medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Single-cell RNA sequencing of human tissue supports successful drug targets.

Emma Dann*[1,2], Erin Teeple*[3], Rasa Elmentaite[2], Kerstin B Meyer[1], Giorgio Gaglia[3], Frank Nestle[4], Virginia Savova[3], Emanuele de Rinaldis[3]+, Sarah A Teichmann[1,2,5] +

Affiliations:

1. Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK
2. Ensocell Therapeutics, BioData Innovation Centre, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA
3. Precision Medicine & Computational Biology, Sanofi Research US, Cambridge, MA, 01701, USA
4. Research & Development, Sanofi, Cambridge, MA, 01701, USA
5. Theory of Condensed Matter, Cavendish Laboratory/Dept Physics, University of Cambridge, JJ Thomson Ave, Cambridge CB3 0HE, UK

*Authors contributed equally
+Co-Corresponding and senior authors: emanuele.derinaldis@sanofi.com, st9@sanger.ac.uk

Abstract

Early characterization of drug targets associated with disease can greatly reduce clinical failures attributed to lack of safety or efficacy. As single-cell RNA sequencing (scRNA-seq) of human tissues becomes increasingly common for disease profiling, the insights obtained from this data could influence target selection strategies. Whilst the use of scRNA-seq to understand target biology is well established, the impact of single-cell data in increasing the probability of candidate therapeutic targets to successfully advance from research to clinic has not been fully characterized. Inspired by previous work on an association between genetic evidence and clinical success, we used retrospective analysis of known drug target genes to identify potential predictors of target clinical success from scRNA-seq data. Particularly, we investigated whether successful drug targets are associated with cell type specific expression in a disease-relevant tissue (cell type specificity) or cell type specific over-expression in disease patients compared to healthy controls (disease cell specificity). Analysing scRNA-seq data across 30 diseases and 13 tissues, we found that both classes of scRNA-seq support significantly increase the odds of clinical success for gene-disease pairs. We estimate that combined they could approximately triple the chances of a target reaching phase III. Importantly, scRNA-seq analysis identifies a larger and complementary target space to that of direct genetic evidence. In particular, scRNA-seq support is more likely to prioritize therapeutically tractable classes of genes such as membrane-bound proteins. Our study suggests that scRNA-seq-derived information on cell type- and disease-specific expression can be leveraged to identify tractable and disease-relevant targets, with increased probability of success in the clinic.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Introduction

Drug discovery begins with the identification of candidate targets, drug-binding molecules whose modulation is hypothesized to be useful for the treatment of disease [1]. The discovery and development of a novel drug for a candidate target progresses in the following steps: target validation, compound screening and lead identification, characterization of mechanism of action, indication(s) selection, safety and efficacy clinical trials, and finally, in successful cases, regulatory approval. Development of a single new drug takes an average of 12-15 years and costs (including concurrent program failures) are estimated to range from 900 million – 2.6 billion USD per success [2,3]. A drug discovery program can fail at each step between early research to regulatory approval, and it is estimated that in >90% of cases failures can be attributed to suboptimal target selection for a given disease, resulting in safety or efficacy issues [4]. Together, these observations point to the need to improve the strategies and the data used in early stages of drug discovery to support the selection of candidate therapeutic targets, to increase the likelihood of clinical success.

Single-cell RNA sequencing (scRNA-seq) data is a particularly promising source of evidence for target selection, providing cell-level resolution of molecular profiles in disease-relevant tissues. Single cell technologies have already been applied extensively to characterize disease biology, in emerging diseases like COVID-19 [5,6], cancer [7–10], and common complex diseases across tissues [11–14]. The rapidly growing body of disease-relevant scRNA-seq data has already begun to inform the development of novel diagnostics and cell-targeting precision therapies [15]. This led us to ask to what extent information on cell type specific expression can boost the selection of promising drug targets.

Retrospective analysis of known drug targets has been used to identify features predictive of target success. Notably, such analyses have shown that targets linked to genetic variants associated with the relevant disease are twice as likely to reach clinical approval as targets with no genetic support [16–18]. These studies greatly impacted decision-making in biotech and pharmaceutical industries. Out of 428 newly FDA-approved drugs from 2013 to 2022, 271 (63%) are backed by direct or indirect human genetic evidence [19,20]. Even though establishing whether this influenced their discovery or development phases is difficult, 250 out of 271 genetics-backed drugs had publicly accessible genetic support before approval.

Given this precedent, in this work we used retrospective analysis to identify potential predictors of target clinical success from scRNA-seq data. We investigated two cell type specific expression modes that are commonly used in scRNA-seq disease analysis and can support target discovery. The modes include cell type specific expression in a disease-relevant tissue (hereafter *cell type specificity*) and cell type specific over-expression in disease patients compared to healthy controls (hereafter *disease cell specificity*). We used a uniform workflow to identify cell type specific and disease cell specific target-disease pairs across 30 complex diseases in 13 disease-relevant tissues using the CZ CellxGene Discover database [21]. We then evaluated how scRNA-seq supported target-disease associations correlate with target success in clinical trials, benchmarking against direct genetic associations as reported from the Open Targets platform [22]. We found that scRNA-seq support significantly increased the odds of clinical success for target-disease pairs and identified a complementary target space to that of direct genetic evidence. These results highlight the value of scRNA-seq data as a key resource, complementary to genetics, to increase probability of clinical success in drug development.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Results

Definition of scRNA-seq support for targets

As a cause or consequence of disease, pathology arises when cells of a particular type develop abnormal traits within a disease-relevant tissue. Safe and effective therapies should precisely target these aberrant cells, without eliciting on-target toxicities in other cells and tissues. Given this need, scRNA-seq data can support target prioritization by identifying genes expressed in a cell type specific manner in tissue from healthy and diseased individuals. We aimed to assess whether cell type specific genes, as identified by scRNA-seq analysis, are more likely to be targets of clinically successful drugs. We considered diseases for which scRNA-seq data was available via the CZ CellxGene Discover database [21]. We defined a disease-relevant (DR) tissue for each disease term. Of the 58 disease terms in the CellxGene database, 30 terms were retained for association analysis, based on availability of data from disease-relevant tissue and overlap with OpenTargets disease annotation terms (see Supplementary Table 1 for a complete list of diseases and reasons to exclude from analysis). The most prevalent diseases were lung and immune disorders (Figure 1A). For each disease term, we collected gene expression count matrices and coarse cell type labels, harmonized using the Cell Ontology [23] (Figure 1B, Supplementary Figure 1, see Methods), for disease-relevant tissue samples from healthy and diseased individuals (Supplementary Table 2).

We next defined two classes of scRNA-seq supported genes for target discovery: (1) cell type specific genes in healthy disease-relevant tissue (*cell type specific*) and (2) genes specifically over expressed in a cell type in tissue from disease patients, compared to healthy tissue (*disease cell specific*) (Figure 1C). We reasoned that drugs targeting *cell type specific* genes inhibit expansion and function of normal cells acquiring aberrant phenotypes in disease. For example, the GLP-1 receptor, targeted by commonly used anti-diabetic drugs, is normally expressed in pancreatic beta cells, which become dysfunctional in disease [13]. Conversely, drugs targeting *disease cell specific* genes suppress aberrant gene programmes directly. For example, inflammatory bowel disease patients are treated with antibodies targeting the tumor necrosis factor (TNF) which is over-expressed in regulatory T cells and other immune subtypes in disease [24].

Enrichment of clinically successful targets in genes with scRNA-seq support

For each disease, we identified cell type specific and disease cell specific genes with highly variable gene (HVG) selection and differential expression (DE) analysis, aggregating mRNA counts across cell types and donors (Figure 1D, see Methods). With this analysis across 30 diseases, we annotated 33654 gene-disease (G-D) pairs as cell type specific and 60851 G-D pairs as disease cell specific (Supplementary Figure 2). To associate scRNA-seq support with clinical success, we extracted information about targets of drugs approved or in trial from the Open Targets platform [1,22,25] (n = 2358 drugs for which the studied diseases are an approved or investigational indication). Across diseases, we annotated 2925 G-D pairs as safe (passed phase I), of which 1646 pairs where also effective (passed phase II), and 601 pairs were also approved (passed phase III) (Supplementary Figure 2, Supplementary Table 3).

We then computed the odds of clinical success, with or without support from scRNA-seq data (Figure 1C, see Methods). Of note, our analyses are disease-specific: we count successful G–D pairs with corresponding scRNA-seq support from analysis of healthy and diseased

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

individuals in the disease-relevant tissue. For example, a gene that is found to be cell type specific in esophagus is not considered as having scRNA-seq support for pulmonary fibrosis.

To enumerate the space of possible G-D pairs, we multiplied the number of diseases considered (N=30) with a "universe" of genes. We define four different universes: all protein-coding genes (N=19620), representing the space of genes that are typically analysed in scRNA-seq data; genes that are antibody-tractable (N=12527) or small molecule-tractable (N=6550) based on Open Targets tractability assessment, representing genes that are tractable by any therapeutic agent; finally, genes already targeted by therapies in clinical trial for any indication (known drug targets, N=936), representing demonstrably druggable proteins (Supplementary Figure 3).

Out of 2925 target-indication pairs which passed at least phase I, 858 were prioritized as either cell type specific or disease cell specific by scRNA-seq analysis (Figure 2A). Considering protein-coding genes, antibody- and small molecule-tractable genes, cell type specific and disease cell specific G-D pairs with scRNA-seq support were always significantly enriched in targets of safe, effective, or approved drugs (Figure 2B, Supplementary Table 4). Out of 2840 protein-coding G-D pairs passing phase I, 356 (12%) were cell type specific in the DR tissue (OR=2.47, p-value = 3.57e-46) and 594 (20%) were disease cell specific (OR=2.34, p-value=4.43e-64). The enrichment of disease cell specific genes in clinically successful targets was the highest amongst antibody-tractable genes. When restricting the analysis to known drug targets, only disease cell specific genes were significantly enriched in effective and approved targets (Figure 2B). This might indicate that specific expression in the disease-relevant tissue is already implicitly used by drug discovery programmes for selecting targets that progress to clinical development. Combining both classes of scRNA-seq support (cell type and disease specific genes) led to significantly higher association with success in phase I and effectiveness (phase II) than each class individually, especially for protein coding and small molecule tractable targets (Figure 2B).

Comparison between scRNA-seq supported and genetic supported targets

We compared genes supported by scRNA-seq with genes associated to the disease by human genetics data, using the Open Targets direct genetic association score [22,26]. Throughout the manuscript, we refer to genes that are prioritized by either genetic association, cell type specificity or disease cell specificity as genes with "omic support". Consistent with previous findings [16,18], genetic-supported genes were strongly associated with clinical success (Figure 2B, OR for approved targets = 5.94, p-value = 1.8e-11). Cell type and disease specific protein-coding genes were as likely to be targets of drugs passing phase I and II as those that have genetic support. In contrast, for targets that are clinically approved (i.e. passed phase III), genetic evidence gave stronger prediction. The identification of genetic evidence as a predictor of clinical success may have biased recent programs toward development of genetically supported drugs, noting that only a subset of the drugs under consideration here were approved in the last 10 years (Supplementary Figure 4).

We observed several differences between scRNA-seq supported targets and targets supported by genetics. Firstly, scRNA-seq supports a larger number of successful target-disease pairs. Amongst the G-D space of safe targets (2925 G-D pairs), 29.3% are scRNA-seq supported, while only 2.3% are directly supported by genetics (Figure 2A). Secondly, we found that different sources of omic evidence support distinct target spaces: only 24% of safe G-D pairs targeted with genetic support overlap with either kind of scRNA-seq evidence (Figure 2C). We tested for association between clinical success and support from both genetic and

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

scRNA-seq, but due to the limited overlap, this analysis likely lacked sufficient statistical power to detect significant differences compared to using genetics alone (Supplementary Figure 5A). Thirdly, genetic and scRNA-seq support were predictive of clinical success in different classes of tractable targets (Supplementary Figure 5B). Genetic support increased chances of approval up to 20-fold for kinases and catalytic receptors but was notably less predictive of success than scRNA-seq support for other classes, such as transporters and rhodopsin-like GPCRs. These classes of genes show high tolerance to loss-of-function mutations (Supplementary Figure 5C), whereas it has been reported that genes associated with GWAS variants are under strong evolutionary constraints [27]. Furthermore, at the compound-level we found that drugs targeting scRNA-seq supported genes are approved or in trial for a significantly higher number of indications, compared to not supported targets (Adjusted $R^2 = 0.167$; p = 2.086e-7, see Methods) (Supplementary Figure 6; Supplementary Table 5). Genetic association was not associated with significantly higher number of indications per drug.

We also observed significant differences when considering the genes with omic support that are not already in clinical development (unexplored supported genes). A large fraction of scRNA-seq supported genes, and especially cell type specific genes, are considered tractable by therapeutic agents (Figure 2D). Across all diseases considered, on average 77% of cell type and disease cell specific genes are antibody tractable, against 51% of genes supported by genetic association (t-test p-value: 5.9-e08). Genetic-supported genes showed a slightly higher average fraction of small molecule tractable genes (40% against 31%, t-test p-value = 0.02), although this was mainly driven by a few diseases (Supplementary Figure 7A). This indicates that scRNA-seq support prioritizes genes with therapeutic potential, especially membrane-bound proteins. This difference between genetic and scRNA-seq support could at least in part be explained by differences in evolutionary constraints: antibody tractable genes have significantly higher tolerance to loss-of-function than non-tractable genes, while small molecule tractable genes are significantly more constrained (Supplementary Figure 7B). This could be due to stronger evolutionary constraints on the sequences of proteins with small molecule binding pockets, as compared to larger, flatter surfaces of protein-protein interaction interfaces [28].

Robustness of association of scRNA-seq support and clinical success

We next tested the robustness of association with clinical success to several parameters used for the definition of genes with scRNA-seq support. Firstly, in our scRNA-seq analysis workflow we do not test for differential expression across all genes, but we pre-select highly variable genes before each comparison (see Methods), as per standard practice for DE analysis [29]. To independently quantify the impact of feature selection before DE analysis, we computed enrichment of successful targets considering only genes selected as highly variable genes for each disease scRNA-seq dataset. DE testing led to significant enrichment of successful targets also within selected HVGs, although with lower odds-ratios (Supplementary Figure 8A). This suggests that both HVG selection and DE testing on scRNA-seq data enrich for successful targets.

Next, we explored the relationship between cell type specificity and differential expression fold change between cell types and disease conditions. Estimated fold changes in gene expression between cell types are higher than those observed in the comparison between disease and healthy states within cell types (Supplementary Figure 8B). Notably, genes significantly over-expressed in a cell type at lower log-fold changes are often ubiquitously highly expressed, while those at higher fold changes are genuinely cell type specific

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

(Supplementary Figure 8C) and more likely to be successful targets (Supplementary Figure 8D, left). Conversely, most disease cell specific genes, including successful clinical targets, are over-expressed in disease patients at low fold changes (Supplementary Figure 8D, right)

According to our definition, disease cell specific genes include both those over-expressed in disease within one or a small subset of cell types and genes over-expressed across multiple cell types. Since the latter category may also be identifiable through bulk expression analysis on whole tissue, we explored whether both tissue-level and cell type-level DE genes contribute to the enrichment of clinically successful targets. To explore this, we aggregated scRNA-seq counts to estimate bulk tissue expression per donor and compared this to genes specifically pinpointed through cell type-aware DE analysis (Supplementary Figure 9A). 74% of disease cell specific successful targets (passing at least phase I) could be identified only with cell type-level DE analysis (Supplementary Figure 9B). In other words, single cell rather than bulk expression data is required to identify most disease cell specific genes. Both tissue-level and cell type-level disease cell specific genes were significantly more likely to be targets of successful drugs (Supplementary Figure 9C). The OR was slightly higher for tissue-level disease markers compared to those only detectable with cell type-aware analysis. This is expected, since bulk expression profiling methods have been incorporated in target discovery pipelines for many years, whilst single cell data has only become available more recently. In addition, we confirmed that drug targets are more strongly enriched in up-regulated genes than down-regulated genes (Supplementary Figure 9A). This aligns with the fact that 890 (73.0%) of 1219 drugs past phase I and 474 (69.5%) of the 695 drugs in phase III or phase IV trials for the diseases in this analysis are categorized as inhibitors, antagonists, degraders, blockers and/or negative regulators of their targets.

We note that our analysis may be constrained by a lack of consistently curated cell type annotations across various scRNA-seq disease datasets. We use cell type labels based on the Cell Ontology [23], leading to broad and possibly inconsistent cell type annotations. The preferred annotation strategy in several data integration studies which re-use public scRNA-seq data is to cluster gene expression profiles in different datasets *de novo* and manually re-annotate clusters [30,31]. We hypothesised that accurate cell type annotations could further improve the ability to prioritize cell type specific genes for target discovery. We explored this hypothesis through analysis of three lung diseases (pneumonia, cystic fibrosis and pulmonary fibrosis) for which curated fine-grained annotations from data integration projects are available in the extended Human Lung Cell Atlas (eHLCA) dataset [30] (Supplementary Figure 10A). We computed cell type specific and disease cell specific genes using Cell Ontology-based annotations and eHLCA fine annotations and compared the enrichment of successful targets between these two gene sets. The gene sets with scRNA-seq support testing on fine or coarse annotations was largely overlapping (Supplementary Figure 10B). The fraction of recovered successful targets and the odds of clinical success were comparable, with slightly increased odds of success by using fine annotations to detect cell type specific genes (Supplementary Figure 10C). For disease specific expression the odds of success were slightly decreased with fine grained annotation, possibly because in this case differences between health and disease may manifest as changes in cell type proportions rather than within-cluster differential gene expression.

