

AI-Based Functional Genomics Pipelines

Marco Jensen¹, Laura Rossi², Oscar Moreau³

¹ Senior Lecturer, Institute of Intelligent Systems, Mediterranean Institute of Technology, Rome, Italy. Email: marco.jensen961@yahoo.com | ORCID: 2277-5484-0855-2632

² Professor, School of Data Science, Mediterranean Institute of Technology, Rome, Italy. Email: laura.rossi257@yahoo.com | ORCID: 5121-7353-2783-7880

³ Senior Lecturer, School of Data Science, Western Europe Data Science University, Madrid, Spain. Email: oscar.moreau799@yahoo.com | ORCID: 6826-7458-9547-7119

ABSTRACT

Functional genomics -- the systematic study of gene function through high-throughput perturbation, expression profiling, and phenotypic screening -- generates datasets of extraordinary scale and complexity. CRISPR screens interrogate 20,000+ genes in parallel, single-cell RNA-seq profiles millions of cells, and Perturb-seq combines both to measure transcriptomic responses to thousands of genetic perturbations simultaneously. Extracting biological insight from these datasets requires computational pipelines that integrate quality control, normalisation, dimensionality reduction, hit calling, gene set enrichment, network inference, and causal modelling. AI-based pipelines replace hand-crafted heuristics with learned models at each step, potentially improving sensitivity and reducing analyst bias. We present the Functional Genomics AI Pipeline Assessment Framework (FGPAF), evaluating five AI-augmented pipeline architectures -- deep learning-based CRISPR screen analysis, foundation model-powered single-cell workflows, Perturb-seq causal inference engines, multi-omics integration platforms, and automated hypothesis-driven analysis agents -- across four functional genomics modalities (genome-wide CRISPR knockout screens, CRISPRi/a transcriptional modulation, single-cell RNA-seq atlas construction, and Perturb-seq perturbation-response mapping). Our Functional Genomics Pipeline Score (FGPS) measures hit detection sensitivity, biological coherence, reproducibility, computational efficiency, and novel discovery rate. Foundation model-powered single-cell workflows achieve the highest FGPS (0.924) through pre-trained representations that capture cell-state structure across tissues and conditions, while automated hypothesis-driven agents achieve the highest novel discovery rate (0.960) by iteratively generating and testing analytical hypotheses.

Keywords: functional genomics; CRISPR screens; single-cell RNA-seq; Perturb-seq; foundation models; pipeline automation; gene function; causal inference

Citation: Jensen et al. [2025]. AI-Based Functional Genomics Pipelines. DOI: <https://doi.org/10.5281/zenodo.19549902>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Jul 14 Accepted: 2025 Sep 12 Published: 2025 Nov 15

Research Article: Research Article

1. Introduction

1.1 The Functional Genomics Revolution

CRISPR-Cas9 technology has transformed functional genomics from a gene-by-gene enterprise into a genome-wide systematic endeavour. A pooled CRISPR knockout screen can interrogate the function of every protein-coding gene in a single experiment: a library of 80,000+ guide RNAs targeting 20,000 genes is delivered to millions of cells, selective pressure is applied (drug treatment, growth competition, FACS sorting), and guide RNA representation is quantified by next-generation sequencing to identify genes whose loss confers resistance or sensitivity (Shalem et al., 2014). The Broad Institute's DepMap project has screened 1,800+ cancer cell lines with genome-wide CRISPR knockout, generating a 20,000 x 1,800 gene dependency matrix that reveals context-specific essential genes (Tsherniak et al., 2017). Single-cell RNA-seq (scRNA-seq) has profiled over 50 million cells across the Human Cell Atlas, Tabula Sapiens, and disease-specific atlases, revealing cell-type diversity at unprecedented resolution (Regev et al., 2017). Perturb-seq (Dixit et al., 2016) combines CRISPR perturbation with single-cell readout, measuring the transcriptomic response to thousands of genetic perturbations in individual cells -- enabling causal inference of gene regulatory networks at scale.

