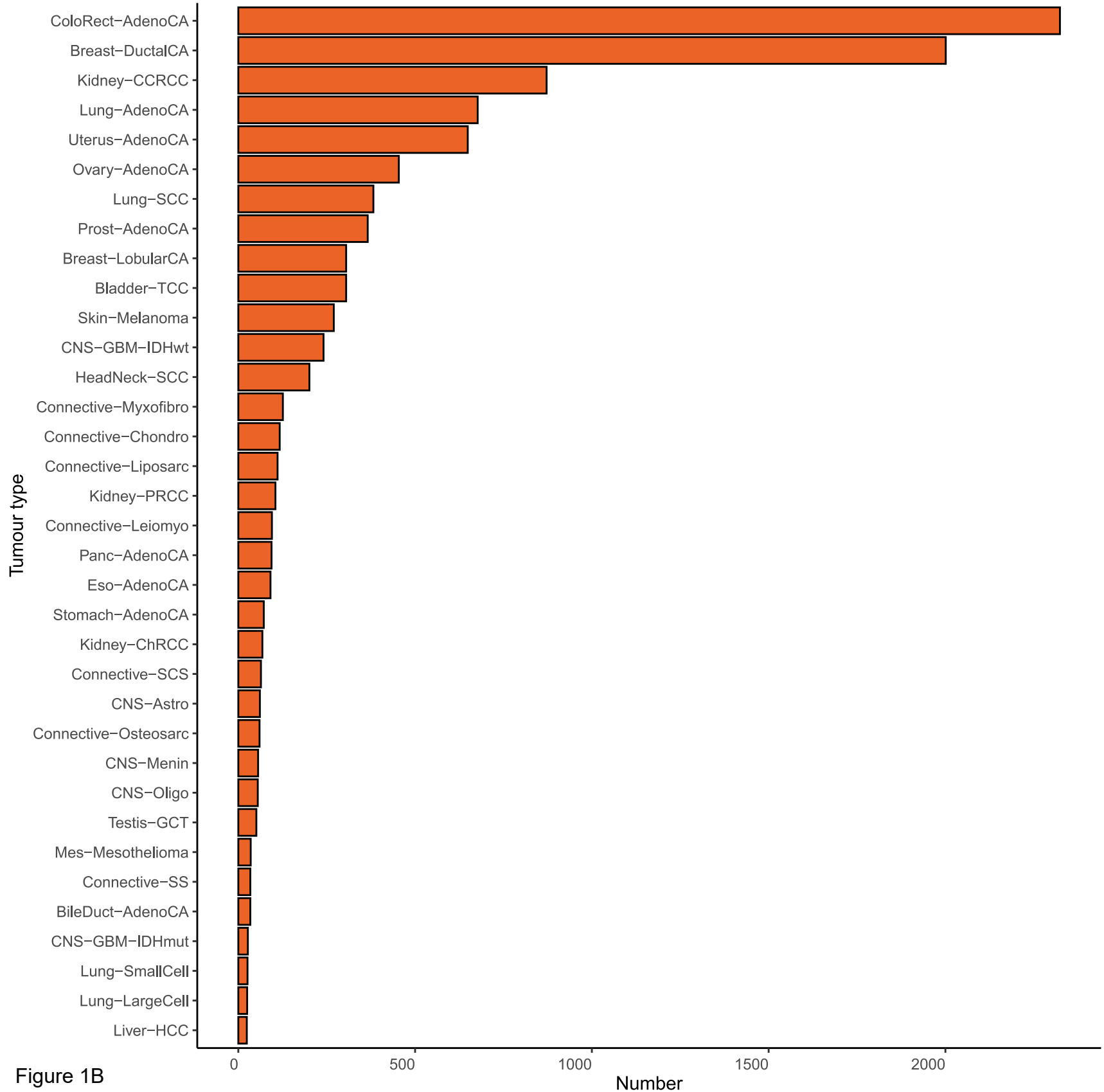
ColoRect−AdenoCA
Breast−DuctalCA
Kidney−CCRCC
Lung−AdenoCA
Uterus−AdenoCA
Ovary−AdenoCA
Lung−SCC
Prost−AdenoCA
Breast−LobularCA
Bladder−TCC
Skin−Melanoma
CNS−GBM−IDHwt
HeadNeck−SCC
Connective−Myxofibro
Connective−Chondro
Connective−Liposarc
Kidney−PRCC
Connective−Leiomyo
Panc−AdenoCA
Eso−AdenoCA
Stomach−AdenoCA
Kidney−ChRCC
Connective−SCS
CNS−Astro
Connective−Osteosarc
CNS−Menin
CNS−Oligo
Testis−GCT
Mes−Mesothelioma
Connective−SS
BileDuct−AdenoCA
CNS−GBM−IDHmut
Lung−SmallCell
Lung−LargeCell
Liver−HCC
Tumour type
0
500
1000
1500
2000
Number

