HistoGWAS: An AI-enabled Framework for Automated Genetic Analysis of Tissue Phenotypes in Histology Cohorts

Shubham Chaudhary[1,2,3], Almut Voigts[1,3,4], Michael Bereket[5], Matthew L. Albert[6], Kristina Schwamborn[7], Eleftheria Zeggini[4,8], Francesco Paolo Casale[1,2,3,@]

[1] Institute of AI for Health, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany

[2] Helmholtz Pioneer Campus, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany

[3] School of Computation, Information and Technology, Technical University of Munich, Garching, Germany

[4] TUM School of Medicine and Health, Technical University of Munich and Klinikum Rechts der Isar, Munich, Germany

[5] Department of Computer Science, Stanford University, Stanford, CA, US

[6] Octant Biosciences, San Francisco, CA, US

[7] Institute of Pathology, TUM School of Medicine and Health, Technical University of Munich, Munich, Germany

[8] Institute of Translational Genomics, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany

[@] Correspondence to francescopaolo.casale@helmholtz-munich.de

Abstract

Understanding how genetic variation affects tissue structure and function is crucial for deciphering disease mechanisms, yet comprehensive methods for genetic analysis of tissue histology are currently lacking. We address this gap with HistoGWAS, a framework that merges AI-driven tissue characterization with fast variance component models for scalable genetic association testing. This integration enables automated, genome-wide assessments of variant effects on tissue histology and facilitates the visualization of phenotypes linked to significant genetic loci. Applying HistoGWAS to eleven tissue types from the GTEx cohort, we identified four genome-wide significant loci, which we linked to distinct tissue histological and gene expression changes. Ultimately, a power analysis confirms HistoGWAS's effectiveness in large-scale histology cohorts, underscoring its transformative potential in studying the effects of genetic variations on tissue and their role in health and disease.

Introduction

Genetic analysis of intermediate phenotypes, such as gene expression and proteomics, has provided insights into the functional mechanisms of loci implicated in human diseases [1] through the identification of key genes and metabolites influenced by disease loci [1–7]. Extending these analyses to medical imaging-derived traits has further illuminated the intermediate effects of genetic loci [8–15]. Despite these advances, comprehensive genetic analyses of the rich tissue and cellular details available in histological images are lacking. This gap represents a significant missed opportunity, especially given the widespread availability of histological data and their relevance to disease progression.

Machine learning and artificial intelligence have revolutionized computational pathology [16], automating diagnostic processes [16,17] and enhancing prognosis through molecular biomarker predictions [18–21]. However, the challenges of conducting genome-wide association studies (GWAS) —particularly the need to assess associations with millions of genetic variants and the modest effect sizes of these genetic signals [22]— have hindered the full exploitation of these technologies in assessing the impacts of germline genetic variants on tissue histology. This limitation highlights an urgent need for new GWAS methodologies that effectively integrate AI within high-throughput statistical testing frameworks.

To address these challenges, we introduce HistoGWAS, an AI-enabled framework for automated genetic analyses of tissue histology phenotypes. HistoGWAS comprises three main components: First, we introduce a novel semantic autoencoding framework that utilizes a pre-trained foundation model to encode tissue images into embeddings, paired with a decoder trained to generate realistic semantic reconstructions from these embeddings (**Figure 1a**). Second, we develop scalable variance component tests to enable association testing between such embeddings and single genetic variants on a genome-wide scale (**Figure 1b**). Lastly, we propose a method to visualize the imaging traits associated with the detected genome-wide significant loci leveraging our decoder on interpolated embeddings [23,24] (**Figure 1c**).

We applied HistoGWAS to eleven tissues from the GTEx cohort, where we identified four independent genome-wide significant signals after multiple testing correction across traits ($P < 3.23 \times 10^{-9}$). We link the identified variants to precise histological changes, molecular pathways and related diseases. Finally, we evaluated the detection power of HistoGWAS across various

sample sizes and variant effect sizes through simulations, providing insights that can guide the design of future studies.

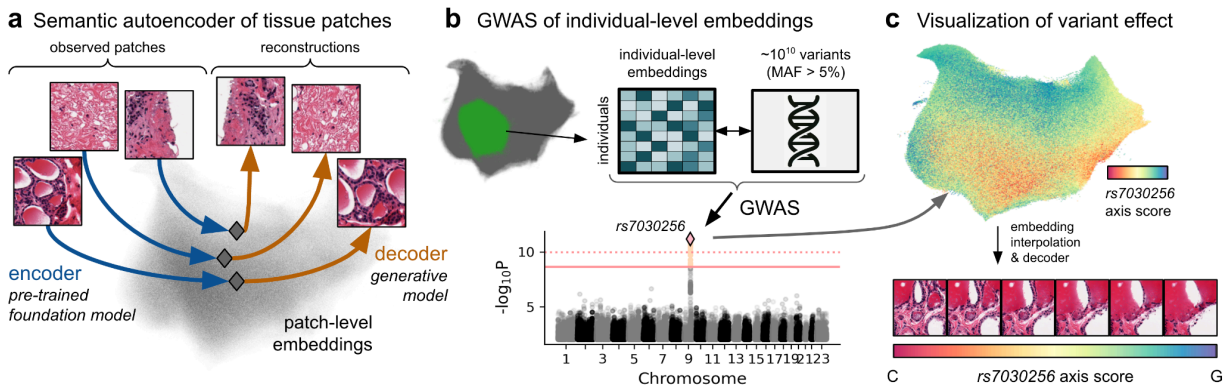

Figure 1 | Overview of the HistoGWAS Workflow. (**a**) A semantic autoencoder encodes slide patches into patch-level embeddings and reconstructs these embeddings back into the image space. We utilize a pre-trained foundation model as the encoder and a generative model as the decoder. (**b**) Illustration of a genome-wide association study (GWAS) with HistoGWAS. Patch-level embeddings from similar patches are averaged within each individual to derive individual-level embeddings, which are then associated with genome-wide variants. The output of this analysis is illustrated as a Manhattan plot, showing the P values for association between the individual-level embeddings and single genetic variants. (**c**) Visualization of the histological changes associated with significant variants (e.g., *rs7030256*) is achieved by projecting embeddings —interpolated along the direction of the genetic effect— back into the image space using the semantic decoder.