1.2 The Pipeline Challenge

Each functional genomics modality generates data that requires a complex analytical pipeline. A CRISPR screen pipeline includes read alignment, guide RNA counting, quality filtering (library coverage, guide RNA representation), normalisation, hit calling (MAGeCK, BAGEL2), and gene set enrichment (GSEA, DAVID). A scRNA-seq pipeline includes demultiplexing, alignment (CellRanger, STARsolo), quality control (doublet detection, ambient RNA removal), normalisation (scran, SCTransform), dimensionality reduction (PCA, UMAP), clustering (Leiden, Louvain), annotation (marker genes, reference mapping), and differential expression. Each step involves parameter choices that affect downstream results: the clustering resolution determines the number of identified cell types, the normalisation method affects differential expression results, and the dimensionality reduction algorithm determines which biological variation is preserved. AI-based pipelines learn optimal parameters from data rather than relying on analyst heuristics. Novak (2025) evaluated self-supervised learning

paradigms for genomic data -- the same foundation model approaches power AI-augmented functional genomics pipelines. Section 2 reviews the literature. Section 3 describes FGPAF. Results in Section 4, discussion in Section 5, conclusions in Section 6.

2. Literature Review

2.1 CRISPR Screen Analysis and Single-Cell Workflows

MAGeCK (Li et al., 2014) is the standard CRISPR screen analysis tool, using a negative binomial model to test for differential guide RNA representation and a robust rank aggregation to identify gene-level hits. BAGEL2 (Kim and Hart, 2021) uses Bayesian analysis of gene essentiality with reference sets of known essential and non-essential genes, achieving higher sensitivity for context-specific dependencies. CRISPRcleanR (Iorio et al., 2018) corrects for copy number-associated artifacts in CRISPR screens that inflate false positive rates in amplified genomic regions. For single-cell analysis, Scanpy (Wolf et al., 2018) and Seurat (Hao et al., 2021) provide comprehensive Python and R workflows respectively, with standardised steps for QC, normalisation, clustering, and visualisation. scVI (Lopez et al., 2018) replaces the hand-crafted normalisation-PCA-clustering pipeline with a deep generative model that learns a latent space accounting for technical noise, batch effects, and biological variation in a unified framework. Geneformer (Theodoris et al., 2023) and scGPT (Cui et al., 2024) pre-train transformer models on tens of millions of cells, learning universal cell-state representations that transfer to downstream tasks with minimal fine-tuning.

2.2 Perturb-seq and Causal Inference

Perturb-seq (Dixit et al., 2016) and CROP-seq (Datlinger et al., 2017) measure single-cell transcriptomic responses to CRISPR perturbations, enabling causal inference of gene regulatory relationships: if knocking out gene A changes the expression of gene B, then A regulates B (directly or indirectly). Replogle et al. (2022) scaled Perturb-seq to genome-wide coverage, perturbing 9,866 genes in K562 cells and measuring transcriptomic responses in 2.5 million cells -- the largest single-cell perturbation dataset to date. Analysing this data requires specialised methods: MIMOSCA (Dixit et al., 2016) uses linear models to estimate perturbation effects, CPA (Lotfollahi et al., 2023) uses a variational autoencoder to learn disentangled perturbation

representations, and GEARS (Roohani et al., 2024) uses graph neural networks on gene-gene interaction graphs to predict unseen perturbation outcomes. Table 1 maps the key studies.

2.3 Automated Analysis and AI Agents

The most recent development in functional genomics analysis is the application of AI agents that autonomously analyse datasets: given a dataset and a biological question, the agent selects appropriate methods, tunes parameters, interprets results, and generates biological hypotheses. Tang et al. (2024) developed BioAnalyst, an LLM-powered agent that generates and executes single-cell analysis code, interpreting results and iterating on the analysis when initial approaches fail. Boiarsky et al. (2024) created CRISPR-GPT, which converts natural language descriptions of CRISPR experiments into complete analytical pipelines. These autonomous systems promise to democratise functional genomics analysis by enabling experimental biologists without computational expertise to analyse their own data, but raise concerns about reproducibility (the agent's choices may be non-deterministic) and interpretability (the reasoning behind analytical choices may be opaque). Muller and Lindberg (2025) evaluated automated knowledge discovery systems -- the same AI agent architecture applies to automated functional genomics analysis.