Figure 1B

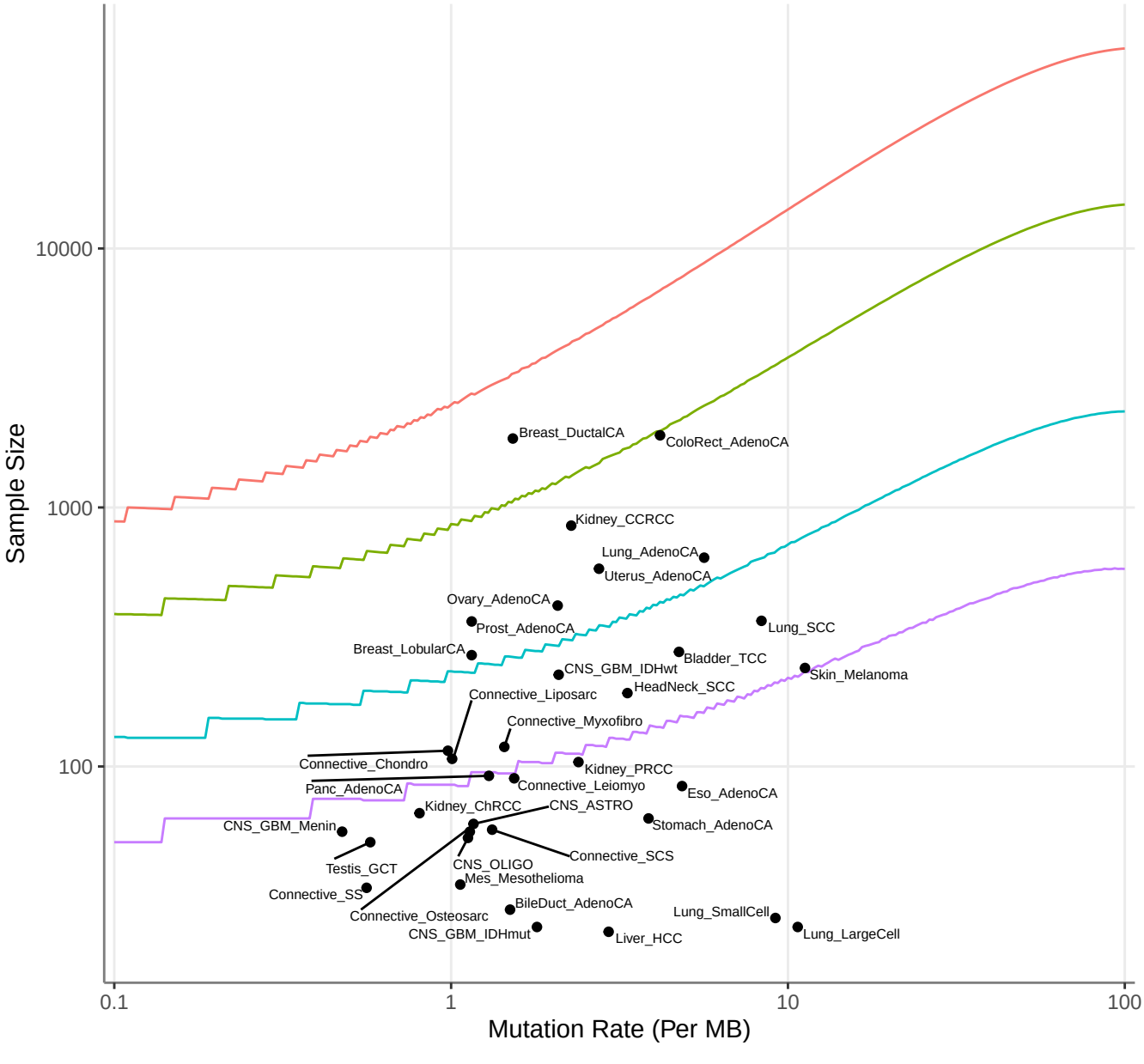
Sample Size
10000
1000
100
0.1
1
10
100
Mutation Rate (Per MB)
Breast_DuctalCA
ColoRect_AdenoCA
Kidney_CCRCC
Lung_AdenoCA
Uterus_AdenoCA
Ovary_AdenoCA
Prost_AdenoCA
Lung_SCC
Breast_LobularCA
Bladder_TCC
CNS_GBM_IDHwt
Skin_Melanoma
HeadNeck_SCC
Connective_Liposarc
Connective_Myxofibro
Connective_Chondro
Kidney_PRCC
Panc_AdenoCA
Connective_Leiomyo
Eso_AdenoCA
Kidney_ChRCC
CNS_ASTRO
CNS_GBM_Menin
Stomach_AdenoCA
Connective_SCS
Testis_GCT
CNS_OLIGO
Mes_Mesothelioma
Connective_SS
BileDuct_AdenoCA
Lung_SmallCell
Connective_Osteosarc
CNS_GBM_IDHmut
Liver_HCC
Lung_LargeCell
Effect Size Above Background
0.01
0.02
0.05
0.1