Results

Histology Semantic Autoencoders for Joint Analysis with Genetic Data

First, we set out to develop an optimized autoencoding strategy for molecular analysis by identifying the most suitable encoder for molecular predictions and building a decoder capable of accurately inverting this process.

We analyzed histology samples from the Genotype-Tissue Expression (GTEx) dataset [2], focusing on eleven tissues with the highest availability of both histology and genetic data (n ≥ 750; **Supplementary Dataset 1**). Following previous studies [25,26], we extracted 192µm x 192µm tissue patches from each whole slide image (**Figure 2a**), resulting in a dataset of 19,901,526 patches across 9,006 slides after rigorous quality control (**Supplementary Dataset 1**, **Methods**).

Starting with the encoder, we determined which of the available pre-trained models is most suited to encode the defined patches into embeddings for genomic analyses (**Figure 2b**). To do so, we assessed how effectively the embeddings from different pre-trained vision models can linearly predict gene expression labels (**Methods**). This comparison encompassed self-supervised contrastive learning methods [26,27], supervised models for cancer type classification [19,28], a histology foundation model trained across images and pathologist descriptions [29], and standard autoencoders directly optimizing image reconstruction [18,20]. Our findings indicate that contrastive learning methods perform best with the RetCCL contrastive model outperforming all others (average Spearman correlation across genes between predicted and observed values of $\rho=0.366\pm0.008$, compared to $\rho=0.346\pm0.008$ for simple contrastive learning, $\rho=0.310\pm0.008$ for cancer classification models, $\rho=0.270\pm0.010$ for the cross-modal foundation model, and $\rho=0.230\pm0.007$ for autoencoders; **Figure 2c**; **Supplementary Figure 1**). We thus selected RetCCL as HistoGWAS encoder.

After selecting our encoder, we aimed to develop a decoder that was capable of reverting patch embeddings back to full resolution images. To this end, we trained conditional generative adversarial networks (GANs) [23,30,31] for each tissue type, conditioning patch image generation on the encoder patch embeddings. Briefly, our GAN framework incorporated a generator network tasked with producing 'fake' patches from patch embeddings and Gaussian noise, paired with a discriminator network trained to distinguish between real and generated images, based on their corresponding patch embeddings (**Figure 2d**, **Methods**). After adversarial training, the generator was able to generate realistic images from embeddings, serving as an effective decoder (**Figure 2e**). Semantic reconstructions from this decoder are stochastic, with varying Gaussian noise inputs and fixed embeddings resulting in stochastic reconstructions retaining key features of the original image (**Figure 2f**).

To quantitatively evaluate the quality of our semantic reconstructions, we assessed the ability to predict gene expression from embeddings of actual vs reconstructed images, where reconstructions were obtained either using HistoGWAS semantic autoencoder or a conventional autoencoder. The correlation coefficient (R^2) for prediction accuracy across genes underscored the superior retention of gene expression signals by our semantic autoencoder, with an R^2 of 0.653, significantly surpassing the conventional autoencoder's R^2 of -0.183 (**Methods, Supplementary Figure 2**).

In summary, we introduced a novel semantic autoencoding strategy that selects an encoder to generate embeddings predictive of molecular features and trains a decoder to invert this process, and demonstrated its superior performance over traditional reconstruction-based autoencoders used in prior histology analyzes [18,20].