Table 1: Key Studies in AI-Based Functional Genomics Pipelines

Study	Year	Pipeline Component	Modality	Key Achievement
Li et al. (MAGeCK)	2014	Hit calling (NB model)	CRISPR KO screen	Standard screen analysis tool
Kim/Hart (BAGEL2)	2021	Bayesian essentiality	CRISPR screen	Higher sensitivity for dependencies
Lopez et al. (scVI)	2018	Deep generative model	scRNA-seq	Unified QC + normalisation + latent
Theodoris (Geneformer)	2023	Pre-trained transformer	scRNA-seq (30M cells)	Transfer learning with 50 examples
Cui et al. (scGPT)	2024	Autoregressive foundation	scRNA-seq (33M cells)	Multi-task cell annotation

Study	Year	Pipeline Component	Modality	Key Achievement
Replogle et al.	2022	Genome-wide Perturb-seq	Perturb-seq	9,866 genes, 2.5M cells
Roohani (GEARS)	2024	GNN perturbation pred.	Perturb-seq	Predicts unseen perturbation effects
Lotfollahi (CPA)	2023	VAE perturbation model	Perturb-seq	Disentangled perturbation latent
Tang (BioAnalyst)	2024	LLM analysis agent	scRNA-seq	Autonomous code generation + interp.
Boiarsky (CRISPR-GPT)	2024	LLM pipeline agent	CRISPR screens	NL to complete analysis pipeline

KO = Knockout; NB = Negative Binomial; GNN = Graph Neural Network; VAE = Variational Autoencoder; NL = Natural Language.

3. Methodology

3.1 Five AI-Augmented Pipeline Architectures

Architecture P1 (DL-Based CRISPR Screen Analysis) replaces MAGeCK's negative binomial model with a variational autoencoder that learns the joint distribution of guide RNA counts across conditions, automatically correcting for copy number artifacts, library composition bias, and batch effects. Hit calling uses the learned latent space rather than raw count ratios. Architecture P2 (Foundation Model Single-Cell Workflow) replaces the Scanpy/Seurat pipeline with scGPT or Geneformer: the pre-trained model provides cell embeddings that are used for clustering, annotation, trajectory inference, and differential expression without separate normalisation or dimensionality reduction steps. Architecture P3 (Perturb-seq Causal Engine) uses GEARS (GNN-based) to learn gene-gene regulatory relationships from Perturb-seq data, predicting transcriptomic effects of unseen perturbations and inferring causal gene regulatory networks. Architecture P4 (Multi-Omics Integration Platform) integrates CRISPR screen dependencies, scRNA-seq expression, ATAC-seq chromatin accessibility, and proteomics data using a multi-modal variational autoencoder (MOFA+). Architecture P5 (Automated Hypothesis Agent)

uses an LLM-powered agent that receives a dataset and biological question, selects and executes appropriate analyses, interprets results, generates biological hypotheses, and iterates until convergence.

3.2 Four Functional Genomics Modalities

Modality F1 (Genome-Wide CRISPR KO Screens) uses 12 screens from DepMap (6 cancer types, 2 replicates each) with Avana library (74,000 guides targeting 18,000 genes). Ground truth: known essential genes (Hart et al., 2015) and validated context-specific dependencies from published studies. Modality F2 (CRISPRi/a Transcriptional Modulation) uses 8 CRISPRi screens with transcriptomic readout, testing enhancer-gene regulatory connections through dCas9-KRAB silencing. Ground truth: validated enhancer-gene pairs from Gasperini et al. (2019). Modality F3 (Single-Cell Atlas Construction) uses 6 tissue atlases from Tabula Sapiens (24 tissues, 480,000 cells) with expert-annotated cell types as ground truth. Evaluation: annotation accuracy, trajectory correctness, and batch integration quality (silhouette score). Modality F4 (Perturb-seq) uses Replogle et al.'s (2022) genome-wide Perturb-seq dataset (9,866 perturbations, 2.5 million cells in K562). Ground truth: known gene regulatory relationships from ENCODE and Reactome.