Figure 1C

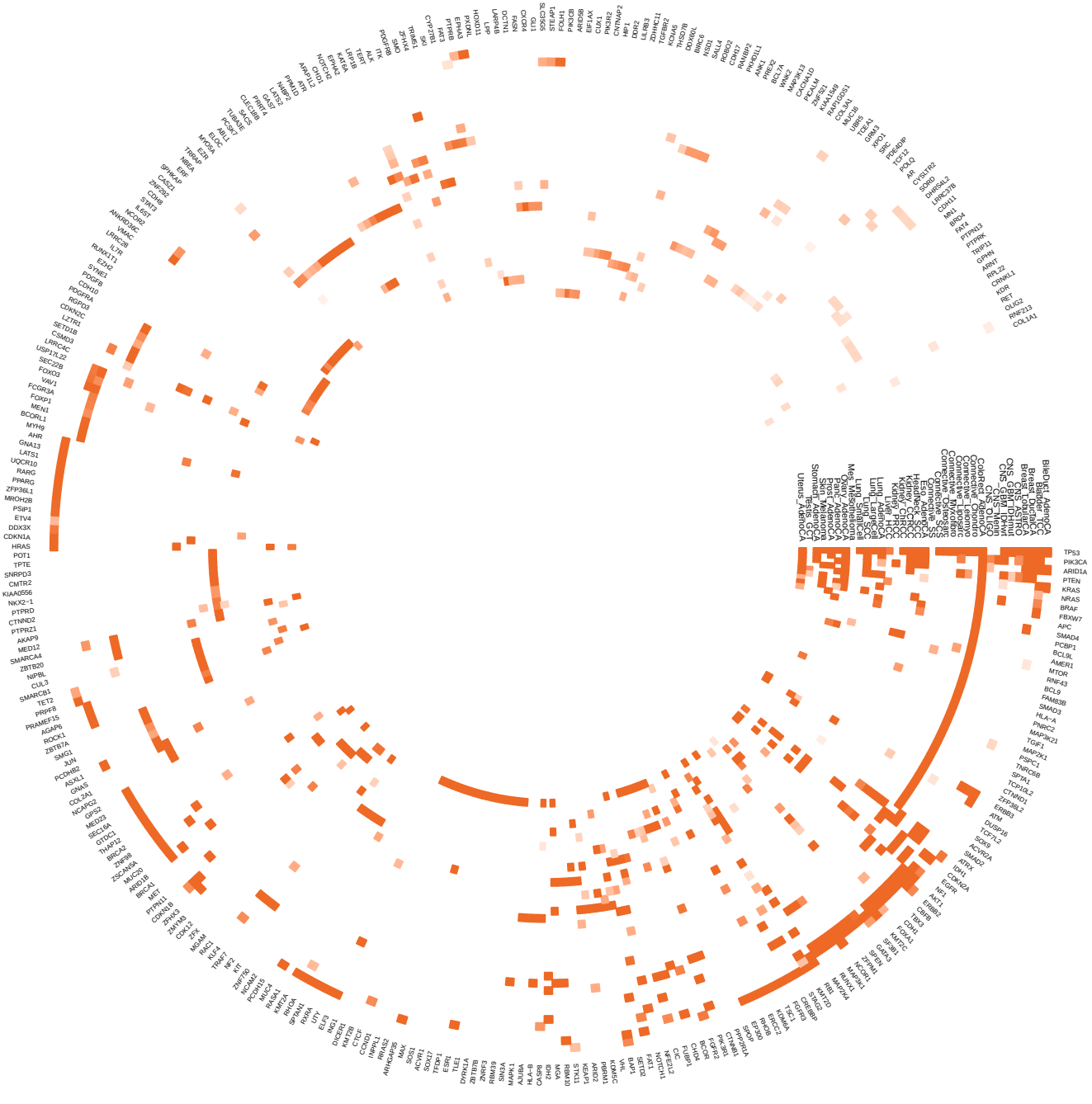

Figure 2A

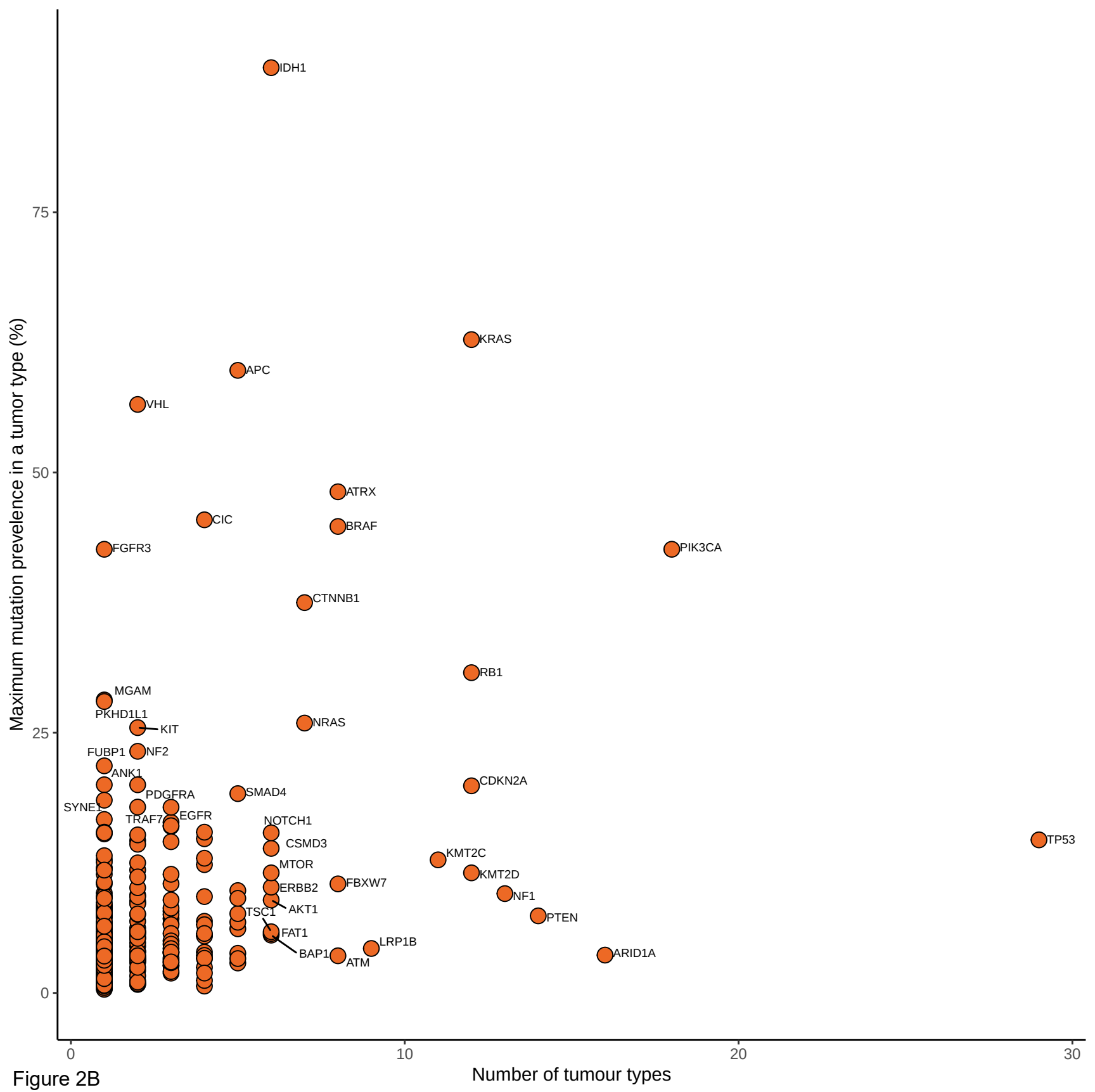
Maximum mutation prevelence in a tumor type (%)
Number of tumour types
0
25
50
75
0
10
20
30
IDH1
KRAS
APC
VHL
ATRX
CIC
BRAF
FGFR3
PIK3CA
CTNNB1
RB1
MGAM
PKHD1L1
KIT
NRAS
FUBP1
NF2
ANK1
CDKN2A
PDGFRA
SMAD4
SYNE1
TRAF7
EGFR
NOTCH1
CSMD3
TP53
KMT2C
MTOR
KMT2D
FBXW7
ERBB2
NF1
AKT1
TSC1
PTEN
FAT1
LRP1B
BAP1
ARID1A
ATM

