PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Building, benchmarking, and exploring perturbative maps of transcriptional and morphological data

Safiye Celik[1]*, Jan-Christian Hütter[2], Sandra Melo Carlos[2], Nathan H. Lazar[1], Rahul Mohan[2], Conor Tillinghast[1], Tommaso Biancalani[2], Marta M. Fay[1], Berton A. Earnshaw[1], Imran S. Haque[1]

1 Recursion, Salt Lake City, Utah, United States of America, **2** Genentech, South San Francisco, California, United States of America

* info@rxrx.ai

OPEN ACCESS

Citation: Celik S, Hütter J-C, Carlos SM, Lazar NH, Mohan R, Tillinghast C, et al. (2024) Building, benchmarking, and exploring perturbative maps of transcriptional and morphological data. PLoS Comput Biol 20(10): e1012463. https://doi.org/10.1371/journal.pcbi.1012463

Editor: Thouis Ray Jones, Broad Institute, UNITED STATES OF AMERICA

Received: February 18, 2023

Accepted: September 6, 2024

Published: October 1, 2024

Data Availability Statement: Code is available at https://github.com/recursionpharma/EFAAR_benchmarking. The datasets we used in this manuscript are publicly available. Here are the links: RxRx3: https://rxrx3.rxrx.ai/downloads GWPS: https://plus.figshare.com/articles/dataset/_Mapping_information-rich_genotype-phenotype_landscapes_with_genome-scale_Perturb-seq_Replogle_et_al_2022_processed_Perturb-seq_datasets/20029387 cpg0016: https://cellpainting-gallery.s3.amazonaws.com/index.html#cpg0016-jump/ cpg0021: https://cellpainting-gallery.s3.

Abstract

The continued scaling of genetic perturbation technologies combined with high-dimensional assays such as cellular microscopy and RNA-sequencing has enabled genome-scale reverse-genetics experiments that go beyond single-endpoint measurements of growth or lethality. Datasets emerging from these experiments can be combined to construct perturbative "maps of biology", in which readouts from various manipulations (e.g., CRISPR-Cas9 knockout, CRISPRi knockdown, compound treatment) are placed in unified, relatable embedding spaces allowing for the generation of genome-scale sets of pairwise comparisons. These maps of biology capture known biological relationships and uncover new associations which can be used for downstream discovery tasks. Construction of these maps involves many technical choices in both experimental and computational protocols, motivating the design of benchmark procedures to evaluate map quality in a systematic, unbiased manner. Here, we (1) establish a standardized terminology for the steps involved in perturbative map building, (2) introduce key classes of benchmarks to assess the quality of such maps, (3) construct 18 maps from four genome-scale datasets employing different cell types, perturbation technologies, and data readout modalities, (4) generate benchmark metrics for the constructed maps and investigate the reasons for performance variations, and (5) demonstrate utility of these maps to discover new biology by suggesting roles for two largely uncharacterized genes.

Author summary

Due to the rapid advancements in genetic perturbation, laboratory robotics, sequencing, and computer vision, more researchers are now generating datasets that capture cellular responses to genetic perturbations. These datasets can be powerful discovery tools for examining known biological relationships and revealing new associations in an unbiased manner when paired with a computational pipeline that can assemble the data into a digestible format. However, the challenge arises from the variety of cellular models, assay

amazonaws.com/index.html#cpg0021-periscope/broad/workspace/profiles/HeLa/.

Funding: The author(s) received no specific funding for this work.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests:. All authors are current or former employees of Recursion Pharmaceuticals, Inc. or Genentech, Inc., and have received real or optional ownership interest in the companies.

designs, terminologies, codebases, and analysis methods involved. In this work we define a unified framework for building and benchmarking perturbative maps, benchmark four different datasets assembled into 18 different maps, explore the impact of different design decisions, and demonstrate how these maps can be used to elucidate gene functions. Our goal is to facilitate comparisons across various technologies and methods by introducing a shared language for the field. The open-source codebase, capable of incorporating new methods, aims to be a resource for researchers developing laboratory or computational methodology. While we caution against definitive recommendations due to numerous variables at play, we hope to stimulate studies directly comparing methods under controlled conditions. Our framework can also help evaluate combining maps across modalities as the field progresses.

Introduction

Advances in genome editing technologies and high-throughput screening capabilities have enabled building perturbative maps through unbiased, large-scale profiling of genetic perturbations. These maps have massive potential to uncover novel biology and, when paired with compound screening, to accelerate drug discovery processes. The increasing prevalence of data generated by genetic perturbation technologies combined with high-dimensional assays [1–6] signals a wider recognition of the power of these data to untangle complex biological mechanisms and to spur advancements in drug discovery.

Recent studies have utilized single-cell pooled screening techniques to create genome-scale perturbation datasets. Replogle et al. [1] used pooled CRISPR interference (CRISPRi) libraries to generate genome-wide perturbation data with single cell RNA-seq as the readout. Several other studies [2–4] used pooled CRISPR-Cas9-mediated gene knockouts with cellular imaging [7] followed by in-situ sequencing of the molecular barcode to assign guides to individual cells and obtain single-cell phenotypic perturbation readouts. While pooled screens are typically resource-efficient and cost-effective, it is not possible to tightly control the number of cells transfected with each specific guide RNA and to account for the interactions between cells that receive different perturbations. In contrast, recent studies [5, 6] utilized array-based screening where highly-automated labs apply thousands of distinct perturbations to separate cell populations in multi-well plates using CRISPR-Cas9 constructs that target individual genes.

These studies each generated map building pipelines and analyses for their proposed datasets. However, to our best knowledge, no study has yet explored different genome-wide perturbative maps in a comprehensive manner using a shared language and consistent benchmarking measures, allowing for a comparison of different maps, particularly the computational pipeline choices for a given perturbation dataset.

Here, we establish a systematic framework and vocabulary (Fig 1), as well as a codebase (github.com/recursionpharma/EFAAR_benchmarking), for constructing and evaluating such maps, which we expect will lead to more comparable analysis and optimization of maps going forward. We describe perturbation signal benchmarks for assessing the effect and consistency of individual perturbations, and biological relationship benchmarks for assessing biological relevance of a map through its ability to recapitulate annotated relationships from large-scale public sources. It is important to note that it is not our goal to establish the optimal analytical choices for each perturbation dataset, but to demonstrate how the benchmarks we present can be used to optimize the analytical choices for each map.

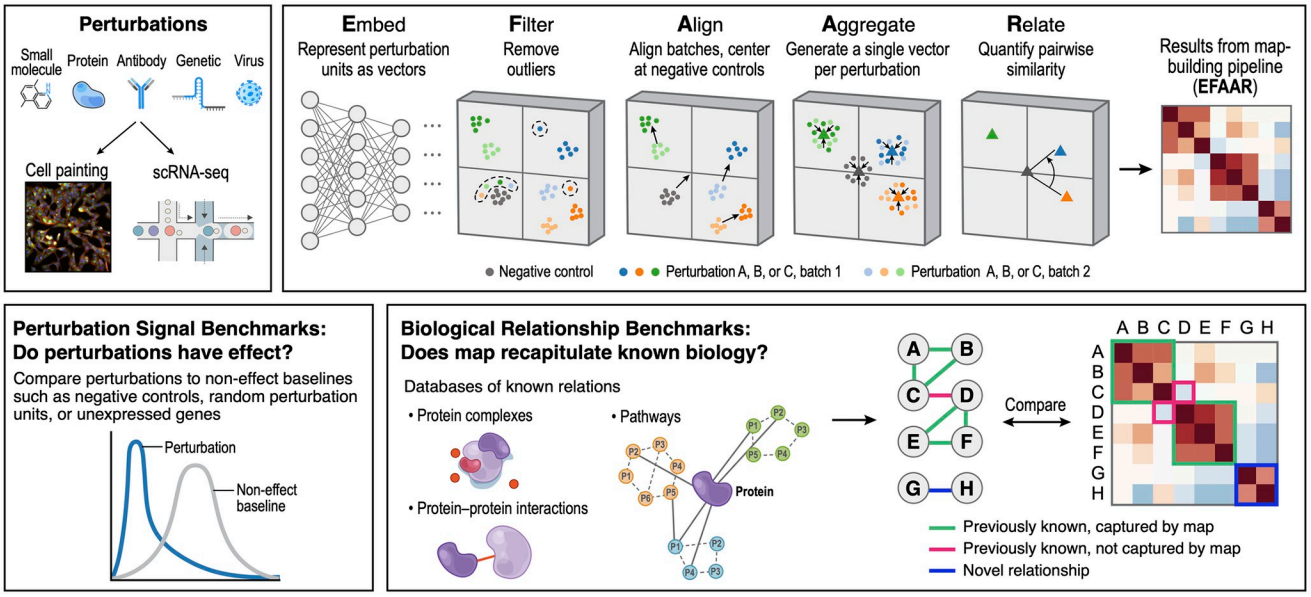

Fig 1. Graphical abstract of the introduced framework.

https://doi.org/10.1371/journal.pcbi.1012463.g001

We construct maps from four recent gene perturbation datasets which perturb at least ∼ 8,000 genes each and have different perturbation modalities, readouts, and experimental setups. We then use our benchmarking framework to compute metrics for multiple computational analysis pipelines on each of these datasets. We use these metrics to explore the strengths and weaknesses of different pipelines in order to provide a deeper understanding of their performance and applicability. An important contribution of our study is the reporting of results from various annotation sources when examining the biological relationship benchmarks. While previous studies focused on recapitulation of protein complexes, we add in protein-protein interactions derived from pathways (Reactome) and signaling cascades (SIGNOR). With the maps we have constructed and analyzed, we provide compelling evidence for a relationship between *C18orf21*, an uncharacterized gene, and *ABT1*, *DKC1*, *POP1*, *POP5*, *RPP30*, and *UTP23*, suggesting that *C18orf21* participates in the RNase mitochondrial RNA processing complex. Similarly, we provide evidence for *C1orf131* interacting with *AATF*, *DDX52*, *NOL6*, *PDCD11*, *PNO1*, *RRP12*, and *UTP20*, supporting its involvement in the small subunit processome.

Results

Map building pipeline

Throughout this paper, we refer to a series of experiments involving genome-scale perturbations as a *perturbation* ***dataset*** and the final set of perturbation embeddings and associations between these as a *perturbative* ***map***. In the main results of this work, we present 18 maps constructed on four datasets.

We call the smallest experimental entity that is measured in a map context a *perturbation unit*. This is a cell in single-cell assays, particularly pooled screens in which each cell receives a distinct barcoded perturbation [1–3, 8]. In the arrayed setting, a perturbation unit is a well containing hundreds of cells treated with the same perturbation [6, 7, 9]. Each perturbation

unit is associated with assay data (e.g., images or gene transcript counts). Building a map which relates perturbations in a meaningful way from these raw assay data requires a number of post-experimental processing steps which we divide into five categories and refer to as an EFAAR pipeline.

- **E**mbedding assay data from each perturbation unit to generate a tractably-sized numeric representation
- **F**iltering perturbation units that do not pass quality criteria
- **A**ligning different batches of perturbation units
- **A**ggregating replicate units representing each targeted perturbation (e.g., a gene)
- **R**elating different perturbations to each other with one or more numeric values

These EFAAR steps may take place in a different order, multiple times (e.g., perturbation units may be filtered pre- and post-embedding), or potentially in a single end-to-end process. We highlight these steps as essential when constructing any perturbative map.