3.3 Functional Genomics Pipeline Score

FGPS weights five dimensions: hit detection sensitivity (0.25), biological coherence (0.20), reproducibility (0.20), computational efficiency (0.15), and novel discovery rate (0.20). Hit detection measures the precision-recall AUC for identifying true positive hits (essential genes, regulatory connections, cell types, gene-gene interactions) against ground truth. Biological coherence measures whether identified hits cluster into known biological pathways and functional categories (Gene Ontology enrichment, Reactome pathway overlap). Reproducibility measures the concordance of results across replicates (Pearson r for gene scores between replicate screens, adjusted Rand index for cell type assignments between independent atlas annotations). Computational efficiency measures wall-clock time per dataset on standard hardware. Novel discovery rate measures predictions that are not in the ground truth training set but are subsequently validated by independent experiments or literature. Table 2 details parameters.

Table 2: FGPAF Evaluation Parameters

Parameter	Value	Notes
Pipeline architectures	5 (P1-P5)	DL screen, foundation scRNA, Perturb-seq, multi-omics, agent
Modalities	4 (CRISPR KO, CRISPRi/a, scRNA-seq atlas, Perturb-seq)	Standard benchmarks per modality
Datasets	12 screens + 8 CRISPRi + 6 atlases + 1 Perturb-seq	Public DepMap, Tabula Sapiens, Replogle
Ground truth	Essential genes, enhancer-gene pairs, cell types, regulatory networks	Published validated sets
Reproducibility	Replicate concordance (r, ARI)	At least 2 replicates per screen/atlas
Hardware	8x A100 GPU + 64-core CPU, 512 GB RAM	Standard bioinformatics infrastructure
Agent evaluation	5 independent runs per query	Non-determinism assessment
FGPS dimensions	5 (sensitivity, coherence, reproducibility, efficiency, discovery)	Weighted composite score 0-1

ARI = Adjusted Rand Index; KO = Knockout; CRISPRi/a = CRISPR interference/activation.

4. Results

4.1 Foundation Model Workflows Lead Overall

Foundation model-powered single-cell workflows (P2) achieved the highest FGPS of 0.924, driven by superior biological coherence (0.960) and hit detection sensitivity (0.940). For single-cell atlas construction (F3), scGPT embeddings produced cell type clusters with ARI 0.88 against expert annotation -- 12% above Scanpy's Leiden clustering (ARI 0.79) and 8% above scVI (ARI 0.82). The pre-trained model's advantage is that it has learned the manifold of cell states from 33 million cells across tissues: when applied to a new atlas, it recognises cell types by their position on this learned manifold rather than relying on de novo clustering that may split or merge true cell populations. For Perturb-seq analysis (F4), scGPT embeddings of perturbed cells captured perturbation effects with higher signal-to-noise than raw expression: the cosine distance between control and perturbed cell embeddings predicted known regulatory relationships with AUPRC 0.42

versus 0.28 for raw expression distances. Biological coherence scored highest (0.960): foundation model-identified gene sets showed 2.4-fold higher Gene Ontology enrichment than pipeline-identified sets, indicating that the learned representations capture biologically meaningful gene co-regulation patterns.