Figure 2B

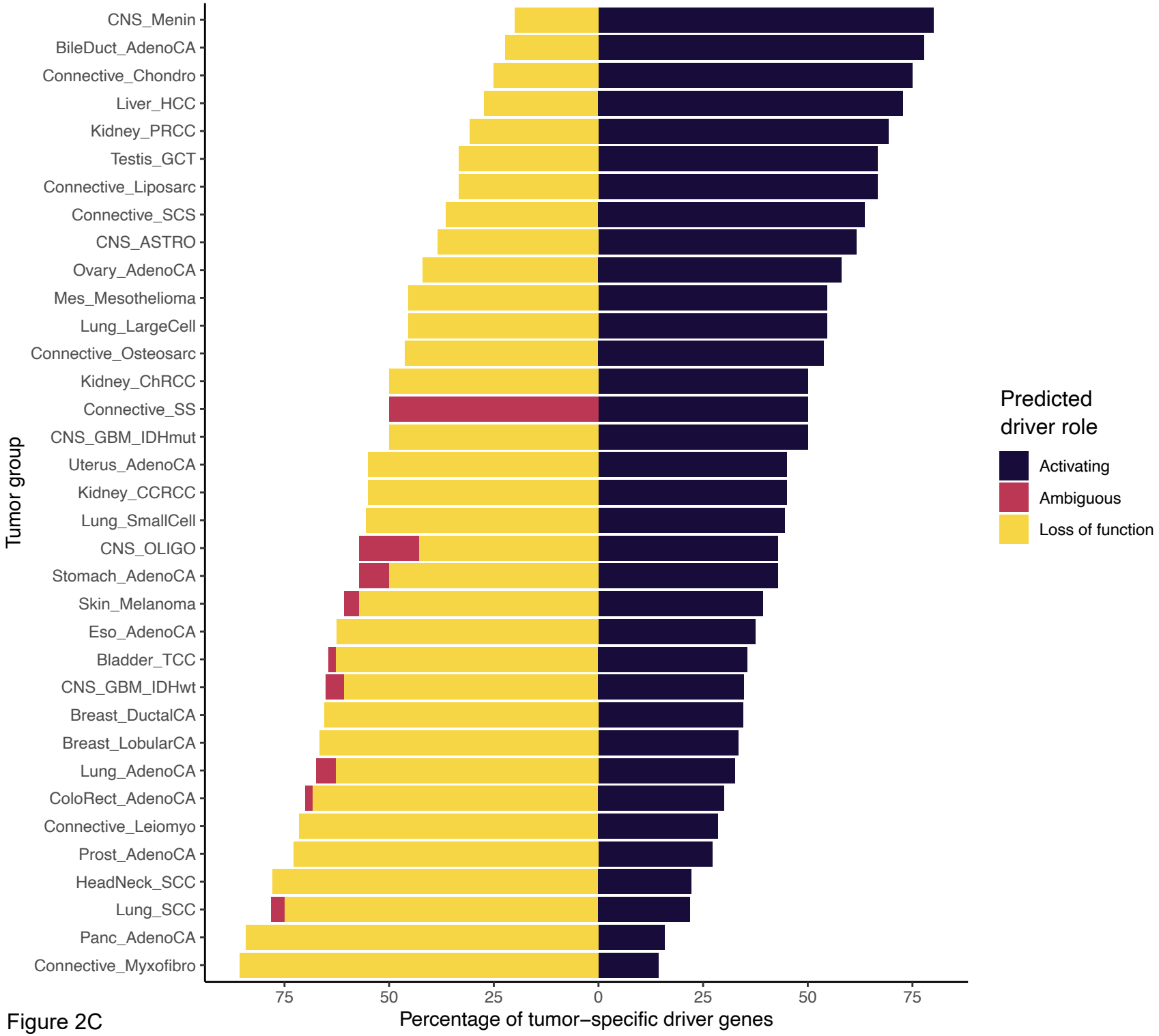
CNS_Menin
BileDuct_AdenoCA
Connective_Chondro
Liver_HCC
Kidney_PRCC
Testis_GCT
Connective_Liposarc
Connective_SCS
CNS_ASTRO
Ovary_AdenoCA
Mes_Mesothelioma
Lung_LargeCell
Connective_Osteosarc
Kidney_ChRCC
Connective_SS
CNS_GBM_IDHmut
Uterus_AdenoCA
Kidney_CCRCC
Lung_SmallCell
CNS_OLIGO
Stomach_AdenoCA
Skin_Melanoma
Eso_AdenoCA
Bladder_TCC
CNS_GBM_IDHwt
Breast_DuctalCA
Breast_LobularCA
Lung_AdenoCA
ColoRect_AdenoCA
Connective_Leiomyo
Prost_AdenoCA
HeadNeck_SCC
Lung_SCC
Panc_AdenoCA
Connective_Myxofibro
Tumor group
75
50
25
0
25
50
75
Percentage of tumor−specific driver genes
Predicted driver role
Activating
Ambiguous
Loss of function