Embedding. This step is aimed at reducing high-dimensional assay data (e.g., 20,000 gene expression values or over one million image pixel values) to a tractable numerical representation that can be used for downstream tasks. For imaging data, embeddings may come from morphological features derived using software like CellProfiler [10] or from intermediate layers of neural networks [5]. For expression data, linear dimensionality reduction methods like principal component analysis (PCA) or non-linear methods based on neural networks may be used [11, 12].

Filtering. In any experimental screening process, some perturbation units will not satisfy pre-defined quality criteria and need to be filtered out. This filtering can occur before or after embeddings are generated. Examples include wells with too high or too low pixel intensity or cells that receive more than one target guide.

Aligning. A *batch effect* is a systematic bias shared by all observations obtained under similar experimental conditions (e.g., microscopy acquisition artifacts, cell donor batch, incubation times) that potentially confound the interpretation of desired biological signal. Aligning across batches can reduce the effects of these unintended variations and bring out more biologically meaningful relationships.

A baseline approach for aligning perturbation units is to use control units in each batch to center and scale features. Another linear method aligning not only the first order statistics but also the covariance structures is Typical Variation Normalization (TVN) [13]. Non-linear methods based on nearest neighbor matching [14, 15] or variational autoencoders have been particularly successful for the alignment of single cell transcriptomic data [11, 12] as well as modeling batch effects in image data [16–18]. Another important technique for image data is Instance Normalization [19] which involves normalizing the features across each channel in an individual sample. This aids in removing bias in feature statistics when training data is divided into small computational batches. In gene expression domain, probabilistic batch effect correction methods that jointly model contributions from genotype and confounding factors have been used to map expression quantitative trait loci (eQTLs) [20]. ComBat and ComBat-seq [21, 22] utilize empirical Bayes models to adjust for batch effects and have been widely used for gene expression data.

Aggregating. There are typically multiple technical or biological replicates representing each perturbation in a given dataset, e.g., the same perturbation may be applied to dozens of wells or hundreds of cells and these must be combined in order to produce a final representation of a perturbation. Coordinate-wise mean and median aggregation are commonly used,

while more advanced methods like the Tukey median [23] may reduce the impact of outliers on the final representation, but increase computational complexity.

Relating. Identifying relationships between biological entities (e.g., gene-gene interactions arising from protein complexes or signaling pathways) is an important use case for maps built based on genetic perturbations. Computing distances (e.g., Euclidean distance) or other similarity or dissimilarity measures (e.g., cosine similarity) between aggregated perturbation representations is commonly used as a proxy for relationships. These, in turn, can also be used to visualize the global structure of perturbations through further dimensionality reduction techniques such as uniform manifold approximation (UMAP) [24] or minimum-distortion embedding (MDE) [25].

Map benchmarking pipeline

To assess maps built using different perturbation technologies, different readout data modalities, and different EFAAR pipelines, it is important to benchmark the resulting perturbation representations both in terms of the effect of the individual perturbations and interactions between perturbations. We call these "perturbation signal benchmarks" and "biological relationship benchmarks", respectively.

Perturbation signal benchmarks. Perturbation signal benchmarks assess the consistency and magnitude of the representations of individual perturbations in a map. We measure consistency by the average cosine of the angle between replicates of a perturbation (Methods: Perturbation signal consistency), and magnitude by the energy distance [26, 27] between the control and perturbation samples (Methods: Perturbation signal magnitude). S1 Fig provides a visual description of the consistency and magnitude measures.

These benchmarks identify the genes whose representation achieves a p-value $< .05$ in the associated statistical test using the unexpressed genes as the baseline for no signal. (Methods: Identification of significance of perturbation signal). The fraction of such genes can be compared between different map processing pipelines (EFAAR parameter choices) and stratified by global annotations like gene expression or functional gene groups. Genes with a significant p-value for both consistency and magnitude receive higher priority when selecting hypotheses, such as identifying targets to pursue in downstream drug discovery tasks.

Biological relationship benchmarks. A typical use case for a map of biology is to discover novel, biologically relevant relationships. The following five annotation sources are used to assess the degree to which each map detects biological relationships. The underlying hypothesis is that if a map can identify known relationships to a high degree, it is an indication that it demonstrates a strong representation of existing biology, and is therefore more likely to accurately represent and uncover novel biological relationships.

- **CORUM**: gene clusters representing protein complexes [28]
- **HuMAP**: gene clusters representing protein complexes [29]
- **Reactome**: protein-protein interactions derived from pathways [30]
- **SIGNOR**: signaling pathway interactions [31]
- **StringDB**: functional protein-protein associations [32]

Recall of annotated pairs is reported for the most extreme 10% of pairwise relationships (Methods: Recall computation). The main premise is that genes whose protein products have related functions or that act in concert will produce readouts that, when properly processed,

display geometric relationships (e.g., similarity). We consider 5% from both tails of the pairwise similarity distribution since negative relationships can indicate opposing functions between genes. A random, uninformative map would achieve a recall of 10%.

Across the five relationship annotation sources we utilize for benchmarking, there are a total of 143,252 unique relationships. 29,600 of these relationships exist in at least two sources, and 54 of them exist in all five sources (S2 Fig).

In this study we analyze maps of gene perturbations, as multiple genome-scale datasets have recently become available [1, 2, 5, 6]. The framework presented here can also be applied to datasets with other types of perturbations. For example, when a map includes both genes and small molecules at a large scale, sources annotating relationships between small molecules and their target genes [33, 34] can be used for benchmarking.

Applications of the map building and benchmarking framework

Datasets. We built and benchmarked transcriptional and morphological maps on four genome-scale CRISPR-based perturbation datasets:

- **RxRx3** from Recursion contains deep neural network embeddings of phenomic images, where CRISPR-Cas9 was applied in an arrayed format to target ∼17,000 genes in primary HUVEC cells [5, 35].
- **GWPS** (Genome-Wide Perturb-Seq) contains single-cell RNA-seq counts as the readout, where CRISPRi was applied to knock down ∼10,000 expressed genes in K562 cells [1].
- **cpg0016** from the JUMP (Joint Undertaking of Morphological Profiling) consortium contains CellProfiler features of Cell Painting [7] images, where CRISPR-Cas9 was applied in an arrayed format to target ∼8,000 druggable genes in the U2OS cell line [6].
- **cpg0021**, contains CellProfiler features of Cell Painting images, where CRISPR-Cas9 was applied in a pooled optical screening format to perturb ∼20,000 genes in HeLa cells [2].

Table 1 presents an overview of the properties of the four datasets. For the three morphology datasets (RxRx3, cpg0016, and cpg0021), details about cell painting protocols are in S1 Table.

EFAAR pipelines. Input data for the EFAAR pipelines were collected through the original sources of the four datasets (Methods: Perturbation data collection). The embedding, filtering, and alignment steps for each dataset and processing pipeline are described below. Aggregation is performed by taking the mean of the aligned embeddings. Relationships are computed using the cosine similarity between the aggregated embeddings.

RxRx3 pipelines.

- Raw: This refers to the 128-dimensional well-level RxRx3 embeddings. They were generated by passing images through a weakly-supervised convolutional neural network (CNN) pretrained on a large set of proprietary Recursion data [36].
- Raw-CS: Raw RxRx3 embeddings are aligned via centering and scaling each feature to the mean and standard deviation of the controls in each batch.
- Raw-TVN: TVN [13] is applied on the RxRx3 embeddings after first centering and scaling features to the mean and standard deviation of the controls globally (not by batch). TVN aligns the data by PCA, centering and scaling globally, and correlation alignment (CORAL) [37] for each batch, all by controls.

Table 1. Properties of the four datasets on which we are applying the map building and benchmarking framework. These statistics specifically pertain to the portions of these datasets utilized in our application (Methods: Perturbation data collection) and not necessarily the full data available from each source. We use original study authors' definition of "expressed" gene whenever applicable (Methods: Filtering genes based on expression).

	RxRx3	GWPS	cpg0016	cpg0021
Cell type	HUVEC	K562	U2OS	HeLa
Perturbation type	CRISPR-Cas9	CRISPRi	CRISPR-Cas9	CRISPR-Cas9
Perturbation unit	well	cell	well	cell*
Readout modality	morphology	transcriptome	morphology	morphology
Total perturbed genes	17,060	9,866	7,977	20,422
Expressed pert. genes	12,250	9,866	6,389	15,490
Mode num guides / gene	6	2	4	4
Guide construct	independent†	paired	pooled‡	independent†
Control guides	intron-targeting	non-targeting	non-targeting	non-targeting
Num non-ctrl samples	1,759,062	1,914,250	42,687	414,445*
Feature source	Deep Learning	RNA-seq counts	CellProfiler	CellProfiler

*For cpg0021, we use pre-aggregated guide-level CellProfiler features. The sample size we report is also based on the guide-level data.

†When each perturbation unit gets only one of multiple guides, the replicate pool for the perturbation signal benchmark computation includes biological replicates (guides) as well as the technical replicates.

‡Screening was in an arrayed format, and all of the guides per gene were pooled in the same well.

https://doi.org/10.1371/journal.pcbi.1012463.t001

- PCA: PCA is applied on the raw embeddings after centering and scaling all features in each batch, and all 128 principal components (PCs) are retained as the embeddings.
- PCA-CS: PCA-transformed embeddings are aligned via centering and scaling each batch by controls.
- PCA-TVN: A similar pipeline to the Raw-TVN, but applied on the PCA-transformed embeddings as the input instead of the raw embeddings.

GWPS pipelines.

- scVI: Raw single-cell expression values are embedded to 128 dimensions using scVI (single-cell Variational Inference) [12]. scVI is a conditional variational auto-encoder providing both embedding and alignment in a single network (Methods: scVI network architecture).
- scVI-CS: scVI embeddings are aligned via centering and scaling each batch by controls.
- scVI-TVN: A similar pipeline to the Raw-TVN in RxRx3, but applied on the scVI embeddings as the input.
- PCA: PCA is applied on the raw RNA-seq counts and the top 128 PCs are retained as the embeddings.
- PCA-CS: The same pipeline as in RxRx3.
- PCA-TVN: The same pipeline as in RxRx3.

cpg0016 pipelines.

- PCA: PCA is applied on the filtered CellProfiler features (Methods: Filtering CellProfiler features for cpg0016) and the top 128 PCs are retained as the embeddings.

- PCA-CS: The same pipeline as in RxRx3 and GWPS.
- PCA-TVN: The same pipeline as in RxRx3 and GWPS.

cpg0021 pipelines.

- PCA: PCA is applied on the CellProfiler features which are pre-aggregated across cells to the guide level, and the top 128 PCs are retained as the embeddings.
- PCA-CS: The same pipeline as in RxRx3, GWPS, and cpg0016.
- PCA-TVN: The same pipeline as in RxRx3, GWPS, and cpg0016.

Here, we report results from pipelines generating 128-dimensional embeddings. S2 Table includes biological relationship benchmark results for larger embedding spaces. The EFAAR choices above were made to analyze the four perturbation datasets as consistently as possible. This exploration is not intended to be all-encompassing; other methods might yield better results for different perturbation datasets.

Benchmarking results. Here we look at the perturbation signal and biological relationship benchmarks for different EFAAR pipelines. We particularly check how sensitive these benchmarks are to the changes in the EFAAR steps. This facilitates the selection of the most effective pipeline for discovering novel relationships.