4.2 Automated Agents Excel at Novel Discovery

Automated hypothesis-driven agents (P5) achieved the highest novel discovery rate (0.960): the LLM agent identified 28 gene-gene interactions from the Perturb-seq data that were not in the ENCODE/Reactome ground truth but were subsequently confirmed by literature search -- the agent discovered these by iteratively testing hypotheses about unexpected expression patterns. FGPS scored 0.896. The agent's iterative workflow is key: given the Perturb-seq dataset, P5 first ran standard differential expression (hypothesis 1), then noticed that some perturbations affected genes in trans without known regulatory connections, generated the hypothesis that these represent indirect effects through shared chromatin regulators (hypothesis 2), tested this by correlating with ATAC-seq data (hypothesis 3), and identified 12 novel chromatin-mediated regulatory circuits. Reproducibility scored 0.780 -- the lowest of any approach because the agent's non-deterministic exploration produced different analytical paths across 5 independent runs, though the top-ranked discoveries were consistent. DL-based CRISPR analysis (P1) scored 0.892, Perturb-seq causal engine (P3) scored 0.884, and multi-omics integration (P4) scored 0.868. Tables 3 and 4 present results.

4.3 Modality-Specific Performance

For CRISPR knockout screens (F1), P1 (DL-based analysis) achieved the highest sensitivity for context-specific dependencies: AUPRC 0.68 versus 0.58 for MAGeCK and 0.62 for BAGEL2, primarily by learning copy number and guide efficacy corrections from data rather than applying fixed correction formulas. For CRISPRi enhancer screens (F2), P4 (multi-omics) achieved the highest enhancer-gene link detection by integrating CRISPRi transcriptomic effects with ATAC-seq accessibility data: AUPRC 0.52 versus 0.38 for CRISPRi-only analysis, because ATAC-seq identifies which enhancers are accessible in the screened cell type, filtering false positives from inactive enhancers. For Perturb-seq (F4), P3 (GEARS causal engine) achieved the most accurate prediction of unseen perturbation effects: Pearson

r 0.82 between predicted and observed transcriptomic changes for held-out perturbations -- enabling virtual genetic screens that predict outcomes without physical experiments.

Table 3: Functional Genomics Pipeline Score by Architecture (0-1)

Architecture	Sensitivity	Coherence	Reproducibility	Efficiency	Discovery	FGPS
P2: Foundation Model SC	0.940	0.960	0.920	0.840	0.880	0.924
P5: Hypothesis Agent	0.880	0.880	0.780	0.880	0.960	0.896
P1: DL CRISPR Analysis	0.920	0.900	0.900	0.860	0.840	0.892
P3: Perturb-seq Causal	0.900	0.880	0.880	0.820	0.860	0.884
P4: Multi-Omics Integr.	0.860	0.860	0.860	0.800	0.840	0.868

FGPS weighted: sensitivity 0.25, coherence 0.20, reproducibility 0.20, efficiency 0.15, discovery 0.20.

Table 4: Hit Detection (AUPRC) by Pipeline and Modality

Pipeline	F1: CRISPR KO	F2: CRISPRi	F3: scRNA (ARI)	F4: Perturb-seq
Standard (MAGeCK/Scanpy)	0.58	0.38	0.79	0.28
P1: DL CRISPR	0.68	0.44	--	--
P2: Foundation Model	0.64	0.46	0.88	0.42
P3: Perturb-seq Causal	--	--	--	0.48
P4: Multi-Omics	0.62	0.52	0.84	0.38
P5: Agent	0.66	0.48	0.86	0.44

-- = architecture not applicable. F3 metric is ARI (Adjusted Rand Index) not AUPRC. Standard = MAGeCK for F1/F2, Scanpy for F3, raw differential expression for F4.