Figure 2C

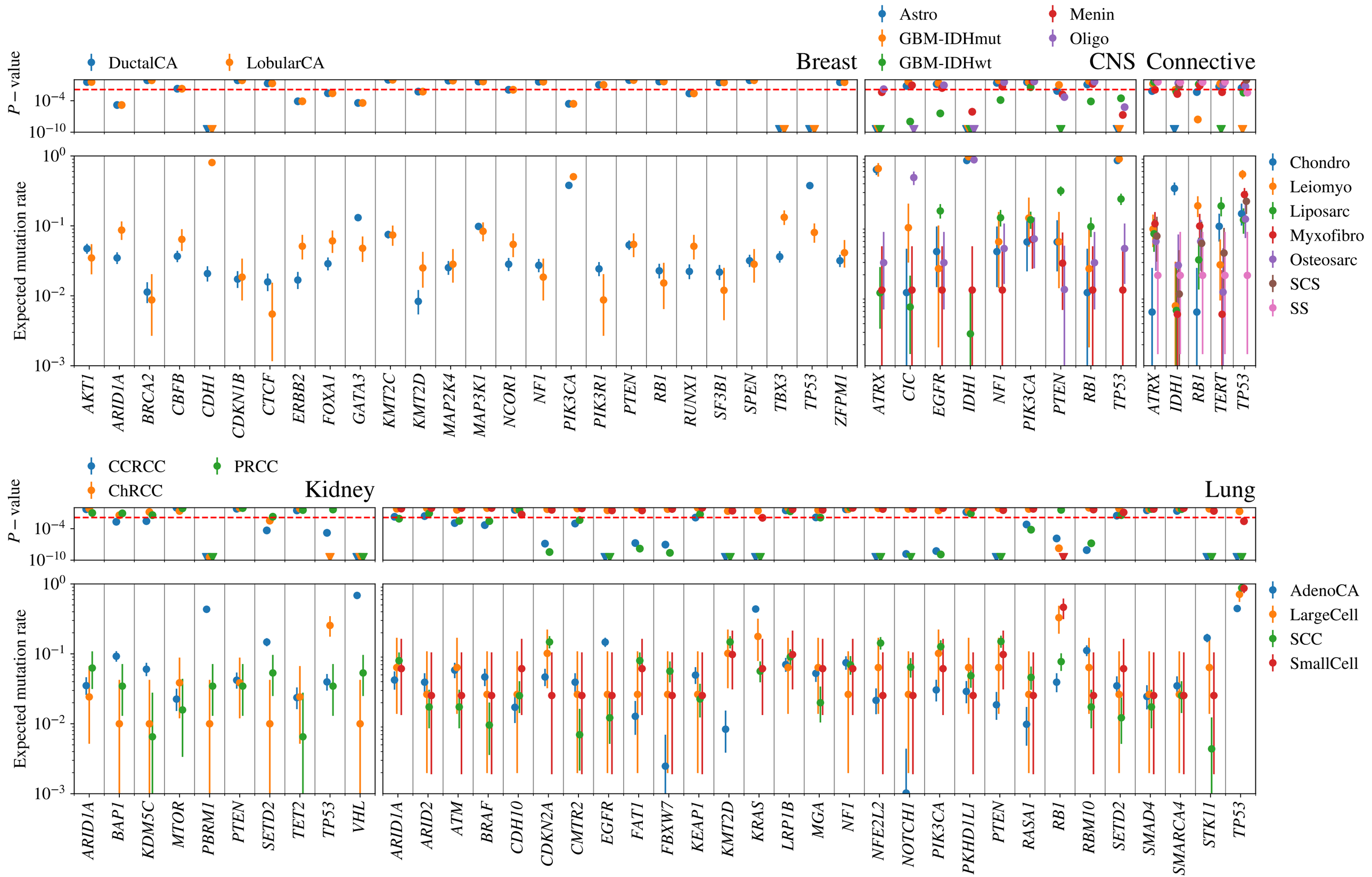
Breast
DuctalCA
LobularCA
CNS
Astro
GBM-IDHmut
GBM-IDHwt
Menin
Oligo
Connective
Chondro
Leiomyo
Liposarc
Myxofibro
Osteosarc
SCS
SS
P − value
Expected mutation rate
AKT1
ARID1A
BRCA2
CBFB
CDH1
CDKN1B
CTCF
ERBB2
FOXA1
GATA3
KMT2C
KMT2D
MAP2K4
MAP3K1
NCOR1
NF1
PIK3CA
PIK3R1
PTEN
RB1
RUNX1
SF3B1
SPEN
TBX3
TP53
ZFPM1
ATRX
CIC
EGFR
IDH1
NF1
PIK3CA
PTEN
RB1
TP53
ATRX
IDH1
RB1
TERT
TP53
Kidney
CCRCC
ChRCC
PRCC
Lung
AdenoCA
LargeCell
SCC
SmallCell
ARID1A
BAP1
KDM5C
MTOR
PBRM1
PTEN
SETD2
TET2
TP53
VHL
ARID1A
ARID2
ATM
BRAF
CDH10
CDKN2A
CMTR2
EGFR
FAT1
FBXW7
KEAP1
KMT2D
KRAS
LRP1B
MGA
NF1
NFE2L2
NOTCH1
PIK3CA
PKHD1L1
PTEN
RASA1
RB1
RBM10
SETD2
SMAD4
SMARCA4
STK11
TP53