Target analysis in diseases with scRNA-seq support

Considering the 24 diseases with at least one target with an approved drug, genetic support was significantly associated with clinical success (targets of effective drugs) for 6 indications, cell type specificity for 10 indications and disease cell specificity for 9 indications (Figure

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

3A, Supplementary Figure 11, Supplementary Table 6). We considered technical factors influencing the variability across diseases in targets supported by scRNA-seq. Firstly, the total number of supported targets correlates with the number of cell types considered in differential expression analysis (Supplementary Figure 12A). For disease cell specific genes, the number of cell types that can be tested is significantly dependent on the number of disease patients in the scRNA-seq cohort ($R^2 = 0.39$, p-value = 1.87e-11). Indeed, we found that with a larger patient cohort we detected more disease cell specific genes (Supplementary Figure 12B). Moreover, when the datasets included at least 10 disease patients, a greater proportion of the supported genes were successful targets (Supplementary Figure 12C). These results support the notion that larger patient cohorts can improve accuracy of detection of disease cell specific targets. Conversely, cell type specific genes appear less dependent on the numbers of donors for the disease-relevant tissue dataset (Supplementary Figure 12B).

As an exemplar disease with high-quality scRNA-seq data, we examined the characteristics of supported targets for systemic lupus erythematosus (SLE). SLE, commonly referred to as lupus, is a chronic autoimmune disease that can affect various organs and tissues. SLE is characterised by auto-antibody production that triggers inflammation and tissue damage. Current therapy options for SLE include broad acting non-steroidal anti-inflammatory drugs, corticosteroids, and immunosuppressants such as methotrexate and azathioprine to control the immune system's activity. In addition, newer cell-targeted biologics like belimumab, which targets B-lymphocyte stimulator protein encoded by *TNFSF13B*, have been approved for treating certain patients with SLE [32,33].

In SLE many genes have been associated to the disease through genetic analyses (Supplementary Figure 2). However, these genes are not significantly enriched for effective drug targets (Figure 3A). Disease cell specific genes point to drugs with systemic immuno-suppressant effects such as paracetamol (targeting *FAAH, PTGS2*), inhibitors of DNA replication (targeting polymerases and tubulin genes), and B cell stimulators (targeting *TNFSF13B, CD40LG*) (Figure 3B). Cell type specific known targets include genes acting in disease-relevant cells, such as toll-like receptors which are involved in autoantibody production in B cells [34]. The unexplored supported genes prioritized by different omic support classes are all enriched in immune-function gene sets. However, we noticed that different data prioritizes genes with distinct molecular function (Supplementary Figure 13A-C). For example, different support classes prioritize different genes involved in interferon gamma signalling: genetic association prioritizes genes encoding for DNA binding proteins and transcription factors in the pathway, including SMAD and IRF transcription factors; disease cell specific genes are induced by interferon signalling downstream in the pathway, including *IFIT* and *ISG* genes. Cell type specific genes include chemokines and membrane bound receptors (e.g. *KLRK1, CMKLR1, IL2RB*) (Supplementary Figure 13D).

As a second example, we examined supported targets in pulmonary emphysema. Pulmonary emphysema is a condition characterized by the gradual destruction of the air sacs (alveoli) in the lungs, resulting in enlarged and rigid air spaces that impair gas exchange [35]. When pulmonary emphysema is coupled with inflammation of the airways, the two conditions are known as chronic obstructive pulmonary disease (COPD). The primary therapy options include bronchodilators, such as short- or long- acting agonists of beta-2-adrenergic receptors that cause the relaxation of airway smooth muscles and anticholinergic medications that inhibit bronchoconstriction [36]. Oral phosphodiesterase protein family inhibitors such as Roflumilast are similarly used to manage smooth muscle relaxation, vasodilatory, and

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

bronchodilatory effects in patients with pulmonary emphysema and COPD. Inhaled corticosteroids may be used as an add-on therapy to reduce local inflammation.

In our analysis, known drug targets were not supported by direct genetic evidence (Figure 3A). Given that pulmonary emphysema is a stage of a progressive lung disease, the absence of robust genetic evidence could be attributed to limited size of patient cohorts at this specific stage of disease. Despite single cell data being available only from 3 patient samples, multiple safe, effective, and approved therapeutic targets were prioritised using our analysis as cell type specific in the disease-relevant tissue (lung) (Figure 3C). For example, angiotensin II receptor (encoded by *AGTR1* gene) antagonist Sacubitril/Valsartan is an effective drug in patients with pulmonary hypertension/emphysema [37], despite it being predominantly used for the treatment of cardiac diseases. Even though *AGTR1* lacked genetic association with lung disease or function, our analysis suggests that *AGTR1* is specifically expressed in lung smooth muscle cells and fibroblasts in scRNA-seq data (Supplementary Figure 14). *AGTR1* presents an example of targets where single cell data analysis might enable interpretability of cell type relevance for disease progression.

We also found that for broad therapeutics that affect a family of genes, single cell data could provide evidence for the most relevant family members based on specificity of expression in the disease-relevant tissues. For example, the non-selective inhibitor Roflumilast targets all phosphodiesterase-4 genes (*PDE4A-D*), however, only *PDE4C* shows selective expression in activated smooth muscle cells and alveolar type 2 cells in the lung (Supplementary Figure 14). Non-selective inhibitors can cause multiple side effects. In the case of Roflumilast, expression of *PDE4B* and *PDE4D* in the sensory nerves is thought to be responsible for nausea side effects [38,39]. Therefore, single cell data can provide rationale for development of selective *PDE4C* inhibitors for the treatment of pulmonary emphysema and other lung conditions associated with hypertension.

Discussion

Lack of efficacy and safety are the leading causes for phase II and III clinical trial failures [40]. Additionally, a promising target may fail to progress to phase I because of multiple reasons. These include inability to establish a mechanistic link between target biology and indication (target validation failure), insufficient promising chemicals, and/or safety risks found during pharmacokinetic and early toxicology studies [4]. Taken together, all these different causes account for the limited probability of a candidate therapeutic target and its cognate drug passing all stages of pre-clinical, clinical research, and regulatory approval (2005-2010 industry average: 5% [4]). Data-driven frameworks in drug discovery can effectively mitigate some of these risks, as demonstrated by the use of genetics data to support target-disease associations [16], but attrition from target ID to clinic remains high [4]. To further increase chances of success, target discovery workflows increasingly access additional information aggregated from pre-clinical data resources, including data from animal models, over-expression in disease-relevant bulk tissue samples, disease pathway analyses, and other bioinformatics resources, as exemplified by the Open Targets Platform [1]. Characterizing the potential impact and biases of different data sources for target credentialing pipelines is critical to push new technologies to translational applications.

Single-cell technologies, along with the growing availability of large, shared single-cell datasets on diseases and healthy controls [21] have opened-up unprecedented opportunities to understand target biology at cellular resolution across disease areas and in diverse patient

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

populations. Single-cell RNA-seq has been applied to investigate pathways driving onset and progression of diseases [41–43], to understand the mechanism of action of different therapeutics [44,45], and to discover biomarkers for patient stratification [46]. This suggests a remarkable depth and breadth of information extractable from scRNA-seq datasets that could support drug discovery.

The goal of this study was to measure how much using single-cell RNA sequencing data from disease-relevant tissues can improve the chances of success for therapeutics by systematically identifying connections between targets and diseases. By aggregating data for 30 diseases affecting 13 tissues, we found that candidate target genes supported by scRNA-seq evidence have approximately three times the chances to lead to clinically successful therapies (Figure 2B).

The association between scRNA-seq support and target clinical success is in line with the fact that human diseases are typically tissue and cell type specific [47]. For example, tissue and cell-type specific eQTLs are enriched for disease-associated SNPs [48–50]. Given the typical timeframes of drug development, it is highly unlikely that any of the targets considered have been initially prioritized or validated using single-cell transcriptomics. While it is possible that other types of tissue-level transcriptomic data have driven decisions in target development, we do not expect these instances to significantly bias the results of our analysis on cell type specific expression. Furthermore, we found that scRNA-seq supported targets were more likely to pass phase I and II than reaching approval. It is possible that cell type specificity is a better indicator of low toxicity than broad efficacy, although this question remains to be further explored.

We compared targets prioritized by scRNA-seq with those prioritized by genetic evidence, which has been highlighted as an important predictor of clinical success [16,18]. Consistent with previous results, for the diseases and target sets included in this analysis, we observed a strong and statistically significant association between direct genetic support for target-disease pairs and clinical development success (Figure 2B). Previous work has highlighted that targets supported by human genetic data are more likely to be successful [16]. It is likely that this has led the pharmaceutical industry to allocate greater resources to development of drugs for these targets and has therefore created a bias amongst the targets in clinical development. However, we also find that direct genetic association support exists only for a subset of target-disease pairs with drugs in clinical development, and scRNA-seq support exists for a larger set of target-disease pairs, with few targets supported by both types of omic evidence (Figure 2A; Supplementary Figure 4). These complementary sets of targets have distinct molecular and druggability characteristics (Figure 2C, Supplementary Figure 5B). For example, we observed that genetic support tends to prioritize evolutionarily conserved genes (Supplementary Figure 5B-C, Supplementary Figure 7B), as previously reported [27]. Loss-of-function-tolerant classes of druggable targets, such as GPCRs and transporters, are instead prioritized by cell type or disease cell specificity, although scRNA-seq data might be biased towards other classes, such as highly expressed genes. We speculate that cell type specificity might prioritize targets of therapies managing symptoms or modulating disease-relevant biological processes parallel to or downstream of genetic causation, which are seldomly prioritized by genetic analysis [19,20]. Importantly, detecting associations between genetic variants and disease requires data from hundreds to thousands of individuals. In our analysis, association between clinical success and scRNA-seq support was drawn from analysis of tissue from tens of individuals, and we show that increasing the

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

size of the scRNA-seq cohort to hundreds of patients increases the fraction of prioritized successful targets even further (Supplementary Figure 12C).

In this study, we considered two distinct patterns of cell type specific expression: cell type specific expression in disease-relevant tissue (*cell type specificity*) and cell type specific over-expression in disease-relevant tissue from disease patients compared to controls (*disease cell specificity*). Both classes of genes were significantly associated with clinical success in several diseases (Figure 3A). Cell type specific targets were less dependent on technical features of the scRNA-seq dataset (Figure 2A, Supplementary Figure 12). This is important because measuring cell type specificity does not require patient data, and this could be computed systematically on open resources such as the Human Cell Atlas Data Portal (data.humancellatlas.org) or the CZ CellxGene database [21].

When considering disease cell specific genes, we found that both genes over-expressed in disease within small subsets of cell types, and genes over-expressed at tissue-level, contribute to the association with clinical success (Supplementary Figure 9). Bulk transcriptomics methods have been used for longer in clinical development pipelines and this is reflected in stronger associations with success, although most disease cell specific successful targets were only identified with cell type-aware analysis. Of note, in this study we define disease cell specificity with naïve cell type-level differential expression analysis, where technical effects are only partially mitigated. We expect that improved experimental design and statistical methods to recover expression differences in scRNA-seq in normal and diseased tissues [51–53] and to distinguish disease-associated cell states [54–56] could further improve the set of target genes and will be highly impactful for target discovery programmes.

Our study is not free of limitations. We rely on the Cell Ontology-based cell type labels [23] provided by data curators upon submission to the CZ CellxGene Discover database. This approach has two primary drawbacks. Firstly, the Cell Ontology's incompleteness may result in labelling rare tissue-specific subpopulations with broad cell type terms. Secondly, inconsistencies may arise as different data curators use the same term for transcriptionally distinct cells or conflicting terms for identical phenotypes. While our label harmonization strategy addresses the latter issue to some extent, it introduces coarser annotations. We anticipate that these issues will be mitigated by increased availability of expertly curated cell type annotations across human tissues, and by unified models for cell type annotation [57]. These will not only enhance the identification of promising drug targets (Supplementary Figure 10) but also facilitate more precise identification of disease-relevant cell types and cellular mechanisms. Additionally, our analysis encompassed both historical and active clinical development data for drug targets, for some of which the ultimate outcomes are still unknown. Finally, we did not account for the similarity between indications, which is important when considering related diseases where genetic association may be lacking for a specific indication (e.g. pulmonary emphysema) but is present for related traits (e.g. lung function).

Looking forward, more sophisticated analyses of cell atlases will boost further drug discovery efforts. For example, analysis of drug target expression patterns across cell types have been used to assess re-purposing potential and on-target toxicities [58]. Methods to infer differentiation trajectories [59,60], cell-cell interactions [61,62], regulatory networks [63], and immune repertoires [64] provide additional unexplored space for novel targets. Furthermore, we envision that high-resolution spatial transcriptomics will provide an added level of insight into drug target relevance based on their expression and disease tissue context

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

[65–67]. Insights on cell and disease cell specific targets gained using high-throughput genomics will inform the design of next generation precision therapeutics, for example antibody-drug conjugates or lipid nanoparticle-mRNA vaccines. Overall, our study provides a framework to assess the potential impact of alternative data analysis methods and modalities on target discovery.

In summary, our work indicates that single-cell data can be a valuable tool for guiding the process of drug target prioritisation and enhancing our understanding of the cellular basis of safe, effective, and approved treatments for diseases.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Main figures

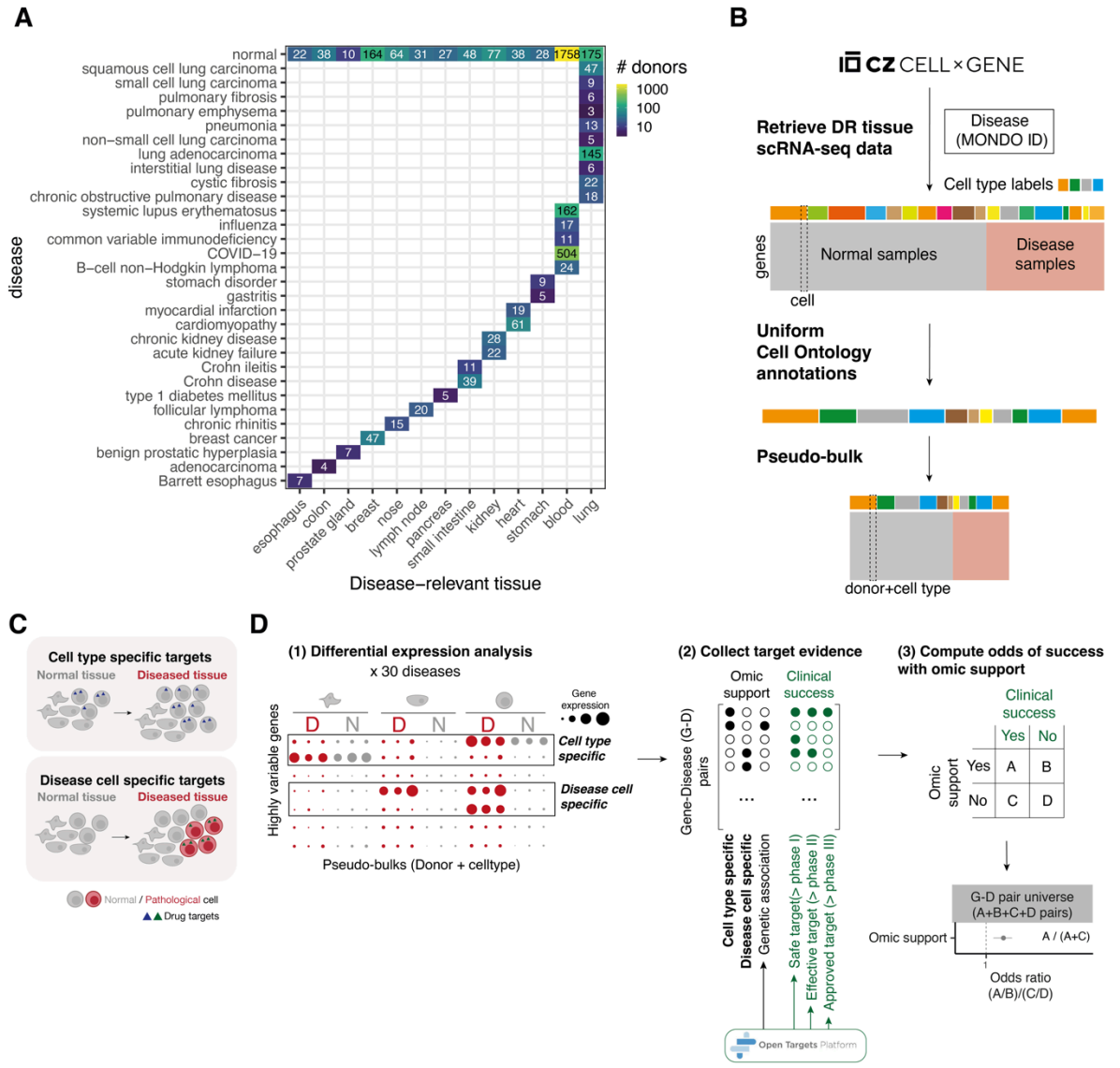


Figure 1: Single-cell dataset selection and pre-processing. (A) Overview of diseases and tissues in scRNA-seq dataset. Table of disease-relevant tissue of samples (x-axis) and disease condition (y-axis) for all scRNA-seq data considered in this study. The number and color of each square indicates the number of individuals for whom scRNA-seq data are available. The availability of data from healthy individuals is shown in the top row (disease condition: normal). (B) Illustration of selection and pre-processing steps for scRNA-seq datasets from CZ CellxGene Discover database (DR: disease-relevant). (C) Illustration of rationale behind scRNA-seq support classes for target discovery: cell types expanding or acquiring aberrant function in disease can be targeted using cell type specific targets. Cells specifically expressing aberrant gene programmes can be targeted with disease cell specific targets (D) Workflow for analysis of association between scRNA-seq support and clinical success. We identify cell type specific and disease cell specific gene-disease pairs through differential expression analysis on pseudo-bulked data from the disease-relevant tissue (1). Data on genetic association and clinical success of targets was collected from the OpenTargets database (2). For each omic support class, we compute the odds ratio for the association between clinical success (passing clinical trials) and different classes of omic support (3).

