Deep generative AI models analyzing circulating orphan non-coding RNAs enable accurate detection of early-stage non-small cell lung cancer

Mehran Karimzadeh[1, *], Amir Momen-Roknabadi[1, *], Taylor B. Cavazos[1, *], Yuqi Fang[1], Nae-Chyun Chen[1], Michael Multhaup[1], Jennifer Yen[1], Jeremy Ku[1], Jieyang Wang[1], Xuan Zhao[1], Philip Murzynowski[1], Kathleen Wang[1], Rose Hanna[1], Alice Huang[1], Diana Corti[1], Dang Nguyen[1], Ti Lam[1], Seda Kilinc[1], Patrick Arensdorf[1], Kimberly H. Chau[1], Anna Hartwig[1], Lisa Fish[1], Helen Li[1], Babak Behsaz[1], Olivier Elemento[2], James Zou[3], Fereydoun Hormozdiari[1, ★], Babak Alipanahi[1, ★], and Hani Goodarzi[4, 5, ★]

[1]Exai Bio Inc., Palo Alto, CA, US
[2]Weill Cornell Medicine, New York, NY
[3]Stanford University, Stanford, CA, US
[4]University of California, San Francisco, CA, US
[5]Arc Institute, Palo Alto, CA, US
★Lead contacts: hani@arcinstitute.org, babaka@exai.bio, and fereydounh@exai.bio
*These authors contributed equally

April 10, 2024

Abstract

Liquid biopsies have the potential to revolutionize cancer care through non-invasive early detection of tumors, when the disease can be more effectively managed and cured. Developing a robust liquid biopsy test requires collecting high-dimensional data from a large number of blood samples across heterogeneous groups of patients. We propose that the generative capability of variational auto-encoders enables learning a robust and generalizable signature of blood-based biomarkers that capture true biological signals while removing spurious confounders (e.g., library size, zero-inflation, and batch effects). In this study, we analyzed orphan non-coding RNAs (oncRNAs) from serum samples of 1,050 individuals diagnosed with non-small cell lung cancer (NSCLC) at various stages, as well as sex-, age-, and BMI-matched controls to evaluate the potential use of deep generative models. We demonstrated that our multi-task generative AI model, Orion, surpassed commonly used methods in both overall performance and generalizability to held-out datasets. Orion achieved an overall sensitivity of 92% (95% CI: 85%–97%) at 90% specificity for cancer detection across all stages, outperforming the sensitivity of other methods such as support vector machine (SVM) classifier, ElasticNet, or XGBoost on held-out validation datasets by more than ∼30%.

Keywords: oncRNA, deep generative AI models, liquid biopsy, lung cancer

Introduction

Lung cancer is the leading cause of cancer mortality in the US, accounting for about 1 in 5 of all cancer deaths (American Cancer Society, 2023). Each year, more people die of lung cancer than of colon, breast, and prostate cancers combined. Early detection of lung cancer improves the effectiveness of treatments and patient survival rates (National Lung Screening Trial Research Team et al., 2011) but

adherence to screening is often low (Lopez-Olivo et al., 2020). Nationally, only 23% of lung cancer cases are diagnosed in the early stages (I–III), when the five-year survival rate is 59%.

Previous attempts for detection of lung cancer through circulating tumor DNA (ctDNA)-based liquid biopsy assays have a low sensitivity (55%–57%) for early-stage disease, when treatments are most effective (Lebow et al., 2023; Cascone et al., 2023). While epigenomic assays have improved upon the overall sensitivity of mutation-based modalities by leveraging the cell-type specificity of DNA methylation (Schrag et al., 2023; Wang et al., 2023) or DNA fragmentation patterns (Mathios et al., 2021; Esfahani et al., 2022), sensitivity for early stage and small tumors remains low due to limited DNA shedding (Phallen et al., 2017).

Reorganization of the chromatin, as commonly observed in cancer cells (Corces et al., 2018), often results in the *de novo* access of the cellular transcriptional machinery to previously inaccessible genomic regions (Hu et al., 2022). Global disruptions in the RNA regulatory machinery in cancer (Perron et al., 2022) may also result in the appearance and stabilization of RNA fragments not commonly observed in normal tissues (Goodarzi et al., 2015). We recently reported the discovery of a class of previously unknown cancer-emergent small RNA (smRNA)s, termed orphan non-coding RNA (oncRNA)s, that arise as a consequence of cancer-specific genomic reprogramming (Fish et al., 2018). OncRNAs are abundant, stable, and actively secreted from living cancer cells into the blood (Wang et al.). We have generated a first-in-kind catalog of over 777,291 oncRNAs across major cancer types (Karimzadeh et al., 2023a). Some oncRNAs, such as *T3p*, exhibit pro-metastatic roles, while others could emerge as a byproduct of reprogrammed RNA metabolism. Contrary to DNA-based assays, oncRNAs do not require cellular death to be released. Active expression and secretion of oncRNAs allows for early detection of cancer subtype stratification in a liquid biopsy setting (Karimzadeh et al., 2023b; Goodarzi et al., 2022).

Since only a fraction of oncRNAs may be present in the volume of a blood draw, smRNA fingerprinting results in sparse patterns from thousands of individual oncRNAs species. Given the zero-inflated nature of oncRNA patterns, the underlying biological variation distinguishing different cancer types or separating cancer from non-cancer may become dominated by technical confounders, such as differences in sequencing depth, RNA extraction, sample processing, and other unknown sources of variation. In addition, often the sample collection process itself involves known sources of variation that should be accounted for, including biological differences between donors (age, sex, BMI, etc.). Therefore, developing a generalizable liquid biopsy assay requires effective strategies for modeling the biological properties of circulating biomarkers of interest and disentangling the technical and biological variation in sequencing data.

In recent years, various classes of neural networks have provided robust and customizable frameworks for guided representation learning. Deep generative models can leverage variational inference (Lopez et al., 2018) or pre-training on masked data (Cui et al., 2023; Chen and Zou, 2023; Rosen et al., 2023) to facilitate a variety of downstream tasks. Given the over-parameterized nature of these networks, a large number of samples is required for the adaptation of these models for clinical genomics applications. Furthermore, within the current framework of these models, explicit encoding of known technical variation (e.g. batch) is necessary, thus limiting the generalizability to new datasets. To overcome these challenges, we developed Orion, a two-arm semi-supervised multi-input variational auto-encoder for a liquid biopsy application using oncRNAs. We showcase the capability of Orion in learning a generalizable pattern of oncRNAs for a variety of applications, including early detection of lung cancer and removing batch effects in the presence of confounded signals.

Results

The liquid biopsy and approach for cancer detection proposed here is the first such effort for using newly annotated lung cancer-emergent and tumor-released oncRNAs as a signature for cancer detection