Figure 2D

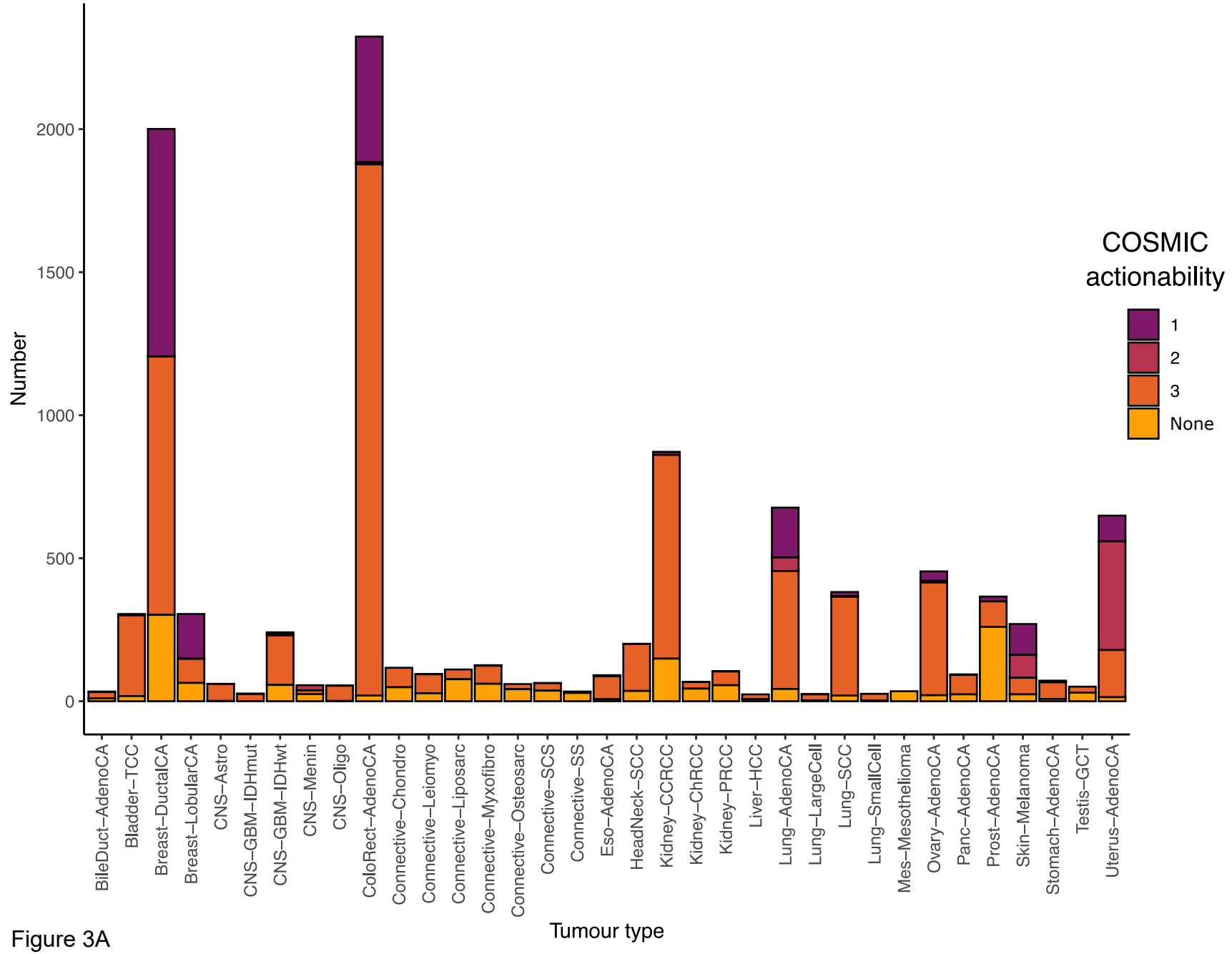
Number
2000
1500
1000
500
0
COSMIC actionability
1
2
3
None
BileDuct–AdenoCA
Bladder–TCC
Breast–DuctalCA
Breast–LobularCA
CNS–Astro
CNS–GBM–IDHmut
CNS–GBM–IDHwt
CNS–Menin
CNS–Oligo
ColoRect–AdenoCA
Connective–Chondro
Connective–Leiomyo
Connective–Liposarc
Connective–Myxofibro
Connective–Osteosarc
Connective–SCS
Connective–SS
Eso–AdenoCA
HeadNeck–SCC
Kidney–CCRCC
Kidney–ChRCC
Kidney–PRCC
Liver–HCC
Lung–AdenoCA
Lung–LargeCell
Lung–SCC
Lung–SmallCell
Mes–Mesothelioma
Ovary–AdenoCA
Panc–AdenoCA
Prost–AdenoCA
Skin–Melanoma
Stomach–AdenoCA
Testis–GCT
Uterus–AdenoCA
Tumour type