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

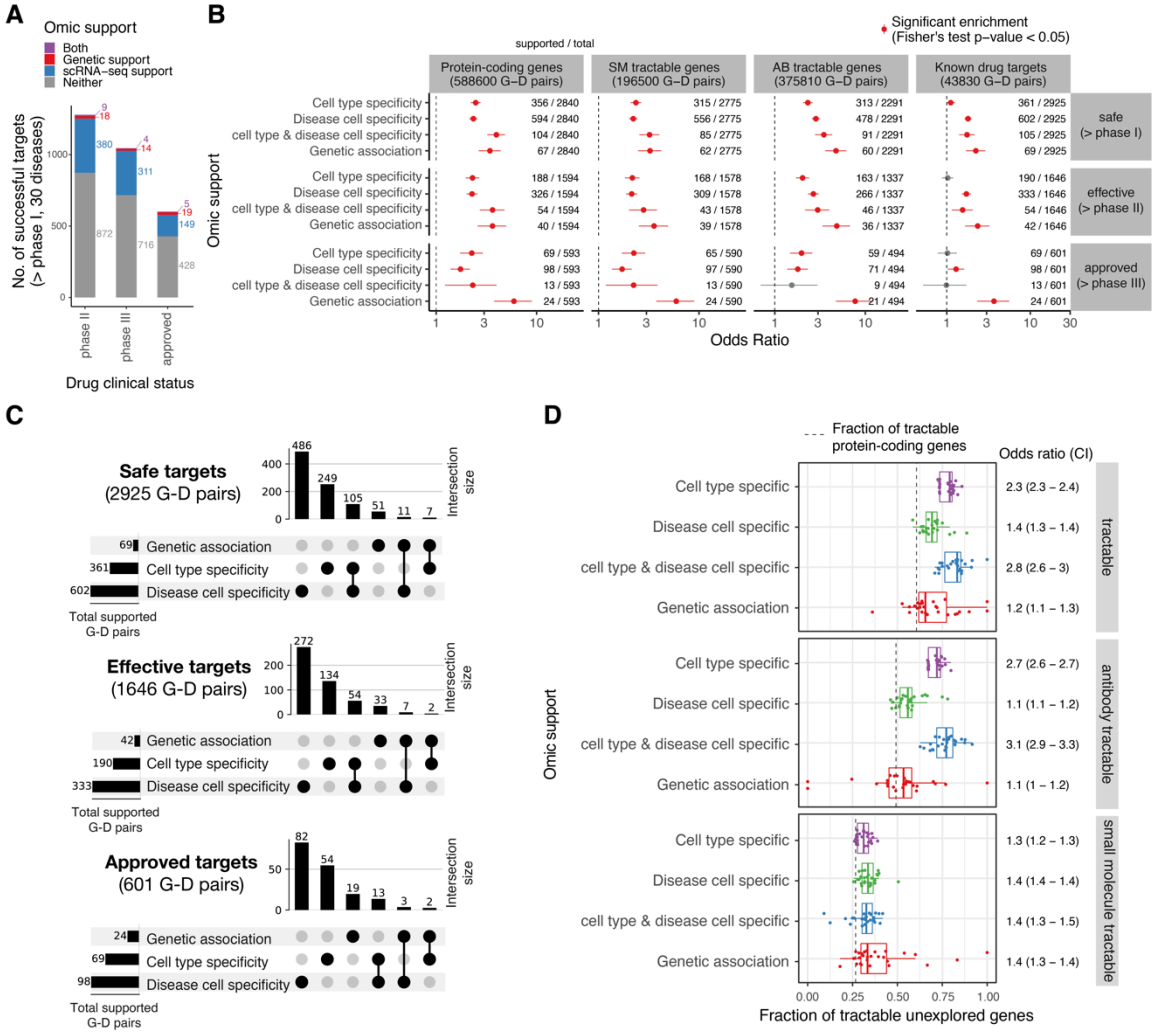


Figure 2: Association between omic-based evidence and target clinical success. (A) Barplot of successful phase I, II and approved target-disease pairs for 30 diseases, colored by type of omic support. Target-disease pairs are grouped by the highest clinical phase reached by the therapeutic agent. (B) Odds ratio (x-axis, in log10 scale) of association between clinical success of a target and different sources of omic support (y-axis). We test association with safe targets (passed phase I, top row), effective targets (passed phase II, middle row) and approved targets (passed phase III, bottom row). Results using different universes of genes are shown in different columns (SM: small molecule, AB: antibody). For each test, the numbers to the right show the number of omic supported over total successful targets. Results are shown considering gene-disease (G-D) pairs for 30 diseases. The error bars denote 95% confidence intervals of the odds ratio. Points in red indicate cases where the enrichment for successful targets was statistically significant (Fisher's exact test p-value < 0.05). The dotted line denotes Odds Ratio = 1 (no enrichment). (C) Upset plots showing the number of successful G-D pairs with omic support (left barplot) and their intersection (top barplot). We show intersection for all safe, effective, and approved targets. (D) Boxplots showing the fraction of unexplored supported genes (not clinically tested drug targets, x-axis) for each class of omic support (y-axis) that are considered tractable based on the Open Targets tractability assessment. Each point represents a disease. Odds ratios and 95% confidence intervals for association between omic support and tractability are shown to the right (considering all protein-coding genes as universe). We distinguish genes that are antibody-tractable, small molecule tractable, or tractable by any class of therapeutic (tractable). The dotted line shows the fraction of tractable genes amongst all protein-coding genes. 27 diseases for which at least one gene had genetic association evidence are shown. In the boxplots, the center line denotes the median; the box limits denote the first and third quartiles; and the whiskers denote 1.5x the interquartile range (IQR).

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

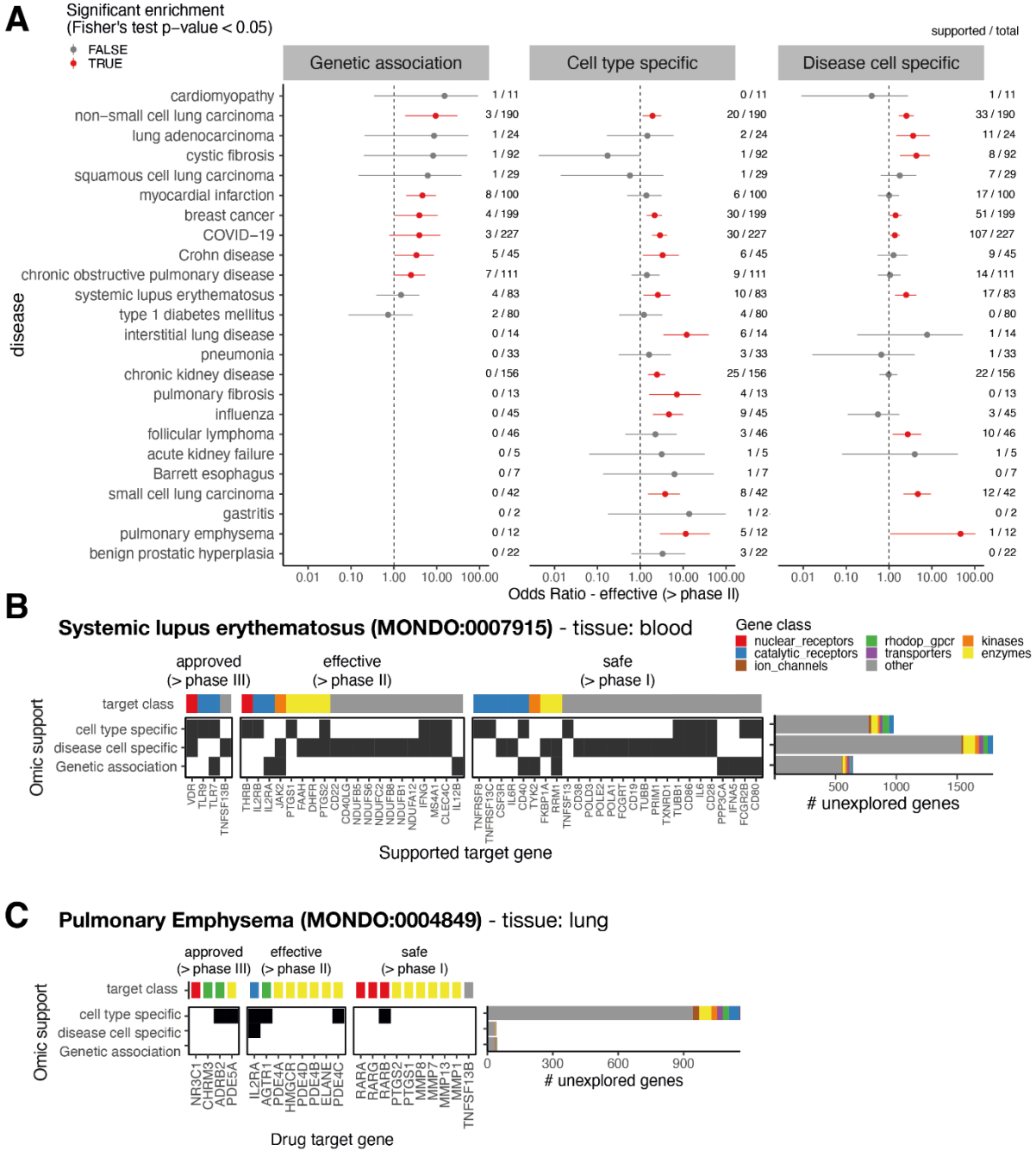


Figure 3: Association between omic support and clinical success stratified by disease. (A) Odds ratio (x-axis, log10 scale) of association between clinical success (effective, passing phase II) of a target and omic support calculated per disease (y-axis). Results are shown for 24 diseases with at least one approved target. Diseases are sorted by odds ratios for association with genetic support. The gene universe used was protein-coding targets. For each test, the numbers to the right show the number of omic supported over total successful targets. The error bars denote 95% confidence intervals of the odds ratio. Points in red indicate cases where the enrichment for successful targets was statistically significant (Fisher's exact test p-value < 0.05). The dotted line denotes Odds Ratio = 1 (no enrichment). (B-C) Supported drug targets for systemic lupus erythematosus (B) and pulmonary emphysema (C). The right barplot shows the number of supported target genes that are not known drug targets in clinical trial (unexplored, including both tractable and non-tractable genes). In (B) only known targets supported by at least one omic class are shown.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Methods

Single-cell RNA-seq data collection from CZ CellxGene Discover platform

To select a set of diseases and scRNA-seq datasets, we downloaded cell- and dataset-level metadata for all *H.Sapiens* datasets from the CZ CellxGene Discover database, using the *cellxgene_census* python API (census version: 2023-07-25) [21]. Disease-relevant (DR) tissues were manually annotated for the 58 disease terms in the database. We excluded datasets profiled with targeted scRNA-seq assays (BD Rhapsody), inDrop and STRT-seq. We further excluded fetal samples, based on Human Developmental Stage Ontology [68], where available, and by manual curation for 12 datasets where stages were annotated as "unknown". 10 disease terms were grouped into 4 broader terms (Supplementary Table 1).

After curation, 30 disease terms were retained for association analysis. Reasons to exclude diseases included: missing overlapping disease terms in Open Targets, missing data from DR tissue, data available from less than 3 donors with the disease, download errors (see Supplementary Table 1 for a complete list of diseases and reasons to exclude from analysis). After selecting suitable datasets, for each disease we downloaded full transcriptome gene expression profiles for all cells from the DR tissue from healthy donors and disease patients, as well as cell type labels (Cell Ontology terms [23]) and sample-level technical metadata (scRNA-seq assay and suspension type, Supplementary Figure 15).

To ensure consistency in granularity of cell type annotations across studies, we implemented a rollup procedure on the Cell Ontology tree, by relabelling cells with parent terms if a given term is a descendant of another term in the dataset (see example outcome in Supplementary Figure 1). For each term, the search for parent terms was limited only to a level of depth in the ontology tree given by the total number of ancestors of the term divided by a factor of 5. For example, if a term had 20 ancestors in the ontology tree, we searched for the 4 closest parent terms in the dataset for relabelling. We recognize that this step reduces the resolution of cell type annotations, yielding broader and partially redundant annotation labels. However, it mitigates the need for batch correction, clustering, and manual cell type annotation across 30 datasets. We defined the cell type labels used after roll-up as *high-level cell type annotations*.

Differential expression analysis and extraction of scRNA-seq supported gene-disease pairs

We identified cell type specific and disease cell specific genes for each disease using differential expression (DE) analysis.

For each disease dataset, we aggregated cell-level gene expression profiles summing counts and size factors (total counts per cell) by donor and high-level cell type annotations (hereafter, pseudo-bulks), following best practice recommendations for DE analysis on scRNA-seq data [69,70]. Only cell types found in at least 3 healthy donors (and 3 disease donors for disease cell specificity analysis) were included in DE testing. To identify cell type specific genes, we selected pseudo-bulks from healthy donors from the disease-relevant tissue and we tested for DE between pseudo-bulks of one cell type against all other cell types. To identify disease cell specific genes, for each cell type we tested for DE between diseased donors and healthy donors. For each test, we selected the top 5,000 highly variable genes amongst considered pseudo-bulks, using the method implemented in the R package *scran* [71]. We tested for differential expression between groups with the *edgeR* quasi-likelihood test [72] using the implementation in the R package *glmGamPoi* [73]. In all tests, we

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

modelled the number of cells per pseudo-bulk as a confounder, as well as suspension type (cell or nuclei) and scRNA-seq assay where possible (when the confounder was not perfectly collinear with the disease label). After DE analysis, we obtained the effect size (log-fold change, logFC) and Benjamini-Hochberg adjusted p-values for each tested gene in each tested cell type.

We annotated a gene-disease (G-D) pair as cell type specific when the gene is significantly over-expressed in at least one cell type compared to all other cell types in healthy disease-relevant tissue (adjusted p-value < 0.01, logFC > 5). The choice of logFC threshold was motivated by the observation that genes significantly over-expressed at lower log-fold changes are often ubiquitously highly expressed, while those at higher fold changes are genuinely cell type specific (Supplementary Figure 8C). We annotated a G-D pair as disease cell specific when the gene is significantly over-expressed in disease in at least one cell type in disease-relevant tissue (adjusted p-value < 0.01, logFC > 0.5). The total number of supported G-D pairs for each disease is shown in Supplementary Figure 2. We annotated a G-D as cell type and disease cell specific if supported by both classes of scRNA-seq support.

Known drug relationships from Open Targets

Open Targets direct association evidence was accessed via download from the Open Targets Platform (version 23.02) [1,25]. Downloads used for this analysis were the 'Diseases' and 'Direct Associations by Type' tables. Experimental Factor Ontology (EFO) disease terms used in Open Targets were mapped to their corresponding term in used in the CellxGene database (MONDO IDs) using the ontology tree available in the Open Biological and Biomedical Ontology Foundry (https://obofoundry.org/ontology/mondo.html). We annotated G-D pairs for which approved or clinical candidate drugs exist using the ChEMBL evidence score from the Open Targets Platform. Briefly, each G-D pair is assigned a score between 0 and 1 based on clinical precedence, then the score is down-weighted by half if the clinical trial has stopped early for negative results (no effect of the drug) or safety and side effects concerns. Following the ChEMBL evidence scoring in Open Targets (https://platform-docs.opentargets.org/evidence#chembl), we classified G-D pairs with a ChEMBL evidence score > 0.1 as safe (> phase I), pairs with score > 0.2 as effective (> phase II), and pairs with score > 0.7 as approved (> phase III). While we do not explicitly exclude gene-disease pairs supported by failed trials, the down-weighting in Open Targets ensured that targets failed in early clinical trials are excluded, and targets failed in phase III were at most classified as passing phase II.

Genetic association

We annotated G-D pairs with genetic support using the genetic direct association score provided in Open Targets, aggregating evidence for association of genes and rare and common variants from several sources (https://platform-docs.opentargets.org/evidence) [1]. We classified as supported by genetics any G-D pair with genetic association score > 0.

Association between omic evidence and clinical success

To test for association between omic evidence (cell type specificity, disease cell specificity, genetic association) and clinical success (passing clinical phase I, II or III) we computed the odds ratio and Fisher exact test p-value under the null hypothesis that the true ratio between the odds of being a successful G-D pair with omic support and of being successful without support is 1. In all association tests, drug indications for clinical success and data for omic

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

support are aligned by disease. To compute odds ratios, 95% confidence intervals and p-values, we used the odds ratio calculation implementation in the python package *scipy* [74].

To enumerate the space of possible G-D pairs for odds ratios analysis, we used the following gene sets as "gene universes": protein-coding genes (N=19620) were obtained from Ensembl v108; antibody-tractable (N=12527) and small molecule-tractable (N=6550) genes, based on the Open Targets' druggability assessment (https://platform-docs.opentargets.org/target/tractability), were obtained from Minikel et al. [18]; Genes targeted by therapies in clinical trial for any indication (known drug targets, N=936) were obtained from Open Targets v23.02; sets of typically druggable targets (Supplementary Figure 5B-C) were obtained from Minikel et al. [18]. Unless otherwise specified, odds ratios shown in the manuscript were computed using protein-coding genes as the gene universe.

Drug-level analysis

We extracted compound-level data from Open Targets for 17,095 drug molecules together with their year of first approval, list of indications, list of targets, and maximum clinical phase using Open Targets "molecule" and "mechanismOfAction" data objects. Among these drugs, we then identified those that had in their approved or investigational indications list any of the 30 diseases considered in the target-level analysis (n = 2358 drugs) and then further narrowed this list of drugs to those in phase II or greater (n = 1219) and phase III or phase IV clinical trials for the 30 diseases considered in this analysis (n=695). Drugs were annotated as having single cell or direct genetic association support for the considered indications if any of their target-disease pairings had this evidence in the preceding target-disease evidence analysis. To examine the number of indications for each drug for one of the 30 diseases in our analysis with genetic or scRNA-seq support, we aggregated Open Targets drug information and counted the total number of approved or investigational indications for each of these drugs.

We used a multiple linear regression model to investigate the possible associations of single cell support, and direct genetic support with the number of indications approved or under investigation per drug, accounting for year of the clinical trial as a confounder (Supplementary Figure 6). To satisfy model assumptions, log(number of indications per drug) was used as the dependent variable to address right-skew in number of indications. Single cell and genetic evidence could be synergistic, so an interaction term was used between these during modelling (Supplementary Table 5).