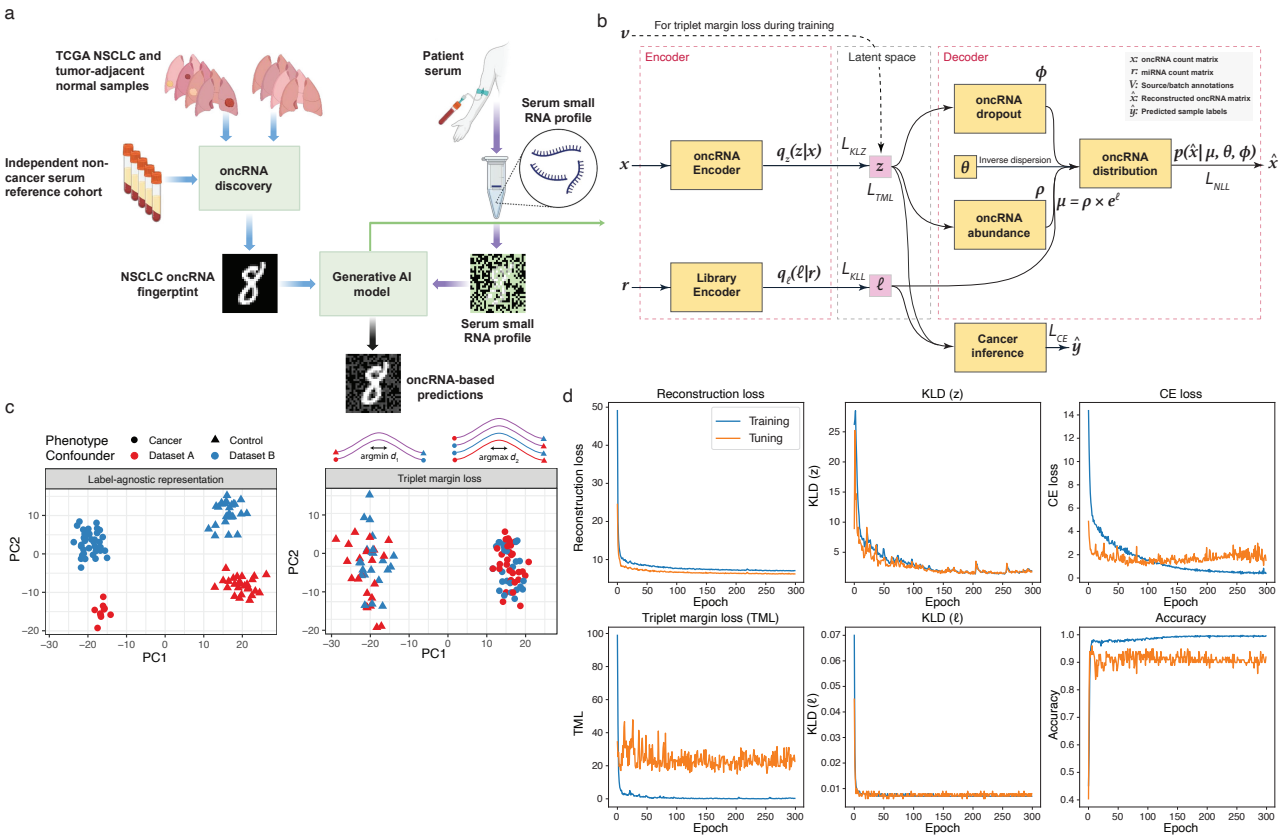


Figure 1: OncRNA-based liquid biopsy platform and Orion architecture. **(a)** We discovered Non-small cell lung cancer (NSCLC) oncRNAs from The Cancer Genome Atlas (TCGA) tissue datasets and investigated them in the blood of patients with NSCLC and non-cancer controls. We showed an analogy depicting NSCLC oncRNA fingerprint as a hand-written digit, serum oncRNA fingerprint as a noisy pattern, and generative AI embeddings as a denoised version. **(b)** Orion architecture requires two input count matrices for oncRNAs ($\boldsymbol{x}$) and endogenous expressed RNAs ($\boldsymbol{r}$). Each input is fed to a standard variational auto-encoder (VAE) where the objective is to learn a joint representation of oncRNA counts under a zero-inflated negative binomial distribution (right). A joint embedding will be used by the cancer inference neural network for classification tasks (bottom right). **(c)** Schematic of triplet margin loss application on simulated data. The left panel shows a label-agnostic embedding, and the right panel shows an embedding with a triplet margin loss constraint to minimize technical variations while preserving biological differences. For each sample, we use positive anchors (same phenotype, different dataset) and negative anchors (different phenotype, any dataset) to minimize or maximize the embedding distance, respectively. **(d)** Loss convergence plots show convergence of 5 of the losses of Orion as well as classification accuracy during training.

from blood. In this approach, using publicly available smRNA-seq data from TCGA (Hammerman et al., 2012; Cancer Genome Atlas Research Network, 2014), first, we discovered a set of oncRNAs; previously un-annotated scarce smRNAs that are selectively expressed in lung tumors versus normal lung tissues. Next, we used the expression of the selected oncRNA features in an in-house dataset of serum samples for cancer detection (Figure 1a, see Methods).

We then developed a deep generative AI model, Orion, for cancer detection using the abundance of cell-free oncRNAs in serum samples (Figure 1b). The proposed model is a generalizable approach that accounts for potential batch and vendor effects and other sources of expression variance that are not related to disease status. By removing these sources of noise, Orion improves the overall accuracy of cancer detection and is generalizable to unseen samples. At a high level, Orion uses variational inference to learn a Gaussian distribution from oncRNA data. We added several additional constraints through cross-entropy (CE) and triplet margin loss (see Methods) to emphasize the task-relevant information (e.g. cancer vs. control) while minimizing the task-irrelevant information (e.g. differences in library size or between sample sources) within the embedding space. A cancer inference neural network

then samples from this distribution to predict labels of interest including detection of cancer or tumor subtype. The model achieves these objectives by minimizing a negative log-likelihood loss based on zero-inflated negative binomial distribution to allow for the relative sparsity of biomarker measurements from the blood. We used 20% of the samples as a held-out validation set and the remaining samples for training within a 10-fold cross-validation setup.

Description of Datasets

NSCLC and tumor-adjacent normal smRNA dataset for oncRNA selection: We used the TCGA smRNA-seq database to identify 255,393 NSCLC-specific oncRNAs through differential expression analysis of NSCLC and non-cancerous tissues (see Methods).

smRNA data: We generated an in-house dataset of serum collected from 1,050 treatment-naive individuals (419 with NSCLC and 631 without a history of cancer). These samples are sourced from two different suppliers, where each supplier provided both cancer and control samples (Table 1, see Methods). We sequenced cell-free smRNA isolated from 0.5 mL of serum to quantify the expression of NSCLC-specific oncRNAs identified in the TCGA data (Figure 1a, see Methods). A total of 237,928 (93.15%) of the selected oncRNAs from tissue samples were detected in at least one of the samples.

Table 1: Sample demographics. Sample size and key demographic aspects of training set and held-out validation set.

Demographics		Training set		Validation set	
		Control	Cancer	Control	Cancer
Sample size	Count, n	506	334	125	85
Age	Mean (SD)	62.18 (11.75)	65.84 (9.60)	61.80 (10.80)	63.85 (10.35)
Sex	Female (%)	238 (47.04%)	125 (37.43%)	50 (40.00%)	40 (47.06%)
Smoking status	Never-Smoked, n (%)	271 (53.56%)	34 (10.18%)	71 (56.80%)	7 (8.24%)
BMI	BMI obese ($\geq$ 30), n (%)	124 (24.51%)	72 (21.56%)	28 (22.40%)	15 (17.65%)
Race	White, n (%)	253 (50.00%)	220 (65.87%)	62 (49.60%)	55 (64.71%)
	Black/African American, n (%)	54 (10.67%)	12 (3.59%)	14 (11.20%)	1 (1.18%)
	Asian, n (%)	15 (2.96%)	4 (1.20%)	3 (2.40%)	0 (0.00%)
	Other/Unknown, n (%)	184 (36.36%)	98 (29.34%)	46 (36.80%)	29 (34.12%)
Ethnicity	Hispanic, n (%)	179 (35.38%)	12 (3.59%)	46 (36.80%)	5 (5.88%)
	Non-hispanic, n (%)	281 (55.53%)	316 (94.61%)	59 (47.20%)	80 (94.12%)
	Other/Unknown, n (%)	45 (8.89%)	6 (1.80%)	19 (15.20%)	0 (0.00%)
Source	Indivumed, n (%)	183 (36.17%)	258 (77.25%)	46 (36.80%)	65 (76.47%)
	MT Group, n (%)	323 (63.83%)	76 (22.75%)	79 (63.20%)	20 (23.53%)

Orion model architecture

To distinguish cases from controls on the basis of their cell-free oncRNA content, we developed **Orion**; a customized, regularized, multi-input, and semi-supervised VAE (Figure 1b). As a VAE, Orion uses variational Bayes objectives to learn the parameters of a zero-inflated negative binomial distribution for expression of each oncRNA. This class of distribution accounts for over-dispersion and low sensitivity which are inherent to blood-based genomic and transcriptomic measurements (Supplementary Figure 1a–c). It has a two-arm architecture, modeling the expression of oncRNAs in one arm and the expression of annotated smRNAs in the other. The latter is used to account for differences in the size of sequencing libraries across samples. Orion also includes additional classification and contrastive learning objectives to accommodate label prediction and remove unwanted confounders in the learned representations (Figure 1b).

The semi-supervised nature of Orion allows its representation learning to capture the biological signal of interest (e.g. cancer detection) while removing unwanted confounders (such as batch effects). The generative capability of Orion during classifier training enables learning a robust pattern of

biomarkers for cancer detection. To ensure that the model learns a biologically grounded representation of the data irrespective of technical confounders, we used contrastive distance metric learning with a triplet margin loss (Figure 1b).

Orion learns a generalizable pattern of cancer-specific oncRNAs from the blood

To evaluate the capability of Orion in cancer detection and its generalizability, we divided our dataset into a held-out 20% and a remaining 80%. For 80% of the data, we trained Orion models in a non-overlapping 10-fold cross-validation setup. During each fold, we identified a subset of TCGA-derived oncRNAs that within the training set, were enriched among the cancer samples compared to control samples of each data source supplier, resulting in an average of 6,376 ± 60 (S.D) oncRNAs per fold. We trained 5 Orion models with different random seeds on each fold and averaged the scores on the test set.

The model achieved area under Receiver-operating characteristic curve (ROC) of 0.97 (95% CI 0.96–0.98) and overall sensitivity of 92% (95% CI 88%–95%) at 90% specificity (Figure 2a). In an identical setup with the same set of oncRNAs for each training fold, SVM classifier (Platt et al., 1999) had an area under ROC of 0.87 (95% CI 0.84–0.89) and overall sensitivity of 61% (95% CI 55%–66%). Other methods such as the commonly used ElasticNet (Zou and Hastie, 2005) model, XGBoost (Chen and Guestrin, 2016), and k-nearest neighbors (k-NN) classifier (Cover and Hart, 1967) also performed worse than Orion (Supplementary Table 1). More importantly, stage I sensitivity (n = 88) was 90% (95% CI 83%–94%) for Orion versus 56% (95% CI 47%–65%) for the SVM classifier at 90% specificity (Figure 2a). Sensitivity for later stages (II, III, and IV with n = 243) was 97% (95% CI 93%–99%) and 63% (95% CI 56%–70%) for Orion and the SVM classifier, respectively (Figure 2b). For detecting tumors smaller than 2 cm (T1a–b, n = 52), Orion achieved a sensitivity of 87% (95% CI 74%–94%) at 90% specificity, while the SVM classifier had a sensitivity of 44% (95% CI 30%–59%) at 90% specificity.

In a bootstrap analysis, AUC of Orion was significantly higher than both the SVM classifier (Δ_{AUC} = 0.1 (95% CI: 0.08–0.13)) and XGBoost (Δ_{AUC} = 0.03 (95% CI: 0.02–0.04), Supplementary Table 1). While AUC of Orion and XGBoost were relatively similar, F_1 score and sensitivity of Orion at 90% specificity were also better for Orion compared to XGBoost (Δ_{F_1} = 0.05 (95% CI 0.02–0.08), $\Delta_{\text{sensitivity}}$ = 9% (95% CI 5%–13%)).