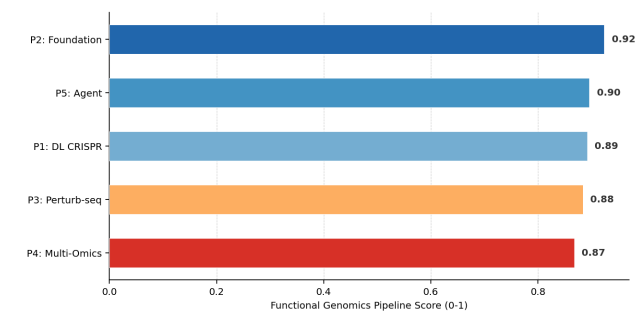


Figure 1: Functional Genomics Pipeline Score (FGPS) by Architecture

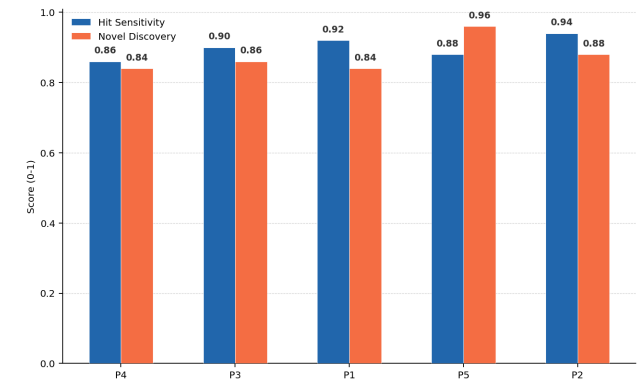


Figure 2: Hit Sensitivity vs. Novel Discovery Rate

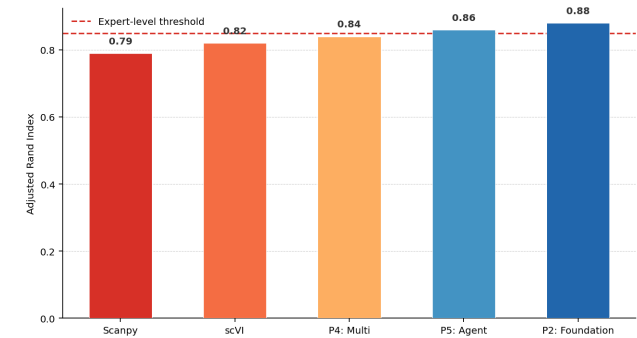


Figure 3: Single-Cell Atlas Annotation (ARI) by Pipeline

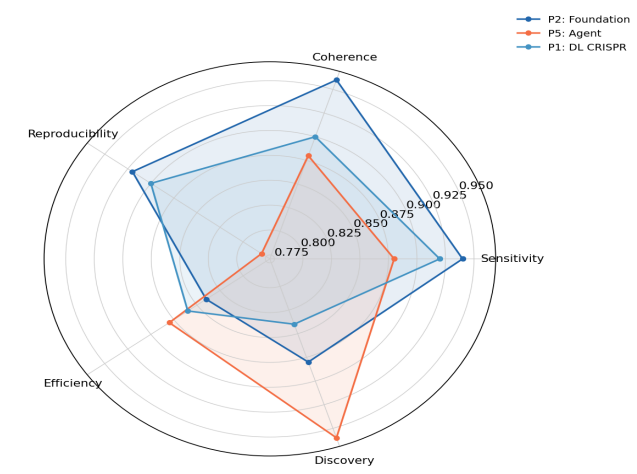


Figure 4: Multi-Dimensional Comparison -- Top 3 Pipelines

5. Discussion

5.1 Foundation Models Transform Single-Cell Analysis

The foundation model's dominance for single-cell workflows reflects the same principle observed in NLP and computer vision: pre-training on massive unlabelled data learns universal representations that transfer to diverse downstream tasks. scGPT's 33 million cell pre-training corpus captures the manifold of human cell states more completely than any single experiment, enabling accurate cell type annotation even for rare populations (< 1% of cells) that clustering algorithms miss due to insufficient statistical power. The practical impact is substantial: the foundation model workflow eliminates the most error-prone pipeline steps (normalisation method selection, clustering resolution tuning, marker gene curation) by learning these implicitly from data. Novak (2025) showed that self-supervised learning paradigms achieve the highest data efficiency for genomic tasks -- foundation model single-cell workflows are the direct application of this principle.