Comparison of fine annotation and ontology-based annotation on lung diseases

To compare gene-disease pairs prioritized with ontology-based annotation and with uniform integration-based annotations, we downloaded the extended Human Lung Cell Atlas (eHLCA) [30] using the CellxGene census API (CellxGene census datasetID: 9f222629-9e39-47d0-b83f-e08d610c7479), selecting normal lung and patient data for 3 diseases (pneumonia, cystic fibrosis and pulmonary fibrosis). These diseases were selected because all scRNA-seq data considered in the ontology-based analysis was included in the eHLCA dataset, therefore allowing us to compare the impact of annotations on matched data. We pseudo-bulked each disease dataset using the finest author-provided annotation (column: *ann_finest_level* in CellxGene metadata) and performed differential expression analysis as described above.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Disease-specific target analysis

To categorize the targets supported by different classes of omic evidence in systemic lupus erythematosus and pulmonary emphysema, we used the annotation of tractable gene classes as defined by Minikel et al. [18]. Gene ontology enrichment analysis was performed using the Enrichr method [75] as implemented in the Python package *GSEApy* [76]. The categorization of IFN-gamma pathway genes into receptors, transcription factors, targets, and secreted proteins (Supplementary Figure 13D) was obtained from OmniPath [77] and Dorothea [78,79].

Data availability

All scRNA-seq data analysed in this study is available via the CZ CellxGene Discover database and CxG Census API (https://chanzuckerberg.github.io/cellxgene-census/, version: 2023-07-25). Data on clinical precedence for known drugs for each target-disease pair, as well as gene-disease genetic association scores, was downloaded from Open Targets (version 23.02, https://platform.opentargets.org/downloads/data). Data on gene tolerance to loss-of-function mutations (LOEUF, loss-of-function observed/expected upper bound fraction) was extracted from gnomAD.v2.1's pLoF metrics by gene data [80] (https://gnomad.broadinstitute.org/downloads). Gene sets used as universes for association analysis are available at https://github.com/emdann/sc_target_evidence/blob/master/data/universe_genes.csv. Processed datasets and analysis outputs are available as supplementary tables and via figshare (doi:10.6084/m9.figshare.25360129).

Code availability

All code to reproduce data downloads, processing and analysis is available at https://github.com/emdann/sc_target_evidence.

Acknowledgements

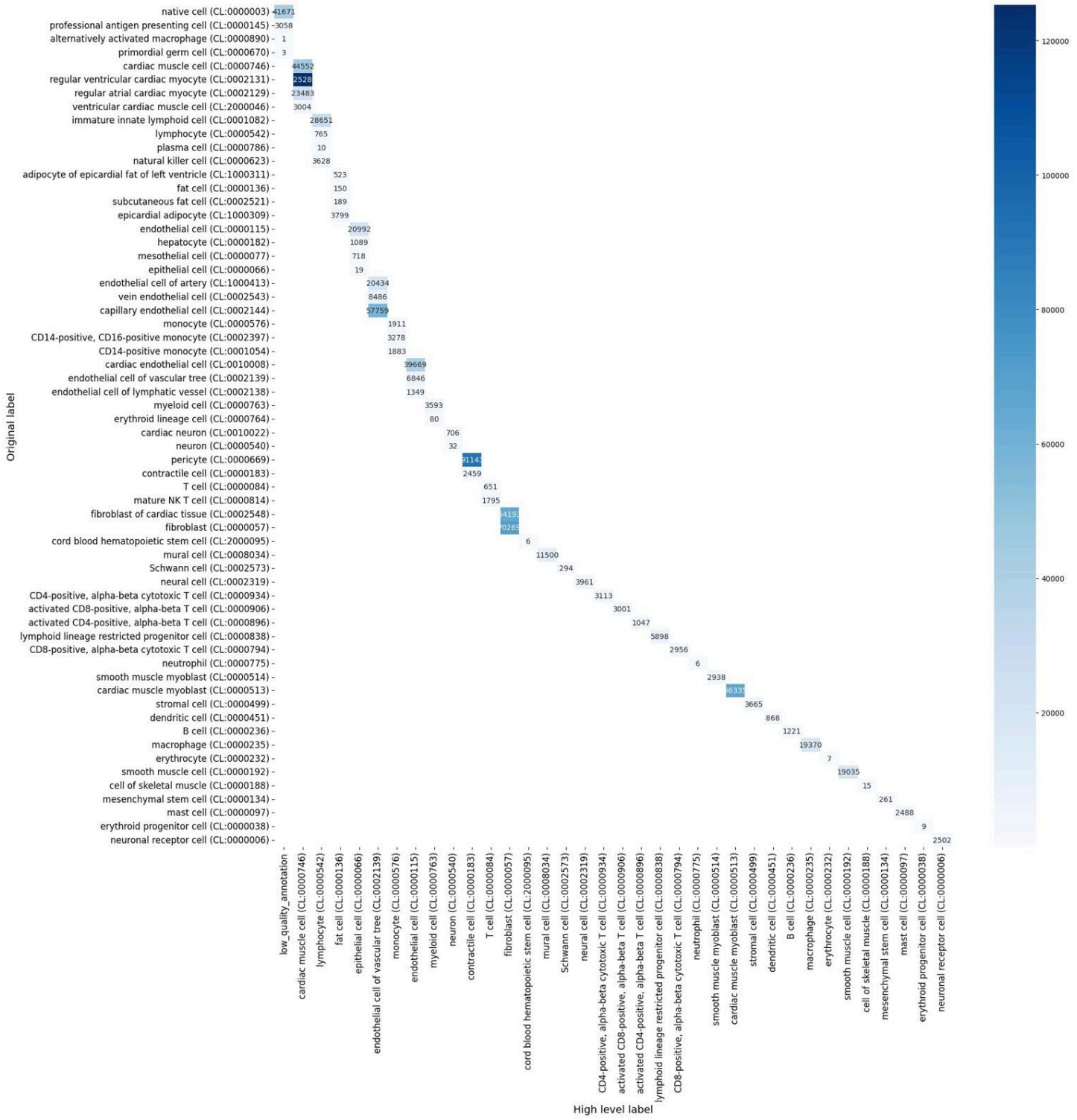
We thank Jeffrey Greves and members of the Teichmann group for valuable discussions on this project. ED, KBM and SAT. acknowledge Wellcome Sanger core funding (WT206194).

Author contributions: ED, ET, RE, GG, VS, EdR and SAT conceptualized the study. ET performed curation of Open Targets data and drug-level data analysis. ED performed curation and processing of scRNA-seq data, differential expression analysis, statistical analysis of association between omic evidence and clinical success, and disease-level target analysis. All authors interpreted the results. ED and ET made the figures. ED, ET, RE and EdR wrote the original manuscript draft. All authors edited and approved the final version of the manuscript. EdR and SAT supervised the work.

Conflicts of interest: ED has consulted for Ensocell Therapeutics. ET, GG, FN, EdR are employees of Sanofi and own Sanofi stock. VS has been leading the application of single-cell biology for drug development at Sanofi since 2018 and owns Sanofi stock. RE is a co-founder and employee of Ensocell Therapeutics. SAT has consulted for or been a member of scientific advisory boards at Qiagen, Sanofi, GlaxoSmithKline and ForeSite Labs. She is a consultant and equity holder for TransitionBio and Ensocell Therapeutics.

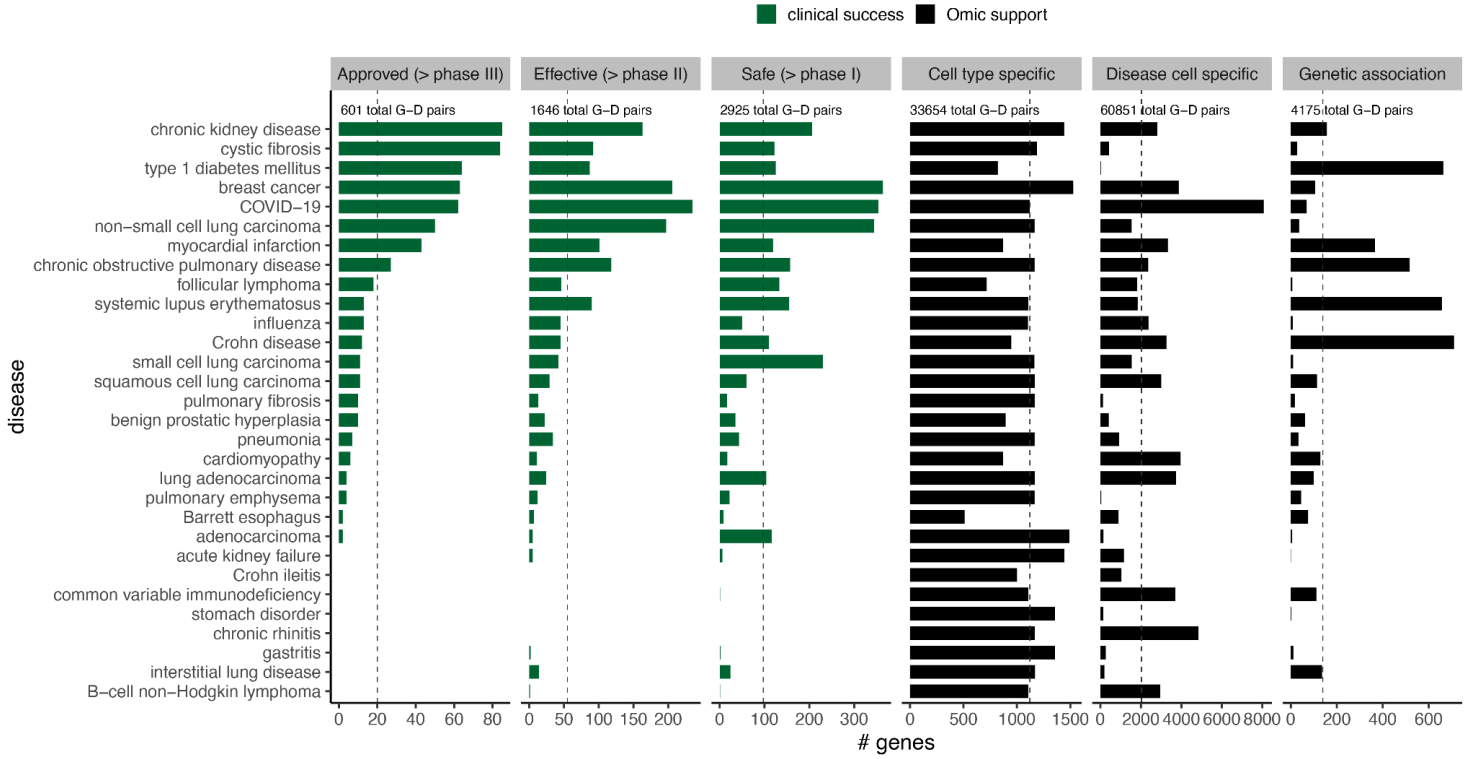
medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Supplementary Figures

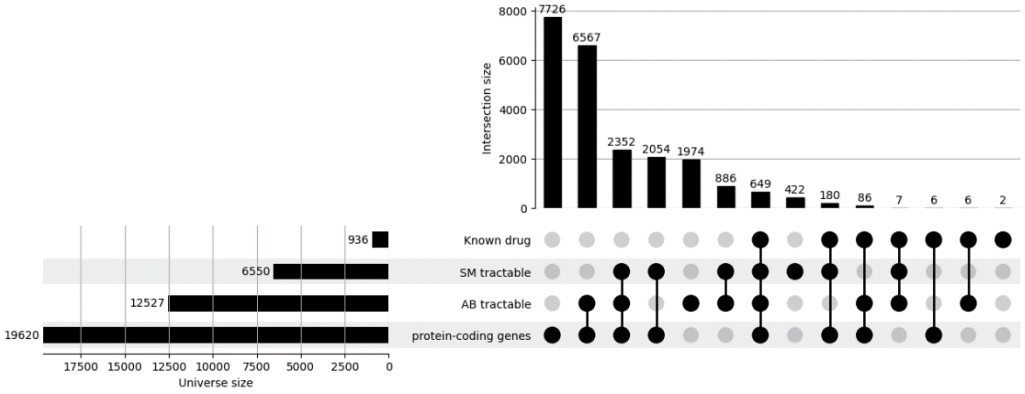


Supplementary Figure 1: Example outcome of harmonisation of cell type annotations based on Cell Ontology. The y-axis shows the original Cell Ontology label used in CZ CellxGene database for the myocardial infarction dataset (disease-relevant tissue: heart) and the x-axis shows the updated label after label harmonisation. The heatmap color and number indicate the number of cells for each label.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

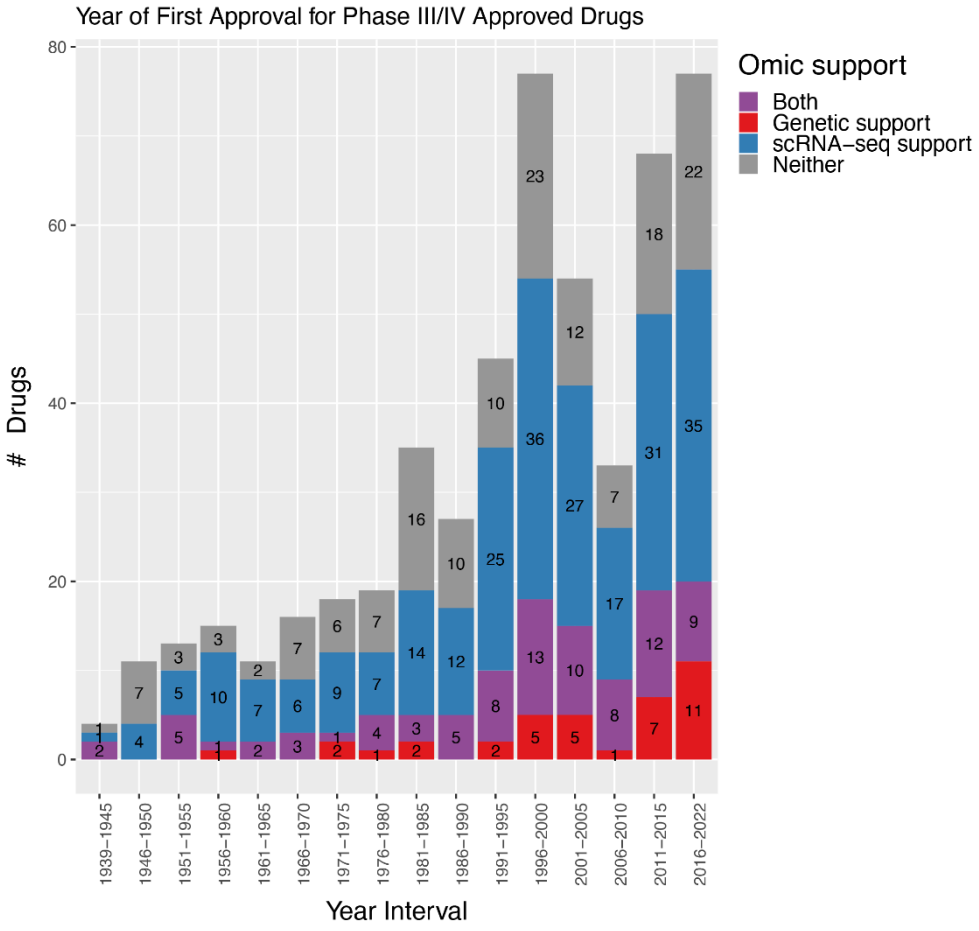


Supplementary Figure 2: Number of clinically successful and supported targets per disease. Barplot of number of gene targets in clinical success groups (green) and with omic support (black) by disease (y-axis). Diseases are ordered by the number of approved (> phase III) targets. The dotted lines denote the mean across diseases. The total number of G-D pairs for each class is reported above the bar plots.



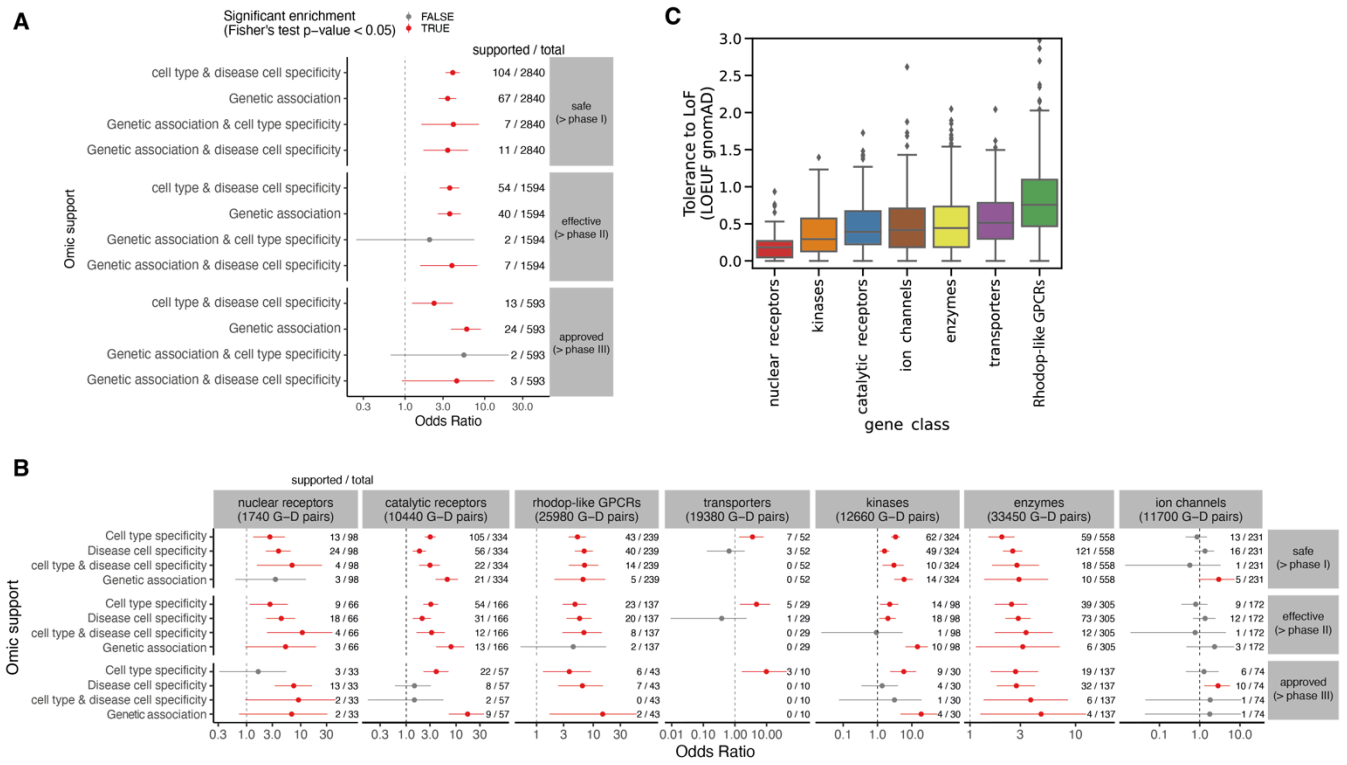
Supplementary Figure 3: Space of analysed genes for odds ratio analysis (gene universes). Upset plot showing total size (left) and intersection size (top) for different gene universes used in the analysis. SM: small molecule; AB: antibody

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .



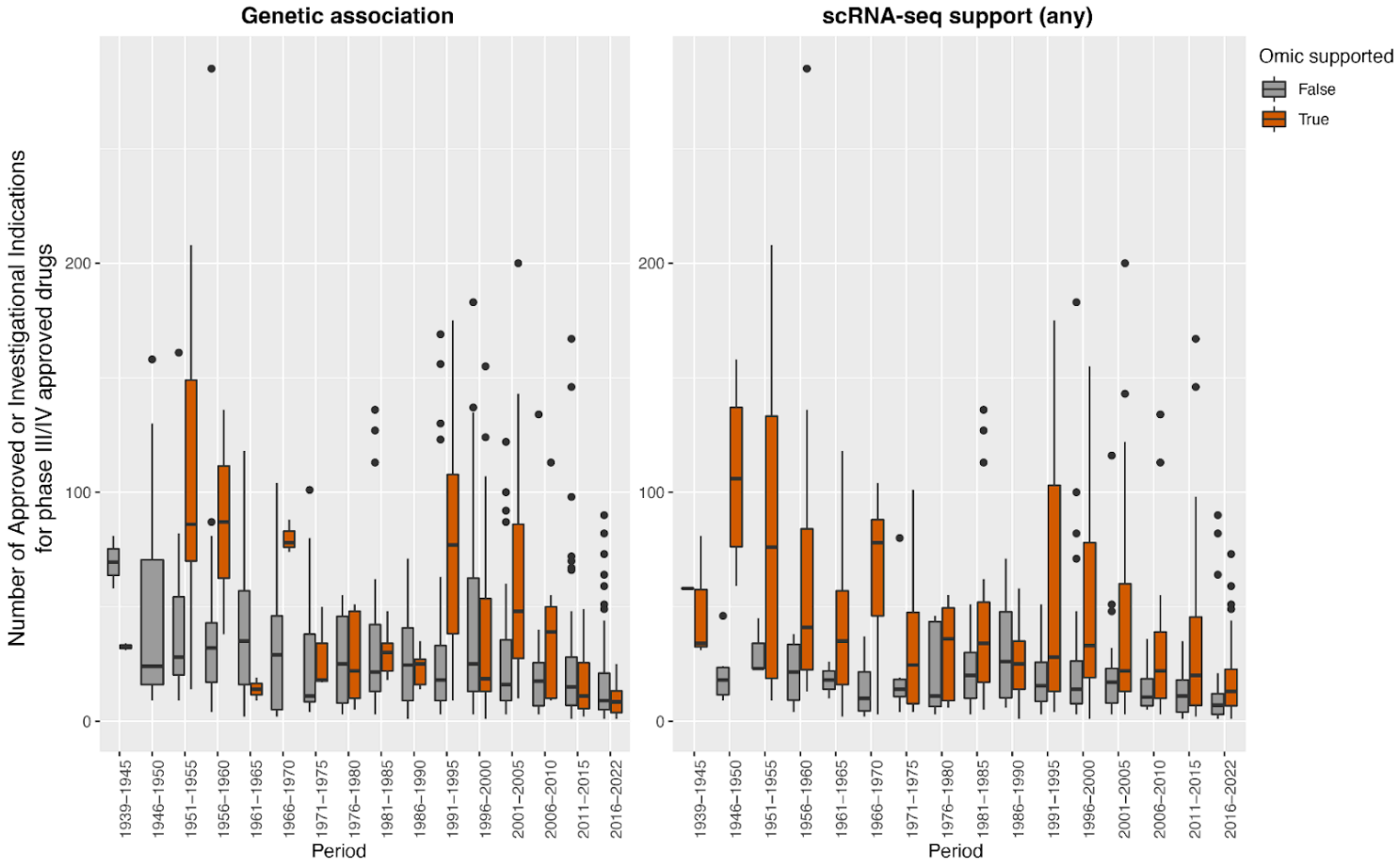
Supplementary Figure 4: Year of approval of considered drugs. Barplot showing year of first approval for drugs in phase III/IV for any of the 30 studied diseases. Color indicates if target-disease pairs for a given drug have scRNA-seq support (blue), genetic association support (red), or both (purple). Drugs without single cell or genetic support are shown in grey.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .



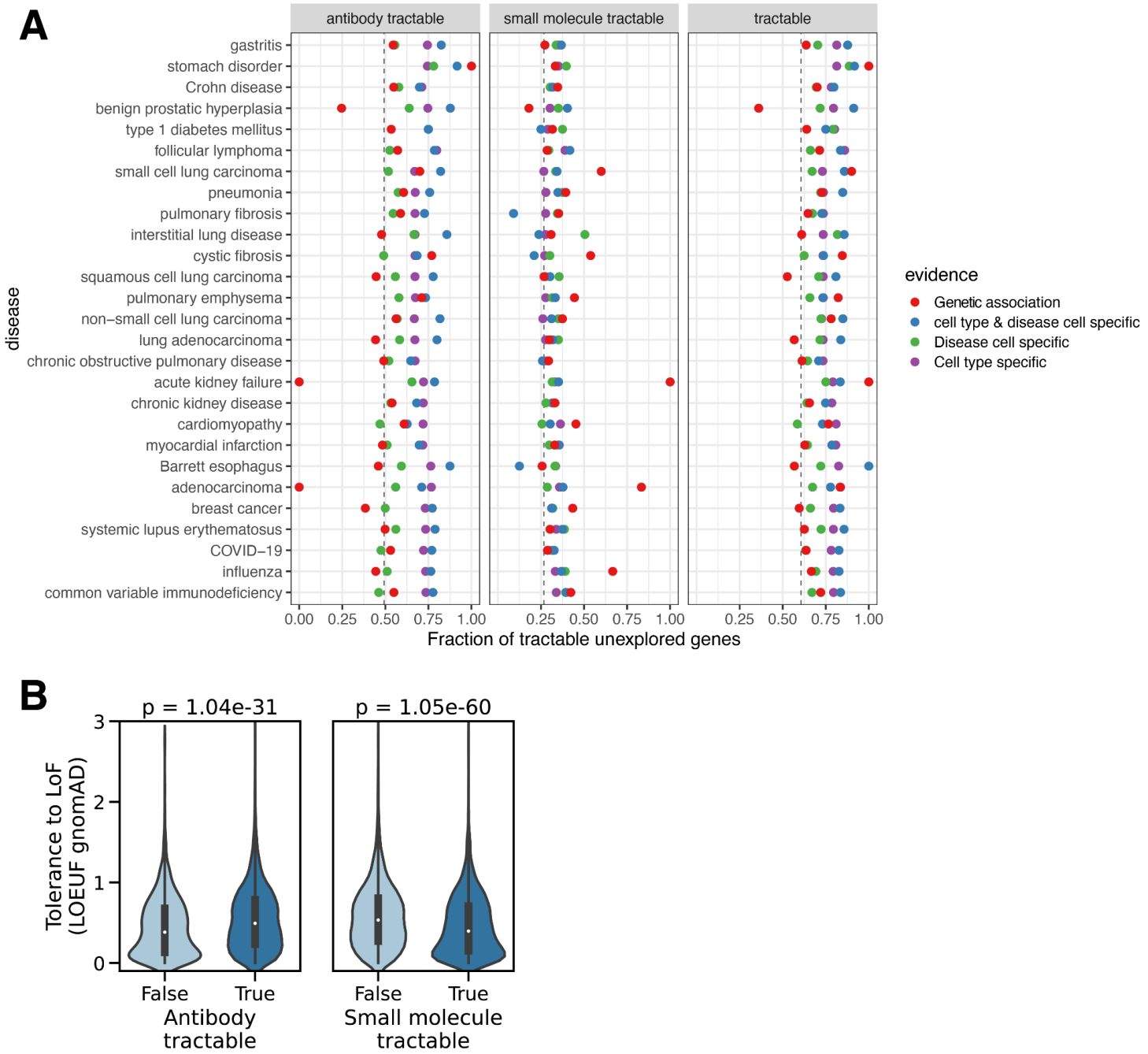
Supplementary Figure 5: Comparison with genetic support. (A) Odds ratio for association with clinical success with combined genetic association and scRNA-seq support. Association is computed using protein-coding genes as the gene universe. (B) Odds ratio for association between omic evidence and clinical success for different classes of druggable proteins. For each test, the numbers to the right show the number of omic-supported targets over total successful targets. The error bars denote 95% confidence intervals of the odds ratio. Points in red indicate cases where the enrichment for successful targets was statistically significant (Fisher's exact test p-value < 0.05). The dotted line denotes Odds Ratio = 1 (no enrichment). (C) Box plot of tolerance to loss-of-function mutations, estimated by Loss-of-function Observed/Expected Upper-bound Fraction (LOEUF) in gnomAD v2.1 (y-axis) for each class of druggable target shown in B (x-axis) (Nuclear receptors: N=46; kinases: N=338; catalytic receptors: N=246; ion channels: N=320; enzymes: N=864; transporters: N=510; GPCRs: N=574). Gene classes are sorted by mean LOEUF score. 15 outlier genes with LOEUF > 3 are not shown. In the boxplots, the center line denotes the median; the box limits denote the first and third quartiles; and the whiskers denote 1.5x the interquartile range (IQR).

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .



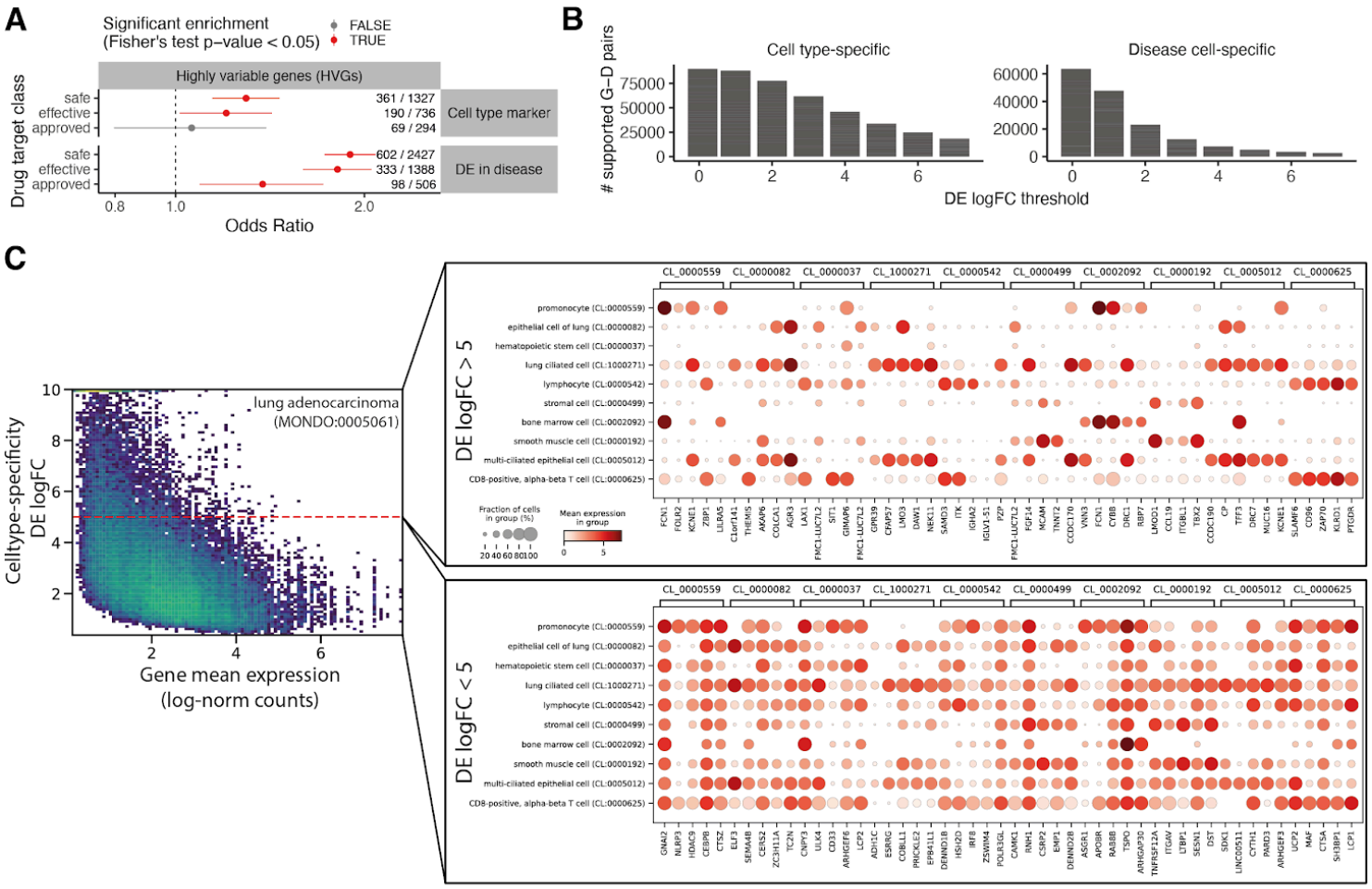
Supplementary Figure 6: Number of approved or investigational indications per drug by year of first approval and scRNAseq evidence support. Boxplots of number of approved or investigational indications (y-axis) for drugs approved (>= Phase III) for the 30 diseases considered in this study. Drugs are stratified by year of first approval (x-axis), and by presence or absence of omic support (fill). The left plot shows the number of indications for drugs supported by genetic association. The right plot shows the number of indications for drugs supported by scRNA-seq (either cell type specific or disease cell specific targets). In the boxplots, the center line denotes the median; the box limits denote the first and third quartiles; and the whiskers denote 1.5x the interquartile range (IQR).

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .



Supplementary Figure 7: Tractability of unexplored targets across diseases. (A) Scatter plot of fraction of tractable unexplored genes (x-axis) for 27 diseases (y-axis) for different classes of omic evidence (color). We consider three categories: antibody tractable, small molecule tractable, and tractable by either class of drugs. Dashed lines represent the fraction of tractable genes across all protein-coding genes. Diseases for which no gene with genetic evidence was found are not shown (n=3). (B) Violin plots of tolerance to loss-of-function mutations, estimated by Loss-of-function Observed/Expected Upper-bound Fraction (LOEUF) in gnomAD v2.1 (y-axis) for each tractable or non-tractable gene considered for analysis in figure 2D (x-axis). The left plot shows LOEUF estimates for antibody tractable genes. The right plot shows LOEUF estimates for small molecule tractable genes. The values on top of each plot show the p-value for Wilcoxon rank-sum test comparing the mean LOEUF between tractable and non-tractable genes (null hypothesis: no difference). In the boxplots, the center dot denotes the median; the box limits denote the first and third quartiles; and the whiskers denote 1.5x the interquartile range (IQR).