We restrict the benchmarks to the perturbed genes that are expressed in the respective cell type (Methods: Filtering genes based on expression). While we attempted to minimize differences between datasets with consistent EFAAR choices, values are not directly comparable across datasets due to inherent differences in cell type, mode of genetic perturbation, sample size, and assay design. However, observed trends for different EFAAR pipelines within a dataset can be informative, revealing strengths and weaknesses for different computational methodologies.

Fig 2 provides the perturbation signal benchmarks (left) and biological relationship benchmarks (right) for the RxRx3 maps we constructed. TVN alignment shows an increased performance for all metrics and for both of the embedding methods (raw and PCA). We believe this is because TVN is more effective at reducing batch-to-batch variation.

Fig 3 shows benchmarks for different EFAAR pipelines on the GWPS dataset. Since this dataset does not include unexpressed gene perturbations, perturbation signal benchmarks cannot be calculated. Similar to the RxRx3 results, using TVN for the alignment step leads to an increased performance for both of the embedding methods (scVI and PCA). Interestingly, the PCA embeddings show a better performance than scVI across all alignment methods.

Figs 4 and 5 show the perturbation signal and biological relationship benchmark values for different EFAAR pipelines on the cpg0016 and cpg0021 datasets, respectively. Again, TVN alignment enhances the recapitulation of annotated gene-gene relationships for the majority of annotation sets (right), but PCA or PCA-CS performs better on perturbation signal benchmarks (left). For cpg0016, an improved batch alignment may have negative effects on perturbation signal benchmarks because the plate layouts are not fully randomized and plate position effects may add to the observed phenotype.

Looking across datasets, benchmark values are lower in the best-performing cpg0016 and cpg0021 maps than in the best-performing RxRx3 and GWPS maps. This could be due to a variety of differences across the datasets (Table 1). For example, although RxRx3, cpg0016, and cpg0021 are all morphology datasets that use CRISPR-Cas9 gene editing, RxRx3 employs intron-targeting guides as the negative controls. This approach may help distinguish between the effects of Cas9-induced cutting and the biological signals resulting from the knockout of

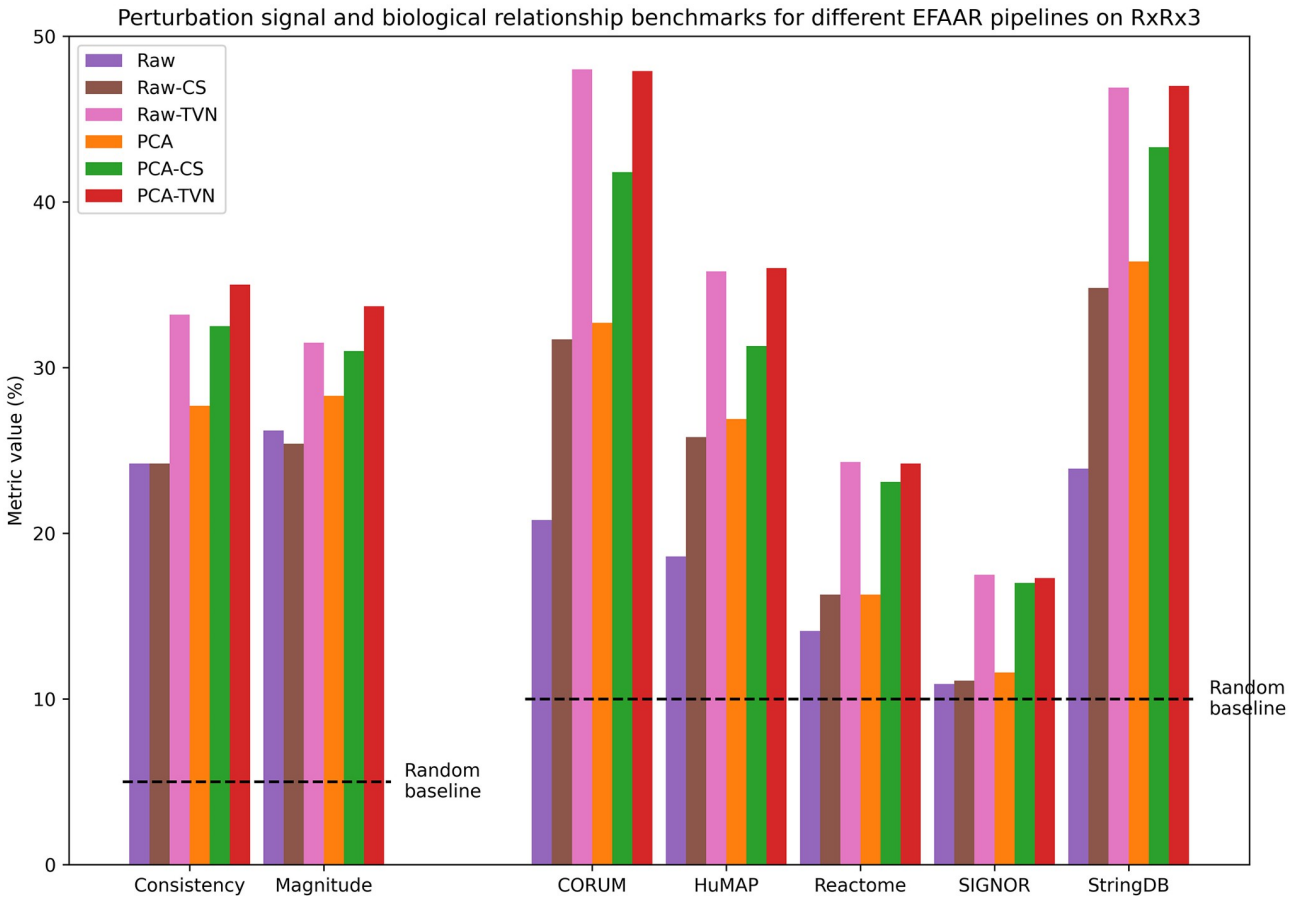

Fig 2. Benchmarking results in maps constructed using RxRx3 and different EFAAR pipelines. Bars for consistency and magnitude (left) show the percentage of perturbations with a significant p-value (< .05). Bars for CORUM, HuMAP, Reactome, SIGNOR and StringDB show the biological relationship benchmarks, i.e., the percentage of annotated relationships falling within the 5% tails (from each side) of the distribution of all pairwise cosine similarities.

https://doi.org/10.1371/journal.pcbi.1012463.g002

the target gene. Alternatively, the single-guide treatments in RxRx3 wells with ∼ 1,500 cells may contain more information or offer more stable representations of the gene perturbations than the single cells from cpg0021 or pooled-guide wells in cpg0016.

We further investigate how sample size influences the efficacy of maps in recapitulating biological relationships by randomly sub-sampling replicates (Methods: Noise sensitivity analyses) for the PCA, PCA-CS, and PCA-TVN maps for all datasets. Increasing the sample size typically leads to an increased recall of known biology and reduced error (measured across different random subsets)(S3–S6 Figs). RxRx3 and cpg0021 show the largest improvement with increased replicates which may be due to these datasets measuring guides separately and thus having a higher diversity across replicates.

Perturbative maps built using different technologies surface different biology. To broadly assess the utility of each dataset, we examined which CORUM protein complexes the maps are able to identify (Methods: Protein complex identification). We applied a stringent p-value cutoff of .01 during our exploration of maps for assessing their utility and using them for identifying poorly characterized cell functions. Examined complexes are not mutually exclusive and may overlap; they may also be subsets of each other.

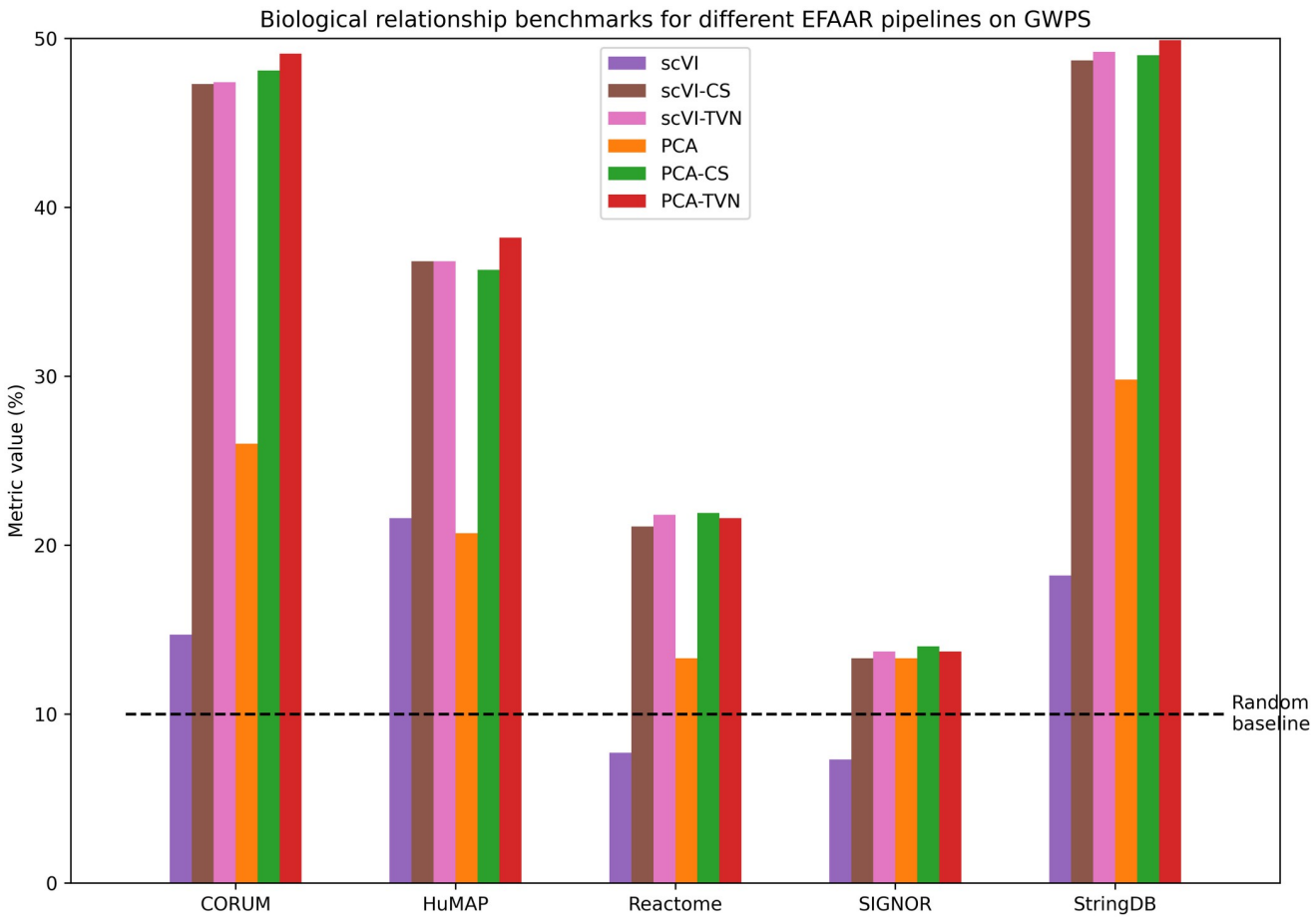

Fig 3. Biological relationship benchmarking results in maps constructed using different EFAAR pipelines on the GWPS dataset. Bar height shows the percentage of annotated relationships which fall within the 5% tails (from each side) of the distribution of all pairwise cosine similarities.

https://doi.org/10.1371/journal.pcbi.1012463.g003

Fig 6A shows the overlaps of significantly identified complexes across the three maps with fully unblinded metadata (GWPS, cpg0016, and cpg0021). We utilized maps built using the PCA-TVN pipeline as these perform best on the biological relationship benchmarks (Figs 3, 4 and 5). GWPS identifies the most complexes as well as the most unique complexes. It also identifies notably more complexes than the other two maps. The largest overlap occurs with complexes shared across GWPS and cpg0016, but not identified by cpg0021.