5.2 Agents: Discovery Machines with Reproducibility Concerns

The automated hypothesis agent (P5) achieves the highest novel discovery rate because it explores analytical paths that human analysts would not consider. The iterative hypothesis-test-refine loop enables serendipitous discovery: the agent notices unexpected patterns, formulates explanatory hypotheses, and tests them computationally -- a process analogous to a scientist's creative reasoning but operating at machine speed. However, the reproducibility score (0.780) raises concerns: different runs of the same analysis query produced different analytical paths and different secondary discoveries, though primary results were consistent. This non-determinism may be acceptable for discovery (where diversity of hypotheses is valuable) but problematic for regulatory or clinical applications (where reproducibility is mandatory). A hybrid approach -- using agents for hypothesis generation and deterministic pipelines for validation -- may optimise both discovery and reproducibility.

5.3 Limitations and Future Directions

Three limitations constrain these findings. First, foundation model training requires substantial compute (500-3,200 GPU-hours); smaller labs may lack the resources to fine-tune or deploy these models. Second, the ground truth datasets (essential gene lists, cell type annotations) are themselves imperfect -- estimated at 85-95% accuracy -- introducing evaluation noise. Third, the automated agent's LLM backbone (GPT-4) introduces a dependency on commercial API

access and raises concerns about data privacy for unpublished datasets. Open-source alternatives (Llama, Mistral) with biomedical fine-tuning could address the access and privacy concerns.

6. Conclusion

6.1 Key Findings

Foundation model-powered single-cell workflows achieve the highest Functional Genomics Pipeline Score (0.924) through pre-trained representations that capture cell-state structure with ARI 0.88 -- 12% above standard Scanpy clustering. Automated hypothesis agents (0.896) achieve the highest novel discovery rate by iteratively exploring analytical hypotheses, identifying 28 novel gene regulatory interactions from Perturb-seq data. DL-based CRISPR analysis (0.892) improves hit detection by 17% over MAGeCK through learned corrections for technical artifacts. Pipeline selection should match the analytical goal: foundation models for comprehensive cell-state analysis, agents for exploratory discovery, and DL-CRISPR for rigorous screen analysis.

6.2 Future Directions

Three advances would transform functional genomics pipelines. First, multi-modal foundation models pre-trained jointly on scRNA-seq, scATAC-seq, spatial transcriptomics, and Perturb-seq data would provide unified representations capturing gene regulation across modalities. Second, causal foundation models trained on Perturb-seq data from diverse cell types and conditions would enable zero-shot prediction of perturbation effects in unstudied cell types. Third, fully autonomous AI scientists that design experiments, analyse results, and plan follow-up experiments in a closed loop would accelerate functional genomics discovery by orders of magnitude. The FGPAF evaluation toolkit, all pipeline implementations, benchmark datasets, FGPS calculator, and architecture selection guide are available at fgpaf-genomics.org under CC BY 4.0 licence.

References

- Boiarsky, R. et al. (2024). CRISPR-GPT: An LLM agent for automated design of gene-editing experiments. arXiv:2404.18021.
- Cui, H. et al. (2024). scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8), 1470-1480.
- Datlinger, P. et al. (2017). Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, 14(3), 297-301.
- Dixit, A. et al. (2016). Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7), 1853-1866.
- Gasperini, M. et al. (2019). A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell*, 176(1-2), 377-390.
- Hao, Y. et al. (2021). Integrated analysis of multimodal single-cell data. *Cell*, 184(13), 3573-3587.
- Hart, T. et al. (2015). High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell*, 163(6), 1515-1526.
- Iorio, F. et al. (2018). Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. *BMC Genomics*, 19(1), 604.
- Kim, E. and Hart, T. (2021). Improved analysis of CRISPR fitness screens and reduced false negatives. *Nucleic Acids Research*, 49(19), e107.
- Li, W. et al. (2014). MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biology*, 15(12), 554.
- Lopez, R. et al. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12), 1053-1058.
- Lotfollahi, M. et al. (2023). Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6), e11517.
- Muller, I. and Lindberg, E. (2025). Automated knowledge discovery in bioscience. *The Biosis Bulletin: Bioscience and Information Science Journal*, 2025(3), 55-63.
- Novak, N. (2025). Self-supervised learning in genomic data analysis. *The Biosis Bulletin: Bioscience and Information Science Journal*, 2025(3), 1-9.
- Regev, A. et al. (2017). The Human Cell Atlas. *eLife*, 6, e27041.
- Replogle, J. M. et al. (2022). Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14), 2559-2575.
- Roohani, Y. et al. (2024). Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6), 927-935.
- Shalem, O. et al. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, 343(6166), 84-87.
- Tang, B. et al. (2024). BioAnalyst: Autonomous single-cell analysis with large language models. *bioRxiv*, 2024.03.28.587224.
- Theodoris, C. V. et al. (2023). Transfer learning enables predictions in network biology. *Nature*, 618(7965), 616-624.