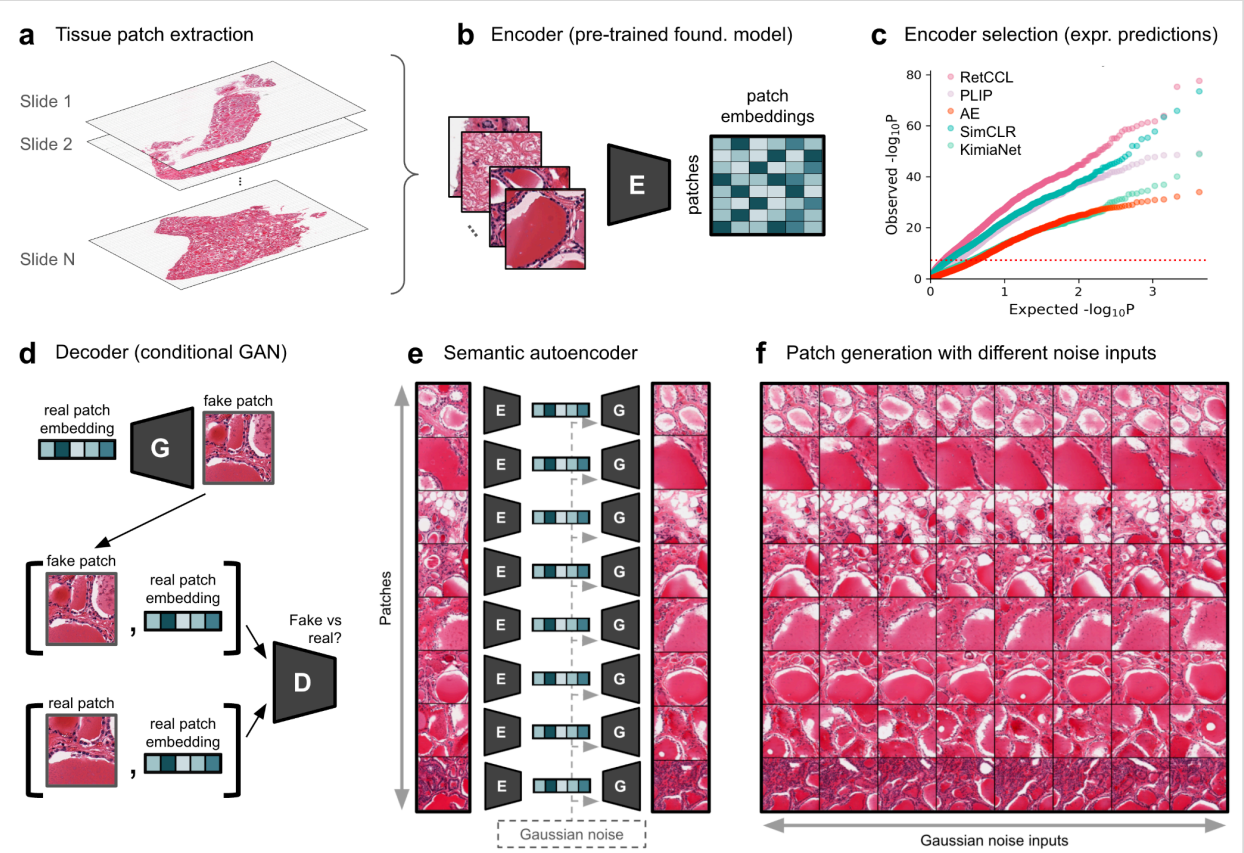

Figure 2: Effective Semantic Autoencoding of Histology Data Using the GTEx Dataset. (**a**) Tissue patches of 192µm x 192µm are extracted from 9,006 whole slide images, totaling 19,901,526 patches after quality control measures (**Methods**, **Supplementary Dataset 1**). (**b**) Patch-level embeddings are derived using a variety of pre-trained models as semantic encoders: The RetCCL contrastive learning model [29], alongside other models such as standard autoencoders [18,20], simple contrastive learning methods (SimCLR) [26,27], a tissue type classification model (KimiaNet) [28], and a pathology foundation model (PLIP) [29]. (**c**) Comparison of models in predicting gene expression is showcased through a QQ plot of P values of association between observed and predicted expression levels across genes (**Methods**). The best model in this comparison is selected as our semantic encoder. (**d**) A conditional generative adversarial network (GAN) setup trains a generator to produce realistic 'fake' patches from embeddings and Gaussian noise, while a discriminator learns to differentiate between real and generated images based on their embeddings, ensuring the generation of high-fidelity reconstructions. (**e**) Encoded patch embeddings and Gaussian noise are fed into the GAN generator, functioning as a decoder, which recreates semantically similar patches. (**f**) Patch generation variability across different patches with varying Gaussian noise inputs highlights the generator's ability to maintain critical features of the original images.

Comprehensive Genetic Association Studies across 11 Tissues Reveal Four Genome-wide Significant Loci

To harness the detailed information available in histological slides and effectively capture the subtle genetic influences that may affect specific subregions within tissues, we focused our genetic analyses on 68 distinct cluster signatures identified through a refined clustering analysis within each tissue (**Methods**, **Supplementary Figure 3, Supplementary Dataset 1**).

First, to validate the biological significance and distinctiveness of these cluster signatures, we conducted correlation analyses between the proportion of patches in each cluster signature and the expression levels of specific genes. This approach uncovered distinct genes and pathways corresponding to different cluster signatures within the same tissue type, revealing the diverse biological functions these signatures represent (**Supplementary Figure 4, Supplementary Dataset 2**, **Methods**).

Next, we employed adapted scalable variance component models to enable genome-wide association testing between individual-level average patch embeddings for each cluster signature and approximately 5 million genetic variants (MAF≥5%), while accounting for patient covariates and population structure (**Methods**). Across 68 cluster signatures, this analysis identified four genome-wide significant loci ($P < 3.23 \times 10^{-9}$; corresponding to 20% FWER), accounting for multiple hypothesis testing correction across variants and cluster signatures using a permutation-based procedure (**Figure 3a-b**, **Supplementary Figure 5**, **Supplementary Dataset 3**, **Methods**). Notably, the same analysis on permuted genotypes yielded calibrated P values with no significant associations (**Methods**, **Supplementary Figure 6**), demonstrating the statistical calibration of the proposed procedure.

The most significantly associated locus, lead intron variant *rs7030256* in the *PTCSC2* gene ($P<6.62 \times 10^{-12}$), was associated with five different signature clusters in thyroid tissue (**Supplementary Figure 5**). Additionally, three significant loci were each associated with a single cluster signatures in sun exposed skin (*rs1432621*; $P<3.94 \times 10^{-10}$), adipose subcutaneous (*rs2770197*; $P<7.39 \times 10^{-10}$), and esophagus mucosa (*rs3766325*; $P<1.64 \times 10^{-9}$), respectively.

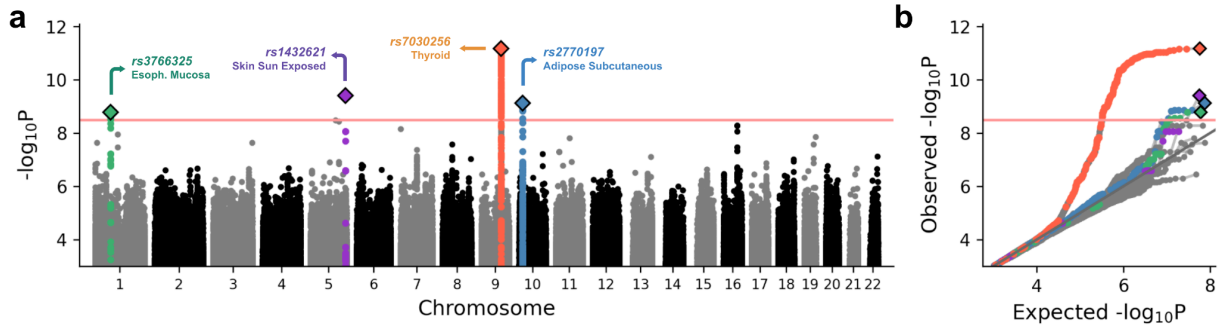

Figure 3 | GWAS Analysis of Histological Embeddings Across 68 Cluster Signatures. (**a**) Manhattan plot showcasing the P values for genome-wide association studies (GWAS) of 68 cluster signatures. The red horizontal lines represent thresholds for multi-trait genome-wide significance, set at family-wise error rates of 20%, determined through a permutation-based procedure (**Methods**). Each genome-wide significant locus is marked with a diamond, while variants in linkage disequilibrium are color-coded to match their lead variant (R^2>0.5). (**b**) QQ plots displaying the P values from genetic association tests, organized by tissue type. Variants are color-coded consistently with (**a**).