Next, we examined complexes identified by all three maps, as well as those uniquely identified by one of the maps. 12 complexes are consistently identified by all three maps. These involve fundamental cellular processes crucial for the proper functioning and regulation of a cell and represent key molecular machinery involved in maintaining cellular integrity, survival, and functionality (S3 Table). To delve deeper, we generated a split heatmap of six of the 12 complexes in GWPS and cpg0016, the two maps identifying the most complexes (Fig 6B). The 20S proteasome shows highly robust within-complex relationships in both transcriptomic and morphological maps, whereas the genes with RNA polymerase II core complex are similar to RNA polymerase II core complex and Mediator complex genes. This data is consistent with the interaction between the Mediator complex and RNA polymerase II [38].

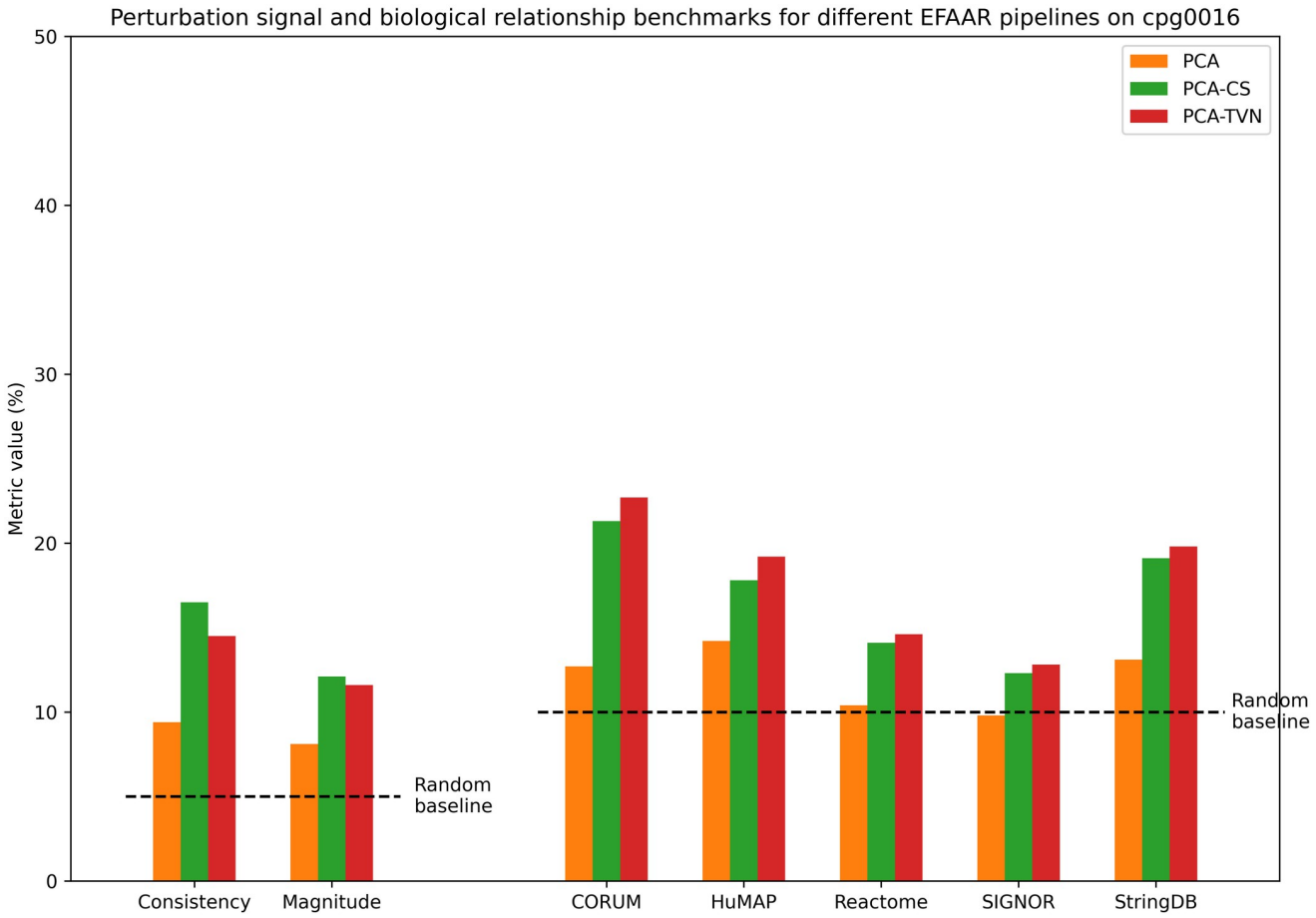

Fig 4. Benchmarking results in maps constructed using different EFAAR pipelines on the cpg0016 dataset. Bars for consistency and magnitude (left) show the percentage of perturbations with a significant p-value (< .05) for each measure. Bars for CORUM, HuMAP, Reactome, SIGNOR and StringDB show the biological relationship benchmarks, i.e., the percentage of annotated relationships falling within the 5% tails (from each side) of the distribution of all pairwise cosine similarities.

https://doi.org/10.1371/journal.pcbi.1012463.g004

Fifty complexes are uniquely identified by the CRISPRi-based GWPS map with a transcriptional readout (S4 Table). A large number of these show associations with chromatin remodeling and transcriptional regulation. One complex is uniquely identified by the cpg0016 map, though the p-value for it is also near the significance level in the cpg0021 map (S5 Table). Four complexes are uniquely identified by the cpg0021 map (S6 Table), which is the largest of the three datasets with morphological readout (Table 1). Emerin complex 25 and Intraflagellar Transport Complex B have roles in maintaining cell shape and structure, thus directly impacting cell morphology.

Next, we examined the cosine similarity structure of the Integrator protein complex which was explored in the GWPS study [1] and is one of the fifty complexes uniquely identified by the GWPS map (S4 Table). The previously uncharacterized gene *C7orf26* was suggested to be a subunit of this complex [1, 29], and was officially renamed *INTS15* in January 2022. We examined how well this complex's structure is captured by the maps we constructed using other perturbation datasets. This complex was identified by RxRx3 (p-value: 3.6e-38) and by GWPS (p-value: 2.2e-25), but not by cpg0021 (p-value: .053). Integrator's identification in the cpg0016 map cannot be assessed since only one of the genes in the complex was screened in the

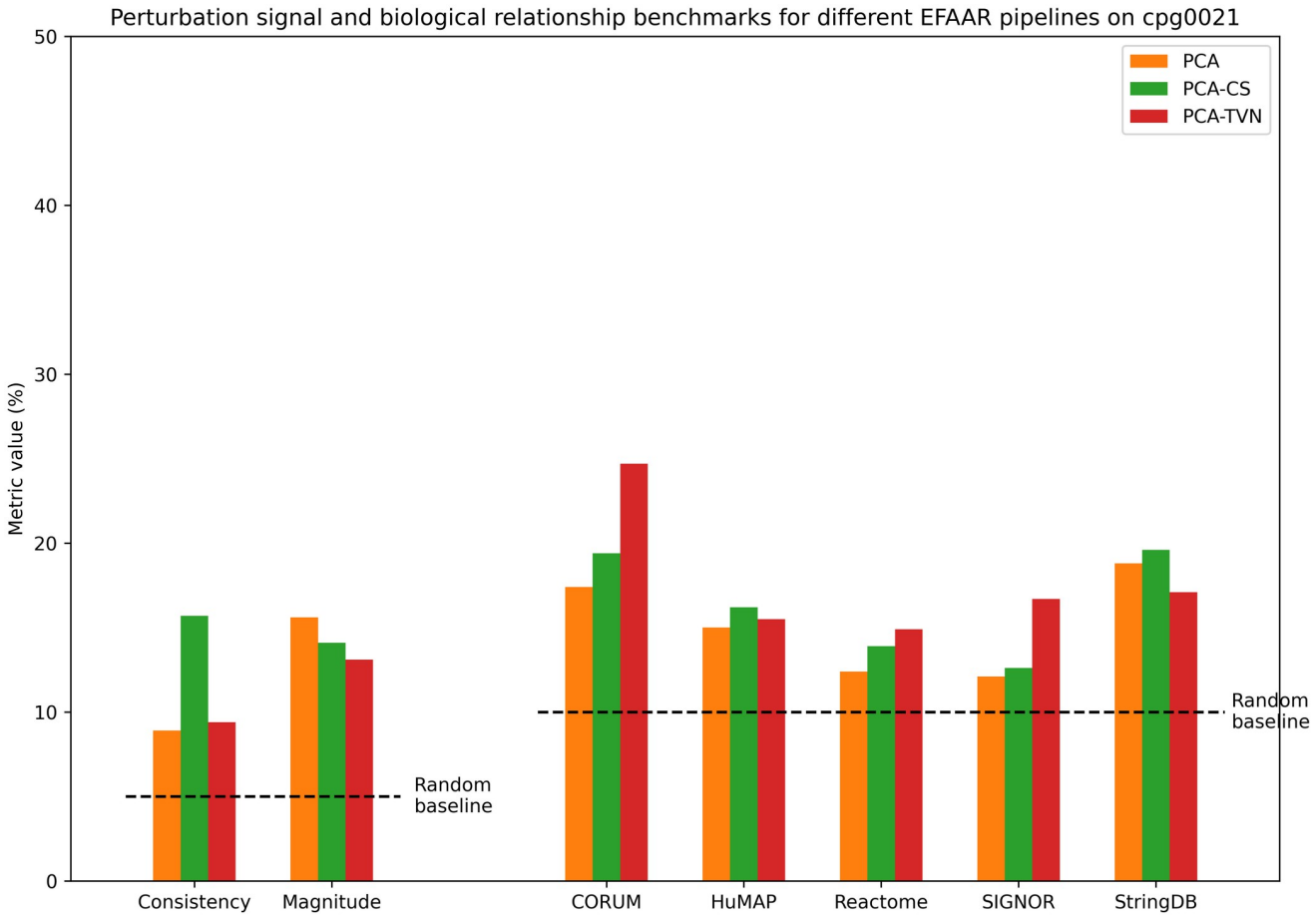

Fig 5. Benchmarking results in maps constructed using different EFAAR pipelines on the cpg0021 dataset. Bars for consistency and magnitude (left) show the percentage of perturbations with a significant p-value (< .05) for each measure. Bars for CORUM, HuMAP, Reactome, SIGNOR and StringDB bars show the biological relationship benchmarks, i.e., the percentage of annotated relationships falling within the 5% tails (from each side) of the distribution of all pairwise cosine similarities.

https://doi.org/10.1371/journal.pcbi.1012463.g005

cpg0016 dataset. Fig 7 represents the cosine similarity heatmap for the Integrator complex in the RxRx3 and GWPS maps. Both of these maps accurately identify the modular structure of the Integrator complex and place C7orf26 in the enhancer module with *INTS10*, *INTS13*, and *INTS14*. S7 Fig represents the heatmap for the cpg0021 map.