To assess the generalizability of Orion, we chose the cutoff corresponding to 90% specificity among the 10-fold cross-validated predictions, and measured various classification metrics on the held-out validation set. Orion demonstrated a strong agreement in performance for the held-out validation set, while XGBoost, ElasticNet, and other model performances were on the lower bound of their 10-fold CV measurements (Figure 2d, Supplementary Table 1). For example, Orion had a consistent specificity of 90% (95% CI (84%–95%) and sensitivity of 92% (95% CI 86%–97%), while XGBoost had 100% specificity at the cost of a lower sensitivity of 33% (95% CI 23%–44%).

As a measure of successful batch effect removal, we expected the model scores for control samples to be similar, and therefore, not distinguish the sample suppliers. Orion had an area under ROC of 0.53 (95% CI 0.47–0.58), suggesting it successfully removed the impact of suppliers, while XGBoost and SVM classifier had higher area under ROCs of 0.59 (95% CI 0.54–0.64) and 0.57 (95% CI 0.52–0.62), respectively.

Given that the control samples in our cohort had an over-representation of individuals without smoking history compared to the cancer samples (54% vs. 10%), we examined the impact of smoking status of samples on model scores. We found that among control samples, Orion validation set score had an area under ROC of 0.6 (95% CI 0.5–0.7) with respect to presence of smoking history, further confirming little variation of the model score for individuals with or without a history of smoking.

To identify the most important oncRNAs for the model, we used SHapley Additive exPlanations (SHAP)(Lundberg and Lee, 2017) average values among model folds. Among the high-SHAP oncRNAs for the model, we observed overlap or vicinity of oncRNAs to some of the genes with significance in

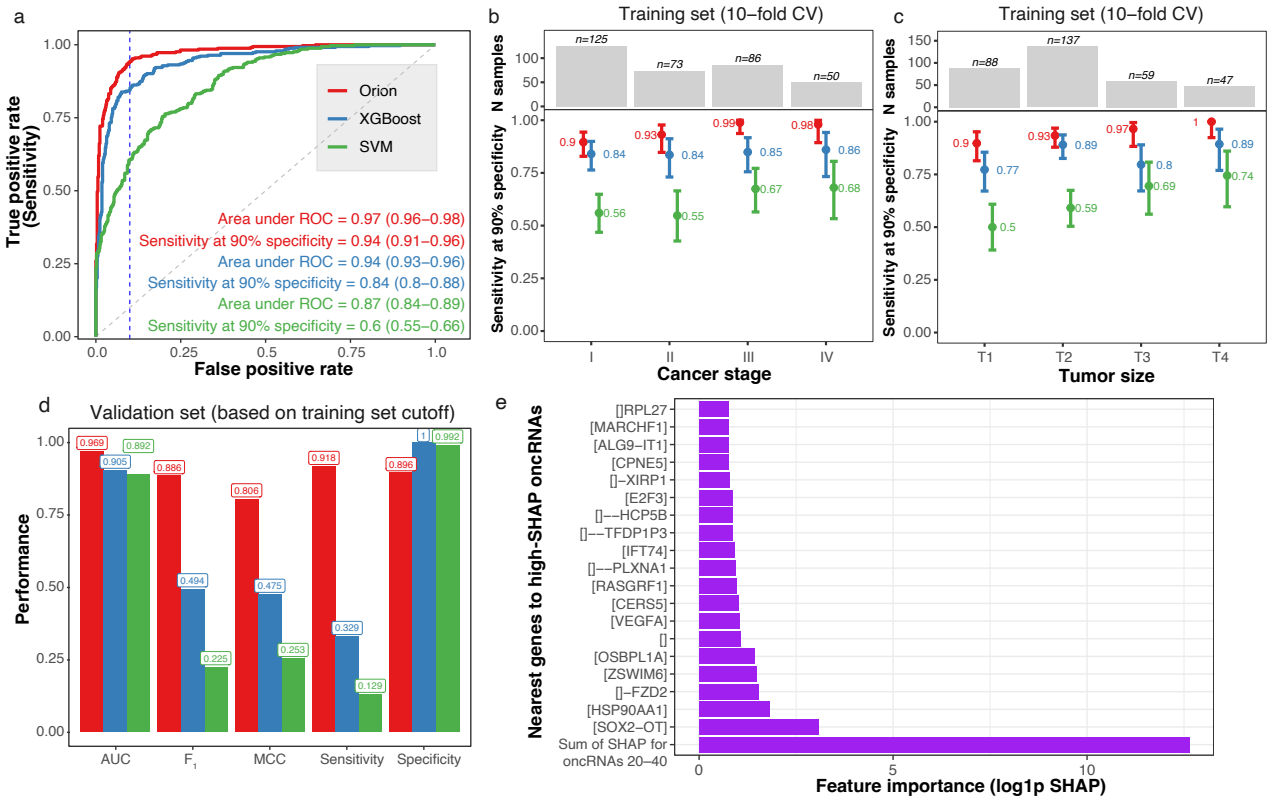


Figure 2: Model performance on training and validation set. (a) The ROC plot on the test set of 10 non-overlapping folds of model training for Orion (red), XGBoost (blue), and SVM classifier (green). The vertical blue line shows specificity at 90%. The text shows the area under ROC and sensitivity at 90% specificity with 95% confidence intervals. **(b)** Sensitivity of the model for tumors of different cancer stages at 90% specificity for Orion (red), XGBoost (blue), and SVM classifier (green). The bar plot shows the number of samples in each category. **(c)** Sensitivity of the model stratified by T score (size) similar to **(b)**. **(d)** Performance measures of binary classification in the held-out validation set. We computed all threshold-dependent metrics (all except area under ROC) based on the cutoff resulting in 90% specificity in the 10-fold cross validated training dataset. The bar height shows the point estimate of area under ROC, F_1 score, Matthew's correlation coefficient (MCC), sensitivity, and specificity. **(e)** Barplot shows log1p of SHAP score (x-axis) for the top 20 oncRNAs (y-axis). Y-axis labels indicate the nearest gene to the oncRNA. The first rows shows the sum of the next 20 oncRNAs (oncRNAs ranked 21st to 40th by their SHAP score). For gene A, [A] indicates overlap, []A indicates 1 kbp distance, []−A indicates 10 kbp distance, []−−A indicates 100 kbp, and [] indicates no genes within 1 Mbp distance.

lung cancer etiology and prognosis. These included SOX2-OT (Dodangeh et al., 2023), HSP90AA1, (Niu et al., 2022; Bhattacharyya et al., 2022), and FZD2 (Tuluhong et al., 2021) (Figure 2e).

To understand the model architecture components of Orion contributing most to high performance and limited batch detection, we performed a series of ablation experiments. We trained multiple models which lacked one or more of Orion's features, such as triplet margin loss, cross entropy loss, reconstruction loss, or generative sampling for computation of the cross entropy loss during training. We found that triplet margin loss allows the model to minimize the impact of the technical variations (Figure 3a). Generative sampling allows the model to achieve higher overall performance and better cross-entropy loss convergence (Figure 3b). Orion's embeddings in the presence of all of its components, particularly with triplet margin loss and generative sampling, result in a better separation of cancer samples from control samples, which allows Orion's classifier to operate on a representation of the data with minimal technical variations (Figure 3c). The presence of different components of Orion, particularly the reconstruction loss, result in a better convergence of the test-set cross entropy loss (Figure 3d).

We hypothesized that training the classifier of the model by sampling from the learned distribution

