Brieflow: An Integrated Computational Pipeline for High-Throughput Analysis of Optical Pooled Screening Data

Matteo Di Bernardo [1,2], Roshan S. Kern [1], Alexa Mallar[1], Andrew Nutter-Upham [1], Paul C. Blainey [4,5,6], Iain Cheeseman[1,3,*]

Affiliations:

[1] Whitehead Institute for Biomedical Research; 455 Main Street, Cambridge, MA 02142.

[2] Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA 02139.

[3] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142.

[4] Broad Institute of MIT and Harvard; 415 Main St., Cambridge, MA 02142.

[5] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02142.

[6] Koch Institute for Integrative Cancer Research at MIT, Cambridge, MA 02142.

*Corresponding author: e-mail: icheese@wi.mit.edu

Abstract

Optical pooled screening (OPS) has emerged as a powerful technique for functional genomics, enabling researchers to link genetic perturbations with complex cellular morphological phenotypes at unprecedented scale. However, OPS data analysis presents challenges due to massive datasets, complex multi-modal integration requirements, and the absence of standardized frameworks. Here, we present Brieflow, a computational pipeline for end-to-end analysis of fixed-cell optical pooled screening data. We demonstrate Brieflow's capabilities through reanalysis of a CRISPR-Cas9 screen encompassing 5,072 fitness-conferring genes, processing more than 70 million cells with multiple phenotypic markers. Our analysis reveals functional gene relationships that were missed in the original study, uncovering coherent biological insights related to mitochondrial function, mRNA processing, vesicular trafficking, and MYC transcriptional control, amongst others. The modular design and open-source implementation of Brieflow facilitates the integration of novel analytical components while ensuring computational reproducibility and improved performance for the use of high-content phenotypic screening in biological discovery.

Background

Functional genomics has been transformed by high-throughput screening technologies that enable the systematic interrogation of gene function. Optical Pooled Screening (OPS) has emerged as a powerful technique for elucidating gene-phenotype relationships by combining pooled genetic perturbations with high-resolution cellular imaging[1]. In a typical OPS workflow, cells are transduced with a pooled library of genetic perturbations (e.g., CRISPR-Cas9), each identified by a unique barcode sequence (Fig. 1a, panel I). Following selection and fixation, cells undergo high-content imaging to capture phenotypic features followed by in situ sequencing to identify barcode sequences (Fig. 1a, panel II).

Unlike traditional pooled screening approaches that rely on simplistic readouts such as cell viability, OPS captures rich, multidimensional phenotypic data while maintaining advantages including increased throughput and reduced reagent consumption. Furthermore, OPS integrates seamlessly into the cell biologist's toolkit, effectively scaling up standard imaging-based analyses to enable systematic investigation of stains, antibodies, and fluorescent reporters at unprecedented throughput. OPS has been successfully deployed for diverse applications, including the discovery of novel antiviral response pathway components[2], identification of cellular factors regulating transcription factor dynamics[3], and the establishment of previously unknown connections between genes and complex morphological and cytoskeletal phenotypes[4].

The scale and complexity of OPS data present formidable computational obstacles. These terabyte-size datasets often contain information from multi-channel fluorescence images of millions of individual cells, each characterized by thousands of phenotypic parameters. The integration of in situ sequencing and phenotypic measurements necessitates accurate matching of barcode identities with corresponding phenotypic features across different imaging modalities with distinct optical characteristics and resolution limitations. Normalization across experimental batches and conditions is essential to minimize technical variation and enable reliable biological comparisons. The implementation of OPS has also been hampered by a lack of standardized, comprehensive codebases. Current approaches typically involve

fragmented analysis pipelines requiring multiple tools and format conversions between processing steps, creating a high barrier to entry for many research groups. This fragmentation leads to inconsistent methodologies between laboratories, limiting the comparability and integration of results across studies. Finally, the immense size of raw image datasets makes conventional data sharing impractical, while incomplete documentation of analysis parameters undermines reproducibility.

To address these challenges, we developed Brieflow, a computational pipeline designed specifically for end-to-end analysis of fixed-cell OPS datasets. Brieflow provides a standardized framework for processing and analyzing optical pooled screening data, offering a critical pathway to incorporate complex cellular morphologies into emerging *in silico* cell models. To demonstrate the improvements of our approach, we benchmarked Brieflow through a comprehensive reanalysis of a large-scale CRISPR-Cas9 screen comprising over 5,000 fitness-conferring genes in HeLa cells[5], illustrating its efficacy in processing and extracting biological insights from terabyte-scale OPS datasets. To enable the advanced interpretation of these datasets, we further develop MozzareLLM, a specialized large language model framework for biological interpretation that automatically analyzes phenotypic clusters for their functional processes and identifies promising gene candidates for experimental validation. By enabling reproducible, consistent analysis across diverse experimental contexts with improved quality and precision, this pipeline represents an essential step toward integrating the morphological dimension of cellular function into comprehensive computational models of cell biology.

Results

Brieflow: An Integrated Framework for OPS Data Analysis

The analysis of OPS datasets is a multi-step process that requires specialized computational tools to extract, integrate, and interpret the multidimensional phenotypic data. To enable the analysis of this data, we developed a computational pipeline that we term Brieflow (Fig. 1a, panel III). Brieflow provides a unified framework that processes data from raw images to generate biological insights. Brieflow employs a modular architecture comprising six integrated components (Fig. 1b): Preprocess, Sequencing-by-Synthesis, Phenotype, Merge, Aggregate, and Cluster. As described in detail below, the Preprocess module converts raw images to analysis-ready image tiles; Sequencing-by-Synthesis identifies in-situ barcodes of genetic perturbations; Phenotype quantifies cellular features from high-content imaging; Merge aligns perturbations and phenotypes using triangle hashing; Aggregate collapses cellular data to the perturbation level; and Cluster identifies phenotypic relationships between perturbations.

Brieflow's modular architecture enables flexible workflows through independent components with well-defined inputs and outputs, allowing execution of either the complete pipeline or individual modules as needed. The implementation leverages Snakemake-based workflow management for automatic parallelization and progress tracking. The architecture comprises four primary components forming a cohesive computational framework: library code, Python scripts for individual rules, Snakemake rules, and Snakemake targets, according to the standardized Snakemake project development template (Fig. 1c, panel I). The modular design separates core analytical functions in the library from execution scripts, allowing researchers to modify specific components without disrupting the overall pipeline. The Brieflow-analysis template repository incorporates a hierarchical, YAML-based configuration system that separates experimental parameters from processing logic in an organized tree structure. The configuration notebooks guide users through parameter selection for each module, and execution scripts are optimized for both local and high-performance computing environments (Fig. 1c, panel II). Together, these components constitute an end-to-end computational framework that standardizes, automates, and streamlines the complex

analytical workflow required for OPS data, enabling reproducible analyses across diverse experimental contexts. As a case study, we use Brieflow to reanalyze the OPS data from Funk et al. 2022 with the Cas9-based targeting of ~5000 fitness conferring genes in human HeLa cells, herein designated as the 'Vesuvius' dataset, with specific findings described for each step below.

Preprocess: Conversion of raw microscopy files to standardized tiled images

The robust analysis of complex OPS datasets first requires standardization of diverse microscopy data formats into a unified, analysis-ready structure. The Preprocess module addresses initial challenges of OPS data analysis by providing tools for data conversion, metadata extraction, and the generation of appropriate illumination correction functions. Illumination correction is crucial because microscopy images typically exhibit non-uniform brightness across the field of view (Extended Data Fig. 2), which could introduce systematic bias in phenotypic measurements. Thus, the final step computes illumination patterns by averaging multiple fields of view across an experimental batch, using all images for a particular channel from a specific plate-well combination. After smoothing with a median filter, each image undergoes correction by division with the illumination function in downstream modules[6]. In our reanalysis of the Vesuvius screen, the Preprocess Module extracted 164,835 tiles of in-situ sequencing images across 11 sequencing rounds, and 57,645 tiles of phenotype images from a total of 2,008 ND2 files. The Preprocess module establishes a flexible foundation for downstream analysis by transforming heterogeneous microscopy data into standardized, efficiently-structured image tiles.

Sequencing-by-synthesis: Accurate identification of genetic perturbations

Accurately linking cellular phenotypes to their causative genetic perturbations requires the precise identification of barcodes, for example corresponding to Cas9 sgRNA guide targeting sequences, within complex in situ sequencing images. The Sequencing-by-synthesis module forms the analytical backbone to detect and decode in situ barcode sequences (Fig. 2a). The Sequencing-by-synthesis workflow encompasses a complete analytical pipeline beginning with precise multi-cycle image alignment to account for

spatial shifts between sequencing rounds, followed by spot detection to identify potential barcode locations, base calling to determine nucleotide sequences from intensity profiles, cell segmentation to define cellular boundaries, and finally barcode-to-cell mapping that associates genetic perturbations with specific cells for downstream phenotypic analysis. This pipeline also generates detailed quality metrics to assess sequencing performance, including read mapping rates, cell assignment statistics, and perturbation distribution analysis, which can help identify areas of low spot signal (Fig. 2b).

The module offers two distinct approaches for identifying sequencing spots: a standard method based on statistical signal detection and a method that utilizes Spotiflow, a deep learning-based spot detection tool trained on fluorescence microscopy data[7]. In our analysis of twenty-five randomly selected tiles from the Vesuvius OPS dataset, we found that Spotiflow consistently detected more spots than the standard method with an appropriate peak threshold (12,554 vs. 9,431 spots on average per tile, respectively) (Fig. 2c). However, the proportion of cells with a mapped gene remained nearly identical between the two approaches (65.1% with the standard method, 64.2% with Spotiflow). In addition, processing with Spotiflow required substantially more time, averaging 47 seconds per tile compared to just 1 second with the standard method (Extended Data Fig. 3). Given these results and the scale of our dataset, we employed the standard statistical method for our reanalysis of the Vesuvius screen. We identified a total of 72,556,135 cells with 81.71% (59,299,224) of cells containing a sgRNA barcode, and 57.53% (41,725,605) successfully assigned to a singular gene target. Together, the Sequencing-by-synthesis module transforms complex multi-cycle imaging data into high-confidence genetic perturbation assignments for each cell.

Phenotype: Comprehensive cellular feature extraction

Capturing the phenotypic impact of genetic perturbations requires the extraction of comprehensive multi-dimensional cellular features from high-content imaging data. To achieve this, the Phenotype module (Fig. 2d) implements a systematic workflow beginning with specimen-specific illumination correction to normalize signal across the field of view, followed by the precise alignment of fluorescence channels to correct for

chromatic aberrations, cellular segmentation to identify nuclear and cellular boundaries, and finally feature extraction to quantify morphological and functional characteristics across multiple cellular compartments.

To achieve robust cellular segmentation for diverse cellular morphologies, Brieflow provides a choice of state-of-the-art deep learning-based methods including Cellpose, which employs vector flow prediction for whole-cell segmentation; MicroSAM, which adapts the Segment Anything Model for microscopy applications; and StarDist, which utilizes star-convex polygons for robust cellular detection[8-10]. In our comparative benchmarking of segmentation approaches on one random imaging tile per plate (8 tiles total), we found that Cellpose exhibited superior segmentation accuracy but required significantly longer processing time per tile (average of approximately 189 seconds)[10], while MicroSAM[8] and StarDist[9] achieved lower accuracies but processed tiles more efficiently (approximately 92 and 25 seconds, respectively). The qualitative evaluation assessed each method's ability to properly delineate cellular boundaries in the absence of ground truth segmentation data (Fig. 2e). Additionally, we evaluated retention rates when merging masks, measuring the percentage of cells containing at least one nucleus, which varied significantly across methods (>90% for Cellpose, <50% for MicroSAM and StarDist) (Extended Data Fig. 4). Based on these segmentation quality metrics, we selected Cellpose for the complete reanalysis of the HeLa cells in the Vesuvius dataset, which present relatively simple morphologies. For applications involving cell lines with more challenging morphological characteristics, we have implemented frameworks for model fine-tuning and parameter optimization, with the other models providing valuable alternatives. The implementation also provides notebook visualization capabilities that enable users to inspect segmentation results qualitatively prior to committing to full-scale processing.