Figure 3A

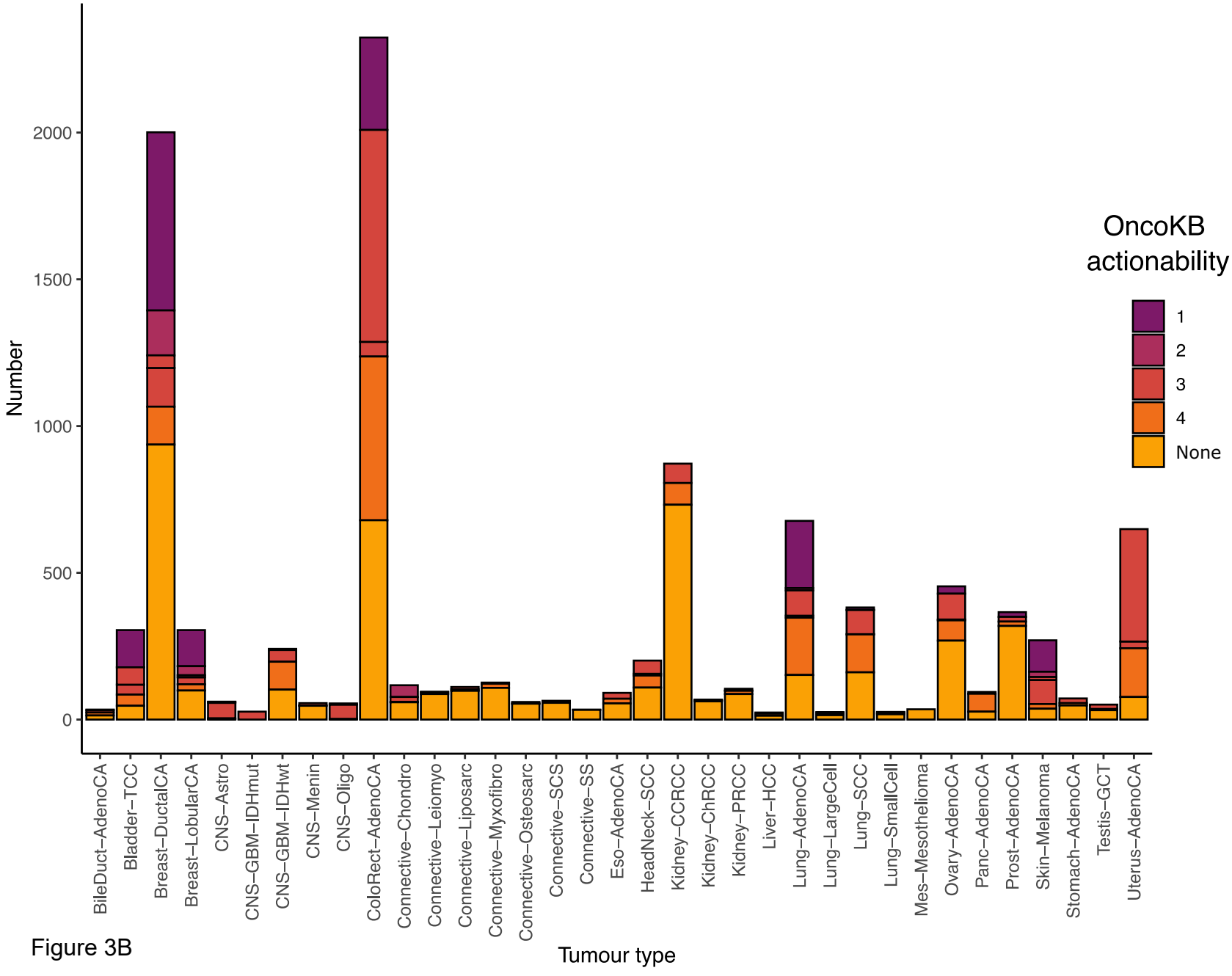
Number
2000
1500
1000
500
0
OncoKB actionability
1
2
3
4
None
BileDuct–AdenoCA
Bladder–TCC
Breast–DuctalCA
Breast–LobularCA
CNS–Astro
CNS–GBM–IDHmut
CNS–GBM–IDHwt
CNS–Menin
CNS–Oligo
ColoRect–AdenoCA
Connective–Chondro
Connective–Leiomyo
Connective–Liposarc
Connective–Myxofibro
Connective–Osteosarc
Connective–SCS
Connective–SS
Eso–AdenoCA
HeadNeck–SCC
Kidney–CCRCC
Kidney–ChRCC
Kidney–PRCC
Liver–HCC
Lung–AdenoCA
Lung–LargeCell
Lung–SCC
Lung–SmallCell
Mes–Mesothelioma
Ovary–AdenoCA
Panc–AdenoCA
Prost–AdenoCA
Skin–Melanoma
Stomach–AdenoCA
Testis–GCT
Uterus–AdenoCA
Tumour type