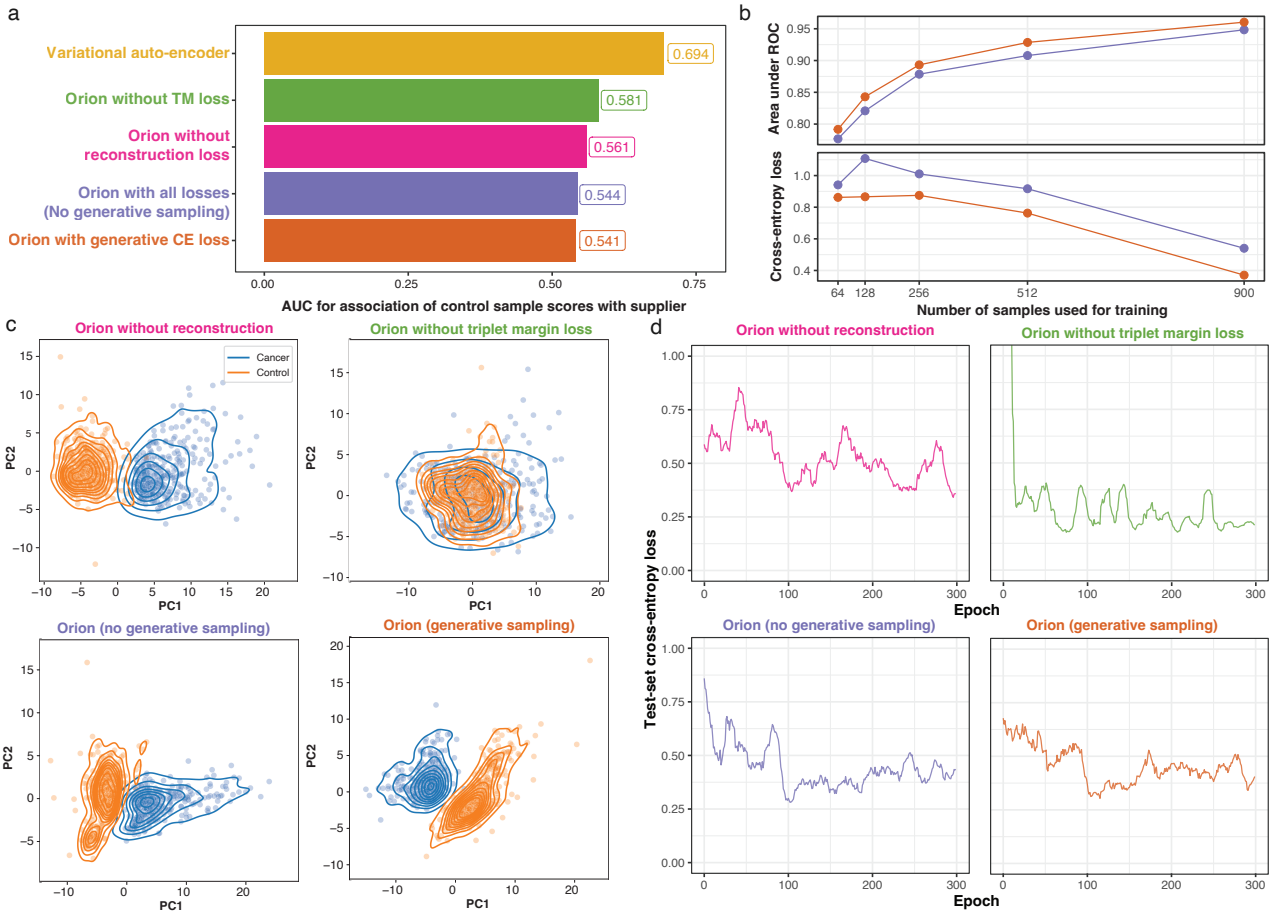


Figure 3: Ablation of Orion components. (a) Area under the ROC of 5 different models when comparing score of the control samples with respect to the sample supplier. **(b)** Area under ROC (top panel) and cross entropy loss (bottom panel) for cancer detection as a function of the number of samples used during training. Orange shows Orion with generative sampling for computation of cross-entropy loss during training, and purple shows Orion without this feature. **(c)** Scatter plots overlaid with kernel density estimates show cancer (blue) and control (orange) samples based on the first two principal components of Orion's embedding space in 4 different conditions. **(d)** Test-set cross entropy loss of the same models.

allows Orion to achieve higher robustness and performance at a smaller sample size. In comparison with an identical architecture where the classifier uses the expected value of the distribution instead of sampling, we observed a significant improvement in convergence and generalizability of the cross entropy loss with smaller sample sizes (Figure 3b–d).

To assess if Orion learns more informative task-relevant embeddings than commonly used methods such as Principal component analysis (PCA) (Pearson, 1901) or Harmony (Korsunsky et al., 2019), we examined how these embeddings compare in downstream tasks. We provided Harmony with the same variables for batch correction as Orion's triplet margin loss (sample supplier and experiment ID). While Orion's key clusters reflect cancer and control labels (Supplementary Figure 2, projected here in UMAP space solely for visualization), the naive representation of Harmony and PCA fail to capture this key biological variability. Next, we trained an XGboost model on the training set and evaluated the performance in cancer detection from the embeddings in the test set. Label-agnostic batch correction of Harmony resulted in loss of biological information and a worse performance than PCA, while Orion outperformed both PCA and Harmony with at least 30% higher sensitivity at 90% specificity (Supplementary Figure 2).

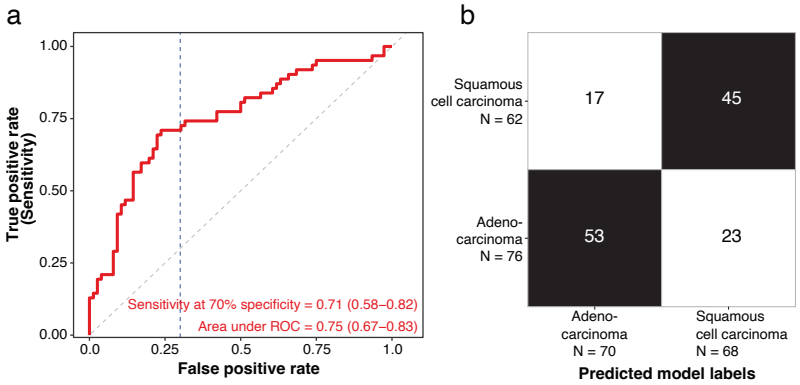


Figure 4: Orion allows distinguishing tumor subtypes from the oncRNA profiles of the blood. (a) ROC plot of Orion for distinguishing squamous cell carcinoma from adenocarcinoma among stage III/IV NSCLC samples. **(b)** Confusion matrix of Orion's subtype prediction at 70% specificity cutoff.

Orion can identify tumor subtype from circulating oncRNAs

In addition to the early detection of cancer signals in patients with NSCLC, understanding tumor histology has major implications in therapy selection and resistance mechanisms. Squamous cell carcinoma transformation of lung adenocarcinoma has been reported to take place after target therapy resistance. Squamous cell carcinoma transformation has been reported to be one of the mechanisms of acquired resistance to epidermal growth factor receptor (EGFR), various tyrosine kinase inhibitors (TKIs) (Park et al., 2019), KRAS inhibitors (Tong et al., 2024), immunotherapies (Hsu et al., 2017), and even spontaneously (Jiang et al., 2019). Traditional methods of stratifying patients to evaluate for squamous cell carcinoma transformation involve repeat biopsies of a lung cancer patient which can lead to severe side effects such as pneumothorax, hemorrhage, and air embolism (Vachani et al., 2022).

We had previously observed that given the tissue-specific landscape of chromatin accessibility in different cancers, oncRNA expression patterns are unique to cancer types and subtypes (Wang et al., 2022), allowing us to detect tissue of origin among different cancer types non-invasively from blood (Karimzadeh et al., 2023a). We hypothesized that biological differences of lung adenocarcinoma and squamous cell carcinoma would also be reflected in cell-free oncRNA content, allowing us to distinguish these major subtypes of NSCLC. While tumor tissues are vastly different from normal tissue, the differences in subtypes of a given tumor are far less substantial. In NSCLC, for example, the agreement of pathologists for different subtypes is approximated to be 0.81 (Stang et al., 2006). As a result, tumor histology subtype prediction is more difficult than cancer detection.

To evaluate our hypothesis, we investigated the potential of distinguishing two major NSCLC subtypes, adenocarcinoma and squamous cell carcinoma, using oncRNAs in blood. For this analysis, we used 20-fold cross-validation to adjust for the reduction in the number of samples given that this is a NSCLC-specific task. For later stage tumors (stages III/IV), Orion achieved an area under ROC of 0.75 (95% CI: 0.67–0.83) and a sensitivity of 71% (95% CI: 56%–84%) at 70% specificity in distinguishing squamous cell carcinoma from adenocarcinoma samples in serum samples (Figure 4).

Discussion

Variational inference serves as the backbone of a plethora of deep generative models, particularly for single-cell genomics applications (Lopez et al., 2018). The flexibility of these models allows for reference building through transfer learning (Lotfollahi et al., 2022) or modeling specific perturbations through contrastive learning (Weinberger et al., 2023). However, when biological signals are weak or scarce, as is the case in liquid biopsies where we are in search of a needle in the haystack, technical confounders that

are due to differences in sequencing platforms or data sources become more pronounced. As a result, without any intervention, the naive representation learning may regress out the signal of interest, as was the case with PCA and even the state-of-the art batch correction method, Harmony (Antonsson and Melsted, 2024) (Supplementary Figure 2). Representation learning, therefore, is rarely used for clinical genomics applications. Instead, classical regularized supervised learning methods (e.g. ElasticNet) are adopted, which are able to resolve the p (number of features) $>> n$ (number of samples) problem by finding an adequate balance between the number of features the model utilizes and the individual weight of each feature. While these methods have been extensively applied in clinical genomics and liquid biopsy, they fail to model non-linear interactions among the input features and the higher-order patterns in the data.

Here we sought to leverage representation learning for obtaining an abstract low-dimensional embedding of cell-free oncRNAs. We hypothesized that a deep generative AI model can augment the downstream classifier to learn robust and generalizable patterns of cancer-specific oncRNAs. This approach not only reduces the number of features by approximately 300 fold, but it can also enhance the number of unique samples the classifier is trained on through generative sampling, essentially converting $p \gg n$ to a favorable $n \gg p$. A key aspect to the success of our approach is tailoring the process of representation learning through the addition of contrastive learning (Ishfaq et al., 2018) (Figure 3). Inherently, these objectives are in contradiction, one enforcing the latent distribution to preserve all sources of data variation, while the other imposes a constraint to remove unwanted variations. As a result, these two objectives meet at the balancing minima of a sacrifice in reconstruction at the gain of emphasizing the biological differences among the samples.

Apart from the detection of early-stage lung cancers, another large unmet need is the lack of diagnostic tools with sufficient sensitivity to detect residual disease after surgery. The ability to detect minimum residual disease (MRD) is important in guiding risk stratification, tailoring adjuvant therapies, and preventing relapse. Surgery is considered the standard treatment with curative intent for early-stage NSCLC, whereas in locally advanced cases (stages IIIA and IIIB), neoadjuvant therapy (NAT) may be used to downstage the tumor prior to surgery. However, recurrence after resection even after NAT is common. Five-year survival rates of 68%–92% of stage I, 53%–60% of stage II, and 13%—36% of stage III NSCLC patients indicate significant risk of recurrence and death after surgery (Goldstraw et al., 2016).

Here we demonstrated the success of our approach in training a model that not only achieved superior performance for cancer detection, but also exhibited generalizability to held-out datasets. Contrary to other methods, Orion scores remained unchanged among samples coming from different sources or with different smoking histories, underscoring the robustness of our model. The performance of Orion for the prediction of tumor subtypes from the blood, despite the lack of clear ground truths in histopathological calls, represents a first step in addressing this task. Given that the pathologist agreement for this task is itself around 80% (Stang et al., 2006) and the observation that our model improved by increasing the number of samples in the training set (Figure 3b), a larger dataset with molecularly-assigned labels could provide an opportunity for liquid histology applications beyond cancer detection using Orion.