To enable feature extraction for the phenotypic analysis, Brieflow implements a pure Python CellProfiler-inspired emulator[11] that integrates algorithms from multiple computational packages. This provides consistent behavior across computing environments while extracting hundreds of cellular features for each channel by mask combination, including morphological parameters, intensity distributions, and texture measurements[11]. Our modular design also enables future integration of enhanced

feature extraction methods as they mature. In our reanalysis of the Vesuvius dataset, the Phenotype Module successfully segmented 69,065,844 cellular objects. Feature extraction generated 1,651 morphological measurements per cell. These features provide robust quantitative metrics for specific cellular properties, including the intensity and distribution of the stains, antibodies, and markers used for the optical pooled screen. The phenotype module also generates quality assessment metrics that evaluate feature consistency, distribution across the experiment, and visualization as plate heatmaps, facilitating identification of spatial trends or anomalies (Fig. 2f). Together, the Phenotype module converts raw fluorescence images into rich quantitative profiles that capture the morphological consequences of genetic manipulations, informing on the functional contribution of each corresponding gene product.

Merge: Integration of genetic and phenotypic data

Connecting a genetic perturbation with its phenotypic effect necessitates precise spatial registration between different imaging magnification captured during the optical pooled screening workflow. The Merge module integrates data from the Sequencing-by-synthesis and Phenotype modules, creating a unified dataset that links genetic perturbations with their corresponding phenotypic effects (Fig. 3a). This integration step uses a sequential triangulation hashing approach on cellular positions to accurately map between different imaging modalities before merging the data generated by the Sequencing-by-synthesis and Phenotype modules, and sequentially deduplicating multiple matches based on relevant parameters.

In our reanalysis, spatial registration in the Merge module between Sequencing by synthesis and Phenotype outputs successfully paired 47,701,367 of cells between modalities, with an average of 87.65% (41,729,395) containing a perturbation identity. This represents approximately a 32% reduction from the initial 70+ million cells identified in the upstream modules. Although alternative registration approaches exist, the triangle hashing method balances efficiency with computational feasibility for large-scale datasets, although integration of higher-quality registration approaches may be warranted for smaller scale screens. Comprehensive visualizations enable the rapid assessment of the spatial distribution of matched cells and quality metrics across the

plate, which often reveal that merging is reduced at well edges (Fig. 3b). Thus, the Merge module creates a unified dataset that directly links genetic perturbations to their corresponding phenotypic signatures at single-cell resolution.

Aggregate: Robust Gene-Level Embeddings

Optical pooled screening approaches obtain phenotype information from thousands of individual cells for each perturbation. Although this generates rich single-cell data that is useful for some applications, aggregating these measurements into perturbation-level profiles provides improved statistically meaningful biological insights regarding gene function. The Aggregate module transforms single-cell perturbation data into robust gene-level embeddings suitable for downstream analysis (Fig. 3c). This module allows the user to split their dataset into different subpopulations using any classifier as appropriate, before employing a standard cell feature processing pipeline with steps to filter data, align to reduce variation across experimental batches, and aggregate data to the perturbation level[12]. The critical alignment step applies a Typical Variance Normalization (TVN) transformation that leverages negative controls to substantially reduce batch effects while preserving phenotypic signals essential for distinguishing between perturbations[13]. These batch-normalized principal components are then aggregated to the perturbation level to create embeddings used for cluster analysis that reveals functional relationships between genetic manipulations. Separately, the original interpretable features (such as nuclear size and stain intensities) are aggregated to the perturbation level to provide quantitative metrics that directly inform our understanding of which specific cellular processes drive the observed functional gene clusters.

In our reanalysis with the Aggregate Module, batch effect correction was highly effective, with the median p-value for batch-associated variation (as measured by ANOVA-F) increasing from highly significant values (6.9×10^{-12} for interphase cells) to non-significant levels (0.54), indicating successful removal of technical variation while preserving biological signal (Fig. 3d). The final aggregated interphase dataset contained 5,299 distinct perturbations with a median of 5,431 cells per perturbation and 795 aligned principal component features. In total, this generated 28,711,561 cells compared to 31,046,046 cells in the original Vesuvius analysis. This reduction from the

initial 41.7 million cells identified in the Merge module was primarily due to the filtering of cells with multiple barcode associations, ensuring high-confidence perturbation-phenotype relationships in downstream analyses. In sum, the Aggregate module transforms single-cell data into two complementary perturbation-level datasets: standardized principal component embeddings that serve as the foundation for clustering analysis, and aggregated interpretable feature values that enable quantitative assessment of specific cellular phenotypes for downstream biological interpretation.

Cluster: Identification of functional relationships

The analysis of OPS data is built on a fundamental cell biological principle: disrupting genes that operate in the same pathway or complex result in similar morphological phenotypic consequences when perturbed. By analyzing quantitative patterns across hundreds of morphological features, this creates phenotypic “fingerprints” that can be compared to reveal functional biological relationships. The Cluster module employs advanced computational approaches to transform the high-dimensional embeddings from the Aggregate module into interpretable clusters and identify functional relationships between genes by grouping genetic perturbations with similar phenotypic profiles (Fig. 3e). At the core of the Cluster module is a dimensionality reduction and clustering pipeline built on PHATE (Potential of Heat-diffusion for Affinity-based Transition Embedding)[14] and Leiden community detection[15], including the ability to evaluate a range of optimal Leiden resolutions (Fig. 3f). To objectively evaluate clustering quality, the module includes a benchmarking framework using precision-recall metrics against reference databases of both pairs of genes, such as the STRING protein-protein interactions[16], and groups of genes, such as CORUM protein complexes[17] and KEGG pathways[18].

In our reanalysis, phenotypically clustering the interphase cells at a resolution parameter of 13 identified 251 distinct phenotypic clusters. Benchmarking against reference databases revealed that 30.7% and 15.5% of clusters were enriched in at least one CORUM or KEGG cluster, respectively (compared to the original Vesuvius clustering, in which 24.3% and 13.5% of the 221 clusters displayed CORUM or KEGG gene enrichment). Shuffled controls yielded negligible enrichment (0.3% for CORUM,

0% for KEGG). Thus, the Cluster module translates high-dimensional phenotypic profiles into biologically meaningful gene groups that illuminate pathway relationships and functional connections.

Interactive visualization and exploration dashboard

Finally, the Brieflow framework incorporates an interactive visualization dashboard that facilitates both quality control assessment and exploratory analysis of optical pooled screening data (Fig. 4a). This dashboard comprises of four displays: Experimental Metadata for information relating to how the screen was conducted, Configuration for information on parameters used in Brieflow, Quality Control for assessment of data integrity, and Cluster Exploration for visualization of gene relationships and interpretation of image montages for particular genes. The dashboard additionally integrates with MozzareLLM, a specialized large language model framework for biological interpretation described in detail below. All of the finalized outputs of the reanalysis of the Vesuvius screen described here can be accessed stably at http://screens.wi.mit.edu/aconcagua. This interactive capability substantially reduces the technical barriers to exploring high-dimensional data, democratizing access to complex OPS results for experimental biologists who may lack computational expertise.

MozzareLLM: Automated biological interpretation and prioritization

Although benchmarking against established biological datasets, such as CORUM and KEGG, provides confidence in cluster quality, extracting novel biological insights from these clusters remains a significant challenge relying on time consuming and imperfect manual evaluation. Large language models (LLMs) have recently emerged as powerful tools for analyzing gene sets and interpreting biological pathways[19], but their application to identifying novel or uncharacterized genes with potential functional significance has been limited. To address this gap, we developed MozzareLLM, a specialized LLM integration framework for automating the interpretation of gene-phenotype relationships and prioritizing candidates for functional investigation.

MozzareLLM implements a structured prompt engineering approach that directs language models to analyze cluster membership, identify dominant biological

processes, and systematically categorize genes based on their annotation status and potential for novel functional discoveries. We evaluated and prompt-tuned MozzareLLM performance across three different sets of clusters that were evaluated experimentally in papers from 2021[20], 2022[5], and 2025[21] using leading LLMs from Anthropic, Google, and OpenAI. With our finalized prompt, Claude 3.7 Sonnet demonstrated superior accuracy in identifying correct biological functions across gene clusters. For example, in the cluster 197 of the original analysis (Funk et al. 2022), Claude correctly identified "RNA modification and processing" and accurately noted its relation to the expected m6A mRNA modification pathway, while models from other providers incorrectly categorized it as having "No coherent biological pathway."

When applied to the reanalysis of the HeLa screen, MozzareLLM with Claude 3.7 Sonnet identified distinct "high confidence" biological processes for 99 of the 251 clusters (Extended Data Fig. 6a). These high confidence clusters contained, on average, 5.2 genes with established functions but potential novel roles in the identified processes, alongside a mean of 1.3 completely uncharacterized genes per cluster (Extended Data Fig. 6b). Thus, MozzareLLM provides a powerful complement to Brieflow for identifying functional insights from large-scale OPS datasets.

Novel biological connections identified through reanalysis of the HeLa screen

Our reanalysis of phenotypic data from the Vesuvius CRISPR-Cas9 screen in HeLa cells using Brieflow revealed multiple novel biological connections not identified in the original study. This improved analysis was driven by a combination of improvements in segmentation, barcode calling, aggregation, and clustering as part of the Brieflow pipeline. With the assistance of MozzareLLM to facilitate a robust analysis of the gene clusters, we identified multiple biological insights with compelling potential for mechanistic follow-up.

Most strikingly, our revised analysis using Brieflow also revealed a coherent functional clustering of mitochondrial processes that was entirely absent from previous analyses of this dataset. Brieflow identified three coherent mitochondrial clusters (Clusters 10, 61, and 87) containing a total of 99 genes with clear functional specialization (Extended Data Fig. 7a). This finding is particularly notable given that our

screen employed only markers of cellular architecture without specific membrane or mitochondrial staining. While mitochondrial gene perturbations exhibited relatively subtle phenotypic deviations from nontargeting controls, our improved pipeline analysis and clustering successfully resolved functional relationships that were previously obscured (Extended Data Fig. 7b). Cluster 10 represents the most comprehensive mitochondrial network, encompassing 53 genes involved in respiratory chain assembly and oxidative phosphorylation, including components from all major respiratory chain complexes, mitochondrial ribosomal proteins, mitochondrial translation factors, and mitochondrial quality control machinery. Cluster 87 contains the respiratory chain complex assembly with particular enrichment for complex I components (NDUFAF4, NDUFB5, NDUFA9, NDUFB9) and mitochondrial ribosomal proteins, whereas Cluster 61 includes the oxidative phosphorylation machinery and mitochondrial protein synthesis components. The original analysis of the Vesuvius dataset distributed these genes across 41 different clusters without substantial enrichment for mitochondrial ribosome subunits or respiratory chain components (Extended Data Fig 7c). MozzareLLM analysis of these clusters identified several genes with potential novel mitochondrial roles, including INTS13 and INTS14 components of the integrator complex in Cluster 87, and multiple actin cytoskeleton regulators (CFL1, WDR1, CAPZB, CAP1) in Cluster 61 that may coordinate cytoskeletal dynamics with mitochondrial positioning or quality control. The presence of KRAS in Cluster 87 provides additional evidence for emerging connections between oncogenic signaling and mitochondrial function, consistent with recent reports linking KRAS activity to mitochondrial metabolism and respiratory chain regulation[22]. This improved resolution of mitochondrial functional networks demonstrates how the enhanced analysis and clustering in Brieflow can reveal previously obscured biological organization within complex phenotypic datasets, even when the experimental readouts are not specifically designed to capture the cellular processes being perturbed.