Biological Characterization of Recovered Genetic Effects on Tissue Histology

To visualize the histological changes linked to each of the identified genetic loci, we utilized our semantic decoder. Briefly, employing the linear mixed model used for association testing, we determined the direction of the genetic effect for each variant in the patch embedding space, termed the 'genetic effect axis'. We then interpolated along this axis and decoded the resulting embeddings back into the image space using our semantic decoder, which illustrated a continuum of histological changes driven by the genetic variants (**Methods**). Additionally, to probe the biological implications of each locus, we analyzed their effects on gene expression levels, investigated pathways enriched in top associated genes, and examined previously reported associations of each locus with global traits and diseases (**Methods**).

The top variant *rs7030256* of the locus identified in thyroid tissue is an intron variant in the Papillary Thyroid Cancer Susceptibility Candidate 2 gene (*PTCSC2*, **Figure 4a**), a thyroid-specific lncRNA gene associated with thyroid health [32]. Notably, this signal colocalizes with genetic variants previously identified in GWAS of hypothyroidism in the FinnGen dataset [33–35] (**Figure 4a**). Histological assessment of images generated through our interpolation procedure revealed an increase in colloid and enlarged follicles associated with allele G along the *rs7030256* axis (**Figure 4b-c**), indicative of potential goiter development. Furthermore, inflammatory cells were found within the colloid. Conversely, allele C was characterized by a

white frame around the colloid, suggesting increased resorption and thyroid activity (**Figure 4b-c; Supplementary Figure 7**). These changes were confirmed in real whole-slide images at lower resolutions (**Figure 4d**). Notably, the *rs7030256* locus also showed associations with changes in the expression levels of 128 genes (Bonferroni-adjusted P < 0.05; **Methods**; **Supplementary Dataset 3**), substantiating its relevance to thyroid physiology. Among the associated genes, we identify *DIO1*, a known regulator of the T3/T4 ratio and hormone secretion [36], genes implicated in hypothyroidism such as *GLIS3-AS1* and *DUOXA2* [37,38], and several genes linked to Papillary Thyroid Cancer, including *LINC01384*, *GREB1L*, *LINC01315*, *LMO3*, and *MUC1* [39–43]. Pathway enrichment analysis highlighted hedgehog signaling and the estrogen response as key pathways among associated genes (**Methods**; **Supplementary Dataset 3**), both known for their impact on thyroid cancer progression [44,45].

In the tissue from the esophagus mucosa, the lead variant of the detected locus is *rs3766325*, an intron variant in the Interferon-Induced Protein 44 gene (*IFI44*), recognized as an immune biomarker in head and neck squamous cell carcinoma [46] and specific infectious diseases [47]. Histologically, allele A of *rs3766325* was associated with notable enlargement of epithelial cell nuclei in the mucosal layer of the esophagus (**Supplementary Figure 7-8**), indicative of post-injury regeneration [48]. Gene expression analysis identified 30 genes associated with the *rs3766325* genotype (Bonferroni-adjusted P < 0.05, **Methods**). Genes upregulated with allele A, such as *LAPTM4B* and *MELK*, were linked to enhanced cell proliferation, whereas downregulated genes, including *MAL* and *CRCT1*, inhibited this process [49–52], which is in line with the observed histological differences. Finally, pathway enrichment analysis was also consistent with these observations, highlighting upregulation of *E2F* target [53] and *MYC* transcription factor targets pathways [54], and downregulation of the p53 tumor suppressor gene [55] with allele A (**Supplementary Dataset 3**).

For lower leg tissue from sun-exposed skin, the intergenic lead variant *rs1432621* on chromosome 5 was histologically characterized by allele G association with increased collagen density in the dense irregular connective tissue (**Supplementary Figure 7-8**). Although no single gene expression was significantly associated with the *rs1432621* genotype after multiple hypothesis testing correction, pathway enrichment analysis in top 50 associated genes revealed relevant pathways, including UV response, glycolysis, and epithelial mesenchymal transition mechanisms [56–58] (**Supplementary Dataset 3**). In subcutaneous adipose tissue, the allele A of the lead variant *rs2770197* was histologically linked to an increase in vacuole size and cell

membrane deterioration, indicators of tissue degradation (**Supplementary Figure 7-8)**. Top 50 associated genes enriched for adipogenesis, fatty acid metabolism, estrogen response (Early), and *mTORC1* Signaling, all playing crucial roles in adipocyte appearance and obesity [59–61] (**Supplementary Dataset 3**).