While the adaptation of deep learning models in clinical genomics is in its early days, our results establish a strong case for the potential of generative AI in advancing the applications of liquid biopsy, as well as liquid histology. The combination of our liquid biopsy platform for profiling a stable, abundant, and cancer-specific biomarker—oncRNAs—and our generative AI model which is compatible with blood-based measurements, provides a novel opportunity for filling a clinical gap in sensitive and early cancer detection and monitoring.

Competing interests

The authors are either employees, shareholders, or stock option holders of Exai Bio, Inc.

Methods

Dataset

Here, we used an in-house dataset of serum collected from 1,050 treatment-naive individuals sourced from two different suppliers: Indivumed (Hamburg, Germany; 229 controls and 323 NSCLC cases) and MT Group (Los Angeles, CA; 402 controls and 96 NSCLC cases). Each supplier also collects samples from multiple sites. The dataset included 157 stage I, 93 stage II, 106 stage III, and 63 stage IV NSCLC cases. We used RNA isolated from 0.5 mL of serum from each donor to generate and sequence smRNA libraries of each sample at an average depth of 19.8 ± 5.8 million 50-bp single-end reads. The NSCLC samples included 222 samples with adenocarcinoma, 160 samples with squamous cell carcinoma, and 37 samples with unknown histological type (Table 1). Despite the challenges of collecting samples from healthy seniors without smoking history, NSCLC and control arms included both smoker and non-smoker samples and similar distribution with respect to age, sex, and body mass index (BMI). Given the imbalance of individuals with smoking history among cases and controls, we observed that the Orion model score did not vary as a function of smoking history among control samples.

Orion architecture

Orion is a variational auto-encoder (Kingma and Welling, 2013), adapting scVI (Lopez et al., 2018) with additional input, connections, and objectives for removing known sources of technical variation as well as performing regression or classification tasks. Let $\boldsymbol{x}_i \in \mathbb{Z}_+^d$ and $\boldsymbol{r}_i \in \mathbb{Z}_+^m$ denote counts for d oncRNAs and m endogenous highly-expressed smRNAs for the i-th sample, respectively. Moreover, let $\boldsymbol{y}_i \in \{0,1\}^b \times \mathbb{R}^t$ and $\boldsymbol{v}_i \in \mathbb{Z}_+^c$ denote the b binary and t real targets (cancer status) and the c known confounders (sample source, processing batch, etc), respectively.

The core idea is that there are linear and nonlinear dependencies between different oncRNAs, e.g., they are generated due to disruption in the same pathway hence their counts are correlated. Therefore, we will be able to project the space of $\mathcal{X}$ — that can be very high-dimensional — onto a low-dimensional latent space $\mathcal{Z}$ using a mapping $f_z : \mathcal{X} \to \mathcal{Z}$ (called oncRNA encoder), while capturing the essence of variation in $\mathcal{X}$. This means that we could find a mapping $g : \mathcal{Z} \to \mathcal{X}$ (called decoder), such that $\hat{\boldsymbol{x}} = g(\boldsymbol{z}) = g(f(\boldsymbol{x}))$ is approximately the same as $\boldsymbol{x}$, e.g., $||\boldsymbol{x} - \hat{\boldsymbol{x}}||_2^2$ is small. In variational auto-encoders instead of deterministically mapping $\boldsymbol{x}$ to $\boldsymbol{z}$, we map $\boldsymbol{x}$ to a (usually Gaussian) distribution $q_{\boldsymbol{z}}(\boldsymbol{z}|\boldsymbol{x})$. When reconstructing $\boldsymbol{x}$, we sample from $\boldsymbol{z} \sim q_{\boldsymbol{z}}(\boldsymbol{z}|\boldsymbol{x})$ and using this sample, we can generate a distribution for reconstructed $\boldsymbol{x}$ as $\hat{\boldsymbol{x}} = p_{\boldsymbol{x}}(\boldsymbol{x}|\boldsymbol{z})$.

A common source of variation in transcriptomic data originates from the total sequenced RNA. An oncRNA might not be observed for two reasons: either it does not exist and is not secreted or it is indeed in blood but due to low-volume blood sampling or limited sequencing, it has not been picked up in the experiment. We assume that $\boldsymbol{z}$ will take care of the former, but for the latter effect, $\ell \in \mathbb{R}$ is another unobserved random variable that accounts for input RNA level and library sequencing depth. Here, since oncRNA counts $\boldsymbol{x}$ are usually small and unsuitable for computing library size — unlike scVI — we use a set of endogenous highly-expressed RNAs and an additional encoder $f_\ell : \mathcal{R} \to \ell$ to compute a normal distribution $q_\ell(\ell|\boldsymbol{r})$ as a proxy for the log of library size. In other words, the library size is log-normal with priors originating from the log of mean and variance of $\sum_m \boldsymbol{r}_i$ in a given min-batch. As a result, ℓ shows a strong correlation with the total number of oncRNA reads, even though it is not derived from oncRNAs (Supplementary Figure 1a–b).

Similar to gene counts across cells in single-cell RNA-seq data, any oncRNA is observed in only a few samples and its counts are mostly zeros, also called zero-inflated. We assume the non-zero counts follow a negative binomial distribution. Inspired by scVI (Lopez et al., 2018), we model the oncRNAs count as a conditional zero-inflated negative binomial (ZINB) distribution $p(\boldsymbol{x}|\boldsymbol{z},\ell)$, where $\boldsymbol{z} \in \mathbb{R}^k$, $k \ll d$ is the latent embedding of $\boldsymbol{x}$.

Orion decoders learn the zero-inflation parameter $\boldsymbol{\phi_i}$ through $f_\phi : \mathcal{Z} \to \boldsymbol{\phi}$ and the transcription scale parameter ρ_i through $f_\rho : \mathcal{Z} \to \boldsymbol{\rho}$. f_ρ involves a softmax step, enforcing representation of the expression of each oncRNA as a fraction of all expressed oncRNAs.

In the Gamma-Poisson representation of the negative binomial distribution, $\boldsymbol{\mu} = \boldsymbol{\rho_i} \times e^{\ell_i}$ will provide the shape parameter of the Gamma distribution, and input-independent learnable parameter $\boldsymbol{\theta}$ will represent the inverse dispersion.

In short, to train Orion:

1. We learn a low-dimensional Gaussian distributions $q_{\boldsymbol{z}}(\boldsymbol{z}|\boldsymbol{x})$ and $q_\ell(\ell|\boldsymbol{r})$, so that zero-inflated negative binomial distribution $q_{\boldsymbol{x}}(\boldsymbol{x}|\boldsymbol{z},\ell)$ has the generative capability of producing realistic *in silico* oncRNA profiles. To do so:

   (a) We minimize

   $$L_{\text{KLZ}} = D_{\text{KL}}\left(q_{\boldsymbol{z}}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right),$$

   where D_{KL} is the Kullback-Leibler divergence (Kullback and Leibler, 1951) and $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0}, I)$ is the prior distribution for $\boldsymbol{z}$.

   (b) We minimize

   $$L_{\text{KLL}} = D_{\text{KL}}\left(q_\ell(\ell|\boldsymbol{r})||p(\ell|\boldsymbol{r})\right),$$

   where $p(\ell|\boldsymbol{r})$ is the prior log-normal distribution for ℓ. Unlike $\boldsymbol{z}$, the prior distribution for ℓ is different from batch to batch and its log-mean and log-standard deviation are computed based on values of $\boldsymbol{r}$ in each mini-batch $\mathcal{B}$.

2. We minimize the reconstruction loss by minimizing the negative log-likelihood of a zero-inflated negative binomial distribution describing the distribution of the input oncRNA data:

   $$L_{\text{NLL}} = -\sum_i \log p_{\boldsymbol{x}}(\boldsymbol{x}_i|\boldsymbol{\mu}_i, \boldsymbol{\theta}_i, \boldsymbol{\phi}_i),$$

   where $\boldsymbol{\mu}_i$ is the product of the softmax of f_ρ (representing transcription scale of each oncRNA) and e^{ℓ_i}; and $\boldsymbol{\theta}_i$, $\boldsymbol{\phi}_i$ represent inverse dispersion and zero-inflation probability (Lopez et al., 2018), respectively (Figure 1b).

3. We use contrastive learning (triplet margin loss) to minimize the impact of known confounders $\boldsymbol{v}$ on $\boldsymbol{z}$. For example, this ensures that all the cancer samples from different sources are projected in proximity of each other (see Triplet Margin Loss section).

   (a) Minimize the distance between samples that have the same label (e.g. all cancer samples or all control samples) but are from a different confounder group (e.g. source, supplier, etc.) in the oncRNA embedding space $\boldsymbol{z}$

   (b) Maximize the distance between samples that have different labels.

$$L_{\text{TML}} = \frac{1}{w \times c} \sum_{i \in \mathcal{B}} \sum_{(i,j,j') \in \mathcal{T}_i} \max(||\boldsymbol{z}_i - \boldsymbol{z}_j||_2^2 - ||\boldsymbol{z}_i - \boldsymbol{z}_{j'}||_2^2 + \alpha, 0),$$

4. We use supervised learning such that the low-dimensional embeddings $\boldsymbol{z}$ are used for regression (smooth L_1-loss (Girshick, 2015)). For classification, we minimized the cross-entropy loss L_{CE} to predict the provided sample labels during training (e.g. cancer vs. control)

We minimize the summation of these 5 losses with weights as hyperparameters:

$$L_{\text{Orion}} = \lambda_1 L_{\text{KLZ}} + L_2 L_{\text{KLL}} + \lambda_3 L_{\text{NLL}} + L_4 L_{\text{TML}} + \lambda_5 L_{\text{CE}}$$

Triplet Margin Loss

For each sample i, we sample ω triplets for each confounder $\boldsymbol{v}_i^c$ as follows:

1. Randomly pick a "positive" anchor $j \neq i$ such that they share the same classification label $\boldsymbol{y}_i = \boldsymbol{y}_j$, but do not share the same confounder $\boldsymbol{v}_i^c \neq \boldsymbol{v}_j^c$.
2. Randomly pick a "negative" anchor $j' \neq i$ such that they do not share the same classification label $\boldsymbol{y}_i \neq \boldsymbol{y}_{j'}$.
3. Add (i, j, j') to $\mathcal{T}_i$, the set of triplets for i.