Our analysis also uncovered functionally coherent clusters in other cellular processes. Cluster 5 revealed a robust mechanistic connection between ribosome biogenesis and multiple incompletely characterized proteins. This cluster contains a strong ribosomal processing signature, but also includes members that were not

detected in the original analysis of the Vesuvius data set implicating C18orf21 (RMP24) and NEPRO (RMP64) as playing critical roles in 40S ribosome biogenesis (Fig. 4a). Indeed, our recent work in parallel to this analysis revealed that RMP24 functions as a previously unidentified RNase MRP-specific protein subunit in the nucleolus, and demonstrated that depletion of either RMP24 or RMP64 specifically impairs 40S ribosomal subunit production, causing defects in rRNA processing and polysome formation[23]. MozzareLLM correctly classified C18orf21/RMP24 as an "uncharacterized" gene and NEPRO/RMP64 as having a "novel role" in this cluster, effectively identifying functional relationships that we confirmed through separate detailed cellular experimentation. Analysis of morphological features revealed that genes within this cluster exhibit remarkably consistent phenotypic signatures, with uniformly reduced nuclear and cellular morphometric parameters and elevated tubulin-associated measurements across both established ribosome biogenesis components (POP5, RPP30) and newly implicated genes (C18orf21, NEPRO) (Extended Data Fig. 8a). The observed cellular size reduction is consistent with disrupted growth and protein synthesis when translation is inhibited (Extended Data Fig. 8b). This shared phenotypic signature provides independent validation that C18orf21 and NEPRO belong functionally within this ribosome biogenesis network, reinforcing our biochemical characterization of their roles in 40S subunit assembly.

The gene clusters generated by Brieflow also uncovered novel associations between functional processes, connecting core features of cell biology in unexpected ways. Analysis of Cluster 60 revealed genes associated with both mRNA processing and vesicular trafficking, containing multiple components of the exocyst complex (EXOC1, EXOC4, EXOC5, EXOC7, EXOC8) and established vesicular trafficking machinery alongside CCR4-NOT complex components (CNOT1, CNOT4, CNOT10, CNOT11). This diverged from the original analysis that distributed these genes across unrelated clusters. The CCR4-NOT complex is well-characterized for its role in mRNA deadenylation and post-transcriptional regulation, but has no established connection to membrane trafficking systems[24]. Analysis of phenotypic parameters from Cluster 60 revealed elevated cellular tubulin levels across all genes in the cluster, providing morphological evidence for a shared role in tubulin homeostasis (Extended Data Fig.

9b). The CCR4-NOT complex has been shown to control tubulin homeostasis by degrading tubulin mRNAs when cells sense excess free tubulin[25], but a role for the Exocyst complex in tubulin assembly or microtubule dynamics would represent a novel functional role for this complex.

Our analysis of Cluster 95 revealed a focused network of genes that expands our understanding of MYC-dependent transcriptional control. While the original analysis distributed these genes across broader clusters focused on general transcriptional regulation (particularly cluster 121 containing numerous transcriptional regulators), our improved clustering approach identified a more specific functional network. This refined cluster links established MYC-associated chromatin modifiers (ZMYND8, KAT2A, SETD2) with two additional genes, TMEM44 and TOX4, that may represent novel components of MYC-dependent transcriptional control (Extended Data Fig. 10a). MozzareLLM classified TOX4 as having a "novel role" and TMEM44 as an "uncharacterized" gene in this context. The presence of TMEM44 in this MYC-focused cluster is particularly compelling, as recent research has shown that the antisense transcript TMEM44-AS1, a lncRNA, forms a direct positive feedback loop with MYC, where MYC transcriptionally induces TMEM44-AS1 expression through super-enhancer binding, while TMEM44-AS1 reciprocally activates MYC signaling[26]. Similarly, the inclusion of TOX4 is intriguing, as this component of the PTW/PP1 phosphatase complex regulates RNA polymerase II phosphorylation—a process critical for transcriptional control[27]. This connection gains additional significance given evidence that MYC-transformed cells develop dependency on phosphorylation pathways involving PTW/PP1 complex components, which regulate both spliceosome activity and transcriptional elongation[28]. Genes within this cluster exhibit consistently decreased DAPI intensity measurements and increased nuclear eccentricity upon disruption (Extended Data Fig. 10b). These consistent phenotypic changes support the functional coherence of this MYC-associated network.

Together, these new insights highlight the power of the Brieflow pipeline, which provides enhanced resolution and reveals previously unappreciated connections between cellular machineries. MozzareLLM facilitated the rapid and systematic evaluation of these clusters, highlighting potential functional relationships that might

have been overlooked even by experienced cell biologists conducting manual review. By combining advanced computational clustering with automated interpretation of gene relationships, this approach significantly accelerates the extraction of biological insights from complex OPS datasets and enables more efficient prioritization of candidates for experimental validation.

Discussion

Although technologies such as Perturb-seq[29] have benefited from extensive methodological refinement through the widespread use of gene expression data, optical pooled screening represents an emerging data type with substantial room for methodological innovation compared to more established functional genomics approaches. The unique analytical challenges inherent to OPS—including multi-modal data integration across different imaging modalities, complex image processing requirements, and the extraction of biologically meaningful features from high-dimensional cellular measurements—necessitate specialized computational approaches that continue to evolve rapidly. These challenges present significant opportunities for innovation in computational methods tailored specifically to the complexities of image-based functional genomics.

Brieflow represents a significant advancement in the analysis of OPS data, addressing key bottlenecks that have limited the widespread adoption of this powerful technique. Although experimental OPS platforms have become increasingly accessible, the analytical bottleneck has emerged as a critical limitation, with many researchers lacking the computational infrastructure or expertise required to extract meaningful biological insights from their high-dimensional datasets. Although existing OPS analysis approaches have made valuable contributions, they often focus on specific analytical challenges rather than providing integrated, end-to-end workflows, with some relying on specialized software environments that limit scalability and broader adoption (Table 1). By providing a unified, modular framework for end-to-end analysis, Brieflow eliminates the need for fragmented pipelines and format conversions, substantially reducing the technical barriers to entry for researchers interested in applying OPS to their biological questions. Our reanalysis of the large-scale “Vesuvius” optical pooled screen demonstrates Brieflow's capabilities in handling large-scale datasets to reveal enhanced biological insights while maintaining high data quality and analytical rigor.

The standardization and computational reproducibility provided by Brieflow will be particularly important as the adoption of OPS approaches continues. Through Snakemake-based workflow management and comprehensive configuration documentation, the Brieflow pipeline enables researchers to share not only their results

but also the exact methodologies used to generate them, enhancing transparency and facilitating collaborative research. The open-source implementation of Brieflow, with its well-documented interfaces and extensible design, provides a foundation for community-driven improvement and customization. The modular architecture enables rapid development of new capabilities—for example, swapping in a custom segmentation method or extending feature extraction to novel cellular structures can be accomplished in just a few days while maintaining the pipeline's robust organization and validation framework. This positions Brieflow as a central backbone for OPS analysis that can evolve alongside the field through community contributions while maintaining backward compatibility with established workflows. Furthermore, the standardized outputs produced through Brieflow constitute valuable resources for the development of machine learning models on functional genomics data. As perturbational datasets capture a broader spectrum of cellular phenotypes, the analysis of multiple OPS experiments processed through a common framework enables the creation of comprehensive training datasets with enhanced generalizability and predictive power. These standardized datasets will significantly advance computational cell modeling efforts, allowing researchers to build increasingly accurate in silico predictions of cellular behaviors that integrate multi-dimensional phenotypic data across diverse genetic perturbations.

Finally, the addition of MozzareLLM for automated biological interpretation represents an innovative approach to extracting actionable insights from high-dimensional phenotypic data. By leveraging large language models specifically oriented towards biological knowledge extraction, this component accelerates the transition from data to hypothesis formation, a critical step in the scientific discovery process.

Code Availability

All code is publicly available. The core Brieflow pipeline is available via GitHub at https://github.com/cheeseman-lab/brieflow. The specific implementation for the HeLa OPS dataset analysis described in this paper (including benchmarking experiments between methods) is available at https://github.com/cheeseman-lab/aconcagua-analysis, while the data used to train the mitotic classifier is available in a separate Brieflow-analysis based repository at https://github.com/cheeseman-lab/logan-analysis. MozzareLLM, designed to scale to different types of gene grouping analyses beyond its integration with Brieflow, is available as a standalone tool at https://github.com/cheeseman-lab/mozzarellm. Comprehensive documentation for Brieflow, including tutorial videos and detailed implementation instructions, can be accessed at https://brieflow.readthedocs.io. For researchers interested in analyzing optical pooled screening data, we recommend first consulting this documentation before creating a new project based on the Brieflow-analysis template to begin their OPS analysis. Any questions, issues, or code bugs should be reported at https://github.com/cheeseman-lab/brieflow/issues.

Data Availability

The raw dataset (>20TB) used in this study can be provided upon reasonable request by contacting the corresponding authors. A stable interactive interface for exploring the results of the HeLa screen analysis, including visualizations of phenotypic clusters and gene relationships, quality control plots, details on the Brieflow run conducted, and experimental metadata, is publicly accessible at http://screens.wi.mit.edu/aconcagua. Processed data including feature matrices, cluster assignments, and MozzareLLM analysis results are available at https://doi.org/[].

Acknowledgements

This work was supported by grants from the NIH (R35GM126930 to I.M.C., R01HG009283 to P.C.B.), and the Chan Zuckerberg Initiative to I.M.C. and P.C.B. M.D. is supported in part by an NSF GRFP fellowship. This research was also supported by the Whitehead Innovation Initiative and the Massachusetts Life Science Center Data

Science Internship Program. A.M was supported by the MIT Undergraduate Research Opportunities Program. We extend our gratitude to David Feldman for the original implementation of the pipeline. We thank Luke Funk for useful information on how the screen was conducted and in-depth discussions on how it was analyzed, and both Luke Funk and Kuan Chung Su for conducting the HeLa screen. We acknowledge members of the Blainey and Cheeseman laboratories for their valuable discussions regarding pipeline development. We are particularly grateful to members of the Whitehead Institute Information Technology Core for their assistance with pipeline acceleration and implementation on high-performance computing infrastructure. We also thank Alán F. Munoz and members of the Broad Imaging Platform for their insights regarding CellProfiler implementation. Additionally, we acknowledge Albert Dominguez Mantes for his thoughtful discussions concerning the implementation of Spotiflow.

Author contributions

M.D. and R.K. conceptualized the work. M.D., R.K., A.M., and A.N-U. implemented the software. M.D., I.C., and P.C.B. drafted the manuscript. All co-authors reviewed and edited the final manuscript.

Competing Interests

P.C.B. is a consultant to or holds equity in 10X Genomics, General Automation Lab Technologies/Isolation Bio, Next Gen Diagnostics, Cache DNA, Concerto Biosciences, Stately Bio, Ramona Optics, Bifrost Biosystems, and Amber Bio. His laboratory has received research funding from Calico Life Sciences, Merck, and Genentech for work related to genetic screening.

Methods

System Design and Implementation Architecture

Brieflow follows a standardized Snakemake project development template with four principal components: library code containing core analytical functions, Python scripts for individual rules, Snakemake rules defining execution logic, and targets specifying pipeline outputs. This architecture separates core analytical logic from execution scripts, enabling modification of specific components without disrupting the overall pipeline (Extended Data Fig. 1).