Evidence for the roles of uncharacterized genes *C18orf21* and *C1orf131*. Motivated by the shared cosine similarity structure of the Integrator complex in the GWPS and RxRx3 maps and a suggested role for the recently characterized *C7orf26*, we sought to determine whether there are any other uncharacterized genes whose roles might be elucidated using these maps. We focused on two uncharacterized genes: *C18orf21* and *C1orf131*, and examined their top connections in the PCA-TVN maps of RxRx3 and GWPS.

Six genes are in the top 25 most cosine similar genes to *C18orf21* in both GWPS and RxRx3 maps (Fig 8A): *ABT1*, *DKC1*, *POP1*, *POP5*, *RPP30*, and *UTP23*. For these, we performed a gene set enrichment analysis on Gene Ontology (GO) biological process, cellular component, and molecular function annotations from MSigDB [39]. This analysis identified terms associated with various RNA-related processes and complexes (Fig 8B). Key roles include the formation and function of ribonucleoprotein complexes (GOCC_SNO_S_RNA_CONTAINING_

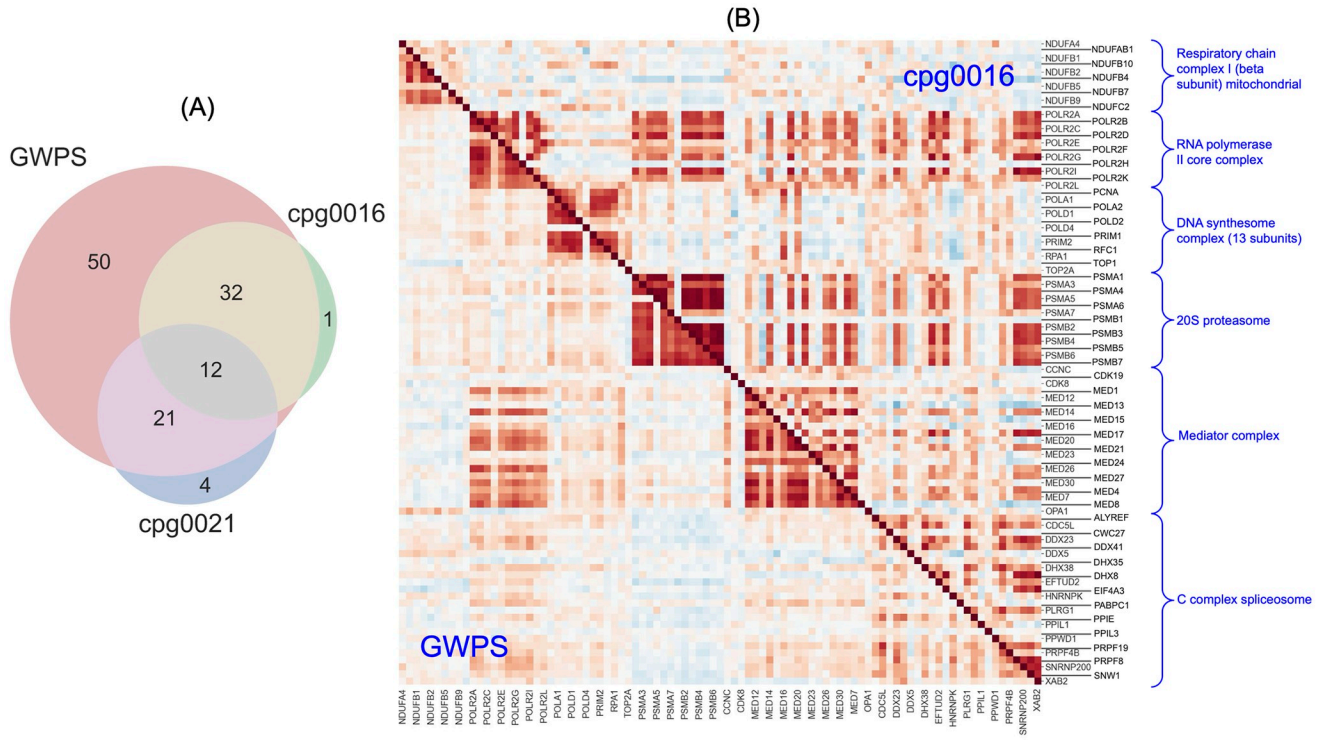

Fig 6. Comparative analysis of biology surfaced by different maps. (A)Venn diagram of the intersection of the CORUM protein complexes captured by the PCA-TVN maps from each of GWPS, cpg0016, and cpg0021. There are 153 evaluated complexes (those with at least ten expressed genes) for GWPS, 83 for cpg0016, and 169 for cpg0021. (B) A split cosine similarity heatmap of the genes in six non-overlapping complexes out of the 12 identified by all of the GWPS, cpg0016, and cpg0021 maps. Below the diagonal represents similarities for the GWPS map, and above the diagonal represents similarities for the cpg0016 map.

https://doi.org/10.1371/journal.pcbi.1012463.g006

RIBONUCLEOPROTEIN_COMPLEX, GOCC_MULTIMERIC_RIBONUCLEASE_P_COMPLEX) and critical RNA maturation steps (GOBP_RNA_5_END_PROCESSING, GOBP_RNA_CAPPING, GOBP_TRNA_5_LEADER_REMOVAL). These six genes are also involved in ribosome biogenesis (GOBP_RIBOSOME_BIOGENESIS) and exhibit endonuclease activities (GOMF_RIBONUCLEASE_P_ACTIVITY). The involvement in the nucleolus (GOCC_NUCLEOLUS) further underscores these genes' role in RNA processing and structural organization.

The six genes connected to *C18orf21* in both RxRx3 and GWPS are also enriched for the ribonuclease mitochondrial RNA processing (RNase MRP) complex in CORUM [28] (p-value: 2e-6), with three of the six genes (*POP1*, *RPP30*, and *POP5*) annotated in this complex. Recent studies have also suggested *C18orf21*'s role in the RNase MRP complex [40, 41]. Our findings reinforce these observations.

Additionally, we observed a strong overlap of interactions across the RxRx3 and GWPS maps for another uncharacterized gene, *C1orf131*. Seven genes are among the top 25 in both RxRx3 and GWPS (Fig 8C): *AATF*, *DDX52*, *NOL6*, *PDCD11*, *PNO1*, *RRP12*, and *UTP20*. The gene set enrichment analysis (Fig 8D) identified several GO terms highlighting roles these genes play in ribosome biogenesis and RNA processing. Notably, terms such as GOCC_SMALL_SUBUNIT_PROCESSOME and GOBP_RIBOSOMAL_SMALL_SUBUNIT_BIOGENESIS emphasize the involvement in the formation and processing of the ribosomal small subunit (SSU) processome, a key component in ribosome assembly. Additionally, the terms GOBP_RIBOSOME_BIOGENESIS, GOBP_RIBONUCLEOPROTEIN_COMPLEX_BIOGE

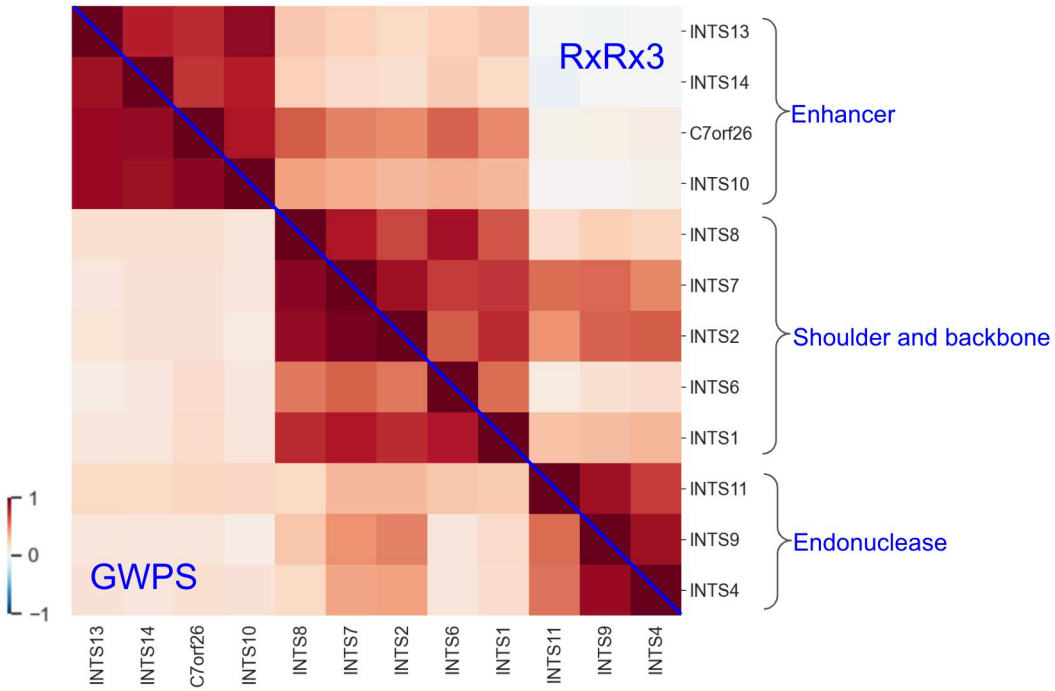

Fig 7. A split cosine similarity heatmap of the Integrator complex subunits from the RxRx3 and GWPS maps. Above the diagonal represents similarities for the RxRx3 map, and below the diagonal represents similarities for the GWPS map. There are three main clusters visible in both, which correspond to the three main modules of the Integrator complex.

https://doi.org/10.1371/journal.pcbi.1012463.g007

NESIS, GOCC_PRERIBOSOME, and GOCC_NUCLEOLUS indicate participation in early ribosome assembly stages within the nucleolus. *C1orf131*, along with these seven genes, has been suggested as a structural component of the small ribosomal subunit (SSU) processome [1, 42, 43]. The evidence shared by the morphological and transcriptional maps presented here corroborate a role for *C1orf131* in the SSU processome.

Although these maps suggest roles for poorly characterized genes, further biochemical research is recommended to validate these hypotheses.

Discussion

In this work we present a general framework for systematically constructing whole-genome perturbative maps of biology and benchmarking their performance globally with publicly-available annotation datasets. As an application of this framework, we utilize several analytical choices to construct 18 maps from four perturbation datasets that use three distinct data types: single-cell transcriptomic data treated with CRISPR interference (GWPS), arrayed phenotypic screening with CRISPR-Cas9 knockout perturbations (RxRx3 and cpg0016), and pooled optical screening with CRISPR-Cas9 knockouts (cpg0021). Additionally, we provide standard benchmarks to assess each map's perturbation consistency, magnitude, and recapitulation rate of known biology. Next, we compared which areas of biology are captured by the best-performing map for each dataset, and present evidence for the roles of two poorly annotated genes, *C18orf21* and *C1orf131*. Our results corroborate recent studies [1, 28, 40–43] by demonstrating connections to the same set of well-characterized genes across powerful maps from two datasets (RxRx3 and GWPS).