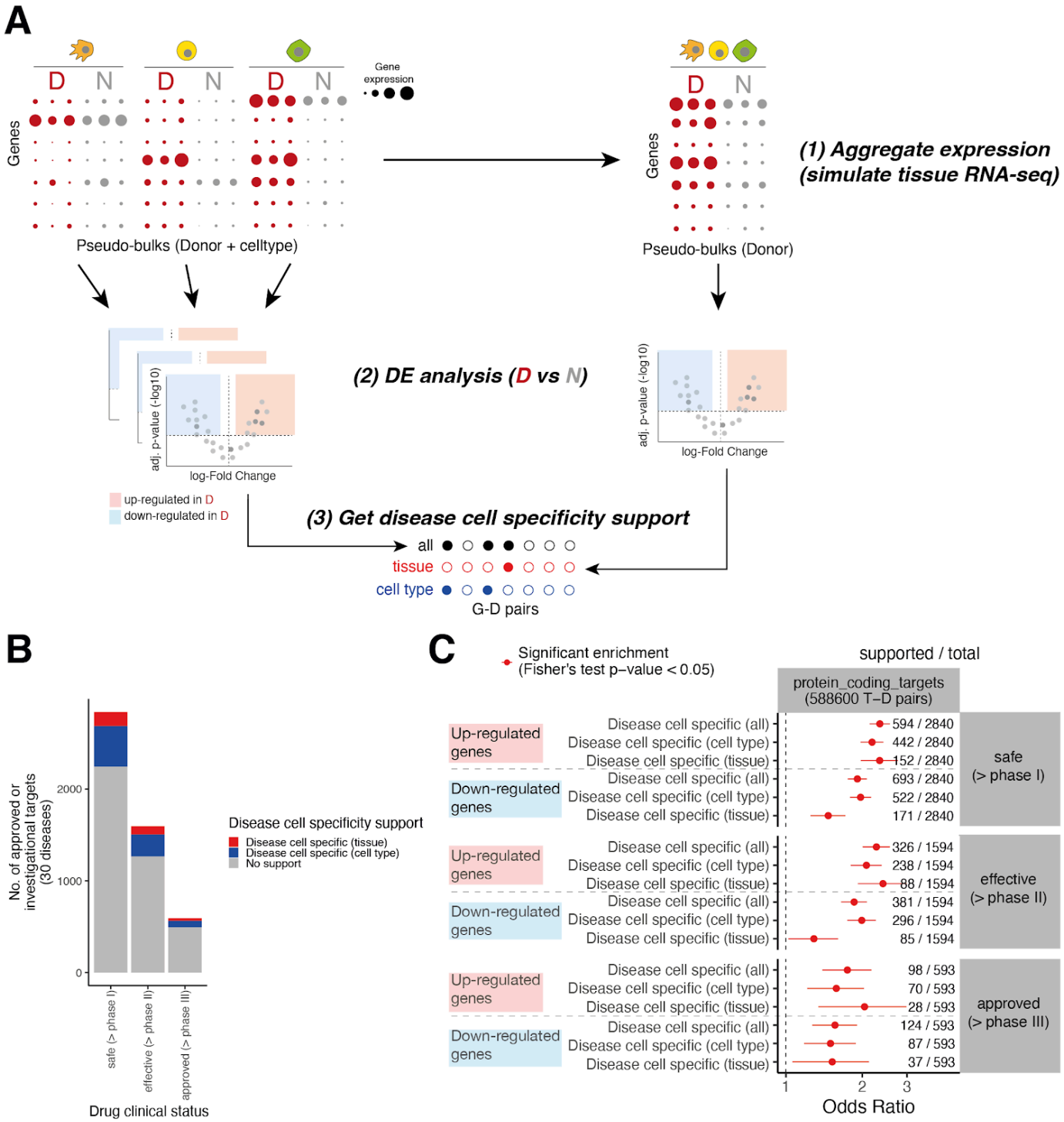
medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .



medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

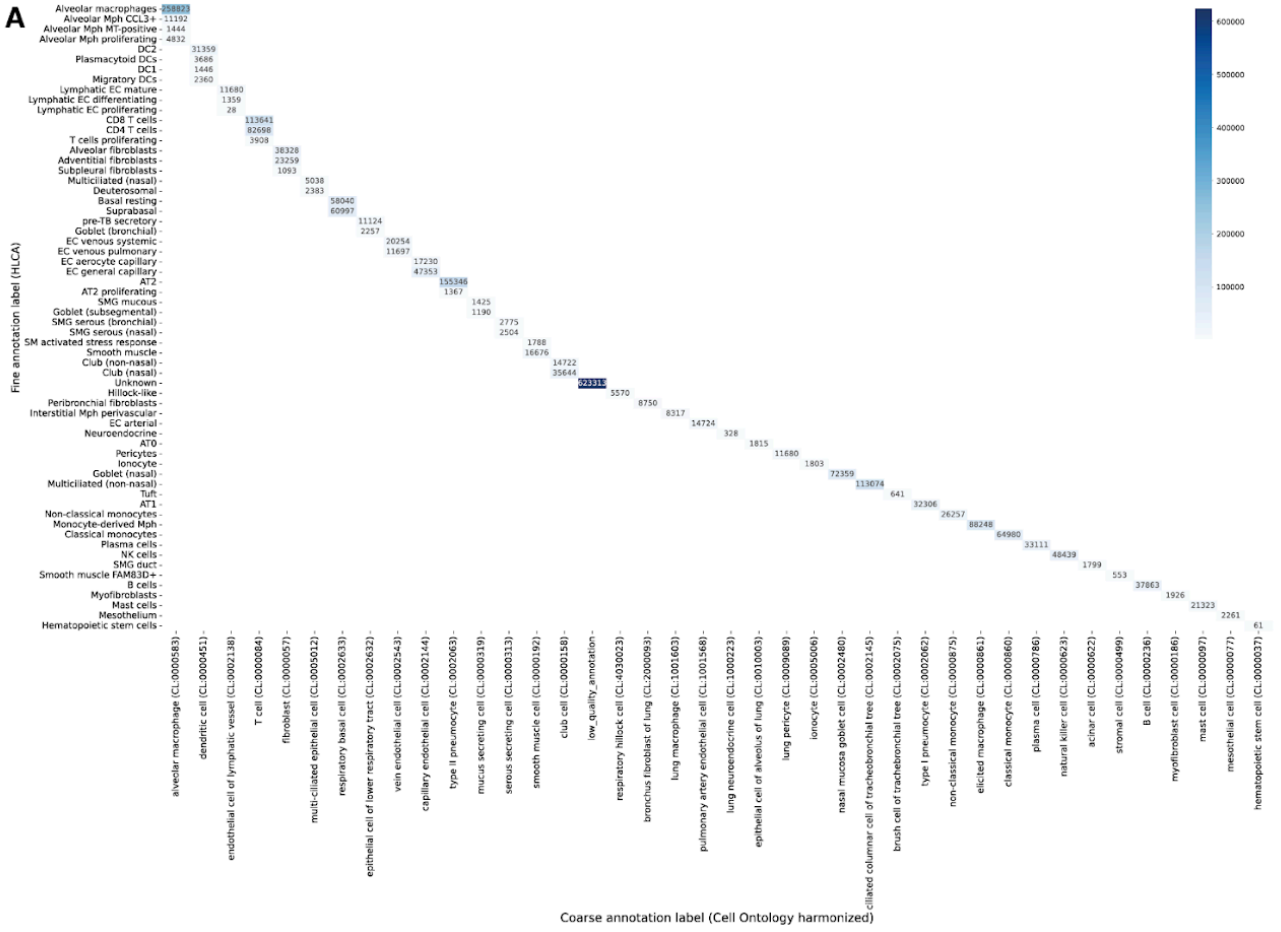
Supplementary Figure 8: Analysis of parameters for definition of targets with scRNA-seq support. (A) Odds ratio (x-axis, in log10 scale) of association between target clinical success (y-axis) and scRNA-seq support for the target, computed from highly variable genes in scRNA-seq datasets of disease-relevant tissue. For each test, the numbers to the right show the number of omic supported targets over total successful targets. (B) Barplot of number of supported G-D pairs with increasing log-Fold Change (logFC) threshold on differential expression (DE) analysis results, for cell type specific genes (left) and disease cell specific genes (right). (C) Example from lung adenocarcinoma scRNA-seq data showing cell type specificity of candidate target genes at high DE log-fold changes. The left scatterplot shows the mean expression (log-normalized counts, x-axis) and DE log-fold change for one-vs-all test (y-axis) used for cell type specificity analysis for each significantly over-expressed gene (1% FDR). The dotplots to the right show the expression, in terms of mean (color) and cell fraction (size) for 5 randomly selected cell type specific genes detected in 10 lung cell types (the cell ontology term is indicated on top of the plots). The top plot shows significant genes with logFC > 5 and the bottom plot shows significant genes with logFC < 5. (D) Odds ratio (x-axis, in log10 scale) of association between clinical success (y-axis) of a target and scRNA-seq support defined using an increasing threshold for DE log-fold change (y-axis). The dotted blue line denotes the threshold selected for analyses throughout this study. For each test, the numbers to the right show the number of omic supported targets over total successful targets. The error bars denote 95% confidence intervals of the odds ratio. Points in red indicate cases where the enrichment for successful targets was statistically significant (Fisher's exact test p-value < 0.05). The dotted line denotes Odds Ratio = 1 (no enrichment).

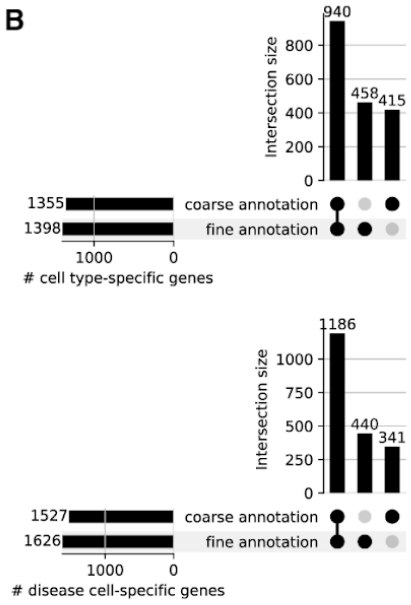
medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

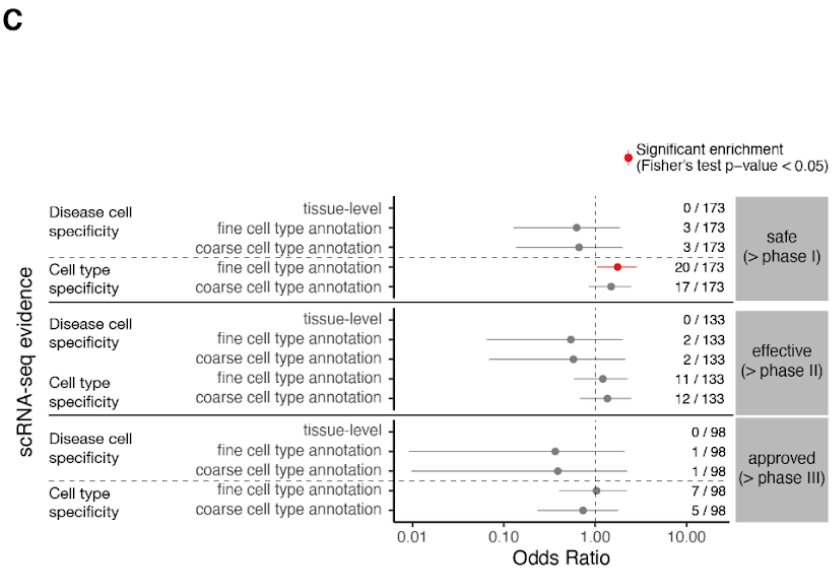


Supplementary Figure 9: Cell type-level and tissue-level differential expression analysis for disease cell specificity. (A) Illustration of strategy to compare genes identified as disease specific with cell type-level or tissue-level differential expression analysis between normal (labelled *N*) and diseased (labelled *D*) samples. For each disease, we compare gene expression between healthy and diseased tissue either per cell type (left panel, cell type-level) or summed across all cell types (right panel, tissue-level). Differential expression in any of these categories is classed as disease cell specific support. (B) Barplot showing the number of successful targets at different clinical stages (x-axis) annotated as disease cell specific at the cell type level (blue) or the tissue level (red). (C) Odds ratio (x-axis, in log10 scale) of association between clinical success of a target and scRNA-seq support (y-axis) selected using up- or down-regulated genes (based on DE analysis log-Fold Change and adjusted p-value > 0.01) with tissue or cell type level analysis, as defined in (A). Results are shown considering gene-disease pairs for 30 diseases. We test association with safe targets (passed phase I, top row), effective targets (passed phase II, middle row) and approved targets (passed phase III, bottom row). For each test, the numbers to the right show the number of omic supported targets over total successful targets. The error bars denote 95% confidence intervals of the odds ratio. Points in red indicate cases where the enrichment for successful targets was statistically significant (Fisher's exact test p-value < 0.05). The dotted line denotes Odds Ratio = 1 (no enrichment).

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

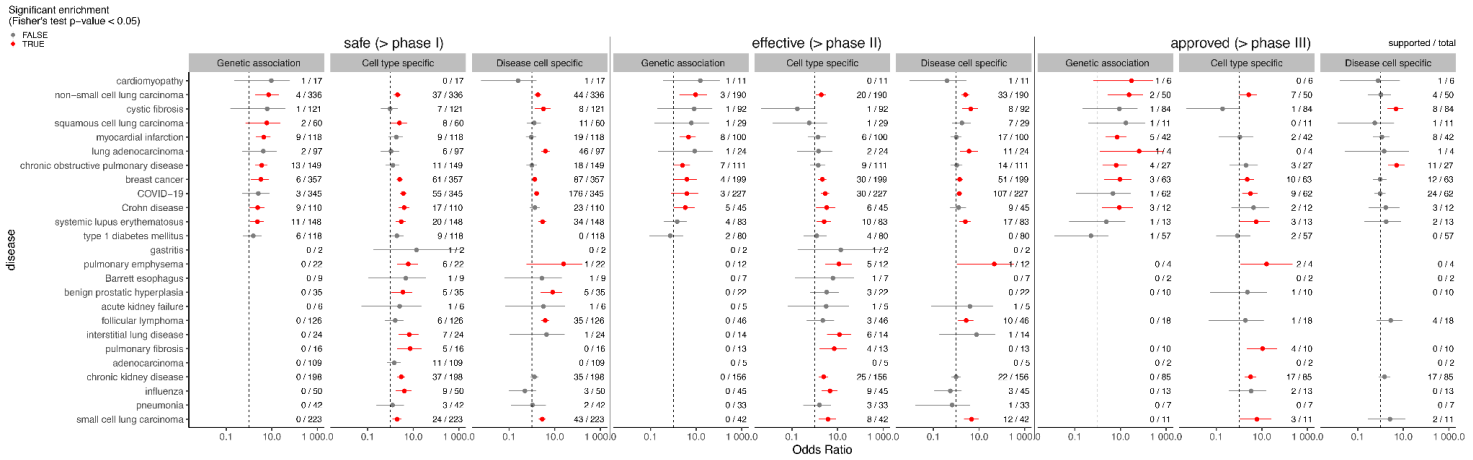






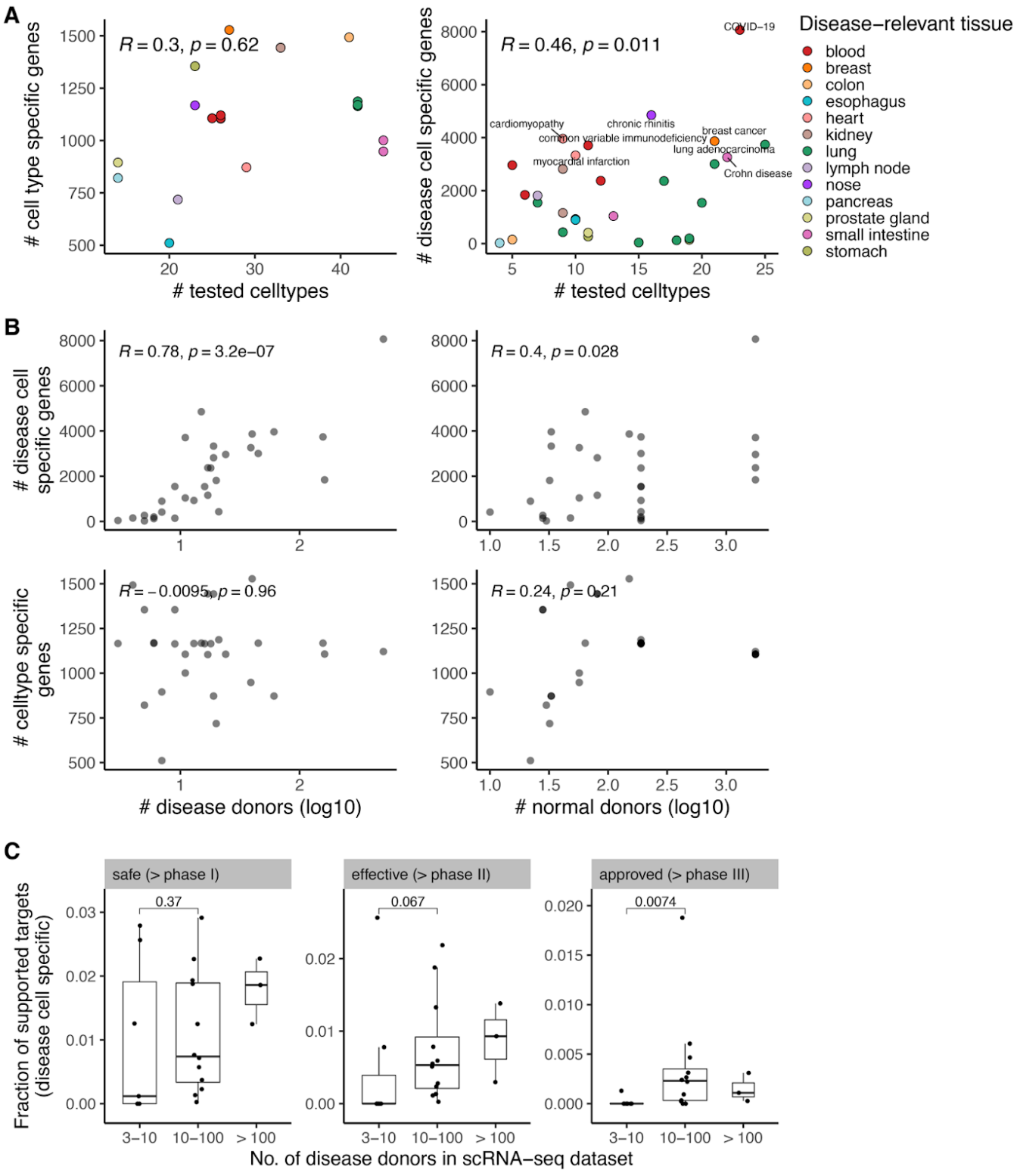
medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Supplementary Figure 10: Impact of fine cell type annotations on scRNA-seq support for target discovery. (A) Confusion table matching the number of cells with coarse annotation labels (uniformed Cell Ontology labels in CZ CellxGene database, x-axis) and fine integration-based annotations on the extended Human Lung Cell Atlas (eHLCA) dataset [30] (CellxGene census datasetID: 9f222629-9e39-47d0-b83f-e08d610c7479) (y-axis). (B) Upset plots showing the total size (left bars) and size of intersections (top bars) of genes prioritized for 3 lung diseases (pulmonary fibrosis, cystic fibrosis, pneumonia) using fine annotations from eHLCA or coarse annotations from CZ CellxGene database. We compare genes prioritized by cell type specificity (top plot) and by disease cell specificity (bottom plot). (C) Odds ratio (x-axis, in log10 scale) of association between clinical success of a target and scRNA-seq support (y-axis) computed using fine or coarse cell type annotations. For disease cell specificity, we also considered genes prioritized by tissue-level analysis, as described in Supplementary Figure 9A. Results are shown considering gene-disease pairs for 3 lung diseases sampled in the eHLCA dataset (pulmonary fibrosis, cystic fibrosis, pneumonia). We test association with safe targets (passed phase I, top row), effective targets (passed phase II, middle row) and approved targets (passed phase III, bottom row). Protein-coding genes were used as gene universe. For each test, the numbers to the right show the number of omic supported targets over total successful targets. The error bars denote 95% confidence intervals of the odds ratio. Points in red indicate cases where the enrichment for successful targets was statistically significant (Fisher's exact test p-value < 0.05). The dotted line denotes Odds Ratio = 1 (no enrichment).