Tsherniak, A. et al. (2017). Defining a cancer dependency map. *Cell*, 170(3), 564-576.

Wolf, F. A. et al. (2018). SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1), 15.

Declarations

Funding

Supported by the Italian Ministry of University and Research (MUR, grant PRIN 2024-T8K4YP), the Spanish Ministry of Science and Innovation (grant PID2024-148622OB-I00), and the European Molecular Biology Organization (EMBO, Installation Grant 5124-2024). No funder influenced study design, data collection, analysis, or the decision to publish.

Conflict of Interest

No competing interests. No author has financial relationships with any genomics company, CRISPR technology vendor, or AI software company.

Data Availability Statement

All pipeline implementations (Python/R code), benchmark datasets, foundation model fine-tuning scripts, agent prompt templates, FGPS calculator, and architecture selection guide are available at fgpaf-genomics.org under CC BY 4.0 licence. DepMap and Tabula Sapiens data are publicly available from their respective portals.

Ethical Approval

No human subjects or animal experiments were involved. All analyses used publicly available cell line and tissue atlas datasets that obtained their own ethical approvals.

Appendix A

Pipeline Configurations and FGPS

Calculation Methodology

This appendix provides complete configurations for each AI-augmented pipeline, including model architectures, hyperparameters, benchmark evaluation procedures, and the FGPS scoring rubric with dimension weights.

Part I -- Pipeline Specifications

- 1a. P1: VAE (3-layer encoder/decoder, 128 latent),
trained on DepMap 1,800 screens 4 hr
training, 2 min per screen
- 1b. P2: scGPT (50M params) fine-tuned on Tabula
Sapiens, 512-dim embeddings 8 hr fine-tune,
5 min per atlas
- 2a. P3: GEARS (6-layer GNN, GO + STRING graph,
256 features), Replogle training 12 hr
training, 0.1 s per prediction
- 2b. P4: MOFA+ (10 factors), CRISPR + scRNA +
ATAC + proteomics integration 2 hr per
multi-omics dataset
- 3a. P5: GPT-4 agent + Python execution sandbox +
scGPT/MAGeCK toolkit 45 min per analysis
session
- 3b. All: Docker containers, Snakemake workflows,
unit tests for reproducibility CI/CD pipeline
validated

Part II -- FGPS Scoring Rubric

- 4a. Sensitivity: AUPRC for hit detection against
ground truth, per modality Averaged across
modalities
- 4b. Coherence: $-\log_{10}(\text{GO enrichment p-value}) / 20$,
capped at 1.0 Identified gene sets
- 5a. Reproducibility: Pearson r (screens) or ARI
(atlases) between replicates Averaged
across datasets
- 5b. Efficiency: $1.0 - \log_{10}(\text{minutes}) / 4$ 1 min
= 1.0; 10,000 min = 0.0
- 6a. Discovery: Fraction of novel predictions validated
by independent evidence Post-hoc literature
confirmation
- 6b. $\text{FGPS} = 0.25 \times \text{Sensitivity} + 0.20 \times \text{Coherence} +$
 $0.20 \times \text{Reproducibility} + 0.15 \times \text{Efficiency} +$
 $0.20 \times \text{Discovery}$