While we attempted to minimize computational pipeline differences across the maps through consistent EFAAR choices, these maps utilize different experimental methods, cell

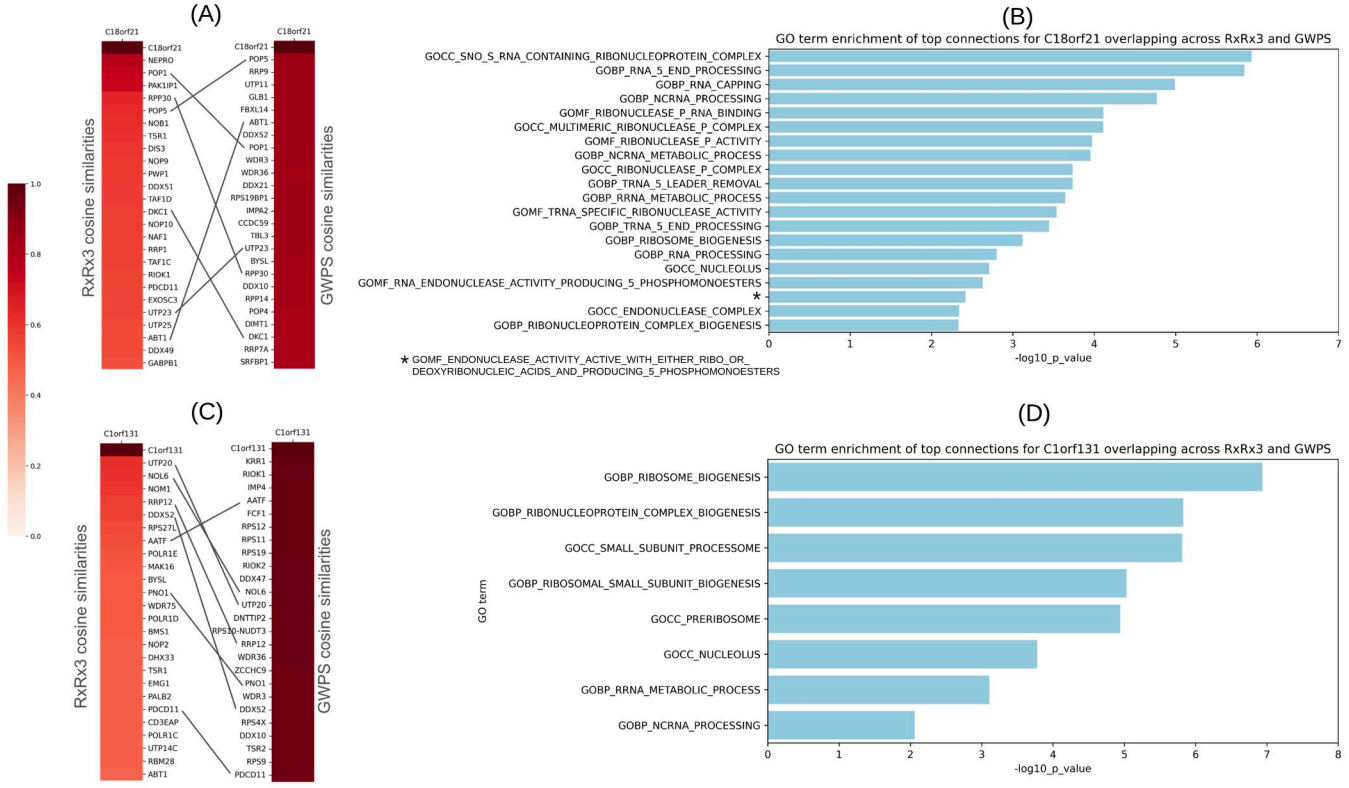

Fig 8. Analysis of strongest gene connections to uncharacterized genes *C18orf21* and *C1orf131*. (A) Top 25 strongest connections to *C18orf21* and the associated cosine similarities in each of the RxRx3 and GWPS maps. Overlapping six genes across are connected by lines between the two heatmaps. (B) GO enrichment results for the six genes that are in the overlap of the top 25 strongest connections to *C18orf21* in the RxRx3 and GWPS maps. Bar lengths represent the Bonferroni-corrected -log10(p-value) from a hypergeometric test. (C, D) Similar data is presented for *C1orf131*.

https://doi.org/10.1371/journal.pcbi.1012463.g008

types, and study designs, all of which impact the benchmarks. For instance, cpg0016 and RxRx3 both involve arrayed assays, but cpg0016 incorporates multiple guide RNAs targeting the same gene in a single well, whereas RxRx3 adds a single guide per well and subsequently aggregates the guides together. Employing a single guide allows for more control and dataset clean-up, while targeting multiple guides may result in higher knockdown and more consistent replicates per gene perturbation. Additionally, the delivery of guides can be performed in a large pool (cpg0021 and GWPS) or through lab automation techniques in an arrayed format (cpg0016 and RxRx3). While pooling is generally more cost-efficient, mixing cells with different treatments may prevent or suppress effects that are visible in arrayed screens from cell-cell interactions in a larger co-localized population.

Previous work for conceptualizing a pipeline for perturbative map processing includes Pycytominer (github.com/cytomining/pycytominer) [44]. Pycytominer is open-source software aimed at standardizing the analysis of high-dimensional morphological features, and it has been utilized by several studies [2, 6, 45] in conjunction with CellProfiler [10] or DeepProfiler [46]. Pycytominer was designed specifically for image-based features and concentrates on individual perturbations rather than assessing the recapitulation of relationships at a map scale.

The framework presented here can be used for any large-scale biological map building and benchmarking effort regardless of data type and can be expanded to include settings where additional perturbation types (small molecules, proteins, antibodies, viruses, etc.) or assay

variables (growth conditions, reagent timing, etc.) are assessed. Moreover, exploration of genome-wide data using these tools can reveal subtle sources of bias that would not be visible at a smaller scale. For example, a recent publication using the framework provided here, unveiled a proximity bias effect, where CRISPR knockouts display unexpected similarities to knockouts of genetically unrelated genes on the same chromosome arm [47].

While our framework can enable building, benchmarking, and exploring maps based on perturbation data from different modalities, we recognize that specific biological questions may require further analysis for validation. Perturbation experiments with transcriptomic readouts, like Perturb-seq, have the unique strength of identifying subsets of interpretable features driving relationships. However, this direction was not pursued in this paper, due to the aim of establishing a unified approach widely applicable across different readout technologies. For a deeper analysis of transcriptomic data, we recommend resources such as MAST [48] and SCEPTRE [49].

Our study reveals important computational and biological insights and suggests opportunities for further exploration. Testing various methods more thoroughly at each stage of EFAAR could improve map quality. Our code repository is open for community contributions, including new map construction and evaluation methods, perturbation datasets, and benchmark sources. Going forward, as more genome-scale datasets are generated and published, utilizing and improving the EFAAR framework and codebase for consistent analyses could lead to new discoveries. As predictions from these maps are validated through orthogonal experimental techniques and annotation sets are improved, analyses using the tools presented here will enable a finer understanding of the relative strengths of different perturbation modalities and techniques. Moreover, as methods to combine different data types are developed, this framework can be used to standardize and evaluate multi-modal explorations. Such methods present an exciting new frontier in understanding cellular biology at the whole-genome scale.

Materials and methods

Perturbation data collection

For RxRx3, we downloaded the embeddings on six-channel images of wells from https://rxrx3.rxrx.ai/downloads. These embeddings were generated by passing images through a weakly-supervised convolutional neural network (CNN) pre-trained on a set of proprietary Recursion data, and the activation values from an intermediate layer of the network were utilized as the 128-dimensional embedding of each image. The model was trained to be partially resilient to batch effects as it attempts to place perturbations from different batches into the same class [36]. The first layer in the CNN applies channel intensity normalization.

For GWPS, we downloaded the raw pre-filtered single-cell RNA-seq count data for K562 cells from https://plus.figshare.com/articles/dataset/_Mapping_information-rich_genotype-phenotype_landscapes_with_genome-scale_Perturb-seq_Replogle_et_al_2022_processed_Perturb-seq_datasets/20029387.

For cpg0016, we downloaded the metadata from https://zenodo.org/records/7661296/files/jump-cellpainting/metadata-v0.5.0.zip?download=1 and the CellProfiler features corresponding to the images from the CRISPR plates from https://cellpainting-gallery.s3.amazonaws.com/index.html#cpg0016-jump/. We filtered CellProfiler features based on image intensity and cell count (Methods: Filtering CellProfiler features for cpg0016) as part of the EFAAR pipeline.

For cpg0021, we downloaded guide-level normalized HeLa CellProfiler features corresponding to the images from the DMEM (Dulbecco's Modified Eagle Medium) plates at https://cellpainting-gallery.s3.amazonaws.com/index.html#cpg0021-periscope/broad/

workspace/profiles/HeLa/. These features were previously summarized to the level of guide-batch pairs, which is the input to our analysis.

Filtering genes based on expression

For RxRx3 and cpg0016, we utilized HUVEC and U2OS RNA-seq data generated using standard techniques. Unexpressed genes are defined as those with zFPKM [50] less than −3 in normalized bulk RNA-seq data. For cpg0021, we employed the same expression collection strategy as outlined by Ramezani et al. [2]— we used HeLa expression data from the Broad Institute's Dependency Map (DepMap) [51], defining unexpressed genes as those with a TPM (transcripts per million) of zero. For GWPS, we treated all perturbed genes as expressed since Replogle et al. [1] reported perturbation of the expressed genome in this dataset.

scVI network architecture

The network is conditioned on the batch information (in this case, the sequencing channel) to correct for batch effects while identifying a lower-dimensional embedding space. We employed an embedding network that has one hidden layer with 256 nodes and a final latent representation with 128 nodes. The decoder network has the same number of latent and hidden nodes and does not share weights with the encoder. For the larger models reported in S2 Table, we also utilized twice as many nodes in the hidden layer as in the latent representation. "scvi-tools" package (version 1.1.2) [52] was utilized in our implementation of scVI models. We used default parameters for model training. The only parameters we configured are batch key, hidden layer size, and latent layer size.

Filtering CellProfiler features for cpg0016

We employed the Elliptic Envelope [53] method with a contamination rate of 1% to identify and eliminate outliers in features related to image intensity, specifically targeting columns containing "ImageQuality_MeanIntensity" in their names. Elliptic Envelope uses covariance estimation to fit a Gaussian distribution to the features. A fixed set of outliers are removed based on a level set of the density of the estimated distribution.

Subsequently, an outlier detection was applied on cell count, by retaining only the wells containing between 50 and 350 (inclusive) in the "Cytoplasm_Number_Object_Number" and "Nuclei_Number_Object_Number" columns. S8 Fig shows the histogram of the values for "Cytoplasm_Number_Object_Number" and "Nuclei_Number_Object_Number" prior to this filtering step.

Lastly, image features were dropped by removing columns whose names start with "Image_". The features that start with "Image_" in CellProfiler refer to measurements and properties that are calculated at the whole image level, and they do not hone in on specific objects or cells within the image. Despite the intensity variation correction implemented by the authors of the cpg0016 data before segmentation, we believe that image features could capture batch effect-related signals.

Benchmark data collection

For StringDB, we downloaded v12.0 protein links from https://string-db.org/cgi/download and classified links with a score $\geq$ 950 as the known relationships.