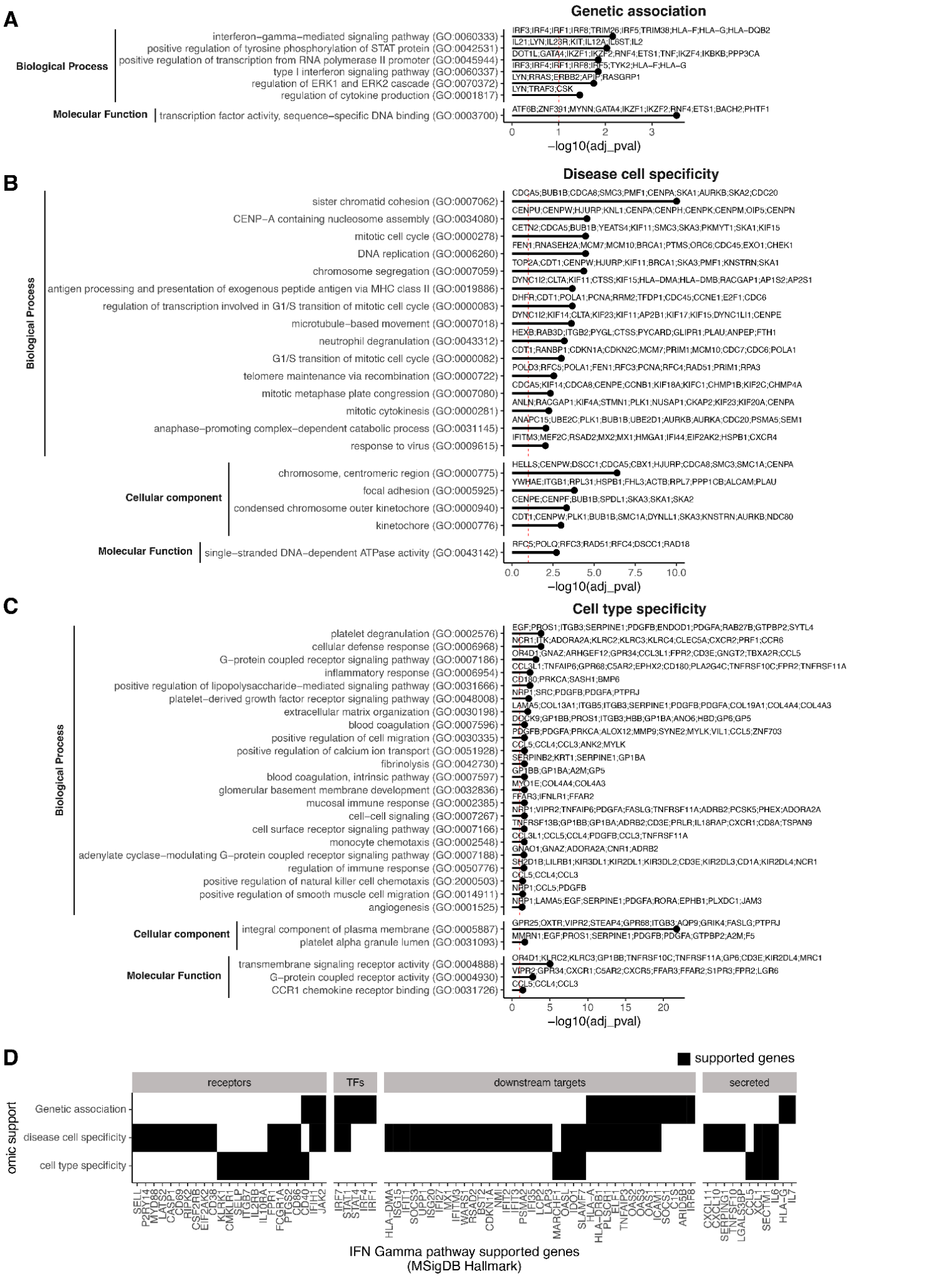
Supplementary Figure 11: Association between omic support and clinical success stratified by disease. Odds ratio (x-axis, in log10 scale) of association between clinical success of a target and scRNA-seq support (y-axis) computed stratifying by disease. Results are shown for 22 diseases with at least 1 approved target. Diseases are sorted by odds ratios for association with genetic support. The gene universe used was protein-coding targets. For each test, the numbers to the right show the number of omic supported targets over total successful targets. The error bars denote 95% confidence intervals of the odds ratio. Points in red indicate cases where the enrichment for successful targets was statistically significant (Fisher's exact test p-value < 0.05). The dotted line denotes Odds Ratio = 1 (no enrichment).

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .



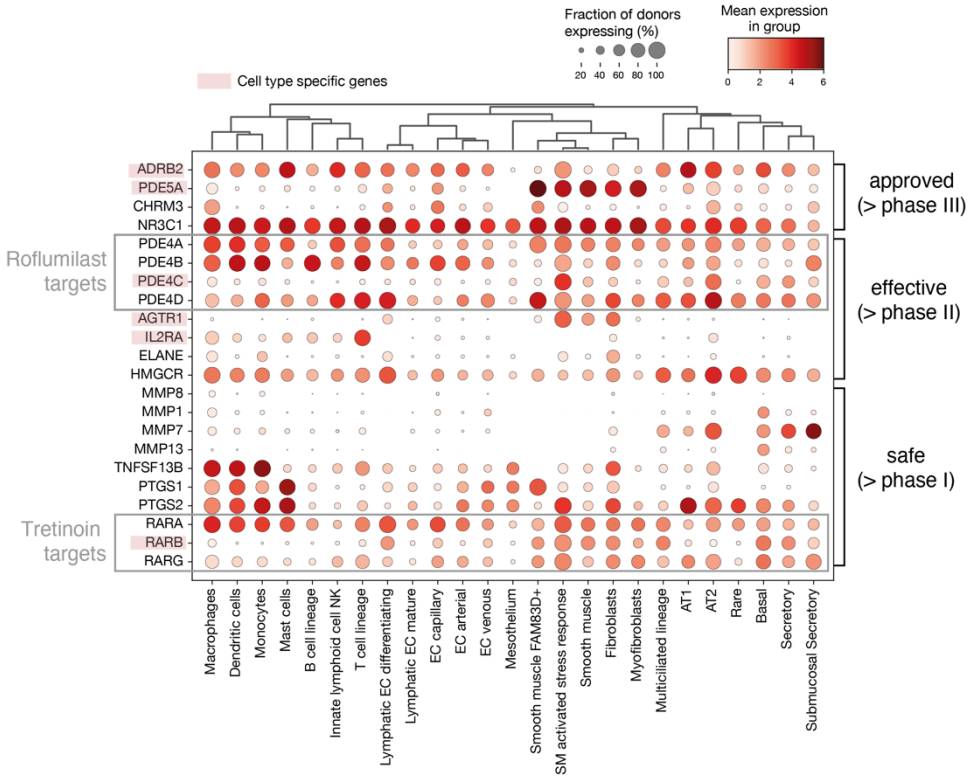
Supplementary Figure 12: Variability in scRNA-seq supported targets between diseases. (A) Scatterplots showing the number of tested cell types in disease-relevant tissue (x-axis) against the number of identified cell type specific (left) and disease cell specific (right) genes. Dots are colored by disease-relevant tissue. Pearson's correlation coefficient and p-value for permutation test are shown on top. (B) Scatterplots showing the number of disease donors (left column) and control donors (right column) in scRNA-seq dataset for each disease against the number of identified cell type specific (bottom row) and disease cell specific (top row) genes. Pearson's correlation coefficient and p-value for permutation test are shown on top. (C) Boxplots showing the fraction of known clinical targets supported by disease cell specificity (y-axis) for different diseases grouped by size of disease donors cohort (x-axis). The p-value for Wilcoxon Rank Sum test comparing small and medium sized cohorts is reported on top. Fractions of safe (> phase I, left), effective (> phase II, center) and approved targets (> phase III, right) are shown. In the boxplots, the center line denotes the median; the box limits denote the first and third quartiles; and the whiskers denote 1.5x the interquartile range (IQR).

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .



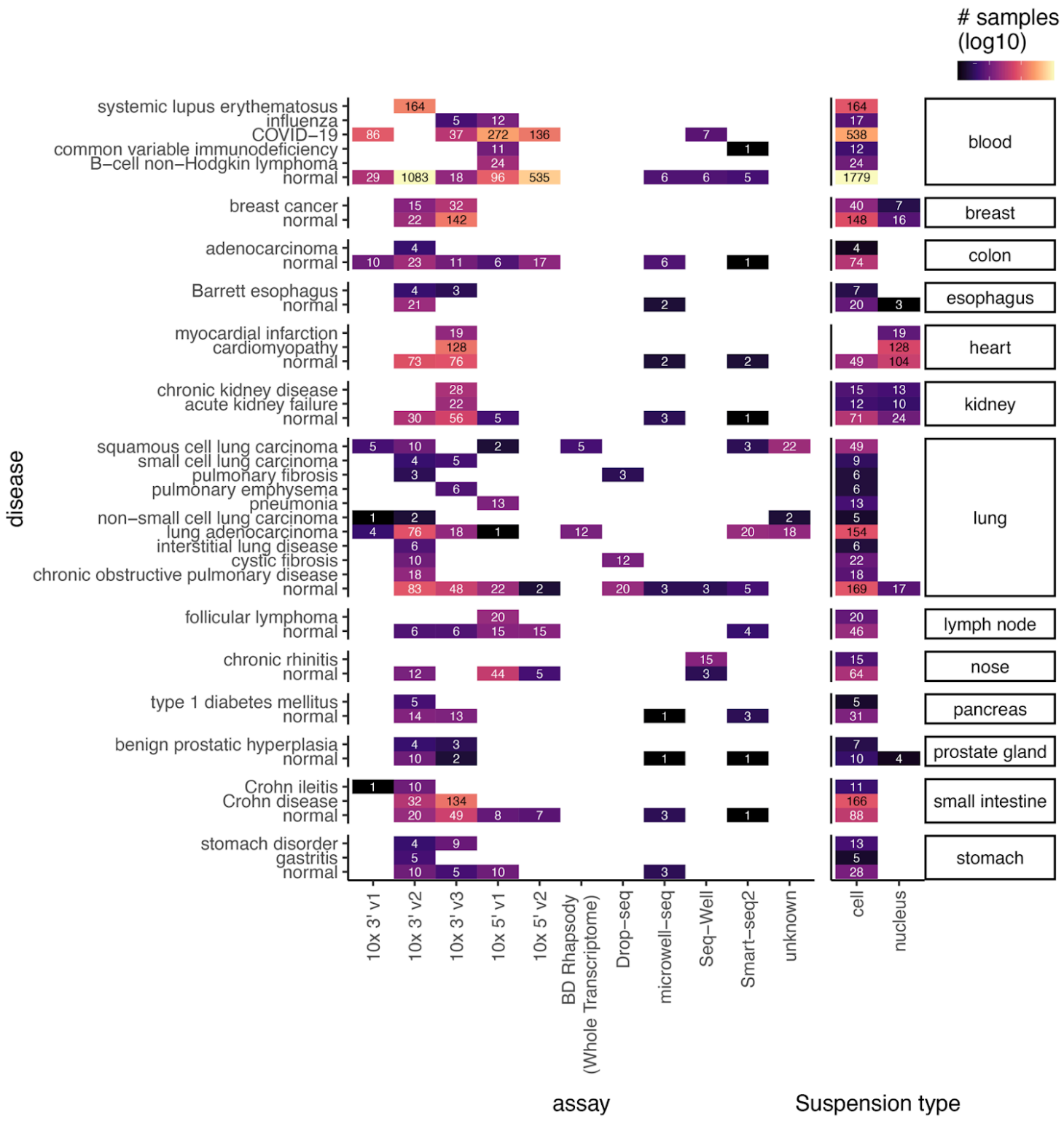
medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Supplementary Figure 13: Functional analysis of supported targets for systemic lupus erythematosus (SLE). (A-C) Gene Ontology (GO) enrichment analysis on unexplored supported genes in SLE (excluding known drug targets). In each graph, we show significantly enriched (adj. p-value < 0.01) GO terms (y-axis) sorted by adjusted p-value (x-axis, negative log10). Terms are grouped by Gene Ontology class (biological process, cellular component, molecular function). For each term, a sample of up to 10 genes associated to the term are shown. We show terms enriched in genetic supported genes (A), disease cell specific genes (B) and cell type specific supported genes (C); (D) Binary table displaying genetic, disease cell, or cell type specificity support for IFN Gamma pathway genes in pulmonary emphysema. The filled bars denote whether the evidence exists (black) or does not exist (white) for each gene. IFN Gamma pathway genes were derived from the MSigDB Hallmark database (only genes supported by at least one class of omic evidence are shown). Genes were categorized into four functional groups (receptors, transcription factors (TFs), targets, and secreted proteins) using OmniPath [77] and Dorothea [78,79]. Each bar represents the presence of genetic, disease cell, or cell type marker evidence.



Supplementary Figure 14: Pulmonary emphysema drug target gene expression in healthy human lung atlas. Dotplot of expression of safe, effective and approved drug target genes for treatment of pulmonary emphysema (y-axis) across cell types found in human lung tissue (x-axis). Dot color denotes the mean expression (log-normalized counts) in a cell type across donors. Dot size denotes the fraction of donors in which the gene in expressed. Lung cells are annotated using curated labels from the Human Lung Cell Atlas [30]. Boxes indicate targets for either Roflumilast (phosphodiesterase-4 inhibitor) or Tretinoin (all-trans retinoic acid). Genes highlighted in red show genes classified as *cell type specific* by DE analysis.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .



Supplementary Figure 15: Technical metadata for disease scRNA-seq datasets. Heatmap showing the scRNA-seq assay and suspension type (x-axis) for samples of different tissues and diseases (y-axis). Heatmap color and annotated numbers denote the number of samples analysed for each group. Diseases are grouped by disease-relevant tissue.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Supplementary Tables

Supplementary Table 1: Table of diseases available in CZ CellxGene database considered for study
[disease] name of disease used in study
[disease_ontology_id] MONDO identifier for disease used in study
[disease_relevant_tissue] Manually curated annotation for disease-relevant tissue
[disease_name_original] Name of disease found in CZ CellxGene database
[disease_ontology_id _original] MONDO identifier for disease found in CZ CellxGene database
[reason2exclude] if not NA, description of reason to exclude disease from final analysis

Supplementary Table 2: Sample-level metadata for scRNA-seq datasets from CZ CellxGene database used in study
[assay] scRNA-seq protocol
[tissue] original tissue annotation
[tissue_general] high-level mapping of a tissue
[suspension type] indicates whether cells or nuclei were isolated
[disease] disease condition of donor
[dataset_id] Identifier for dataset in CellXGene Census
[donor_id] Identifier for donor in dataset
[development_stage_ontology_term_id] Human Developmental Stages ontology term for age of donor
[sample_id] sample identifier (donor, assay, tissue)
[disease_name_original] name of disease found in CZ CellxGene database
[disease_ontology_id _original] MONDO identifier for disease found in CZ CellxGene database
[disease_ontology_id] MONDO identifier for disease used in study
[disease_relevant_tissue] Manually curated annotation for disease-relevant tissue

Supplementary Table 3: Table of target-disease pairs with annotation of clinical success and omic support
[gene_id] Ensembl ID for gene
[disease_ontology_id] MONDO identifier for disease
[disease] name of disease
[gene_name] gene name
[gene_class] annotation of tractable gene classes
[genetic_association] OpenTargets genetic association score
(https://platform-docs.opentargets.org/evidence#evidence-data-sources)
[known_drug] OpenTargets known drug score (https://platform-docs.opentargets.org/evidence#evidence-data-sources)
[is_druggable, is_safe, is_effective, is_approved] clinical status for each gene-disease pair
[GWAS_evidence] is gene-disease pair supported by genetic association
[ct_marker_evidence] is gene-disease pair supported by cell type specificity
[disease_evidence] is gene-disease pair supported by disease cell specificity
[ct_marker_and_disease_evidence] is gene-disease pair supported by cell type and disease cell specificity
[disease_evidence_celltype] is gene-disease pair supported by disease cell specificity (celltype-level)
[disease_evidence_tissue] is gene-disease pair supported by disease cell specificity (tissue-level)

Supplementary Table 4: Results of association analysis between omic support and clinical success across diseases
[odds_ratio] Odds ratio of association between evidence and clinical success
[ci_low] 95% confidence interval of odds ratio (bottom)
[ci_high] 95% confidence interval of odds ratio (top)
[pval] Fisher exact test p-value for enrichment (alternative hypothesis: odds ratio higher than 1)
[n_success] Number of successful gene-disease pairs
[n_insuccess] Number of not successful gene-disease pairs
[n_supported_approved] Number of successful gene-disease pairs supported by omic evidence
[n_supported] Total number of gene-disease pairs supported by omic evidence
[evidence] omic support class (all_sc_evidence indicates cell type and disease cell specific genes)
[clinical status] Clinical success class
[universe] Name of considered gene universe
[universe_size] Number of genes in gene universe

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Supplementary Table 5: Results of multiple linear regression model predicting log(number of investigational or approved indications of a drug) from its year of first approval, drug target-disease support by any single cell evidence, and drug target-disease support by any direct genetic association.