The data structure employs standardized file formats including TIFF for image data, TSV and parquet for tabular data, and PNG for evaluation outputs. Consistent naming conventions facilitate automated data organization and rule matching, while clearly separating raw, intermediate, and analysis data. The pipeline implements multi-level parallelization across sample, well, and tile levels, with GPU acceleration available for computationally intensive tasks.

Brieflow-analysis, provided as a separate template repository, incorporates a hierarchical YAML-based configuration system organized in a tree structure that separates experimental parameters from processing logic. Users create a new repository from this template for each screen analysis, with Brieflow integrated as a git submodule to ensure consistent versioning while enabling project-specific customization. The repository includes comprehensive testing procedures to verify correct installation across computing environments.

Brieflow implements multi-level parallelization across sample, plate, well, and tile levels, with GPU acceleration available for computationally intensive tasks such as deep learning-based segmentation. The framework includes specialized execution scripts that optimize job submission for different computing environments, processes plates in sequence to avoid overwhelming computational resources, consolidating multiple small tasks into fewer, larger jobs. The system includes configuration profiles for different execution environments that specify resource allocation, job submission parameters, and environment variables, and makes use of the Snakemake plugin catalog for use of Brieflow across different user-specific interfaces.

The framework supports interchangeable components through specialized branches designed for different experimental contexts, enabling community contributions in specific areas such as spot calling, segmentation, or feature extraction. Standardized interfaces between modules ensure that modifications to specific components can be implemented without disrupting the overall workflow.

Preprocess Module

The Preprocess module converts proprietary ND2 microscopy files to standardized TIFF format while extracting comprehensive metadata including spatial information, optical parameters, and experimental identifiers. The module processes diverse file organizations including well-based image storage or individual tile-based acquisitions, combined or separated imaging rounds, and single-channel or multi-channel file structures.

Configuration parameters are set through interactive notebooks where users specify all ND2 files to be processed. This automatically generates a comprehensive file mapping all combinations of plates, wells, tiles, and channels, eliminating the need for hardcoded values in downstream analysis. Illumination correction is implemented through batch-level correction field calculation, computing illumination patterns by averaging multiple fields of view across an experimental batch. Users can specify a sample fraction (ranging from 0-1) to control the proportion of images used for correction field estimation, with our implementation using the complete dataset. After smoothing with a median filter, each image undergoes correction by division with the illumination function in downstream modules. The implementation leverages multi-threading capabilities to accelerate processing of large image datasets.

Sequencing By Synthesis Module

The Sequencing by synthesis module processes in situ sequencing images through a sequential workflow: alignment of sequencing cycles, illumination correction optimized for sequencing signal detection, application of a Laplacian-of-Gaussian filter to enhance

spots, and computation of intensity standard deviation to identify consistent signals across cycles.

Alignment supports multiple methods, including reference-based alignment using DAPI as a fiduciary marker (if imaged in each round) and signal-based alignment using the mean sequence signal. Our implementation used the latter method, which requires fewer imaging channels but maintains high registration accuracy. The module offers an upsample factor parameter to enable sub-pixel alignment when needed, though we found that standard integer-pixel resolution was sufficient. To account for sub-pixel alignment errors, the module applies maximum filtering with a configurable width (set to 3 pixels in our analysis) that dilates sequencing signals slightly, ensuring robust signal capture despite minor registration imperfections.

For barcode detection, two distinct approaches are implemented: a standard method based on statistical signal detection and a deep learning-based method using Spotiflow. The standard approach (which we implemented with a peak threshold of 350) first applies a Laplacian-of-Gaussian filter (with kernel width $\sigma = 1$ pixel) to enhance spots across sequencing channels. After applying this filter to all sequencing cycles, the standard deviation across cycles and mean across channels are computed to identify consistent signals. Local maxima are then identified within a defined neighborhood (default width of 5 pixels), with each peak assigned a score representing the difference between the maximum and minimum values in that neighborhood. This approach effectively captures spots that maintain consistent signal across multiple sequencing cycles while filtering out noise and non-specific background. The Spotiflow method provides an alternative approach, utilizing a pre-trained U-Net model to predict spot locations from a selected sequencing cycle across all four base channels, followed by merging to exclude overlapping signals based on a configurable minimum pixel distance parameter. For benchmarks, Spotiflow was run with a probability threshold of 0.3, using the first cycle for spot detection, and a minimum distance of 1 pixel between detected spots.

Base calling implements a statistical approach that considers channel cross-talk and signal-to-noise ratios. Each called base receives a quality score reflecting confidence in the call, calculated as:

$$Q = 1 - \frac{\log_2(2 + I_{\text{second}})}{\log_2(2 + I_{\text{first}})}$$

where I_{first} and I_{second} are the intensities of the highest and second-highest channels, respectively. The module corrects for varying base intensities using either median-based or percentile-based normalization (we selected the median-based approach, based on better correction on A-C crosstalk, Extended Data Fig. 3), with an option to restrict this correction to reads within cells, thereby excluding aberrant signals.

For segmentation of cells in sequencing images, multiple approaches are supported including Cellpose, MicroSAM, StarDist, and Watershed. We employed Cellpose with the cyto3 model with a nuclei diameter of 9.44 pixels and cell diameter of 18.85 pixels using base C of the last cycle (due to high background signal). Cell association assigns called barcodes to cells based on their spatial location, with configurability to assign reads to either cellular or nuclear compartments, supporting T7-based methods where DNA barcodes are localized to the nucleus. The module implements optional filtering of unreliable barcode assignments, including minimum quality scores for base calls (set to 0 in our analysis) and reference-based error correction against a library of expected barcodes (unused in our analysis).

Phenotype Module

The Phenotype module extracts cellular measurements through multiple processing steps: specimen-specific illumination correction, alignment of fluorescence channels to correct for chromatic aberrations, and segmentation of nuclear and cellular boundaries. The implementation supports both single-step and multi-step alignment strategies, with options for custom offsets when specific channels require targeted adjustment.

In alignment with previous methodologies, phenotype images were first maximum intensity projected to compress z-slices into a single plane. For cell segmentation, the module implements a primary workflow to identify nuclear and cellular boundaries, followed by a derived segmentation step that extracts cytoplasmic regions by subtracting nuclear masks from cellular masks. Multiple segmentation approaches are supported, including deep learning-based methods such as Cellpose, MicroSAM, and StarDist, each optimized for different cellular morphologies and imaging

conditions. Our implementation used Cellpose with the cyto3 model, with nuclei diameter of 41.25 pixels and cell diameter of 72.27 pixels, and default threshold parameters.

For feature extraction, Brieflow implements a pure Python CellProfiler-inspired emulator that integrates algorithms from multiple computational packages including scikit-image, mahotas, and SciPy. This implementation was necessitated by integration challenges between Java-based CellProfiler and Brieflow's Snakemake-based workflow architecture. The emulator extracts fundamental morphological parameters, intensity-based features across multiple channels, texture measurements through Haralick and PFTAS algorithms, and advanced shape descriptors including Zernike moments and Feret diameters. Given that the original Vesuvius dataset analysis robustly identified biologically relevant perturbation clusters, this feature extraction methodology remained largely unchanged, with a notable extension to include dedicated cytoplasmic feature extraction.

To assess the fidelity of this emulator approach, a comprehensive benchmarking comparison was conducted across eight phenotype tiles using a preliminary pipeline developed by the CellProfiler team (cp_measure) as part of their ongoing efforts to provide a pure Python implementation. This comparative analysis revealed a bimodal correlation pattern between methodologies. Robust and interpretable intensity measurements including integrated intensity, maximum intensity, and standard deviation exhibited strong agreement between methods (Extended Data Fig. 4e). Similarly, global shape descriptors such as Hu moments demonstrated exceptional correlation between implementations (Extended Data Fig. 4f), with correlation coefficients consistently high across all six moments. Conversely, features sensitive to local variations, particularly correlation features and radial distribution features, showed minimal agreement between implementations (Extended Data Fig. 4g, h). Notably, in several feature categories, the cp_measure implementation produced consistently flat or zero values, suggesting incomplete functionality within this developmental package.

From an efficiency perspective, the emulator completed feature extraction in a mean of 2.5 minutes per tile, while cp_measure required over 220 minutes (3.7 hours) per tile on average, rendering the latter impractical for high-throughput analyses of

datasets comprising thousands of tiles and millions of cells. This benchmarking strategy was not designed to evaluate the relative biological informativeness of different implementations, but rather to assess compatibility with established CellProfiler methodologies while maintaining computational feasibility for large-scale screening. Brieflow's modular pipeline architecture was specifically engineered to accommodate future integration of improved feature extraction methods as they mature, facilitating seamless incorporation of advancements in computational cell biology while preserving analytical throughput.

For user-specified channels, specialized foci detection and quantification is implemented using edge enhancement to enable counting and characterization of subcellular punctate structures, which we configured for the GH2AX channel (index 2). Feature extraction operates at three distinct cellular compartments—nucleus, cell, and cytoplasm—allowing precise characterization at each level. The module also computes correlation and colocalization metrics between channels, including Pearson correlation, overlap coefficient, and Manders' colocalization coefficients. For spatial relationships between cells, the module calculates neighbor metrics including counts of adjacent cells and percent boundary contact.

Merge Module

The Merge module integrates data from Sequencing by synthesis and Phenotype modules through spatial registration between imaging modalities. The foundation of the merge process is accurate alignment that identifies corresponding fields of view between datasets, leveraging metadata from the Preprocess module.

The implementation employs a Delaunay triangulation strategy to create hash-based descriptors for cell patterns, enabling robust matching even with partial cell correspondence. Each triangle serves as a feature descriptor, capturing spatial relationships between neighboring cells. The nine-edge hash method extracts vector displacements for edges connected to each triangle, creating a distinctive signature that can be matched across imaging modalities despite differences in magnification and optical characteristics.

The merge process uses a multi-step alignment algorithm that iteratively refines the transformation model as more matching points are identified. Initial alignment begins with a small set of candidate tile-site pairs, either specified by the user or automatically determined. Our implementation used six predefined tile pairs, providing starting points for the iterative alignment. The algorithm evaluates each candidate pair by computing a score based on the overlap of triangulation patterns and a determinant value that measures scaling consistency.

To control alignment quality, the pipeline offers configurable parameters for determinant range, score threshold, and distance threshold. Our implementation used a determinant range of [0.06, 0.065], ensuring the scaling factor remained within acceptable bounds. The matching score threshold was set to 0.1, requiring a minimum level of triangle pattern overlap to consider a match valid. For cell-level matching, a distance threshold of 2 pixels controlled the maximum distance between transformed coordinates for cells to be considered matches.

With spatial registration established, individual cells between modalities are linked based on their spatial coordinates, applying the optimized transformation parameters. The matching process incorporates distance-based thresholds to filter low-confidence associations. A deduplication pipeline resolves cases where multiple cells from one modality match a single cell in the other modality, applying a hierarchical strategy that first selects the best Sequencing by synthesis match for each phenotype cell (based on whether a cell has a known barcode), then selects the best phenotype match for each remaining Sequencing by synthesis cell (based on highest minimum channel intensity).

Quality assessment evaluates alignment quality using determinant range checks to verify scaling consistency and scoring parameters to assess matching accuracy. After cell matching, match rates for both modalities are calculated, reporting the percentage of cells successfully integrated into the final dataset.

Aggregate Module

The Aggregate module transforms single-cell perturbation data into robust gene-level embeddings through a systematic workflow that enables analysis of specific cellular

subpopulations. The module begins with data splitting, separating cells into distinct classes based on cell cycle state or morphological phenotypes using a flexible framework that supports any scikit-learn compatible classifier, loaded as a serialized dill file through an abstract base class.