At the end of this process, each sample will have $|\mathcal{T}_i| = \omega \times c$ triplets picked for it, where ω is a hyperparameter set to 16.

During training we add a cost function that moves samples from different sources or processing batches that share the same label (e.g. cancer samples from different sources) closer to each other, while moving samples with different labels (e.g. cancer samples from non-cancer samples) further apart:

$$L_{\text{TML}} = \frac{1}{w \times c} \sum_{i} \sum_{(i,j,j') \in \mathcal{T}_i} \max(||\boldsymbol{z}_i - \boldsymbol{z}_j||_2^2 - ||\boldsymbol{z}_i - \boldsymbol{z}_{j'}||_2^2 + \alpha, 0),$$

where α is a hyperparameter that enforces what should be the minimum difference of distances between a sample and its positive and negative anchors in the latent space, and it is set to $\alpha = 1$.

Model Parameters

On its default mode used in this study, Orion has 1 hidden layer for encoding oncRNAs with 1,500 hidden units, 1 hidden layer for encoding library size from endogenous RNAs with 1,500 unit, an embedding space of $d = 50$ latent variables for learning the Gaussian distribution underlying the oncRNA data, an embedding space of $s = 1$ latent variable for learning the library size distribution from endogenous RNAs, and one hidden layer for decoding oncRNA data from the latent distribution. We used dropout ($p = 0.5$), L_2 regularization (L_2= 2). The classification layer has 1 hidden layer of size 25, mapping the 50 normalized latent values to generative predictions for each class.

Orion encoders have a hidden layer of size 1,500 and map X to parameters of $\boldsymbol{z}_d$ with 50 dimensions and map Q to parameters of $\boldsymbol{z}_s$ with 1 dimension.

The model performs classification through a 2-layer perceptron head. The input of the classification head comes from the batch-normalized product of oncRNAs and library size embeddings, i.e., $\boldsymbol{z} \times \ell$. uring training, we sample from $q_{\boldsymbol{z}}(\boldsymbol{z}|\boldsymbol{x})$ $\eta = 100$ times for each data point to improve model robustness and sensitivity to noise. At test time, we use the deterministic expected values of $\boldsymbol{z}$ and ℓ.

Identifying oncRNAs

To identify a set of orphan non-coding RNAs, we utilized smRNA-sequencing data from 10,403 tumor and 679 adjacent normal tissue samples from TCGA spanning 32 unique tissue types. Quality control was applied to the GRCh38-aligned BAM files to remove reads that were < 15 base pairs or were considered low complexity based on a DUST score > 2 (Schmieder and Edwards, 2011). Additionally, we removed reads that mapped to chrUn, chrMT, or other non-human transcripts. After filtering, we identified *de novo* smRNA loci by merging all reads across the 11,082 TCGA samples and performing peak calling on the genomic coverage to identify a set of smRNA loci that were < 200 base pairs. This resulted in 74 million distinct candidate loci for feature discovery.

For discovery of lung tumor-specific oncRNAs, we restricted to lung tumors ($n = 999$) and all adjacent normals ($n = 679$) and filtered the candidate loci for those that appeared in at least 1% of samples resulting in 1,293,892 smRNAs. We then used a generalized linear regression model to identify those smRNAs that were significantly more abundant in lung tumors compared to normal tissues. Our model adjusted for age, sex, and principal components to capture the global smRNA expression variability across tissues and batches. After multi-testing correction we restricted to suggestively significant smRNA features (FDR $q < 0.1$) that were enriched in lung tumors (OR > 1) resulting in ∼260k lung-tumor associated oncRNAs for downstream applications in serum.

Training and evaluation strategy

Our dataset included a total of 1,257 samples obtained from 1,050 patients, with 183 samples having been sequenced more than once. We used 20% of the patients as a held-out validation set, ensuring an equal representation of suppliers, histological subtype (adenocarcinoma and squamous cell carcinoma), and patient cohort (NSCLC or control) among the training and held-out validation sets.

Within the training set, we used a similarly stratified 10-fold cross-validation to select the oncRNAs and train the model on the training set. Each data split ensured samples of the same patient were either in the training or test splits. We reported the performance measures only for one sample of each patient. We train 5 models per fold, each trained with a different random seed. The score of the test set of each fold was averaged over these 5 models. The training set performance measures are based on the held-out set of each fold. For the held-out validation set, we use the average of the 50 models (5 models for each of the 10 folds). We defined the model cutoff based on the cross-validated scores of the training set and reported the performance for the held-out validation set using that cutoff.

Feature selection

We used the The Cancer Genome Atlas (TCGA) smRNA-seq database to identify 255,393 NSCLC-specific oncRNAs as previously described (Karimzadeh et al., 2023c). Each tissue sample expressed a mean of 37,115 ± 14,457 S.D. of these oncRNAs. After processing serum samples for the present study, 237,928 (93.16%) of these oncRNAs were detected in at least one sample.

Within each fold of the training set, we identified oncRNAs present in at least 2% of the training set samples provided by each supplier. Additionally, for training set samples of each supplier, we identified oncRNAs that were over-represented in the cancer samples (log odds ratio > 0). Within each training fold, we selected oncRNAs passing these criteria in both of the suppliers (MT Group and Indivumed). Among the features passing these criteria, we performed 8 rounds of XGBoost classification within the training set, each time setting aside oncRNAs with non-zero Gini impurity index as a measure of feature importance. This resulted in obtaining an average of 6,376 ± 60 oncRNAs in each model fold and a total of 14,014 oncRNAs identified in at least 1 fold.

Benchmarks

Training other models

We used normalized oncRNA counts by dividing $\boldsymbol{x}_i$ by the total number of highly-expressed small RNA reads $\boldsymbol{r}_i$ as a surrogate of the the sequencing library depth:

$$\frac{1,000 \times \boldsymbol{x}_i}{\sum_{m=1}^{m} \boldsymbol{r}_{i,m}},$$

where $\boldsymbol{r}_{i,m}$ is the counts of m-th smRNA for sample i. We used scikit-learn's `StandardScaler` on the training set of each fold, and applied it on test-set or held-out set for utilizing the model. We used scikit-learn's `LogisticRegressionCV` to identify the best set of hyperparameters in a 2-fold cross-validation setup within the training set. The hyperparameters included L_1 ratios `[0, .1, .5, .7, .9, .95, .99, 1]` and the default C parameters. The best hyperparameters were provided to a scikit-learn `LogisticRegression` model for training on the entire training set. ElasticNet models used identical oncRNAs and samples as Orion. For other models including XGBoost, SVM classifier, and k-NN, we used the default parameters.

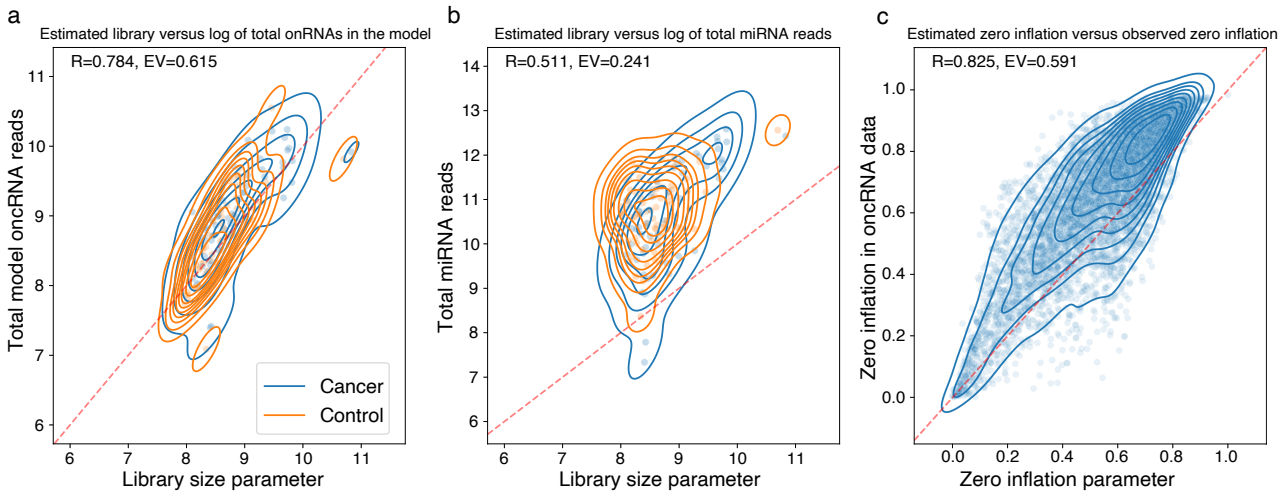
Embedding benchmarks

In this study, Orion has an embedding space with a multi-variable Gaussian with a dimension of 50. We used 50 principal components from the same oncRNAs (scaled to total miRNA content). We fed the PCA matrix to harmony, specifying sample source and experiment ID as `batch_key` parameter. These are the same variables that we used to guide triplet margin loss.