Supplementary Table 6: Results of association analysis between omic support and clinical success for each disease (gene universe: protein-coding genes)
[odds_ratio] Odds ratio of association between evidence and clinical success
[ci_low] 95% confidence interval of odds ratio (bottom)
[ci_high] 95% confidence interval of odds ratio (top)
[pval] Fisher exact test p-value for enrichment (alternative hypothesis: odds ratio higher than 1)
[n_success] Number of successful gene-disease pairs
[n_insuccess] Number of not successful gene-disease pairs
[n_supported_approved] Number of successful gene-disease pairs supported by omic evidence
[n_supported] Total number of gene-disease pairs supported by omic evidence
[evidence] omic support class (all_sc_evidence indicates cell type and disease cell specific genes)
[clinical status] Clinical success class
[disease_ontology_id] MONDO identifier for disease
[disease] name of disease
[disease_relevant_tissue] Manually curated annotation for disease-relevant tissue

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

References:

1. Ochoa D, Hercules A, Carmona M, Suveges D, Baker J, Malangone C, et al. The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. Nucleic Acids Res. 2023;51: D1353–D1359. doi:10.1093/nar/gkac1046

2. Moffat JG, Vincent F, Lee JA, Eder J, Prunotto M. Opportunities and challenges in phenotypic drug discovery: an industry perspective. Nat Rev Drug Discov. 2017;16: 531–543. doi:10.1038/nrd.2017.111

3. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. J Health Econ. 2016;47: 20–33. doi:10.1016/j.jhealeco.2016.01.012

4. Morgan P, Brown DG, Lennard S, Anderton MJ, Barrett JC, Eriksson U, et al. Impact of a five-dimensional framework on R&D productivity at AstraZeneca. Nat Rev Drug Discov. 2018;17: 167–181. doi:10.1038/nrd.2017.244

5. Barmada A, Handfield L-F, Godoy-Tena G, de la Calle-Fabregat C, Ciudad L, Arutyunyan A, et al. Single-cell multi-omics analysis of COVID-19 patients with pre-existing autoimmune diseases shows aberrant immune responses to infection. Eur J Immunol. 2023; e2350633. doi:10.1002/eji.202350633

6. Sungnak W, Huang N, Bécavin C, Berg M, Queen R, Litvinukova M, et al. SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. Nat Med. 2020;26: 681–687. doi:10.1038/s41591-020-0868-6

7. Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su M-J, Melms JC, et al. A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. Cell. 2018;175: 984-997.e24. doi:10.1016/j.cell.2018.09.006

8. Kildisiute G, Kalyva M, Elmentaite R, van Dongen S, Thevanesan C, Piapi A, et al. Transcriptional signals of transformation in human cancer. Genome Med. 2024;16: 8. doi:10.1186/s13073-023-01279-z

9. Li R, Ferdinand JR, Loudon KW, Bowyer GS, Laidlaw S, Muyas F, et al. Mapping single-cell transcriptomes in the intra-tumoral and associated territories of kidney cancer. Cancer Cell. 2022;40: 1583-1599.e10. doi:10.1016/j.ccell.2022.11.001

10. Liu X, Jin S, Hu S, Li R, Pan H, Liu Y, et al. Single-cell transcriptomics links malignant T cells to the tumor immune landscape in cutaneous T cell lymphoma. Nat Commun. 2022;13: 1158. doi:10.1038/s41467-022-28799-3

11. Bolton C, Smillie CS, Pandey S, Elmentaite R, Wei G, Argmann C, et al. An integrated taxonomy for monogenic inflammatory bowel disease. Gastroenterology. 2022;162: 859–876. doi:10.1053/j.gastro.2021.11.014

12. Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, et al. Single-cell transcriptomic analysis of Alzheimer's disease. Nature. 2019;570: 332–337. doi:10.1038/s41586-019-1195-2

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

13. Segerstolpe Å, Palasantza A, Eliasson P, Andersson E-M, Andréasson A-C, Sun X, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. Cell Metab. 2016;24: 593–607. doi:10.1016/j.cmet.2016.08.020

14. Perez RK, Gordon MG, Subramaniam M, Kim MC, Hartoularos GC, Targ S, et al. Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. Science. 2022;376: eabf1970. doi:10.1126/science.abf1970

15. Rood JE, Maartens A, Hupalowska A, Teichmann SA, Regev A. Impact of the Human Cell Atlas on medicine. Nat Med. 2022;28: 2486–2496. doi:10.1038/s41591-022-02104-7

16. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, et al. The support of human genetic evidence for approved drug indications. Nat Genet. 2015;47: 856–860. doi:10.1038/ng.3314

17. King EA, Davis JW, Degner JF. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. PLoS Genet. 2019;15: e1008489. doi:10.1371/journal.pgen.1008489

18. Minikel EV, Painter JL, Dong CC, Nelson MR. Refining the impact of genetic evidence on clinical success. bioRxiv. 2023. p. 2023.06.23.23291765. doi:10.1101/2023.06.23.23291765

19. Ochoa D, Karim M, Ghoussaini M, Hulcoop DG, McDonagh EM, Dunham I. Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs. Nat Rev Drug Discov. 2022;21: 551. doi:10.1038/d41573-022-00120-3

20. Rusina PV, Falaguera MJ, Romero JMR, McDonagh EM, Dunham I, Ochoa D. Genetic support for FDA-approved drugs over the past decade. Nat Rev Drug Discov. 2023;22: 864. doi:10.1038/d41573-023-00158-x

21. CZI Single-Cell Biology Program, Abdulla S, Aevermann B, Assis P, Badajoz S, Bell SM, et al. CZ CELL×GENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. bioRxiv. 2023. doi:10.1101/2023.10.30.563174

22. Ghoussaini M, Mountjoy E, Carmona M, Peat G, Schmidt EM, Hercules A, et al. Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. Nucleic Acids Res. 2021;49: D1311–D1320. doi:10.1093/nar/gkaa840

23. Diehl AD, Meehan TF, Bradford YM, Brush MH, Dahdul WM, Dougall DS, et al. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. J Biomed Semantics. 2016;7. doi:10.1186/s13326-016-0088-7

24. Smillie CS, Biton M, Ordovas-Montanes J, Sullivan KM, Burgin G, Graham DB, et al. Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. Cell. 2019;178: 714-730.e22. doi:10.1016/j.cell.2019.06.029

25. Koscielny G, An P, Carvalho-Silva D, Cham JA, Fumis L, Gasparyan R, et al. Open

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Targets: a platform for therapeutic target identification and validation. Nucleic Acids Res. 2017;45: D985–D994. doi:10.1093/nar/gkw1055

26. Mountjoy E, Schmidt EM, Carmona M, Schwartzentruber J, Peat G, Miranda A, et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. Nat Genet. 2021;53: 1527–1533. doi:10.1038/s41588-021-00945-5

27. Mostafavi H, Spence JP, Naqvi S, Pritchard JK. Systematic differences in discovery of genetic effects on gene expression and complex traits. Nat Genet. 2023;55: 1866–1875. doi:10.1038/s41588-023-01529-1

28. Andreani J, Guerois R. Evolution of protein interactions: from interactomes to interfaces. Arch Biochem Biophys. 2014;554: 65–75. doi:10.1016/j.abb.2014.05.010

29. Lun ATL, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Research. 2016. p. 2122. doi:10.12688/f1000research.9501.2

30. Sikkema L, Ramírez-Suástegui C, Strobl DC, Gillett TE, Zappia L, Madissoon E, et al. An integrated cell atlas of the lung in health and disease. Nat Med. 2023;29: 1563–1577. doi:10.1038/s41591-023-02327-2

31. Suo C, Dann E, Goh I, Jardine L, Kleshchevnikov V, Park J-E, et al. Mapping the developing human immune system across organs. Science. 2022;376: eabo0510. doi:10.1126/science.abo0510

32. Navarra SV, Guzmán RM, Gallacher AE, Hall S, Levy RA, Jimenez RE, et al. Efficacy and safety of belimumab in patients with active systemic lupus erythematosus: a randomised, placebo-controlled, phase 3 trial. Lancet. 2011;377: 721–731. doi:10.1016/s0140-6736(10)61354-2

33. Panush RS. A phase III, randomized, placebo-controlled study of belimumab, a monoclonal antibody that inhibits B lymphocyte stimulator, in patients with systemic lupus erythematosus. Year B Med. 2012;2012: 17–18. doi:10.1016/j.ymed.2012.09.010

34. Fillatreau S, Manfroi B, Dörner T. Toll-like receptor signalling in B cells during systemic lupus erythematosus. Nat Rev Rheumatol. 2021;17: 98–108. doi:10.1038/s41584-020-00544-4

35. Taraseviciene-Stewart L, Voelkel NF. Molecular pathogenesis of emphysema. J Clin Invest. 2008;118: 394–402. doi:10.1172/JCI31811

36. Pahal P, Avula A, Sharma S. Emphysema. StatPearls Publishing; 2023. Available: https://www.ncbi.nlm.nih.gov/books/NBK482217/

37. De Simone V, Guarise P, Zanotto G, Morando G. Reduction in pulmonary artery pressures with use of sacubitril/valsartan. J Cardiol Cases. 2019;20: 187–190. doi:10.1016/j.jccase.2019.08.006

38. McIvor RA. Future options for disease intervention: important advances in phosphodiesterase 4 inhibitors. Eur Respir Rev. 2007;16: 105–112.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

doi:10.1183/09059180.00010504

39. Spina D. PDE4 inhibitors: current status. Br J Pharmacol. 2008;155: 308–315. doi:10.1038/bjp.2008.307

40. Harrison RK. Phase II and phase III failures: 2013-2015. Nat Rev Drug Discov. 2016;15: 817–818. doi:10.1038/nrd.2016.184

41. Jagadeesh KA, Dey KK, Montoro DT, Mohan R, Gazal S, Engreitz JM, et al. Identifying disease-critical cell types and cellular processes by integrating single-cell RNA-sequencing and human genetics. Nat Genet. 2022;54: 1479–1492. doi:10.1038/s41588-022-01187-9

42. Jackson HW, Fischer JR, Zanotelli VRT, Ali HR, Mechera R, Soysal SD, et al. The single-cell pathology landscape of breast cancer. Nature. 2020;578: 615–620. doi:10.1038/s41586-019-1876-x

43. Van Galen P, Hovestadt V, Wadsworth M II, Hughes T, Griffin GK, Verga JA, et al. Single-cell RNA-seq reveals AML cellular hierarchies relevant to clinical outcomes and immunity. Blood. 2018;132: 542–542. doi:10.1182/blood-2018-99-113502

44. Dominguez CX, Müller S, Keerthivasan S, Koeppen H, Hung J, Gierke S, et al. Single-cell RNA sequencing reveals stromal evolution into LRRC15+ myofibroblasts as a determinant of patient response to cancer immunotherapy. Cancer Discov. 2020;10: 232–253. doi:10.1158/2159-8290.CD-19-0644

45. Imai Y, Kusakabe M, Nagai M, Yasuda K, Yamanishi K. Dupilumab effects on innate lymphoid cell and helper T cell populations in patients with atopic dermatitis. JID Innov. 2021;1: 100003. doi:10.1016/j.xjidi.2021.100003

46. Sun D, Guan X, Moran AE, Wu L-Y, Qian DZ, Schedin P, et al. Identifying phenotype-associated subpopulations by integrating bulk and single-cell sequencing data. Nat Biotechnol. 2022;40: 527–538. doi:10.1038/s41587-021-01091-3

47. Hekselman I, Yeger-Lotem E. Mechanisms of tissue and cell-type specificity in heritable traits and diseases. Nat Rev Genet. 2020;21: 137–150. doi:10.1038/s41576-019-0200-9

48. Liu X, Li YI, Pritchard JK. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell. 2019;177: 1022-1034.e6. doi:10.1016/j.cell.2019.04.014

49. Westra H-J, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013;45: 1238–1243. doi:10.1038/ng.2756

50. Gamazon ER, GTEx Consortium, Segrè AV, van de Bunt M, Wen X, Xi HS, et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. Nat Genet. 2018;50: 956–967. doi:10.1038/s41588-018-0154-4

51. Ahlmann-Eltze C, Huber W. Analysis of multi-condition single-cell data with latent embedding multivariate regression. bioRxiv. 2023. doi:10.1101/2023.03.06.531268

52. Boyeau P, Regier J, Gayoso A, Jordan MI, Lopez R, Yosef N. An empirical Bayes

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

method for differential expression analysis of single cells with deep generative models. Proc Natl Acad Sci U S A. 2023;120: e2209124120. doi:10.1073/pnas.2209124120

53. Missarova A, Dann E, Rosen L, Satija R, Marioni J. Sensitive cluster-free differential expression testing. bioRxiv. 2023. doi:10.1101/2023.03.08.531744

54. Dann E, Henderson NC, Teichmann SA, Morgan MD, Marioni JC. Differential abundance testing on single-cell data using k-nearest neighbor graphs. Nat Biotechnol. 2022;40: 245–253. doi:10.1038/s41587-021-01033-z

55. Skinnider MA, Squair JW, Kathe C, Anderson MA, Gautier M, Matson KJE, et al. Cell type prioritization in single-cell data. Nat Biotechnol. 2021;39: 30–34. doi:10.1038/s41587-020-0605-1

56. Dann E, Cujba A-M, Oliver AJ, Meyer KB, Teichmann SA, Marioni JC. Precise identification of cell states altered in disease using healthy single-cell references. Nat Genet. 2023;55: 1998–2008. doi:10.1038/s41588-023-01523-7

57. Xu C, Prete M, Webb S, Jardine L, Stewart BJ, Hoo R, et al. Automatic cell-type harmonization and integration across Human Cell Atlas datasets. Cell. 2023;186: 5876-5891.e20. doi:10.1016/j.cell.2023.11.026

58. Kanemaru K, Cranley J, Muraro D, Miranda AMA, Ho SY, Wilbrey-Clark A, et al. Spatially resolved multiomics of human cardiac niches. Nature. 2023;619: 801–810. doi:10.1038/s41586-023-06311-1

59. Van den Berge K, Roux de Bézieux H, Street K, Saelens W, Cannoodt R, Saeys Y, et al. Trajectory-based differential expression analysis for single-cell sequencing data. Nat Commun. 2020;11: 1201. doi:10.1038/s41467-020-14766-3

60. Gayoso A, Weiler P, Lotfollahi M, Klein D, Hong J, Streets A, et al. Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells. Nat Methods. 2023. doi:10.1038/s41592-023-01994-w

61. Dimitrov D, Türei D, Garrido-Rodriguez M, Burmedi PL, Nagai JS, Boys C, et al. Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. Nat Commun. 2022;13: 3224. doi:10.1038/s41467-022-30755-0

62. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. Nat Protoc. 2020;15: 1484–1506. doi:10.1038/s41596-020-0292-x

63. Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat Methods. 2020;17: 147–154. doi:10.1038/s41592-019-0690-6

64. Suo C, Polanski K, Dann E, Lindeboom RGH, Vilarrasa-Blasi R, Vento-Tormo R, et al. Dandelion uses the single-cell adaptive immune receptor repertoire to explore lymphocyte developmental origins. Nat Biotechnol. 2023. doi:10.1038/s41587-023-01734-7

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

65. Kuppe C, Ramirez Flores RO, Li Z, Hayat S, Levinson RT, Liao X, et al. Spatial multi-omic map of human myocardial infarction. Nature. 2022;608: 766–777. doi:10.1038/s41586-022-05060-x

66. Sun C, Wang A, Zhou Y, Chen P, Wang X, Huang J, et al. Spatially resolved multi-omics highlights cell-specific metabolic remodeling and interactions in gastric cancer. Nat Commun. 2023;14: 2692. doi:10.1038/s41467-023-38360-5

67. Rocque B, Guion K, Singh P, Bangerth S, Pickard L, Bhattacharjee J, et al. Technical optimization of spatially resolved single-cell transcriptomic datasets to study clinical liver disease. Res Sq. 2023. doi:10.21203/rs.3.rs-3307940/v1

68. fbastian, Niknejad A, Mungall C, Echchiki A, Matentzoglu N, Caron A, et al. obophenotype/developmental-stage-ontologies: August 2023 release. Zenodo; 2023. doi:10.5281/ZENODO.592936

69. Squair JW, Gautier M, Kathe C, Anderson MA, James ND, Hutson TH, et al. Confronting false discoveries in single-cell differential expression. Nat Commun. 2021;12: 5692. doi:10.1038/s41467-021-25960-2

70. Crowell HL, Soneson C, Germain P-L, Calini D, Collin L, Raposo C, et al. Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. Nat Commun. 2020;11: 6077. doi:10.1038/s41467-020-19894-4

71. Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 2016;17: 75. doi:10.1186/s13059-016-0947-7

72. Lun ATL, Chen Y, Smyth GK. It's DE-licious: A recipe for differential expression analyses of RNA-seq experiments using quasi-likelihood methods in edgeR. Methods Mol Biol. 2016;1418: 391–416. doi:10.1007/978-1-4939-3578-9_19

73. Ahlmann-Eltze C, Huber W. glmGamPoi: fitting Gamma-Poisson generalized linear models on single cell count data. Bioinformatics. 2021;36: 5701–5702. doi:10.1093/bioinformatics/btaa1009

74. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17: 261–272. doi:10.1038/s41592-019-0686-2

75. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44: W90-7. doi:10.1093/nar/gkw377

76. Fang Z, Liu X, Peltz G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. Bioinformatics. 2023;39. doi:10.1093/bioinformatics/btac757

77. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat Methods. 2016;13: 966–967. doi:10.1038/nmeth.4077

medRxiv preprint doi: https://doi.org/10.1101/2024.04.04.24305313; this version posted April 5, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

78. Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. 2019;29: 1363–1375. doi:10.1101/gr.240663.118

79. Badia-i-Mompel P, Vélez Santiago J, Braunger J, Geiss C, Dimitrov D, Müller-Dott S, et al. decoupleR: ensemble of computational methods to infer biological activities from omics data. Bioinform Adv. 2022;2. doi:10.1093/bioadv/vbac016

80. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581: 434–443. doi:10.1038/s41586-020-2308-7