For our implementation, we generated an external dataset and manually annotated 560 mitotic cells and 3,821 interphase cells to create a robust training set. We used an XGBoost classifier that distinguishes between mitotic and interphase cells, trained on 70% of this annotated data using 514 features extracted from a Brieflow analysis on the DAPI channel alone, with no normalization applied. When evaluated on a held-out test set (30% of the data), the classifier achieved high accuracy (99.4%), with excellent recall for both mitotic (97.6%) and interphase (99.7%) cells (Extended Data Fig. 5). Following classification, the pipeline can further subdivide datasets by channel combinations of interest, enabling analysis of phenotypic effects specific to particular cellular structures or markers.

The filter step implements a comprehensive strategy addressing common data quality issues. Query-based filtering allows removal of specific cell subsets based on configurable criteria; our implementation filtered to cells with a single gene mapping. Perturbation filtering restricts analysis to cells with assigned genetic perturbations, while missing value filtering drops columns and rows with a configurable proportion of missing data (we used thresholds of 0.05 for columns and 0.01 for rows). The module offers optional imputation of remaining missing values with K-nearest neighbors (enabled in our implementation). The Local Outlier Factor algorithm identifies outlier cells with anomalous marker patterns for removal, with a configurable contamination parameter (set to 0.01).

Batch alignment addresses technical variation by first organizing samples by experimental batch (defined in our case by plate and well identifiers), then applying Principal Component Analysis (PCA) to reduce dimensionality while preserving variation. Users can configure either the target variance to preserve (we used 0.99) or a specific component count. The critical Typical Variation Normalization (TVN) step leverages non-targeting control samples to standardize data, first centering and scaling based on control distributions, then aligning the covariance of all cell data based on the

control covariance structure. This generates globally representative principal components at the single-cell level that capture essential morphological feature relationships while removing technical variation.

The final aggregation stage combines cellular measurements into perturbation-level representations through two complementary approaches. First, batch-normalized principal components are aggregated to the perturbation level to create embeddings used for clustering analysis that reveals functional relationships between genetic manipulations. Second, the original interpretable features (such as nuclear size and stain intensities) are aggregated to the perturbation level to provide quantitative metrics that directly inform understanding of which specific cellular processes drive observed functional clusters. The aggregation supports multiple methods including mean and median approaches (we selected median). Quality assessment tools are integrated throughout the module, generating comprehensive reports of missing value patterns and feature distributions before and after batch correction. The implementation leverages efficient data structures and lazy evaluation to handle large datasets, with support for parallel processing and memory optimization for high-performance computing environments.

Cluster Module

The Cluster module identifies functional relationships between genes by grouping genetic perturbations with similar phenotypic profiles. This final analytical stage transforms the high-dimensional embeddings from the Aggregate module into interpretable clusters that reveal biological pathways and functional connections.

At the core of the Cluster module is a dimensionality reduction and clustering pipeline built on PHATE (Potential of Heat-diffusion for Affinity-based Transition Embedding) and Leiden community detection. PHATE preserves both local and global structure while reducing high-dimensional data to a visually interpretable two-dimensional representation. This approach leverages a diffusion-based metric that captures continuous phenotypic transitions better than linear methods like PCA or nonlinear methods like t-SNE. The PHATE embeddings serve as input to the Leiden

algorithm, which identifies communities of genes with similar phenotypic signatures, with resolution parameters controlling cluster granularity.

To objectively evaluate clustering quality, the module implements a comprehensive benchmarking framework that tests gene clusters against established biological databases. The evaluation employs two complementary approaches: pairwise gene relationship analysis using STRING protein-protein interactions, and group-level enrichment analysis using CORUM protein complexes and KEGG pathways. For pairwise evaluation, the system identifies true positive interactions when genes from known interacting pairs appear in the same cluster, while false positives and negatives are determined by mismatched clustering relative to known interactions. This yields standard precision and recall metrics, with an optional adjusted precision calculation that accounts for incomplete knowledge in reference databases.

For group-level evaluation, the module applies Fisher's exact test with Benjamini-Hochberg correction to assess enrichment of known gene complexes or pathways within each cluster. The system filters reference complexes to ensure robustness, requiring at least three genes present in the dataset and at least two-thirds of complex components represented. An additional filter removes larger complexes that share more than 10% of gene pairs with smaller ones, preventing redundant enrichment calls. The pipeline performs parallel analysis against randomly shuffled controls to establish baseline performance and confirm that identified patterns represent true biological signal rather than technical artifacts.

The Cluster module provides extensive visualization capabilities, generating PHATE embeddings that allow intuitive exploration of gene relationships in two-dimensional space. Cluster assignments are highlighted through color coding, enabling rapid identification of gene communities. Additional visualizations include cluster size distribution plots to assess clustering granularity, enrichment pie charts showing the proportion of clusters supported by different reference databases, and precision-recall curves to identify optimal clustering resolutions. These visualization tools facilitate both quality control assessment and biological interpretation of the clustering results.

MozzareLLM

As part of our comprehensive Brieflow pipeline for optical pooled screening analysis, we implemented MozzareLLM, a specialized framework that leverages large language models to analyze phenotypic clusters and identify promising gene candidates for experimental validation. MozzareLLM employs a systematic analytical approach for gene function classification and prioritization within phenotypic clusters.

The system first processes cluster membership data through a structured prompt engineering methodology that guides large language models to identify the dominant biological pathway represented in each gene cluster. This identification step applies stringent criteria for pathway confidence assessment, distinguishing between high confidence pathways with substantial literature support, medium confidence pathways with partial support, and low confidence or incoherent pathways with minimal functional coherence.

Following pathway identification, MozzareLLM categorizes genes into three mutually exclusive functional groups. Established pathway genes represent well-documented members with clear functional roles in the identified pathway, supported by multiple publications demonstrating direct involvement. Uncharacterized genes have minimal to no functional annotation in published literature, characterized by limited experimental validation and few publications focused on the gene. Novel role genes have established functions in other pathways that may contribute to the identified pathway in previously unrecognized ways.

For both uncharacterized and novel role genes, the system assigns a prioritization score ranging from 1 to 10 based on specific criteria. Uncharacterized genes receive higher scores when they have virtually no studies or unknown molecular function, with lower scores assigned to genes with partial evidence for pathway involvement. Novel role genes receive higher scores when compelling rationale exists for a previously unrecognized role with minimal existing literature, while lower scores indicate that existing data already supports involvement in the pathway.

Our implementation supports integration with multiple large language model providers including OpenAI, Anthropic, and Google, allowing researchers to select the most appropriate model for their analytical needs. The framework incorporates specialized prompts optimized for biological knowledge extraction, with conservative

prioritization approaches that emphasize high-confidence discoveries with clear experimental follow-up potential.

The MozzareLLM methodology produces standardized outputs that include the dominant biological process, pathway confidence level, gene classifications with prioritization scores, and rationales explaining why particular genes merit investigation. This structured approach facilitates efficient experimental planning by highlighting the most promising candidates for functional validation, particularly those representing potential novel pathway components or genes with unexpected functions.

HeLa Screen Generation and Analysis

The Vesuvius HeLa screen dataset comprised a CRISPR-Cas9 screen targeting 5,072 fitness-conferring genes with four sgRNAs per gene and 250 non-targeting control sgRNAs. Following library transduction into HeLa cells with integrated doxycycline-inducible Cas9, cells underwent puromycin selection for 4 days. Cas9 expression was then induced with doxycycline for 78 hours, identified as the optimal timepoint for maximizing observable phenotypes while minimizing dropout of knockout cells. The dataset consisted of eight 6-well plates with 45 wells containing fixed-cell imaging data capturing four cellular markers (DNA, DNA damage response, actin, and microtubules). More comprehensive details on how the screen was conducted are available in the original publication[5].

References

1. Feldman, D. et al. Optical Pooled Screens in Human Cells. *Cell* **179**, 787-799.e717 (2019).
2. Carlson, R.J., Leiken, M.D., Guna, A., Hacohen, N. & Blainey, P.C. A genome-wide optical pooled screen reveals regulators of cellular antiviral responses. *Proceedings of the National Academy of Sciences* **120**, e2210623120 (2023).
3. Ramezani, M. et al. A genome-wide atlas of human cell morphology. *Nat Methods* **22**, 621-633 (2025).
4. Labitigan, R.L.D. et al. Mapping variation in the morphological landscape of human cells with optical pooled CRISPRi screening. *eLife* **13** (2024).
5. Funk, L. et al. The phenotypic landscape of essential human genes. *Cell* **185**, 4634-4653.e4622 (2022).
6. Singh, S., Bray, M.A., Jones, T.r. & Carpenter, A.e. Pipeline for illumination correction of images for high-throughput microscopy. *Journal of Microscopy* **256**, 231-236 (2014).
7. Mantes, A.D. et al. (bioRxiv, 2024).
8. Archit, A. et al. Segment Anything for Microscopy. *Nat Methods* **22**, 579-591 (2025).
9. Schmidt, U., Weigert, M., Broaddus, C. & Myers, G. (eds. A.F. Frangi, J.A. Schnabel, C. Davatzikos, C. Alberola-López & G. Fichtinger) 265-273 (Springer International Publishing, 2018).
10. Stringer, C. & Pachitariu, M. Cellpose3: one-click image restoration for improved cellular segmentation. *Nat Methods* **22**, 592-599 (2025).
11. Stirling, D.R. et al. CellProfiler 4: improvements in speed, utility and usability. *BMC Bioinformatics* **22**, 433 (2021).
12. Celik, S. et al. Building, benchmarking, and exploring perturbative maps of transcriptional and morphological data. *PLOS Computational Biology* **20**, e1012463 (2024).
13. Ando, D.M., McLean, C.Y. & Berndl, M. (bioRxiv, 2017).
14. Moon, K.R. et al. Visualizing structure and transitions in high-dimensional biological data. *Nat Biotechnol* **37**, 1482-1492 (2019).
15. Traag, V.A., Waltman, L. & van Eck, N.J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* **9**, 5233 (2019).
16. Szklarczyk, D. et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research* **51**, D638-D646 (2023).
17. Steinkamp, R. et al. CORUM in 2024: protein complexes as drug targets. *Nucleic Acids Research* **53**, D651-D657 (2025).
18. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a model of the real world. *Nucleic Acids Research* **53**, D672-D677 (2025).
19. Hu, M. et al. Evaluation of large language models for discovery of gene set function. *Nat Methods* **22**, 82-91 (2025).
20. Wainberg, M. et al. A genome-wide atlas of co-essential modules assigns function to uncharacterized genes. *Nat Genet* **53**, 638-649 (2021).

21. Schaffer, L.V. et al. Multimodal cell maps as a foundation for structural and functional genomics. *Nature*, 1-10 (2025).
22. Martin, T.D. et al. A Role for Mitochondrial Translation in Promotion of Viability in K-Ras Mutant Cells. *Cell Reports* **20**, 427-438 (2017).
23. Smith, E.M., Ly, J., Haug, S. & Cheeseman, I.M. (bioRxiv, 2025).
24. Vedeler, A. & Bramham, C.R. Editorial: Coordination of mRNA Transport and Translation With Vesicle and Organelle Trafficking and Dynamics. *Front Cell Dev Biol* **9**, 800136 (2021).
25. Höpfler, M. et al. Mechanism of ribosome-associated mRNA degradation during tubulin autoregulation. *Molecular Cell* **83**, 2290-2302.e2213 (2023).
26. Tian, J. et al. TMEM44 as a Novel Prognostic Marker for Kidney Renal Clear Cell Carcinoma is Associated with Tumor Invasion, Migration and Immune Infiltration. *Biochem Genet* **62**, 1200-1215 (2024).
27. Lee, J.-H., You, J., Dobrota, E. & Skalnik, D.G. Identification and Characterization of a Novel Human PP1 Phosphatase Complex. *J Biol Chem* **285**, 24466-24476 (2010).
28. Cossa, G. et al. Localized Inhibition of Protein Phosphatase 1 by NUAK1 Promotes Spliceosome Activity and Reveals a MYC-Sensitive Feedback Control of Transcription. *Molecular Cell* **77**, 1322-1339.e1311 (2020).
29. Dixit, A. et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866.e1817 (2016).
30. Kudo, T. et al. Multiplexed, image-based pooled screens in primary cells and tissues with PerturbView. *Nat Biotechnol*, 1-10 (2024).