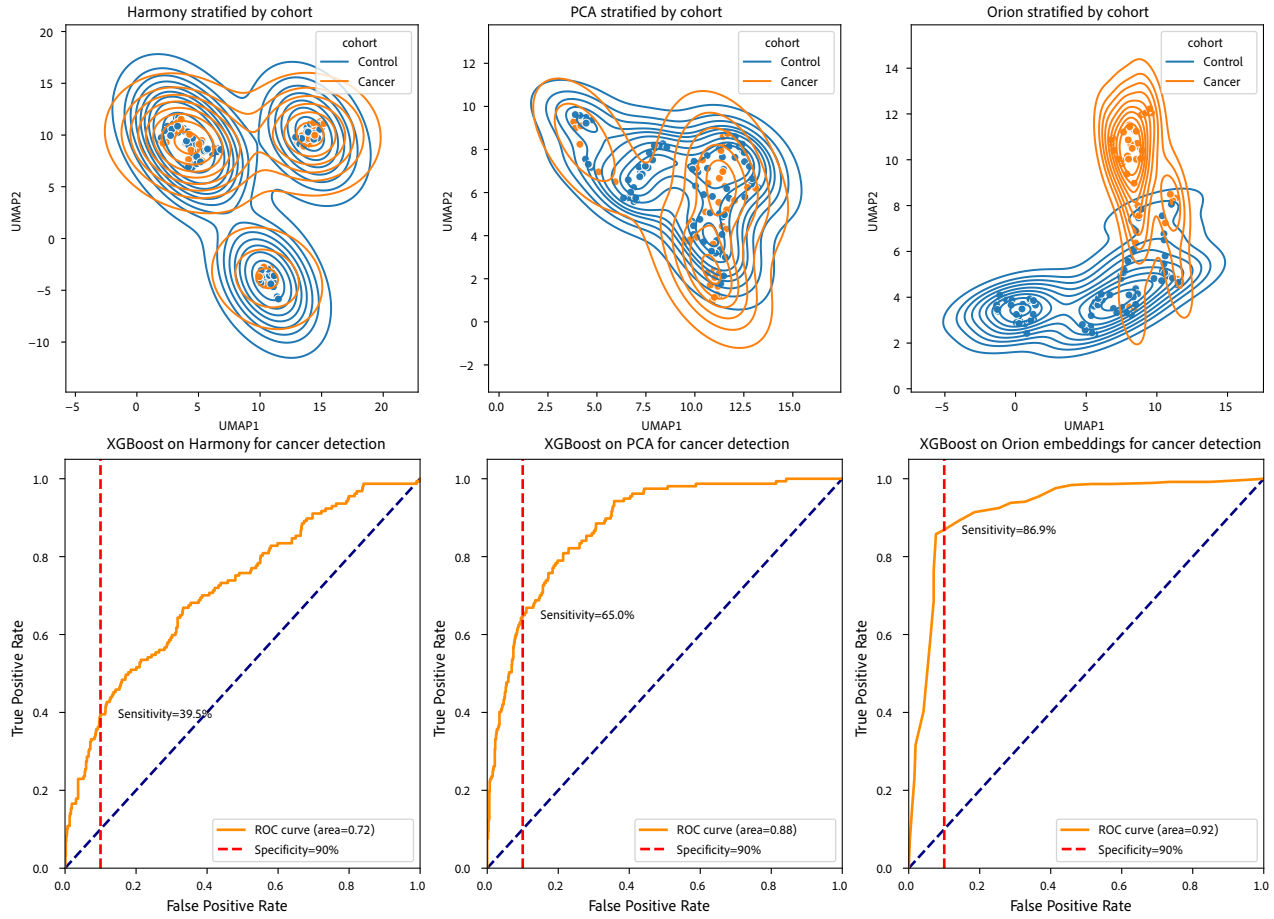
We used Orion's embeddings from the training set and the same subset of PCA and harmony for training XGBoost models to predict cancer with default parameters. We applied the model on Orion's embeddings from the test as well as the PCA and harmony for the same subset of samples (Supplementary Figure 2).

Supplementary Table 1: Performance of Orion, ElasticNet, support vector machines (SVM) classifier, and k-NN within the training set (10-fold CV) and held-out validation set. For the training set, we chose the cutoff based on the threshold closest to 90% specificity. We applied the same cutoff on the validation set. Values indicate the point estimate and 95% confidence intervals.

Method	AUC	Sensitivity	Specificity	F_1 score	Dataset
Orion	**0.97 (0.96–0.98)**	**0.94 (0.92–0.97)**	0.90 (0.87–0.93)	**0.90 (0.88–0.92)**	Training set
XGBoost	0.94 (0.93–0.96)	0.85 (0.82–0.89)	0.90 (0.87–0.93)	0.85 (0.82–0.88)	Training set
ElasticNet	0.93 (0.91–0.95)	0.81 (0.77–0.85)	0.90 (0.88–0.93)	0.83 (0.79–0.86)	Training set
SVM	0.87 (0.84–0.89)	0.61 (0.55–0.66)	0.90 (0.87–0.93)	0.69 (0.64–0.73)	Training set
KNN	0.80 (0.77–0.83)	0.53 (0.48–0.58)	**0.89 (0.86–0.91)**	0.62 (0.58–0.67)	Training set
Orion	**0.97 (0.95–0.99)**	**0.92 (0.86–0.97)**	0.90 (0.84–0.95)	**0.89 (0.83–0.93)**	Validation set
XGBoost	0.90 (0.86–0.94)	0.33 (0.23–0.44)	**1.00 (1.00–1.00)**	0.50 (0.38–0.61)	Validation set
ElasticNet	0.92 (0.88–0.96)	0.09 (0.04–0.16)	**1.00 (1.00–1.00)**	0.17 (0.07–0.27)	Validation set
SVM	0.89 (0.85–0.93)	0.13 (0.06–0.21)	0.99 (0.97–1.00)	0.23 (0.12–0.34)	Validation set
KNN	0.80 (0.73–0.86)	0.31 (0.21–0.41)	0.98 (0.94–1.00)	0.46 (0.34–0.56)	Validation set



Supplementary Figure 1: Orion properly estimates ZINB parameters. (a) Scatter plot overlaid with kernel density estimates show the estimated library size parameter (x-axis) estimated through the endogenous highly expressed smRNA input, compared to log of the total number of oncRNAs in the input matrix (y-axis). Orange shows control samples, while blue shows cancer samples. **(b)** Similar to **(a)** but y-axis represents the log of the total number of miRNA reads. **(c)** Estimated zero-inflation of each oncRNA (x-axis) compared to the fraction of the samples expressing that oncRNA (y-axis).



Supplementary Figure 2: Preserving biological signal during batch effect removal. Top panels show the UMAP of embeddings from harmony, PCA, and Orion (test set embedding). The bottom panel shows the result of training an xgboost classifier to detect presence of cancer from the top panel embeddings.

References

American Cancer Society. Lung cancer statistics. https://www.cancer.org/cancer/types/lung-cancer/about/key-statistics.html, 2023. Accessed: 2023-01-04.

National Lung Screening Trial Research Team, Denise R Aberle, Amanda M Adams, Christine D Berg, William C Black, Jonathan D Clapp, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409, August 2011.

Maria A Lopez-Olivo, Kristin G Maki, Noah J Choi, Richard M Hoffman, Ya-Chen Tina Shih, Lisa M Lowenstein, Rachel S Hicklen, and Robert J Volk. Patient adherence to screening for lung cancer in the US: A systematic review and meta-analysis. *JAMA Network Open*, 3(11):e2025102, November 2020.

Emily S Lebow, Narek Shaverdian, Jordan E Eichholz, Leah B Kratochvil, Megan McCune, et al. ctDNA-based detection of molecular residual disease in stage I-III non-small cell lung cancer patients treated with definitive radiotherapy. *Frontiers in Oncology*, 13:1253629, 2023.

Tina Cascone, Gozde Kar, Jonathan D Spicer, Rosario García-Campelo, Walter Weder, Davey B Daniel, David R Spigel, Maen Hussein, Julien Mazieres, Julio Oliveira, et al. Neoadjuvant durvalumab alone or combined with novel immuno-oncology agents in resectable lung cancer: the phase II NeoCOAST platform trial. *Cancer Discovery*, 13(11):2394–2411, 2023.

Deb Schrag, Tomasz M Beer, Charles H McDonnell, Lincoln Nadauld, Christina A Dilaveri, Robert Reid, Catherine R Marinac, Karen C Chung, Margarita Lopatin, Eric T Fung, et al. Blood-based tests for multicancer early detection (PATHFINDER): a prospective cohort study. *The Lancet*, 402 (10409):1251–1260, 2023.

Zhoufeng Wang, Kehui Xie, Guonian Zhu, Chengcheng Ma, Cheng Cheng, Yangqian Li, et al. Early detection and stratification of lung cancer aided by a cost-effective assay targeting circulating tumor DNA (ctDNA) methylation. *Respiratory Research*, 24(1):1–9, 2023.

Dimitrios Mathios, Jakob Sidenius Johansen, Stephen Cristiano, Jamie E Medina, Jillian Phallen, Klaus R Larsen, Daniel C Bruhm, Noushin Niknafs, Leonardo Ferreira, Vilmos Adleff, et al. Detection and characterization of lung cancer using cell-free DNA fragmentomes. *Nature Communications*, 12 (1):5060, 2021.