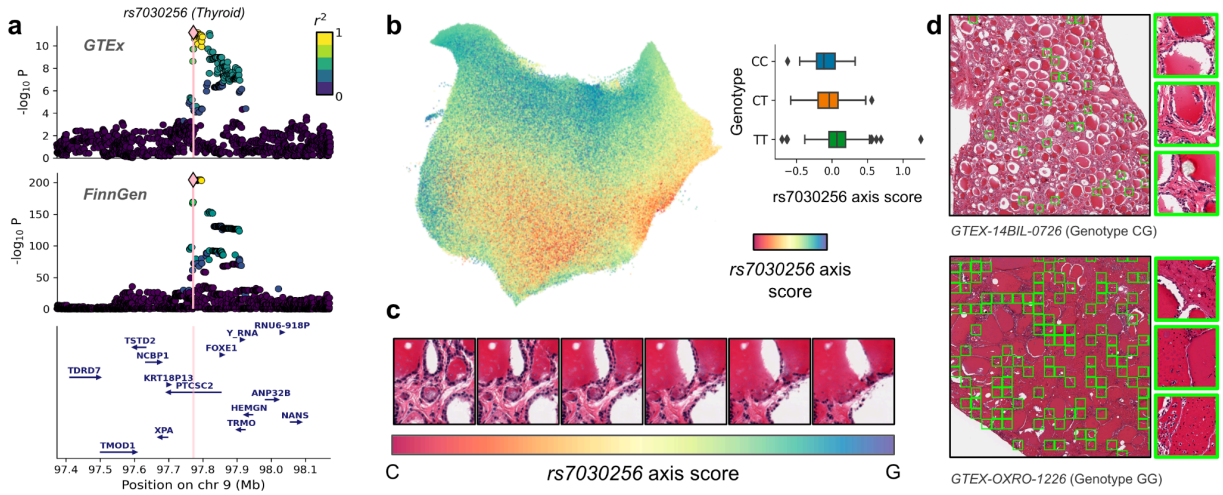

Figure 4 | Effect of Intronic Variant *rs7030256* on Thyroid Histology. (**a**) Dual LocusZoom plots illustrating the colocalization of signals for *rs7030256*: the upper plot is derived from GTEx Thyroid Cluster Signature 2, while the lower plot presents findings from the FinnGen GWAS of strict autoimmune hypothyroidism [35]. Nearby genes are also shown to highlight the genomic context of *rs7030256*. (**b**) UMAP of patch embeddings in thyroid, colored by the inferred *rs7030256* axis scores (**Methods**). An associated box plot displays the variation in these scores across genotypes. (**c**) Histological images demonstrating allele effects by projecting interpolated embeddings —along the direction of the *rs7030256* axis— back into the image space using the semantic decoder. (**d**) Whole slide images showcasing phenotypes associated with different alleles of *rs7030256*. Patches closely linked to specific alleles (defined by the top/bottom 5% percentile of the *rs7030256* axis score) are highlighted in green. For each allele, three patches are magnified to provide a detailed view of the histological differences.

Power Analysis Demonstrates HistoGWAS' Effectiveness and Scalability

To gauge HistoGWAS' efficacy for advancing histogenetics studies, we undertook a power analysis, simulating cohorts with up to 10,000 individuals. Briefly, we simulated trait embeddings with influences from covariates, a genetic variant and Gaussian noise (**Methods**). After confirming calibration of P values under the null (no simulated genetic effects) (**Supplementary Figure 9**), we assessed statistical power of detecting the simulated causal variant with HistoGWAS at $P<5x10^{-8}$, for varying sample sizes and genetic effect sizes.

In scenarios similar to our study's scale (650-1000 individuals), we had high power (>95%) for variants explaining at least 0.2% of the variance in the embeddings space (**Figure 5a**). Notably, as we increased sample size, the ability to detect subtler genetic effects significantly improved: we could detect variants that explain ≥0.1%, ≥0.05% and ≥0.02% with 2,000, 5,000 and 10,000 individuals, respectively. The heatmap in **Figure 5b** shows the sample sizes needed to achieve the desired statistical power to detect genetic effects with a given magnitude, serving as a guide for the design of future histogenetic cohorts.

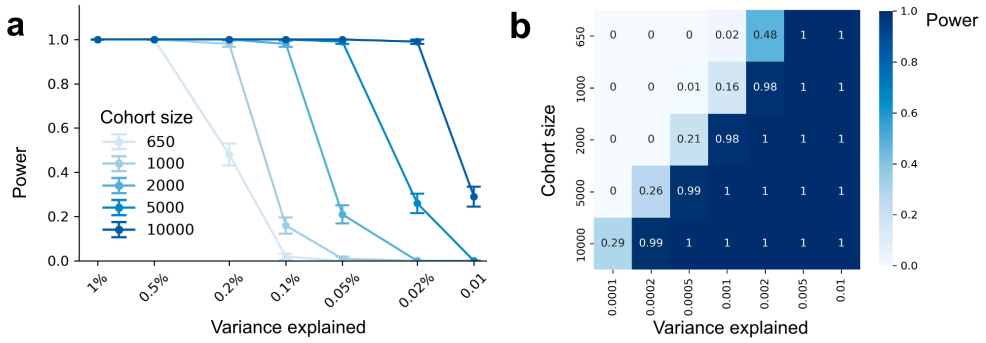

Figure 5 | Power Analysis of HistoGWAS Across Varying Sample Sizes and Genetic Effect Sizes. (**a**) Average statistical power of HistoGWAS for identifying single genetic variants, varying the variance explained by the genetic variant and the cohort size. Standard errors are derived from 100 simulation seeds. (**b**) Heatmap illustrating the average statistical power at various combinations of variance explained (rows) and cohort size (columns).

Discussion

We introduce HistoGWAS, an integrated AI-enabled framework for automated and scalable genetic analysis of tissue phenotypes leveraging large histology cohorts. This is achieved through the integration of a novel semantic autoencoder for semantic compression of tissue phenotypes with scalable variance component models enabling association testing between single genetic variants and the derived imaging representations. This setup allows for the visualization of genetic effects on tissue structures and enables the examination of tissue changes mediating the effect of genetic loci on disease risk, thereby extending genetic analyses of intermediate phenotypes to tissue histology for the first time [1–7].

Applying HistoGWAS to data from eleven tissue types within the GTEx cohort yielded four significant loci ($P<2.29\times10^{-9}$). A particularly noteworthy finding is the identification of the

PTCSC2 locus in thyroid tissue ($P<6.62e^{-12}$), a thyroid-specific lncRNA gene linked to thyroid health [32]. The top variant *rs7030256* of the locus is characterized by pronounced histological changes including the increase in colloid and enlargement of follicles, consistent with non-toxic multinodular goiter. The locus colocalizes with a previously detected genetic signal for hypothyroidism and general thyroid issues [35], which demonstrates the ability of HistoGWAS to uncover disease mechanisms by leveraging genetic and histological data.

Moreover, while the GTEx cohort's unique composition poses challenges for independent validation, power analysis confirms that HistoGWAS is exceptionally effective in large-scale histology cohorts, demonstrating its potential to pioneer the new field of histo-genetic analyses. Here, focusing on germline genetic effects, we enable the search of causal histology features to understand disease, distinguishing our analysis from previous descriptive approaches [18–21]. When considering this alongside the generality and scalability of the framework—utilizing a pretrained foundation model, simple decoding, and scalable association testing—HistoGWAS is poised to become the gold standard for detailed characterization of genetic effects on tissue structure and function and their role in disease progression.

Finally, the approach of integrating advanced AI tools for the encoding and decoding high-content phenotypes with multivariate statistical analysis of latent spaces is applicable to a wide array of high-content phenotypic modalities and use cases. For example, the approach implemented in HistoGWAS can facilitate the study of genetic effects on morphological changes visible through non-invasive medical imaging [8–14], and enable the assessment of the impacts of genetic and chemical perturbations in high-content screens of in vitro systems [62]. This is particularly relevant in advanced systems like organoids, enabling a detailed description and visualization of complex phenotypic changes resulting from induced perturbations.