Table Legends

Feature	PERISCOPE[3]	OpticalPooledScreens[1]	PerturbView[30]	Brieflow
End-to-End Integration	?	?	?	?
Unified Framework	?	?	?	?
Analysis Documentation	Comprehensive	Minimal	Minimal	Comprehensive
Workflow Management	Manual	Snakemake	Snakemake	Snakemake
Modular Architecture	?	?	?	?
Community Extensibility	?	?	?	?
Primary Focus	CellProfiler-based analysis	Basic phenotype/SBS	Tissue spatial alignment	Complete OPS analysis
Data Integration	Multiple separate pipelines	Fragmented workflow	Specialized use case	Single integrated pipeline
Analytical Scope	Complete analysis, no interpretation	Incomplete pipeline	Specialized application	Complete analysis and interpretation

Table 1: Comparison of existing optical pooled screening analysis pipelines and their capabilities. Features are compared across four computational frameworks for OPS data analysis. Periscope relies on CellProfiler-based architecture requiring integration of multiple separate analytical pipelines. OpticalPooledScreens provides basic phenotype and sequencing analysis components but lacks comprehensive integration and documentation for end-to-end workflows. PerturbView focuses specifically on tissue spatial alignment applications with limited broader analytical scope. Brieflow provides the first unified framework encompassing complete analysis from raw data to biological interpretation. Extensive benchmarking of alternative pipelines was not feasible due to limited accessibility and the time-intensive nature of adapting existing frameworks to perform the comprehensive end-to-end analysis demonstrated with Brieflow.

Figure Legends

Fig. 1. Brieflow: An integrated computational framework for optical pooled screening analysis.

a. Schematic of optical pooled screening workflow. Cells expressing inducible Cas9 are transduced with sgRNA library, perturbed, and subjected to both phenotypic imaging and in situ sgRNA sequencing, followed by computational analysis using Brieflow.

b. The Brieflow pipeline architecture comprising six integrated modules: Preprocess converts raw microscopy files to standardized Python readable formats; Sequencing by Synthesis identifies guide barcodes; Phenotype extracts cellular features; Merge integrates genotype and phenotype data using triangulation patterns; Aggregate condenses single-cell data to perturbation-level statistics; and Cluster identifies functional relationships between perturbations.

c. Implementation architecture showing the relationship between core components. Left: The Brieflow codebase with modular structure. A high level Snakefile imports rules and targets from each module. Rules run individual Python scripts, which call on functions defined in lib. Right: The complementary Brieflow-analysis template. A configuration notebook allows the user to test and implement parameters, which are automatically written into the configuration file at the conclusion of the testing process. These are employed by the configuration scripts, which call the Snakefile with appropriate parameters relating to the module being run.

Fig. 2. Sequencing by Synthesis and Phenotype modules for high-throughput analysis of optical pooled screening data.

a. Workflow of the Sequencing by synthesis module. In situ sequencing images undergo multi-cycle alignment, spot detection, base calling, cell segmentation, and barcode-to-cell mapping, generating a comprehensive output table with cellular barcode assignments with per-base and per-barcode mapping statistics.

b. Quality assessment metrics visualizing sequencing performance for plate 1. Top: Heatmap showing fraction of cells mapping to a singular gene symbol, enabling

identification of areas with low spot signal (circled). Bottom: Distribution of cells corresponding to genetic perturbations.

c. Comparison of spot detection methods, representative tile shown. Spotiflow detected more spots per tile than the standard statistical method with a threshold of 350 (12,554 vs. 9,431) but resulted in comparable gene mapping efficiency (64.2% vs. 65.1).

d. Phenotype module workflow for extracting cellular measurements from high-content imaging data. Multi-channel fluorescence images undergo illumination correction, alignment, segmentation, and feature extraction, quantifying morphological and functional characteristics across cellular compartments.

e. Benchmarking of segmentation approaches, representative panel shown. Cellpose achieved superior segmentation with HeLa cells, evident through qualitative inspection.

f. Quality assessment visualizations of phenotypic measurements for plate 1. Top: Minimum Phalloidin intensity used for cytoplasmic segmentation. Bottom: Cell count distribution, demonstrating correlation between staining anomalies and decreased segmentation efficiency (circled).

Fig. 3. Merge, Aggregate, and Cluster modules for integrating genetic and phenotypic data.

a. Merge workflow integrating phenotypic features with genetic perturbation data. Phenotypic measurements (left matrix) and barcode mapping statistics (right matrix) are combined using a sequential process of triangle hashing based on spatial location for registration, merging of corresponding cells, and deduplication to resolve multiple matches, resulting in a unified dataset (bottom) that links genetic perturbations with their associated phenotypic signatures.

b. Quality assessment visualizations of Merge performance for plate 2. Left: Heatmaps showing the distribution of successfully merged cells across six representative wells, with brighter yellow regions indicating higher merge efficiency. Right: Merge coverage visualization highlighting cells with minimum intensity value of zero, which are deprioritized during the deduplication process. These predominantly occur at tile boundaries due to optical edge effects in the microscopy system.

c. Aggregate workflow for transforming single-cell data into perturbation-level embeddings. The module takes high-dimensional single-cell feature matrices (top matrix) and splits cells into distinct subpopulations of interest, removes low-quality measurements, outliers, and cells with anomalous marker patterns, and standardizes feature distributions using non-targeting controls as reference. Finally, to generate standardized gene-level representations (bottom matrix), cellular measurements are combined into perturbation-level statistics.

d. Quality assessment visualizations demonstrating batch effect correction through the Aggregate module. Violin plots show the distribution of the mean tubulin feature and a representative principal component across experimental batches for all cells before (top) and after (bottom) application of Typical Variation Normalization.

e. Cluster module workflow for identifying functional relationships between genetic perturbations. High-dimensional perturbation embeddings undergo dimensionality reduction using PHATE to generate a two-dimensional representation that preserves both local and global structure. The analytical framework incorporates multiple complementary approaches for functional interpretation, including annotations from reference databases, pairwise comparisons of gene relationships, and group-level pathway enrichment analysis.

f. Cluster evaluation showing precision-recall performance at different Leiden resolution parameters using the CORUM benchmark. Points represent clustering results at various resolution values, with higher resolutions yielding more granular clusters with higher precision but lower recall, and lower resolutions producing fewer clusters with higher recall but lower precision. Precision measures the proportion of gene pairs in the same cluster that belong to the same protein complex, using an adjusted calculation that only counts false positives when both genes are known to belong to different complexes. This avoids penalizing potential novel interactions absent from reference databases. The red circle highlights resolutions 11-15, representing an optimal balance between precision and recall.

Fig. 4. Interactive visualization dashboard and MozzareLLM for biological interpretation of OPS data.

a. Schematic of the interactive visualization dashboard integrated within Brieflow. The dashboard comprises four displays: Experimental Metadata for information relating to how the screen was conducted, Configuration for information on parameters used in Brieflow, Quality Control for assessment of data integrity, and Cluster Exploration (displayed) for visualization of gene relationships, interpretation of gene montages, and intuitive understanding of MozzareLLM output.
b. MozzareLLM workflow for automated biological interpretation of phenotypic clusters. The system processes cluster data (left) through a structured prompt engineering approach that feeds into a range of language models (center), producing pathway summaries with varying degrees of confidence for clusters with functional annotation of genes within it (right). MozzareLLM identifies dominant biological processes for each cluster, categorizes genes into established pathway members, characterized genes with potential novel functions, and uncharacterized genes, assigning prioritization scores based on evidence strength and discovery potential.

Extended Data Legends

Extended Data Fig. 1: Computational Workflow Schema of the Brieflow Analysis Pipeline

Snakemake generated rulegraph for all Brieflow steps.

Extended Data Fig. 2: Illumination Correction Function for Phenotypic Image Processing

Illumination correction function generated during preprocessing for DAPI phenotype images in well A1 of plate 1 (left) showing lower illumination intensity at well corners (circled), and representative DAPI image from the same well (right).

Extended Data Fig. 3: Spotiflow Performance Benchmarking and Normalization Strategies for Improved Base Calling Accuracy

a. Runtime and memory usage comparison for Spotiflow relative to the standard method across benchmarking comparisons. **b.** Comparison of normalization strategies for

accurate base calling for an example tile. The plots illustrate how raw fluorescence intensity data (top) is transformed using two correction methods: median normalization (middle) and percentile normalization (bottom). The median method corrects cross-talk by constructing a transformation matrix from the median signals where each base has maximum intensity, while the percentile method selectively uses spots with ≥95th percentile relative signal to define channel separation. Both approaches significantly improve discrimination between complementary base pairs (C-A and T-G), enabling more confident base calling as indicated by the colored points showing base assignments.

Extended Data Fig. 4: Comparative Analysis of Nuclei and Cell Segmentation Methods: Detection Efficiency, Accuracy, and Computational Performance

a. Initial nuclei identified by various segmentation methods in phenotype benchmark shows relatively similar counts. **b.** Initial cells identified by various segmentation, showing that cellular body detection varies largely across models. **c.** Coalescing of cell and nuclear measurements (each cell has at least one nuclei) shows large drop in number of finalized cells for MicroSAM and StarDist, relative to Cellpose, which retains most cell-nuclei pairs. **d.** Runtime comparison of segmentation methods. **e.** DAPI nuclear intensity features including integrated intensity, maximum intensity, and standard deviation show strong agreement between emulator and cp_measure implementations. **f.** Cellular Tubulin Hu moment features comparison between emulator and cp_measure implementations demonstrates high correlation across all six moments. **g.** Cellular Phalloidin radial distribution features show minimal agreement between implementations, with substantial variability observed across different feature types. **h.** Cellular Phalloidin-Tubulin colocalization metrics comparison between implementations reveals inconsistent correlation patterns, with notable clustering of values at specific ranges. Each dot represents one of 8,139 cells across the 8 example images analyzed, with different colors indicating distinct feature measurements. The r value shown is the Pearson correlation coefficient, quantifying the linear relationship between the two measurement approaches.

Extended Data Fig. 5: Development and Validation of a Machine Learning Classifier for Mitotic and Interphase Cell Discrimination

a. Training data used to generate a mitotic vs. interphase cell classifier. In the final classifier, all data was used for training. **b.** Confusion matrix on initially held out 30% test set shows high degree of accuracy in classifying interphase and mitotic cells. **c.** Example classification of interphase and mitotic cells on re-analyzed dataset (with DAPI marker). High confidence cells in both interphase and mitotic categories show high accuracy, while low confidence interphase cells show a modest proportion of inaccurately classified mitotic cells.

Extended Data Fig. 6: MozzareLLM-Based Analysis of Interphase Cell Clusters and Gene Function Classification Confidence

a. Results from MozzareLLM analysis of final interphase clustering, with MozzareLLM confidence level displayed across clusters. **b.** Examining high confidence pathways alone, a larger proportion of cell counts per cluster fall into the established gene category, with low number of uncharacterized and slightly higher numbers of novel role genes called per cluster.