Mohammad Shahrokh Esfahani, Emily G Hamilton, Mahya Mehrmohamadi, Barzin Y Nabet, Stefan K Alig, Daniel A King, Chloe B Steen, Charles W Macaulay, Andre Schultz, Monica C Nesselbush, et al. Inferring gene expression from cell-free dna fragmentation profiles. *Nature Biotechnology*, 40 (4):585–597, 2022.

Jillian Phallen, Mark Sausen, Vilmos Adleff, Alessandro Leal, Carolyn Hruban, James White, Valsamo Anagnostou, Jacob Fiksel, Stephen Cristiano, Eniko Papp, et al. Direct detection of early-stage cancers using circulating tumor DNA. *Science translational medicine*, 9(403):eaan2415, 2017.

M Ryan Corces, Jeffrey M Granja, Shadi Shams, Bryan H Louie, Jose A Seoane, et al. The chromatin accessibility landscape of primary human cancers. *Science*, 362(6413), October 2018.

Wei Hu, Yangjun Wu, Qili Shi, Jingni Wu, Deping Kong, Xiaohua Wu, Xianghuo He, Teng Liu, and Shengli Li. Systematic characterization of cancer transcriptome at transcript resolution. *Nature Communication*, 13(1):6803, November 2022.

Gabrielle Perron, Pouria Jandaghi, Elham Moslemi, Tamiko Nishimura, Maryam Rajaee, Rached Alkallas, Tianyuan Lu, Yasser Riazalhosseini, and Hamed S Najafabadi. Pan-cancer analysis of mRNA stability for decoding tumour post-transcriptional programs. *Communications Biology*, 5(1): 851, 2022.

Hani Goodarzi, Xuhang Liu, Hoang CB Nguyen, Steven Zhang, Lisa Fish, and Sohail F Tavazoie. Endogenous trna-derived fragments suppress breast cancer progression via ybx1 displacement. *Cell*, 161(4):790–802, 2015.

Lisa Fish, Steven Zhang, Johnny X Yu, Bruce Culbertson, Alicia Y Zhou, Andrei Goga, and Hani Goodarzi. Cancer cells exploit an orphan RNA to drive metastatic progression. *Nature Medicine*, 24 (11):1743–1751, November 2018.

Jeffrey Wang, Jung Min Suh, Brian J Woo, Albertas Navickas, Kristle Garcia, Keyi Yin, Lisa Fish, Taylor Cavazos, Benjamin Hanisch, Daniel Markett, et al. Systematic annotation of orphan RNAs reveals blood-accessible molecular barcodes of cancer identity and cancer-emergent oncogenic drivers. *bioRxiv*. doi: 10.1101/2024.03.19.585748.

Mehran Karimzadeh, Jeffrey Wang, Taylor B Cavazos, Lee S Schwartzberg, Michael Multhaup, Jeremy Ku, Xuan Zhao, Jieyang Wang, Kathleen Wang, Rose Hanna, et al. Detection of early-stage cancers using circulating orphan non-coding RNAs in blood., 2023a.

Mehran Karimzadeh, Jeffrey Wang, Aiden Sababi, Oluwadamilare I Afolabi, Dung Ngoc Lam, Alice Huang, Diana R Corti, et al. Abstract 5711: Blood-based early detection of non-small cell lung cancer using orphan noncoding RNAs. *Cancer Research*, 83(7_Supplement):5711–5711, April 2023b.

Hani Goodarzi, Albertas Navickas, Jefferey Wang, Kristle Garcia, Mark J Magbanua, Lisa Fish, et al. Abstract PD9-04: Tumor-released circulating orphan non-coding RNAs reflect treatment response and survival in breast cancer. *Cancer Research*, 82(4_Supplement):PD9–04, February 2022.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, December 2018.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, and Bo Wang. scGPT: Towards building a foundation model for single-cell multi-omics using generative AI. *bioRxiv*, pages 2023–04, 2023.

Yiqun T Chen and James Zou. GenePT: A simple but Hard-to-Beat foundation model for genes and cells built from ChatGPT. *bioRxiv*, October 2023.

Yanay Rosen, Maria Brbić, Yusuf Roohani, Kyle Swanson, Ziang Li, and Jure Leskovec. Towards universal cell embeddings: Integrating single-cell RNA-seq datasets across species with SATURN. *bioRxiv*, September 2023.

Peter S Hammerman, Doug Voet, Michael S Lawrence, Douglas Voet, Rui Jing, Kristian Cibulskis, Andrey Sivachenko, Petar Stojanov, Aaron McKenna, Eric S Lander, et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–525, 2012.

Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543, 2014.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Fatemeh Dodangeh, Zahra Sadeghi, Parichehr Maleki, and Jamshid Raheb. Long non-coding RNA SOX2-OT enhances cancer biological traits via sponging to tumor suppressor mir-122-3p and mir-194-5p in non-small cell lung carcinoma. *Scientific Reports*, 13(1):12371, 2023.

Mengyuan Niu, Bin Zhang, Li Li, Zhonglan Su, Wenyuan Pu, Chen Zhao, Lulu Wei, Panpan Lian, Renwei Lu, Ranran Wang, et al. Targeting hsp90 inhibits proliferation and induces apoptosis through akt1/erk pathway in lung cancer. *Frontiers in Pharmacology*, 12:724192, 2022.

Nirjhar Bhattacharyya, Samriddhi Gupta, Shubham Sharma, Aman Soni, Sali Abubaker Bagabir, Malini Bhattacharyya, Atreyee Mukherjee, Atiah H Almalki, Mustfa F Alkhanani, Shafiul Haque, et al. CDK1 and HSP90AA1 appear as the novel regulatory genes in non-small cell lung cancer: a bioinformatics approach. *Journal of Personalized Medicine*, 12(3):393, 2022.

Dilihumaer Tuluhong, Tao Chen, Jingjie Wang, Huijuan Zeng, Hanjun Li, Wangmu Dunzhu, Qiurong Li, and Shaohua Wang. Fzd2 promotes tgf-β-induced epithelial-to-mesenchymal transition in breast cancer via activating notch signaling pathway. *Cancer Cell International*, 21:1–13, 2021.

Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, November 1901.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-Ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, 16(12):1289–1296, December 2019.

Sehhoon Park, Joungho Han, and Jong-Mu Sun. Histologic transformation of ALK-rearranged adenocarcinoma to squamous cell carcinoma after treatment with ALK inhibitor. *Lung Cancer*, 127:66–68, 2019.

Xinyuan Tong, Ayushi S Patel, Eejung Kim, Hongjun Li, Yueqing Chen, Shuai Li, Shengwu Liu, Julien Dilly, Kevin S Kapner, Ningxia Zhang, et al. Adeno-to-squamous transition drives resistance to KRAS inhibition in LKB1 mutant lung cancer. *Cancer Cell*, 2024.

Chia-Lin Hsu, Kuan-Yu Chen, Shuenn-Wen Kuo, and Yih-Leong Chang. Histologic transformation in a patient with lung cancer treated with chemotherapy and pembrolizumab. *Journal of Thoracic Oncology*, 12(6):e75–e76, 2017.

Meng Jiang, Xiaolong Zhu, Xiao Han, Haiyan Jing, Tao Han, Qiang Li, and Xiao Ding. Histologic transformation of non-small-cell lung cancer in brain metastases. *International Journal of Clinical Oncology*, 24:375–384, 2019.

Anil Vachani, Meijia Zhou, Sudip Ghosh, Shumin Zhang, Philippe Szapary, Dheeraj Gaurav, and Iftekhar Kalsekar. Complications after transthoracic needle biopsy of pulmonary nodules: a population-level retrospective cohort analysis. *Journal of the American College of Radiology*, 19(10):1121–1129, 2022.

Jeffrey Wang, Helen Li, Lisa Fish, Kimberly H Chau, Patrick Arensdorf, Hani Goodarzi, and Babak Alipanahi. Discovery and validation of orphan noncoding RNA profiles across multiple cancers in TCGA and two independent cohorts. *Cancer Research*, 82(12_Supplement):3353–3353, 2022.

Andreas Stang, Hermann Pohlabeln, Klaus M Müller, Ingeborg Jahn, Klaus Giersiepen, and Karl-Heinz Jöckel. Diagnostic agreement in the histopathological evaluation of lung cancer tissue in a population-based case-control study. *Lung Cancer*, 52(1):29–36, April 2006.

Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, et al. Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, 40(1):121–130, January 2022.

Ethan Weinberger, Chris Lin, and Su-In Lee. Isolating salient variations of interest in single-cell data with contrastiveVI. *Nature Methods*, 20(9):1336–1345, 2023.

Sindri E Antonsson and Páll Melsted. Batch correction methods used in single cell rna-sequencing analyses are often poorly calibrated. *bioRxiv*, pages 2024–03, 2024.

Haque Ishfaq, Assaf Hoogi, and Daniel Rubin. TVAE: Triplet-based variational autoencoder using metric learning. *arXiv preprint arXiv:1802.04403*, 2018.

Peter Goldstraw, Kari Chansky, John Crowley, Ramon Rami-Porta, Hisao Asamura, Wilfried EE Eberhardt, Andrew G Nicholson, Patti Groome, Alan Mitchell, Vanessa Bolejack, et al. The IASLC lung cancer staging project: proposals for revision of the TNM stage groupings in the forthcoming (eighth) edition of the TNM classification for lung cancer. *Journal of Thoracic Oncology*, 11(1):39–51, 2016.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

Robert Schmieder and Robert Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864, 2011.

M Karimzadeh, TB Cavazos, J Wang, M Multhaup, Y Fang, J Ku, X Zhao, K Wang, R Hanna, OI Afolabi, et al. AI-based early detection and subtyping of non-small cell lung cancer from blood samples using orphan noncoding RNAs. *Journal of Thoracic Oncology*, 18(11):S173, 2023c.