Methods

Preprocessing and Encoding of Tissue Histology Data

Data Preprocessing and Patch Extraction. An initial dataset was curated from the GTEx project, focusing on tissues for which both histological and genetic data was available for a minimum of 750 individuals. This resulted in the inclusion of 11 tissue types, encompassing a total of 9,006 slides (**Supplementary Dataset 1**). We delineated patches measuring 192μm x

192µm for each analyzed slide, employing a pre-established grid. The process involved converting low-resolution slides to grayscale, followed by the application of the binary thresholding *cv2.threshold* function from the Opencv library [63], to differentiate between tissue (foreground) and the absence of tissue (background). This mask was then utilized to select patches containing at least 50% tissue, which were exported as 256 x 256 images (0.75 µm per pixel). This process resulted in a dataset containing 19,901,526 patches from 11 different tissues (see **Supplementary Dataset 1**).

Compared Models for Semantic Encoding. We evaluated embeddings from four distinct pretrained models as candidate semantic encoders for HistoGWAS: (i) RetCCL [29], pretrained on the TCGA dataset using cluster-guided contrastive learning, (ii) SimCLR[26,27], pretrained on the ImageNet dataset employing a standard contrastive learning procedure, (iii) KimiaNet [28], pretrained on TGCA for cancer type classification, (iv) PLIP [29], pretrained on slide images and associated pathologist descriptions from OpenPath through a multimodal contrastive-learning framework. All models were applied to the GTEx dataset without fine tuning. In addition to pretrained models, we trained a classical autoencoder (AE) on GTEx for each of the 11 tissues, which aimed to optimize reconstruction using an L2 reconstruction loss (**Supplementary Information**). This approach has been considered in previous analyses of histology and genomics dataset in the GTEx dataset [18,20]. After defining embeddings for all foreground patches, we conducted Principal Component Analysis (PCA) within each tissue, and reduced embeddings to their leading 64 principal components. This decision was informed by a simulation analysis that evaluated the calibration of P values in genome-wide studies using HistoGWAS under a null model (**Supplementary Figure 10**; see also the power analysis section).

Encoder Performance Evaluation and Selection. Across the five compared models, we aimed to select the model whose embeddings could best predict gene expression levels across the 11 analyzed tissues. Briefly, for each tissue, individual-level embeddings were computed as the mean of patch-level embeddings from the same individual. We then employed a variance component model for each gene, with gene expression as the outcome, and effects from individual-level embedding as random effects. For this analysis, we considered the $\log_{10}$ TPM expression values obtained from the GTEx Analysis V8 Open Access Datasets [64], targeting highly-variable genes within each tissue, identified using the scanpy *highly_variable_genes* function [65]. This model was fitted on 50% of the individuals (training set), and then used to make

gene expression predictions on the remaining 50% (test set). Prediction accuracy was evaluated using Spearman's correlation between observed and predicted gene expression levels across individuals in the test set, with significance determined by Bonferroni-corrected Spearman P values < 0.05. Among the evaluated models, RetCCL emerged as the most accurate in gene expression prediction and was consequently chosen for all further analyses with HistoGWAS (**Figure 2c**, **Supplementary Figure 1**).

Development of Generative Model for HistoGWAS Decoder

To invert the encoding process, we train a generator within the conditional Generative Adversarial Networks (cGAN) framework [30], specifically adapted to condition on patch embeddings from the HistoGWAS encoder. This adaptation enables the generator to create synthetic patch images from observed patch embeddings, while a discriminator learns to differentiate between real and generated images based on those embeddings. For generating high-resolution 256x256 images, we adopted the progressive training strategy from Progressive GANs [23], gradually adding new layers to both the generator and discriminator to capture increasingly finer details during the training process [23].

Architecture of the Generator network. Our model processes a 512-dimensional latent representation z through a convolutional function $G(z)$ to generate an image. While in classic GANs z is sampled from a standard normal, we here condition its generation on 64-dimensional patch embeddings x by assuming z is drawn from a multivariate normal distribution with mean $m_z(x)$ and standard deviation $s_z(x)$, both functions of x. By rewriting $z = m_z(x) + s_z(x) \odot \varepsilon$, with $\varepsilon \sim N(0, 1)$ and $\odot$ denoting the Hadamard product, we can separate the dependency on x from the stochasticity and enable backpropagation via the reparametrization trick [66]. We employ linear layers for both $m_z(x)$ and $s_z(x)$, with $s_z(x)$ using a softplus activation function to ensure a non-negative output. Our implementation adapts the publicly available Progressive GAN architecture from Facebook [67]. For detailed specifications of the generator architecture, we refer to this code base.

Architecture of the Discriminator Module. Similar to [23], our discriminator processes a patch image y, producing a 512-dimensional intermediate representation h through the convolutional function $D(y)$. Given h, the output of the discriminator is scalar for each patch, and is the sum of an unconditional component (a linear layer applied to h) and a conditional component. The

conditional component is derived from the dot product between h and a 512-dimensional encoding of the observed patch embedding, obtained from patch embeddings via a linear layer. This conditioning approach is implemented in BigGAN to condition on per-class learnable encodings [68], but we here replace learnable encodings with a learnable linear function of the observed patch embeddings. For detailed specifications of the discriminator architecture, we refer to the Facebook code base [67].

Progressive training and optimization details. We trained our model using the Wasserstein GAN loss [69] with gradient penalty to improve stability and convergence (λ=10). Following [67], the training initiates with 4x4 pixel image generation, progressively increasing the resolution by adding layers to both generator and discriminator through stages: 8x8, 16x16, 32x32, 64x64, 128x128, and culminating in 256x256 images. Each resolution phase is designated a set number of iterations — 48,000 for the initial and 96,000 for subsequent stages, where a batch of 64 patch images was processed at each iteration. We employed the Adam optimizer [70], with β_1 set to 0, β_2 to 0.99, and a learning rate of 0.01, optimizing both networks iteratively to balance the generator and discriminator learning processes [71].