Figure 3B

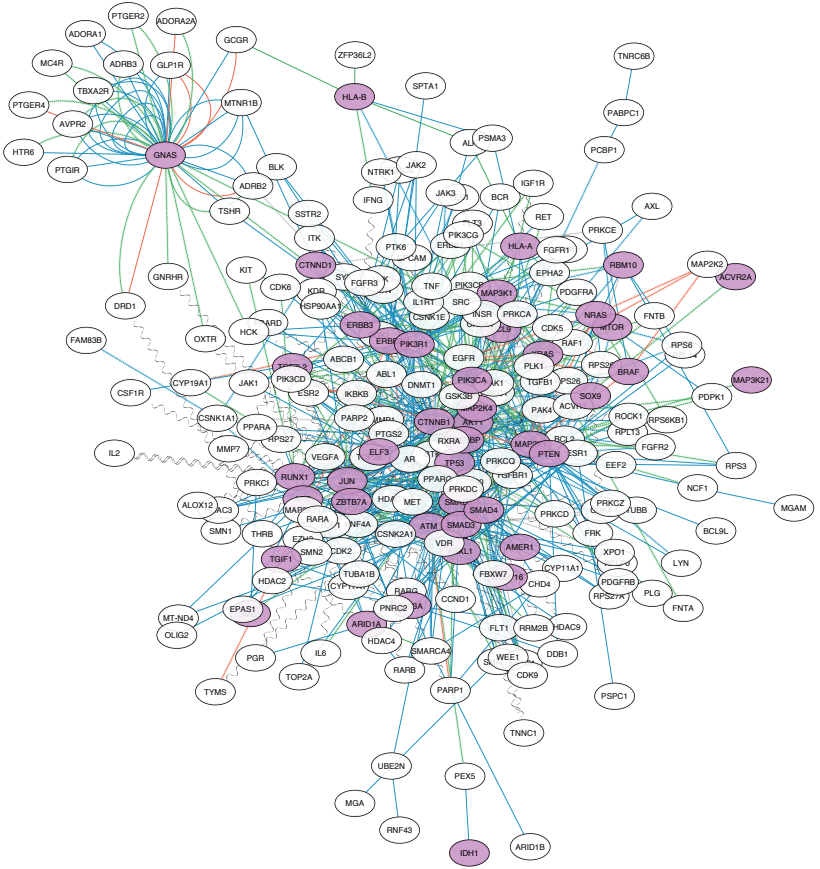
PTGER2
ADORA2A
ADORA1
GCGR
MC4R
ADRB3
GLP1R
TBXA2R
PTGER4
MTNR1B
AVPR2
HTR6
GNAS
PTGIR
BLK
ADRB2
TSHR
SSTR2
ITK
GNRHR
KIT
CDK6
KDR
DRD1
HCK
OXTR
FAM83B
CTNND1
FGFR3
HSP90AA1
ERBB3
ABCB1
PIK3CD
CYP19A1
JAK1
CSF1R
ESR2
IKBKB
CSNK1A1
PPARA
PPARP2
PTGS2
RPS27
MMP7
IL2
VEGFA
ELF3
AR
RUNX1
JUN
PRKCI
ZBTB7A
ALOX12
MET
MAP2K1
RARA
SMN1
THRB
NF4A
CSNK2A1
ATM
SMAD4
SMAD3
PRKCD
PRKCZ
UBB
FRK
BCL9L
TGIF1
SMN2
CDK2
VDR
AMER1
XPO1
LYN
HDAC2
TUBA1B
FBXW7
CYP11A1
CHD4
PDGFRB
PLG
FNTA
MT-ND4
OLIG2
EPAS1
PNRC2
ARID1A
CCND1
FLT1
RRM2B
HDAC9
HDAC4
SMARCA4
IL6
PGR
TOP2A
RARB
WEE1
DDB1
CDK9
TYMS
PARP1
PSPC1
TNNC1
UBE2N
PEX5
MGA
RNF43
IDH1
ARID1B
ZFP36L2
SPTA1
HLA-B
TNRC6B
PABPC1
PCBP1
PSMA3
JAK2
NTRK1
IFNG
JAK3
BCR
IGF1R
RET
AXL
PRKCE
PIK3CG
PTK6
HLA-A
FGFR1
RBM10
MAP2K2
ACVR2A
EPHA2
PDGFRA
TNF
MAP3K1
IL1R1
SRC
CSNK1E
INSR
PRKCA
NRAS
FNTB
CDK5
RAF1
RPS6
EGFR
PIK3R1
PIK3CA
PLK1
BRAF
MAP3K21
TGFB1
SOX9
PDPK1
DNMT1
GSK3B
PAK4
ROCK1
RPS6KB1
CTNNB1
RPL13
RXRA
PTEN
FGFR2
RPS3
EEF2
NCF1
MGAM
BCL9L
LYN
PLG
FNTA
DDB1
PSPC1
TNNC1
UBE2N
PEX5
MGA
RNF43
IDH1
ARID1B

Figure 4