Extended Data Fig. 7: Mitochondrial Network Recovery and Phenotypic Validation

a. Phenotypic heatmap of mitochondrial gene clusters identified by Brieflow. Features were selected by combining the top differential morphological features from each cluster, with hierarchical clustering performed within each cluster block. Phenotypic analysis reveals shared feature patterns across these clusters, validating their common functional signatures. **b.** PHATE visualization of mitochondrial clusters in Brieflow analysis. Distinct mitochondrial clusters were recovered despite modest phenotypic deviations from nontargeting controls in PHATE space. **c.** PHATE visualization from original Vesuvius analysis for the same genes, colored by their original cluster assignments. Legend omitted due to distribution across 41 different clusters.

Extended Data Fig. 8: Phenotypic Validation of Ribosome Biogenesis Cluster a. Phenotypic ranking plot for cellular area measurements across all genes in the screen.

Cluster 5 genes (red dots) show consistently reduced nuclear area compared to non-targeting controls, with established ribosomal components (POP5, RPP30, C18orf21) and newly implicated genes (NEPRO) exhibiting similar phenotypic signatures. **b.** Heatmap of top differential morphological features for Cluster 5 genes. Differential features were identified by comparing cluster genes against all other genes using Mann-Whitney U tests and robust z-score effect sizes, selecting features with the largest absolute differences. Feature values represent median measurements per perturbation following center-scaling normalization on non-targeting controls, with heatmap display using global z-score normalization across all genes in the dataset. Genes within this cluster display coherent phenotypic patterns with elevated tubulin-associated measurements (red) and reduced nuclear morphometric parameters (blue), validating the functional relationship between established and newly identified ribosome biogenesis components.

Extended Data Fig. 9: Functional Refinement and Phenotypic Analysis of Vesicle Trafficking Pathways a. Cluster redistribution for vesicle trafficking genes. Sankey diagram illustrating how genes from original Vesuvius vesicle trafficking clusters (54, 140, 201) were reassigned to more refined clusters in our Brieflow reanalysis. Genes are colored according to their original cluster assignment, with boxes highlighting additional genes incorporated into updated clusters. The refinement reveals distinct vesicular transport modules including separated exocyst complex and CCR4-NOT components. **b.** Phenotypic validation of Cluster 60 showing varying levels of elevated cellular tubulin across both CCR4-NOT components (CNOT1, CNOT4) and exocyst components (EXOC1, EXOC4, EXOC7), supporting their functional coordination through shared tubulin regulatory mechanisms.

Extended Data Fig. 10: Functional Refinement and Phenotypic Analysis of Transcriptional Regulation Pathways a. Cluster redistribution for transcriptional regulation genes. Sankey diagram showing redistribution of genes from original Vesuvius transcriptional regulation clusters (39, 121, 146) into more functionally

coherent groupings. The refined clustering separates distinct transcriptional regulatory modules with greater functional resolution, including the focused MYC-dependent network identified in Cluster 95. **b.** Phenotypic validation of Cluster 95 showing the relationship between nuclear eccentricity and DAPI intensity measurements. Cluster 95 genes exhibit decreased DAPI intensity and increased nuclear eccentricity (trend line r = -0.13), indicating altered chromatin organization characteristic of disrupted transcriptional control networks.

Fig. 1

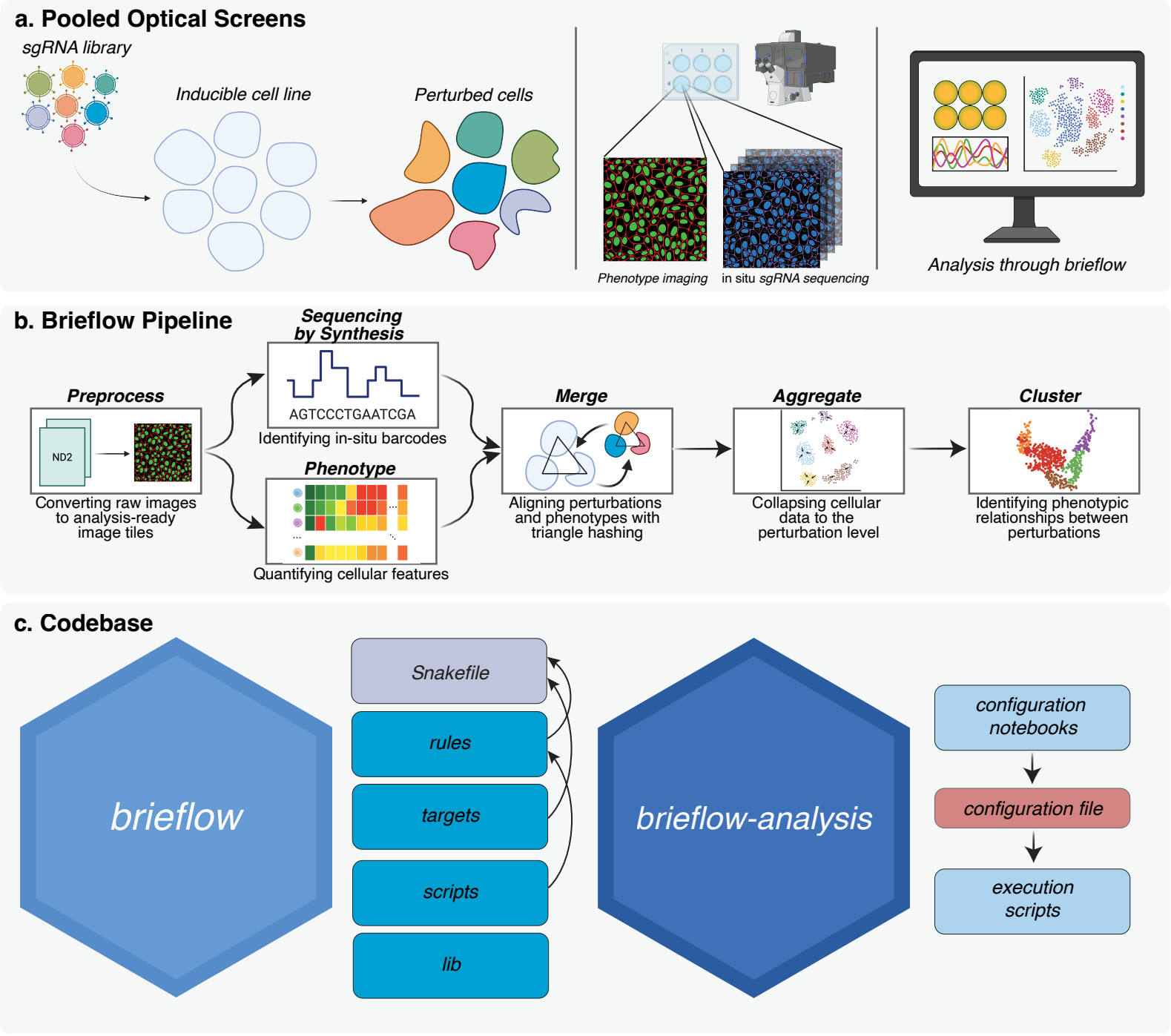

a. Pooled Optical Screens
sgRNA library
Inducible cell line
Perturbed cells
Phenotype imaging
in situ sgRNA sequencing
Analysis through brieflow
b. Brieflow Pipeline
Preprocess
ND2
Converting raw images to analysis-ready image tiles
Sequencing by Synthesis
AGTCCCTGAATCGA
Identifying in-situ barcodes
Phenotype
Quantifying cellular features
Merge
Aligning perturbations and phenotypes with triangle hashing
Aggregate
Collapsing cellular data to the perturbation level
Cluster
Identifying phenotypic relationships between perturbations
c. Codebase
brieflow
Snakefile
rules
targets
scripts
lib
brieflow-analysis
configuration notebooks
configuration file
execution scripts

Fig. 2

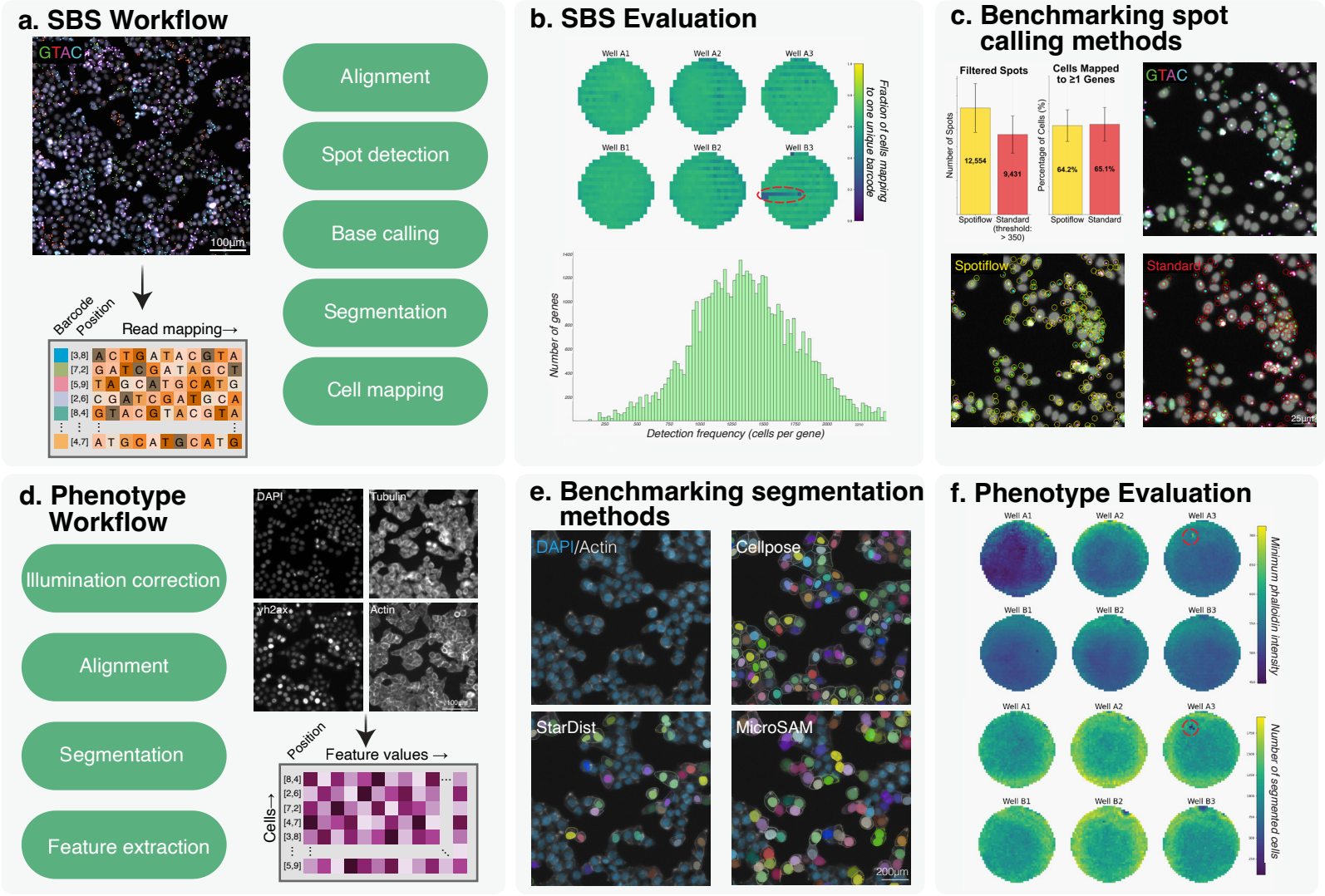
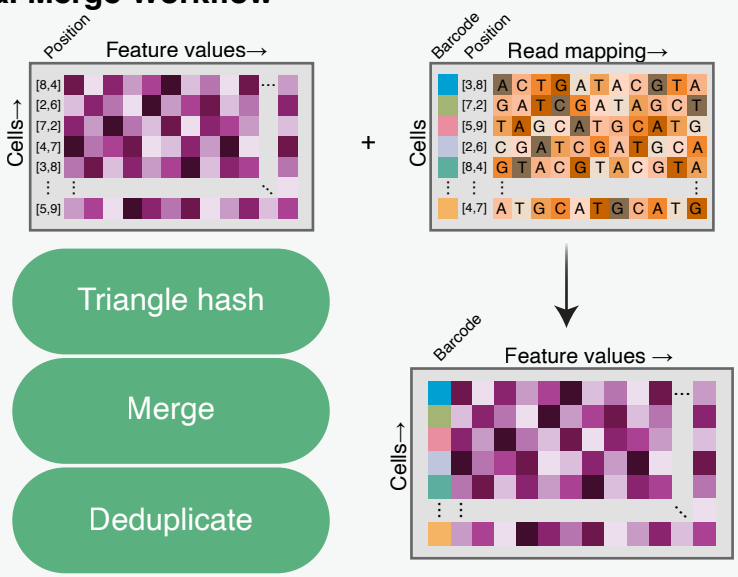
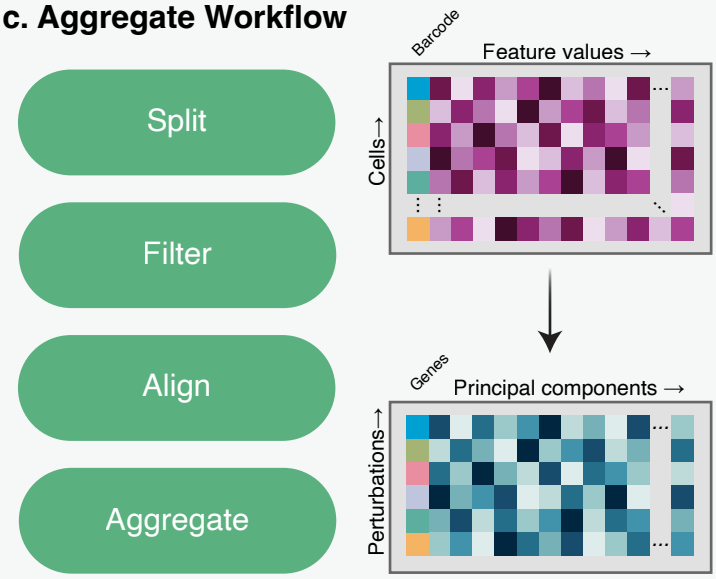