Validation of Semantic Reconstruction Accuracy. Our semantic autoencoder combines the RetCCL model with top 64 principal component extraction for the encoding process, and the cGAN generator for decoding. After visually assessing original and reconstructed patches under varied noise inputs (**Figure 3c**), we proceeded for a quantitative validation using gene expression. Specifically, we compared the quality of gene expression prediction from embeddings of actual versus reconstructed images. Briefly, for each tissue, we fitted a variance component model on 50% of individuals using individual-level embeddings from real patches as inputs. After training, we used this model to predict gene expression for the remaining 50% of the individuals using as inputs individual-level embeddings from either original or reconstructed patches. Consistency of gene expression prediction accuracy from embeddings of actual images versus reconstructed images was quantified as the coefficient of determination (R^2) between prediction test statistics across genes—R^2=1 would indicate full retention of expression signal (**Supplementary Figure 2**). We compared retained expression signals of our semantic autoencoder versus a conventional autoencoder, showcasing a marked improvement (**Supplementary Figure 2**). A full presentation of our evaluation procedure can be found in **Supplementary Information**.

Genetic Analysis of Histological Embeddings

Definition of Histological Cluster Signatures for GWAS. To optimally capture the diverse histological phenotypes within each tissue for genetic analysis, we performed an unsupervised analysis of RetCCL embeddings using the *scanpy* Python module [65]. Briefly, for each tissue, dimensionality was reduced to the 64 leading principal components via Principal Component Analysis (PCA). We then constructed the nearest neighborhood graph (n_neighbors=10), applied Uniform Manifold Approximation and Projection (UMAP) [72], and executed Leiden clustering (resolution=0.5) [73]. These hyperparameters were optimized to discern large, homogeneous clusters showcasing morphological consistency (**Supplementary Figure 3-4**). To ensure clusters had sufficient representation for downstream genetic analyses and to filter out imaging artifacts and tissue impurities, we retained clusters represented by at least 10 patches per slide in at least 650 distinct slides. This approach yielded 68 distinct tissue signature clusters for subsequent analysis (**Supplementary Figure 3-4**), containing 19,901,526 patches (**Supplementary Dataset 1**).

Associating Cluster Signatures with Gene Expression. To validate the histological cluster signatures derived from image analysis, we assessed their correlation with gene expression levels. For each signature, we calculated its abundance within slides—defined as the proportion of patches per signature—and analyzed its association with gene expression using a linear model, with Gaussianized gene expression levels as the dependent variable. P values of association were assessed via a log-likelihood ratio test. QQ plots for each cluster signature, highlighting the top five genes associated with each signature and showcasing examples of histological patches from that signature are displayed in **Supplementary Figure 4**. Full results from this analysis can be found in **Supplementary Dataset 2**.

Association testing framework. To assess genetic associations with histological traits, we employed a variance component within a linear mixed model framework. Briefly, given the genotype vector **g** for a variant of interest across N individuals, the $N \times L$ matrix of individual-level embeddings **X**, and the $N \times K$ covariate matrix **F** of K covariates, we considered the generalized variance component model:

$$\text{link}^{-1}(\mathbf{g}) = \mathbf{F}\boldsymbol{\alpha} + \mathbf{u}, \quad \text{where } \mathbf{u} \sim N(\mathbf{0}, \sigma_X^2\, \mathbf{K}(\mathbf{X})). \qquad (1)$$

Here, the link function depends on the chosen likelihood for genotype values, **α** denotes the covariate effects, and **K**(**X**) is an $N \times N$ cosine similarity-based covariance function, which describes pairwise similarities between individuals based on their histological embeddings **X**. This model enables association testing by assessing $\sigma_X^2 > 0$ through a variance component score test [74], analogous to SKAT models [75,76]. We explored both binomial [77,78] and Gaussian likelihoods for genotype values, opting for the latter due to its computational efficiency and statistical calibration—despite potential P value inflation due to model mismatches. The embedding dimension L represents the number of principal components on RetCCL embeddings, performed separately within each tissue. To select L, we evaluated the calibration of our statistical tests under the null model through simulations (**Supplementary Information**), which indicated $L = 64$ as a reasonable compromise due to increasingly deflated P values with larger numbers of components (**Supplementary Figure 10**). Within this framework, we tested for association between individual-level embeddings of 68 cluster signatures and approximately 5 million genetic variants with minor allele frequency ≥ 5%, while accounting for sex, age, type of death, and the four leading genetic principal components. More info about the selected covariance function, efficient computations in the testing procedure and the choice of the likelihood model can be found in **Supplementary Information**.

Multiple hypothesis testing correction. To account for multiple hypothesis testing across both genetic variants (MAF ≥ 5%) and cluster signatures, while also considering test dependence due to linkage disequilibrium, we employed a permutation-based approach to control the family-wise error rate (FWER). We conducted 100 permutations of the genetic data for each of the 68 cluster signatures, yielding a total of 6,800 genome-wide association analyses under the null hypothesis. By taking the minimum P value across all variants for each permutation, we obtained a distribution of 6,800 minimum P values. We used the 20th percentile of this distribution to empirically determine the 20% FWER threshold (α=0.2). To adjust for multiple testing across the 68 cluster signatures, this threshold was further divided by the number of clusters. This procedure resulted in a stringent P value threshold of $P < 3.23 \times 10^{-9}$.

Characterization of Genome-wide Significant Genetic Loci

Visualization of genetic effects on histology. To illustrate histological alterations linked to significant genetic variants, we implemented a method integrating embedding interpolation with semantic decoding. Initially, we applied a variance component model to establish the direction of

each genetic variant's effect within the embedding space, termed the 'genetic effect axis'. For the interpolation process, we calculated extreme embeddings from patches at both extremes of the path-level projection scores along the genetic effect axis—specifically, we considered average embeddings across the patches in the 1st-5th and the 95th-99th percentiles of the projection score, respectively. Interpolating between these extreme embeddings and translating them back into image space via our semantic decoder allowed us to visualize a continuum of histological transformations associated with the genetic variant's influence. Given the stochastic nature of our decoder, multiple visual interpretations of this semantic interpolation can be generated by varying the input noise, enabling thorough evaluation (**Supplementary Figure 11**). A comprehensive description of this procedure can be found in the **Supplementary Information**.