For CORUM [28] and HuMAP [29], we downloaded the gene clusters representing protein complexes from the online data supplementary to the corresponding study. The pairwise

relationships used for biological relationship benchmark recall computation consist of the union of all pairwise within-complex relationships across all complexes.

For Reactome, we downloaded "Human protein-protein interactions" from "Protein-Protein Interactions derived from Reactome pathways" at https://reactome.org/download-data in February 2021.

For SIGNOR, we downloaded complete *Homo sapiens* data at https://signor.uniroma2.it/downloads.php in February 2021.

Perturbation signal consistency

For a genetic perturbation g, we assume access to a total number of n_g query perturbation units. For each perturbation unit $i = 1, \ldots, n_g$, we have an embedding vector $x_{g,i}$. As the perturbation consistency test statistic, we used avgsim, defined as the mean of the cosine similarities across all pairs of the perturbation unit profiles for g in all batches. Formally,

$$\text{avgsim}_g = \frac{1}{n_g n_g} \sum_{i}^{n_g} \sum_{j}^{n_g} \frac{\langle x_{g,i}, x_{g,j} \rangle}{\|x_{g,i}\| \|x_{g,j}\|}. \tag{1}$$

This consistency value will be zero when all perturbation units are orthogonal, and it will be one when they align perfectly in their orientation.

Perturbation signal magnitude

The energy distance [26, 27] measures how distant the replicate units of a perturbation are from the negative controls, essentially measuring the effect size of the perturbation in a high-dimensional space. For each query perturbation, we compute the distance of the replicate perturbation units' distribution to the control units' distribution using tests derived from energy statistics. Assuming access to two sets of embeddings $x_1, \ldots, x_{n_1}$ (representing query perturbation units) and $y_1, \ldots, y_{n_2}$ (representing negative control units), the energy distance is defined as

$$\text{energy}_g = \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \|x_i - y_j\| - \frac{1}{n_1^2} \sum_{i=1}^{n_1} \|x_i - x_j\| - \frac{1}{n_2^2} \|y_i - y_j\|. \tag{2}$$

This distance value will be zero when the distributions are identical, and positive between non-identical distributions.

Identification of significance of perturbation signal

The genes that achieved a consistency (or magnitude) p-value $< .05$ were considered significant. Perturbation signal benchmark reports the fraction of such genes.

Statistical significance was assessed using a permutation test comparing the test statistic (consistency or magnitude as computed above) of the query gene against an empirical null distribution generated using the corresponding metric for the set of unexpressed genes in each map. A non-parametric test was used to assign a p-value. Parametric tests are not preferred for perturbation signal benchmark metrics because the underlying population of samples do not typically follow a probability distribution from an easily-specified parametric family.

For the consistency metric, average cosine similarity was computed for the unexpressed genes g'_k, $k = 1, \ldots, K$. A p-value was assigned to an expressed gene g by

$$p_g = \frac{\max\{\{\text{avgsim}_{g'_k} \geq \text{avgsim}_g\}, 1\}}{K}. \quad (3)$$

For the magnitude metric, energy distance was computed for the unexpressed genes g'_k, $k = 1, \ldots, K$. A p-value was assigned to an expressed gene g by

$$p_g = \frac{\max\{\#\{\text{energy}_{g'_k} \geq \text{energy}_g\}, 1\}}{K}. \quad (4)$$

For both consistency and magnitude, at least 1,000 unexpressed genes (i.e., $K = 1{,}000$) are required for the null to get a representative p-value.

Recall computation

To assess how well a map recapitulates known biological relationships, we calculated recall measures on known gene-gene relationships as follows. First, we calculated pairwise cosine similarities between the aggregated perturbation embeddings of all perturbed genes in the map. Self-links were excluded since the cosine similarity for these is one and this biases the recall computation. Next, all relationships that fall below the 5th and above the 95th percentiles of this distribution were selected as "predicted links". Then we calculated the recall for each benchmark annotation source as the ratio of the predicted true links over all links. Since different perturbation sets contain different number of genes ("Total perturbed genes" in Table 1), when we compute recall, we adjusted the denominator to only include gene-gene interactions from the benchmark annotation source in which both genes are present in the perturbation dataset. This adjustment is necessary to ensure fairness across different perturbative maps when benchmarking them against biological relationships. We then multiplied the ratios by 100 which shifts the results from ratios to percentages.

Noise sensitivity analyses

Subsampling was done as follows for each dataset. We repeated random selection three times and report mean and standard deviation of the results.

GWPS data contains 267 Gel Bead-in Emulsion (GEM) groups as batches. Each gene readout comes from about 250 cells. The number of GEM groups for each gene can significantly vary from two to nearly all groups. We randomly sampled 50, 100, or 200 cells for each gene across all possible GEM groups for that gene.

RxRx3 data covers $\sim$ 17,000 genes through 88 experiment templates each of which contains readouts from six CRISPR guides for each of $\sim$ 200 genes. Each experiment contains identical nine plates, and each experiment template was screened twice, resulting in a total of 176 experiments. We randomly sampled two, four, or eight plate replicates for each gene from each experiment.

cpg0016 data includes six batches (runs), five of which include all of the $\sim$ 8,000 screened genes, containing a total of about 28 plates. We randomly sampled one, two, or four out of the six runs.

cpg0021 data includes five DMEM plates, each containing readouts from four CRISPR guides for $\sim$ 20,000 genes. We randomly sampled one, two, or four out of the five plates.

Protein complex identification

To evaluate the accuracy of a map in recapitulating established biological gene clusters and to gain insights into its efficacy in capturing various facets of biology, we conducted the following analysis. For each known cluster with at least ten genes in the map, we generated all pairs of genes in the cluster (excluding self-links) as our ground truth for known gene relationships within that cluster. Subsequently, we computed the Kolmogorov-Smirnov (KS) statistic and its associated p-value by comparing the cosine similarity distribution of these pairs to the cosine similarity distribution of those that consist of one gene inside the cluster and another outside it. The latter serves as our "negative" distribution. We then defined the identified clusters as those with a p-value $< .01$ from each map.

Gene Ontology enrichment analysis

We downloaded the "GO: Gene Ontology gene sets" collection (version 2023.2) from the Molecular Signatures Database (MSigDB) under the "C5: ontology gene sets" category (https://www.gsea-msigdb.org/gsea/msigdb/collections.jsp). A hypergeometric test was performed to assess the enrichment of Gene Ontology (GO) terms within the top 25 connections of *C18orf21* or *C1orf131*. Only the terms with at least ten genes in the map were evaluated. To account for the issue of multiple hypothesis testing, Bonferroni correction was applied, ensuring a stringent control of the family-wise error rate. Only those GO terms with a corrected p-value $< .01$ were considered significantly enriched.

Supporting information

S1 Table. Details on cell painting protocols for RxRx3, cpg0016, and cpg0021 datasets. Although cpg0016 and RxRx3 employ the same six stains, the WGA and Phalloidin are imaged in two separate channels in RxRx3 while they are combined in a single channel in cpg0016.
(XLSX)

S2 Table. Biological relationship benchmarks for the GWPS, cpg0016, and cpg0021 maps when using larger dimension sizes for the embedding space.
(XLSX)

S3 Table. The complexes that are significantly identified by all three maps and the significance level for each in different maps. ***** means $< 1e-5$, **** means $< 1e-4$, *** means $< 1e-3$, and ** means $< .01$.
(XLSX)

S4 Table. The complexes that are uniquely identified by the GWPS map and the significance level for each in different maps. ***** means $< 1e-5$, **** means $< 1e-4$, *** means $< 1e-3$, and ** means $< .01$, * means $< .05$, and "ns" means a non-significant p-value ($\geq .05$). Dashes represent the cases where the corresponding map did not have ten expressed genes from the complex, so the p-value was not calculated.
(XLSX)

S5 Table. One complex that is uniquely identified by the cpg0016 map and the significance level for it in different maps. ** means $< .01$, * means $< .05$, and "ns" stands for a non-significant p-value ($\geq .05$).
(XLSX)

S6 Table. The complexes that are uniquely identified by the cpg0021 map and the significance level for each in different maps. ** means $< .01$, * means $< .05$, and "ns" stands for a

non-significant p-value ($\geq$ .05). Dashes represent the cases where the corresponding map did not have ten expressed genes from the complex, so the p-value was not calculated.
(XLSX)

S1 Fig. Visual description of (A) consistency and (B) magnitude measures.
(TIFF)

S2 Fig. Intersection of the interacting gene pairs based on the five different annotation sources utilized in this study. y-axis is in log scale.
(TIFF)

S3 Fig. Noise sensitivity analyses through down-sampling of the replicates for RxRx3. Each bar represents an average of results from three random samples. Dotted line represents the performance of a random, uninformative map.
(TIFF)

S4 Fig. Noise sensitivity analyses through down-sampling of the replicates for GWPS. Each bar represents an average of results from three random samples. Dotted line represents the performance of a random, uninformative map.
(TIFF)

S5 Fig. Noise sensitivity analyses through down-sampling of the replicates for cpg0016. Each bar represents an average of results from three random samples. Dotted line represents the performance of a random, uninformative map.
(TIFF)

S6 Fig. Noise sensitivity analyses through down-sampling of the replicates for cpg0021. Each bar represents an average of results from three random samples. Dotted line represents the performance of a random, uninformative map.
(TIFF)

S7 Fig. Integrator complex heatmap for the cpg0021 PCA-TVN map. The modular structure of the Integrator complex is not visible in this map.
(TIFF)

S8 Fig. (A) Cytoplasm and (B) nuclei object counts for the cpg0016 data.
(TIFF)

Acknowledgments

We would like to thank James Taylor, Renat Khaliullin, Seyhmus Guler, and John Urbanik for their contributions to the development of the EFAAR pipeline and benchmarking methodologies. We would like to thank Leslie Gaffney and Orit Rozenblatt-Rosen for their help with the graphical representation of the EFAAR pipeline. We also would like to thank Dan Maljovec for his invaluable assistance in making the benchmarking repository accessible to the public.

Author Contributions

Conceptualization: Safiye Celik, Jan-Christian Hütter, Sandra Melo Carlos, Nathan H. Lazar, Tommaso Biancalani, Imran S. Haque.

Data curation: Safiye Celik, Jan-Christian Hütter.

Formal analysis: Safiye Celik, Jan-Christian Hütter, Sandra Melo Carlos, Rahul Mohan, Conor Tillinghast.

Funding acquisition: Tommaso Biancalani, Imran S. Haque.

Investigation: Safiye Celik, Jan-Christian Hütter, Sandra Melo Carlos.

Methodology: Safiye Celik, Jan-Christian Hütter, Sandra Melo Carlos, Nathan H. Lazar, Marta M. Fay, Berton A. Earnshaw.

Project administration: Safiye Celik, Imran S. Haque.

Resources: Safiye Celik, Jan-Christian Hütter, Tommaso Biancalani, Imran S. Haque.

Software: Safiye Celik, Jan-Christian Hütter, Sandra Melo Carlos, Rahul Mohan, Conor Tillinghast.

Supervision: Nathan H. Lazar, Tommaso Biancalani, Imran S. Haque.

Validation: Safiye Celik, Jan-Christian Hütter, Nathan H. Lazar.

Visualization: Safiye Celik, Jan-Christian Hütter, Sandra Melo Carlos.