a. SBS Workflow
GTAC
100µm
Barcode
Position
Read mapping→
Alignment
Spot detection
Base calling
Segmentation
Cell mapping
b. SBS Evaluation
Well A1
Well A2
Well A3
Well B1
Well B2
Well B3
Fraction of cells mapping to one unique barcode
Number of genes
Detection frequency (cells per gene)
c. Benchmarking spot calling methods
Filtered Spots
Cells Mapped to ≥1 Genes
Number of Spots
Percentage of Cells (%)
12,554
9,431
64.2%
65.1%
Spotiflow
Standard (threshold: > 350)
Spotiflow
Standard
GTAC
Spotiflow
Standard
25µm
d. Phenotype Workflow
Illumination correction
Alignment
Segmentation
Feature extraction
DAPI
Tubulin
γh2ax
Actin
100µm
Position
Feature values →
Cells→
e. Benchmarking segmentation methods
DAPI/Actin
Cellpose
StarDist
MicroSAM
200µm
f. Phenotype Evaluation
Well A1
Well A2
Well A3
Well B1
Well B2
Well B3
Minimum phalloidin intensity
Number of segmented cells

Fig. 3

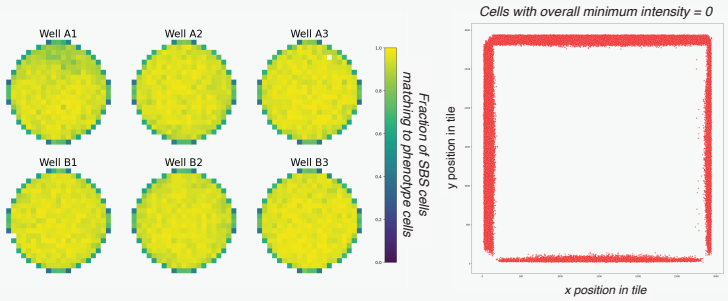
a. Merge Workflow



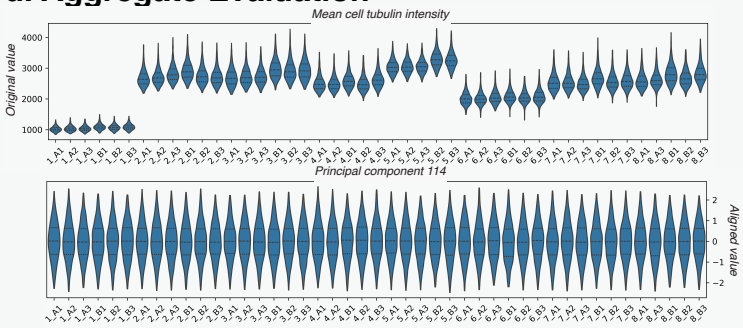
c. Aggregate Workflow



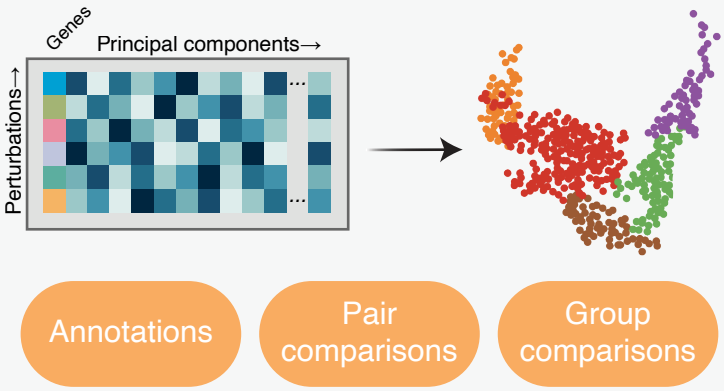
b. Merge Evaluation



d. Aggregate Evaluation



e. Cluster Workflow



f. Cluster Evaluation

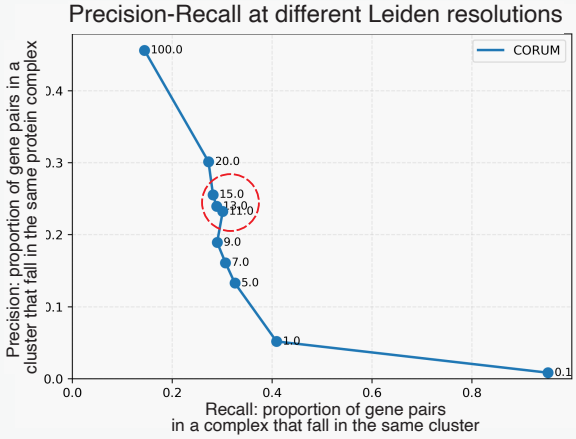


Fig. 4

a. Interactive Dashboard

Experimental Overview
Analysis Overview
Quality Control
Cluster Analysis

Filters

Channel Combo - *Used to subset features during aggregation*
- DAPI_TUBULIN_GH2AX_PHALLOIDIN

Cell Class - *Used to subset single cell data with classifier provided during aggregation*
- all
- Mitotic
- Interphase

Leiden Resolution - *Used in the Leiden clustering algorithm to determine gene clusters*
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15

Cluster Analysis

Click a cluster to see details, panning and zooming is easily done through the top right of the cluster panel

Gene Search: Select a gene...
Cluster Search: Select a cluster...

Cluster Analysis

Click a cluster to see details, panning and zooming is easily done through the top right of the cluster panel

Gene Search: AATF
Cluster Search: 5

Cluster 5: 61 genes — Close Cluster

LLM Cluster Analysis

Dominant Process: Ribosome biogenesis and rRNA processing — Confidence: High

Cluster 5 shows a strong and coherent signature of ribosome biogenesis and rRNA processing, with 34 established genes directly involved in this pathway. The cluster contains numerous components of the small subunit (SSU) processome, RNA helicases, and factors involved in pre-rRNA processing and modification. Five genes remain functionally uncharacterized, with C18orf21 and C1orf131 being virtually unstudied and high-priority candidates for investigation. Additionally, 17 genes with established roles in other pathways may have novel functions in ribosome biogenesis, including NEPRO (Notch signaling), BPTF (chromatin remodeling), and ALDH18A1 (amino acid biosynthesis). The connection between these diverse cellular processes and ribosome biogenesis represents an interesting area for further research.

Established Genes: DHX37, RCL1, DDX18, UTP25, UTP14A, UTP3, UTP20, BUD23, FCF1, IMP4, RRP7A, NOL6, POP7, DDX21, RBIS, USP36, TRMT112, BMS1, NGDN, RPP30, GAR1, TSR2, UTP11, RRP36, TSR1, RRP9, DDX49, NOP14, AATF, NAT10, POP5, PDCD11, KRR1, ESF1

Uncharacterized Genes:

- **NOP9** Limited functional annotation in the literature. While it appears to be involved in ribosome biogenesis based on its name (nucleolar protein), there is minimal experimental validation of its specific function in rRNA processing pathways.
- **NOL10** Very limited functional characterization despite its name suggesting nucleolar localization. No clear experimental evidence defining its specific role in ribosome biogenesis, making it a promising candidate for investigation.
- **C18orf21** Virtually unstudied gene with no functional annotation in the literature. Its presence in this cluster suggests a potential role in ribosome biogenesis that has not been characterized.
- **NOC4L** Limited functional characterization despite its name suggesting nucleolar localization. Its specific role in ribosome biogenesis remains unclear, making it a good candidate for further investigation.
- **C1orf131** Uncharacterized open reading frame with virtually no functional annotation in the literature. Its presence in this ribosome biogenesis cluster suggests a novel function that merits investigation.

Novel Role Genes:

- **NEPRO** Known for its role in Notch signaling and neural progenitor maintenance, but its presence in this cluster suggests a previously unrecognized function in ribosome biogenesis. This connection between developmental signaling and ribosome biogenesis would be novel.
- **BPTF** Well-characterized as a chromatin remodeling factor in the NURF complex, but its potential role in ribosome biogenesis has not been established. May regulate transcription of ribosomal genes through its chromatin remodeling activity.
- **CSNK1A1** Well-studied kinase involved in Wnt signaling and other pathways, but its potential role in ribosome biogenesis is not established. May phosphorylate key factors in the ribosome biogenesis pathway.
- **CBLL1** Known E3 ubiquitin ligase involved in endocytosis and RNA methylation as part of the WMM complex, but a specific role in ribosome biogenesis has not been described. May regulate stability of ribosome biogenesis factors.
- **NOM1** Known to target PPP1CA to the nucleolus, but its direct role in ribosome biogenesis is not well-established. May serve as a link between phosphatase activity and ribosome assembly.

Gene Montages

Select a gene to view (within this cluster): C18orf21

Uniprot Entry: Q32NC0

Uniprot Function: Not available

Select Guide: AGTTATCCAGAACCAACAGC

Channel: DAPI

Channel: PHALLOIDIN

b. LLM gene set evaluator

mozzarellm

Anthropic

Gemini

Open-AI

Cluster analysis:

Pathway summary

Confidence:

Low

Medium

High

Gene analysis:

Established in pathway

Novel role in pathway

Uncharacterized gene