Writing – original draft: Safiye Celik, Jan-Christian Hütter, Sandra Melo Carlos, Nathan H. Lazar, Imran S. Haque.

Writing – review & editing: Safiye Celik, Jan-Christian Hütter, Sandra Melo Carlos, Nathan H. Lazar, Berton A. Earnshaw, Imran S. Haque.

References

1. Replogle JM, Saunders RA, Pogson AN, Hussmann JA, Lenail A, Guna A, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. Cell. 2022;. https://doi.org/10.1016/j.cell.2022.05.013 PMID: 35688146
2. Ramezani M, Bauman J, Singh A, Weisbart E, Yong J, Lozada M, et al. A genome-wide atlas of human cell morphology. bioRxiv. 2023. https://doi.org/10.1101/2023.08.06.552164 PMID: 37609130
3. Sivanandan S, Leitmann B, Lubeck E, Sultan MM, Stanitsas P, Ranu N, et al. A Pooled Cell Painting CRISPR Screening Platform Enables de novo Inference of Gene Function by Self-supervised Deep Learning. bioRxiv. 2023.
4. Funk L, Su KC, Ly J, Feldman D, Singh A, Moodie B, et al. The phenotypic landscape of essential human genes. Cell. 2022; 185(24):4634–4653.e22. https://doi.org/10.1016/j.cell.2022.10.017 PMID: 36347254
5. Fay MM, Kraus O, Victors M, Arumugam L, Vuggumudi K, Urbanik J, et al. RxRx3: Phenomics Map of Biology. bioRxiv. 2023.
6. Chandrasekaran SN, Ackerman J, Alix E, Ando DM, Arevalo J, Bennion M, et al. JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations. bioRxiv. 2023.
7. Bray MA, Singh S, Han H, Davis CT, Borgeson B, Hartland C, et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nature protocols. 2016; 11 (9):1757–1774. https://doi.org/10.1038/nprot.2016.105 PMID: 27560178
8. Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell. 2016; 167(7):1853–1866. https://doi.org/10.1016/j.cell.2016.11.038 PMID: 27984732
9. Cuccarese MF, Earnshaw BA, Heiser K, Fogelson B, Davis CT, McLean PF, et al. Functional immune mapping with deep-learning enabled phenomics applied to immunomodulatory and COVID-19 drug discovery. bioRxiv. 2020.
10. Stirling DR, Swain-Bowden MJ, Lucas AM, Carpenter AE, Cimini BA, Goodman A. CellProfiler 4: improvements in speed, utility and usability. BMC Bioinformatics. 2021; 22(1). https://doi.org/10.1186/s12859-021-04344-9 PMID: 34507520
11. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. Nature communications. 2019; 10(1):1–14. https://doi.org/10.1038/s41467-018-07931-2 PMID: 30674886
12. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nature methods. 2018; 15(12):1053–1058. https://doi.org/10.1038/s41592-018-0229-2 PMID: 30504886

13. Ando DM, McLean CY, Berndl M. Improving phenotypic measurements in high-content imaging screens. BioRxiv. 2017.
14. Haghverdi L, Lun AT, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nature biotechnology. 2018; 36(5):421–427. https://doi.org/10.1038/nbt.4091 PMID: 29608177
15. Polański K, Young MD, Miao Z, Meyer KB, Teichmann SA, Park JE. BBKNN: fast batch alignment of single cell transcriptomes. Bioinformatics. 2020; 36(3):964–965. https://doi.org/10.1093/bioinformatics/btz625 PMID: 31400197
16. Wang ZJ, Lopez R, Hütter JC, Kudo T, Yao H, Hanslovsky P, et al. Multi-ContrastiveVAE disentangles perturbation effects in single cell images from optical pooled screens. bioRxiv. 2023.
17. Sohn K, Lee H, Yan X. Learning structured output representation using deep conditional generative models. Advances in neural information processing systems. 2015; 28.
18. Lotfollahi M, Naghipourfar M, Theis FJ, Wolf FA. Conditional out-of-distribution generation for unpaired data using transfer VAE. Bioinformatics. 2020; 36(Supplement 2):i610–i617. https://doi.org/10.1093/bioinformatics/btaa800 PMID: 33381839
19. Ulyanov D, Vedaldi A, Lempitsky V. Improved Texture Networks: Maximizing Quality and Diversity in Feed-forward Stylization and Texture Synthesis. arXiv. 2017.
20. Stegle O, Parts L, Durbin R, Winn J. A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies. PLoS Computational Biology. 2010; 6(5):e1000770. https://doi.org/10.1371/journal.pcbi.1000770 PMID: 20463871
21. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2006; 8(1):118–127. https://doi.org/10.1093/biostatistics/kxj037 PMID: 16632515
22. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genomics and Bioinformatics. 2020; 2(3). https://doi.org/10.1093/nargab/lqaa078 PMID: 33015620
23. Tukey JW. Mathematics and the picturing of data. In: Proceedings of the International Congress of Mathematicians, Vancouver, 1975. vol. 2; 1975. p. 523–531.
24. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426. 2018;.
25. Agrawal A, Ali A, Boyd S, et al. Minimum-distortion embedding. Foundations and Trends in Machine Learning. 2021; 14(3):211–378. https://doi.org/10.1561/2200000090
26. Szekely GJ. Potential and kinetic energy in statistics. Lecture Notes, Budapest Institute. 1989;.
27. Rizzo ML, Székely GJ. Energy distance. wiley interdisciplinary reviews: Computational statistics. 2016; 8(1):27–38.
28. Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, et al. CORUM: the comprehensive resource of mammalian protein complexes–2019. Nucleic acids research. 2019; 47(D1): D559–D563. https://doi.org/10.1093/nar/gky973 PMID: 30357367
29. Drew K, Wallingford JB, Marcotte EM. hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. Mol Syst Biol. 2021; 17(5):e10016. https://doi.org/10.15252/msb.202010016 PMID: 33973408
30. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The reactome pathway knowledgebase 2022. Nucleic acids research. 2022; 50(D1):D687–D692. https://doi.org/10.1093/nar/gkab1028 PMID: 34788843
31. Licata L, Lo Surdo P, Iannuccelli M, Palma A, Micarelli E, Perfetto L, et al. SIGNOR 2.0, the SIGnaling network open resource 2.0: 2019 update. Nucleic acids research. 2020; 48(D1):D504–D510. https://doi.org/10.1093/nar/gkz949 PMID: 31665520
32. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Research. 2004; 33(Database issue):D433–D437. https://doi.org/10.1093/nar/gki005
33. Zdrazil B, Felix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Research. 2023; 52(D1):D1180–D1192. https://doi.org/10.1093/nar/gkad1004
34. Harding SD, Armstrong JF, Faccenda E, Southan C, Alexander SPH, Davenport AP, et al. The IUPHAR/BPS Guide to PHARMACOLOGY in 2024; 2023.
35. Blucher AS, Celik S, Jensen JD, Taylor J, Cuccarese MF, Cooper JC, et al. Poster: Mapping Biology With a Unified Representation Space for Genomic and Chemical Perturbations to Enable Accelerated Drug Discovery. In: Learning Meaningful Representation of Life Workshop at NeurIPS; 2021.

36. Sypetkowski M, Rezanejad M, Saberian S, Kraus O, Urbanik J, Taylor J, et al. RxRx1: A Dataset for Evaluating Experimental Batch Correction Methods. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops; 2023. p. 4285–4294.

37. Sun B, Feng J, Saenko K. Return of frustratingly easy domain adaptation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 30; 2016.

38. Richter WF, Nayak S, Iwasa J, Taatjes DJ. The Mediator complex as a master regulator of transcription by RNA polymerase II. Nature Reviews Molecular Cell Biology. 2022; 23(11):732–749. https://doi.org/10.1038/s41580-022-00498-3 PMID: 35725906

39. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database hallmark gene set collection. Cell systems. 2015; 1(6):417–425. https://doi.org/10.1016/j.cels.2015.12.004 PMID: 26771021

40. Palukuri MV, Marcotte EM. Super.Complex: A supervised machine learning pipeline for molecular complex detection in protein-interaction networks. PLOS ONE. 2021; 16(12):e0262056. https://doi.org/10.1371/journal.pone.0262056 PMID: 34972161

41. Wainberg M, Kamber RA, Balsubramani A, Meyers RM, Sinnott-Armstrong N, Hornburg D, et al. A genome-wide atlas of co-essential modules assigns function to uncharacterized genes. Nature Genetics. 2021; 53(5):638–649. https://doi.org/10.1038/s41588-021-00840-z PMID: 33859415

42. Singh S, Vanden Broeck A, Miller L, Chaker-Margot M, Klinge S. Nucleolar maturation of the human small subunit processome. Science. 2021; 373 (6560). https://doi.org/10.1126/science.abj5338 PMID: 34516797

43. Dörner K, Ruggeri C, Zemp I, Kutay U. Ribosome biogenesis factors–from names to functions. The EMBO Journal. 2023; 42(7). https://doi.org/10.15252/embj.2022112699 PMID: 36762427

44. Serrano E, Chandrasekaran SN, Bunten D, Brewer KI, Tomkinson J, Kern R, et al. Reproducible image-based profiling with Pycytominer. arXiv. 2023.

45. Way GP, Natoli T, Adeboye A, Litichevskiy L, Yang A, Lu X, et al. Morphology and gene expression profiling provide complementary information for mapping cell state. Cell Systems. 2022; 13(11):911–923.e9. https://doi.org/10.1016/j.cels.2022.10.001 PMID: 36395727

46. Moshkov N, Bornholdt M, Benoit S, Smith M, McQuin C, Goodman A, et al. Learning representations for image-based profiling of perturbations. Nature Communications. 2024; 15(1). https://doi.org/10.1038/s41467-024-45999-1 PMID: 38383513

47. Lazar NH, Celik S, Chen L, Fay MM, Irish JC, Jensen J, et al. High-resolution genome-wide mapping of chromosome-arm-scale truncations induced by CRISPR-Cas9 editing. Nature Genetics. 2024. https://doi.org/10.1038/s41588-024-01758-y PMID: 38811841

48. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biology. 2015; 16(1). https://doi.org/10.1186/s13059-015-0844-5 PMID: 26653891

49. Barry T, Wang X, Morris JA, Roeder K, Katsevich E. SCEPTRE improves calibration and sensitivity in single-cell CRISPR screen analysis. Genome Biology. 2021; 22(1). https://doi.org/10.1186/s13059-021-02545-2 PMID: 34930414

50. Hart T, Komori H, LaMere S, Podshivalova K, Salomon DR. Finding the active genes in deep RNA-seq gene expression studies. BMC Genomics. 2013; 14(1):778. https://doi.org/10.1186/1471-2164-14-778 PMID: 24215113

51. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, et al. Defining a Cancer Dependency Map. Cell. 2017; 170(3):564–576.e16. https://doi.org/10.1016/j.cell.2017.06.010 PMID: 28753430

52. Gayoso A, Lopez R, Xing G, Boyeau P, Valiollah Pour Amiri V, Hong J, et al. A Python library for probabilistic analysis of single-cell omics data. Nature Biotechnology. 2022. https://doi.org/10.1038/s41587-021-01206-w PMID: 35132262

53. Rousseeuw PJ, Driessen KV. A Fast Algorithm for the Minimum Covariance Determinant Estimator. Technometrics. 1999; 41(3):212–223. https://doi.org/10.1080/00401706.1999.10485670