Association with expression levels and global traits and diseases. To characterize genome-wide significant loci, we analyzed their influence on gene expression, biological pathways, and broader traits. Gene expression impact was assessed by testing the association between Gaussianized expression levels of highly variable genes in relevant tissues and the lead locus variant, controlling for sex, age, type of death, and four principal genetic components. Associations were deemed significant at a Bonferroni-adjusted $P < 0.05$. For pathway analysis, we employed Fisher's exact test using the *enrichr* tool within the *gseapy* Python module [79], using the *MSigDB_Hallmark_2020* annotations [80], focusing on the top 50 genes positively or negatively associated. The top five pathways with altered activity, irrespective of P values, were selected for further interpretation. Additionally, the open target genetics interface was used to explore the loci's links with global traits [81].

Power analysis

To assess HistoGWAS's statistical power under diverse scenarios, we performed a power analysis, simulating individual-level embeddings as additive contributions from simulated covariates (sex, age, and genetic principal components), a genetic variant, and Gaussian noise. Associations between these simulated embeddings and genetic variants were evaluated using the HistoGWAS testing framework and power was assessed at FWER 5%. Comprehensive simulation details are provided in **Supplementary Information**. Power estimations varied across cohort sizes (650, 1,000, 2,000, 5,000, 10,000) and the variance explained by the genetic variant (1%, 0.5%, 0.2%, 0.1%, 0.05%, 0.02%, 0.01%), employing 100 random simulation seeds for each scenario.

Use of Artificial Intelligence

In the preparation of this manuscript, we utilized the large language model GPT-4 (https://chat.openai.com/) for editing assistance, including language polishing and clarification of text. While this tool assisted in refining the manuscript's language it was not used to generate contributions to the original research, data analysis, or interpretation of results. All final content decisions and responsibilities rest with the authors.

Declarations

Ethics approval and consent to participate

No new data was generated for this study. The different ethics approval for the GTEx dataset can be found in the corresponding publications [2]. The research conformed to the principles of the Declaration of Helsinki.

Competing interests

MA is an employee of Octant and in the scientific advisory board of HI-Bio. The other authors declare that they have no competing interests.

Acknowledgements

We would like to thank Thomas Schwarz-Romond for feedback on the manuscript. This research was conducted using data from the GTEx database, under dbGaP application number #32009. F.P.C. and S.C. received funding by the Free State of Bavaria's Hightech Agenda through the Institute of AI for Health (AIH). Results from the FinnGen study allowed for the linking of the described gene locus in thyroid and various medical conditions.

Authors' contributions

S.C., M.B. and F.P.C. developed the methods. S.C. carried out the experiments and data analysis. A.V., M.A., K.S., and E.Z. provided critical insights for interpreting the results. K.S. offered expert guidance to describe histological phenotypes. F.P.C. conceived the study and supervised the work. The initial draft was written by S.C., A.V., and F.P.C., with all authors contributing to subsequent revisions and refinements of the manuscript.

Availability of data and materials

Data availability

No primary data were generated for this study. The GTEx v8 dataset is available at https://gtexportal.org/home (under dbGaP protection). All histological slides mentioned in the paper and the figures can be accessed and analyzed under https://www.gtexportal.org/home/histologyPage.

Code availability

- HistoGWAS toolkit: https://github.com/AIH-SGML/HistoGWAS
- RetCCL: https://github.com/Xiyue-Wang/RetCCL
- SimCLR: https://lightning-bolts.readthedocs.io/en/0.3.4/models_howto.html
- KimiaNet: https://github.com/KimiaLabMayo/KimiaNet
- PLIP: https://github.com/PathologyFoundation/plip
- Progressive GAN: https://github.com/facebookresearch/pytorch_GAN_zoo

References

1. Suhre, K., McCarthy, M. I. & Schwenk, J. M. Genetics meets proteomics: perspectives for large population-based studies. *Nat. Rev. Genet.* **22**, 19–37 (2021).
2. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
3. Suhre, K. *et al.* Human metabolic individuality in biomedical and pharmaceutical research. *Nature* **477**, 54–60 (2011).
4. Kastenmüller, G., Raffler, J., Gieger, C. & Suhre, K. Genetics of human metabolism: an update. *Hum. Mol. Genet.* **24**, R93–R101 (2015).
5. Anderson, N. L. & Anderson, N. G. The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1**, 845–867 (2002).
6. Melzer, D. *et al.* A genome-wide association study identifies protein quantitative trait loci

(pQTLs). *PLoS Genet.* **4**, e1000072 (2008).

7. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* **12**, 581–594 (2013).
8. Ji, Y. *et al.* Genome-Wide and Abdominal MRI Data Provide Evidence That a Genetically Determined Favorable Adiposity Phenotype Is Characterized by Lower Ectopic Liver Fat and Lower Risk of Type 2 Diabetes, Heart Disease, and Hypertension. *Diabetes* **68**, 207–219 (2018).
9. Somineni, H. *et al.* Machine learning across multiple imaging and biomarker modalities in the UK Biobank improves genetic discovery for liver fat accumulation. *medRxiv* 2024.01.06.24300923 (2024) doi:10.1101/2024.01.06.24300923.
10. Liu, Y. *et al.* Genetic architecture of 11 organ traits derived from abdominal MRI using deep learning. *Elife* **10**, (2021).
11. Agrawal, S. *et al.* Inherited basis of visceral, abdominal subcutaneous and gluteofemoral fat depots. *Nat. Commun.* **13**, 3771 (2022).
12. Kirchler, M. *et al.* transferGWAS: GWAS of images using deep transfer learning. *Bioinformatics* **38**, 3621–3628 (2022).
13. Sergouniotis, P. I., Diakite, A., Gaurav, K., Birney, E. & Fitzgerald, T. Autoencoder-based phenotyping of ophthalmic images highlights genetic loci influencing retinal morphology and provides epidemiologically informative biomarkers. *bioRxiv* (2023) doi:10.1101/2023.06.15.23291410.
14. Xie, Z. *et al.* iGWAS: image-based genome-wide association of self-supervised deep phenotyping of human medical images. *bioRxiv* (2022) doi:10.1101/2022.05.26.22275626.
15. Meyer, H. V. *et al.* Genetic and functional insights into the fractal structure of the heart. *Nature* **584**, 589–594 (2020).
16. Song, A. H. *et al.* Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering* **1**, 930–949 (2023).

17. Generating highly accurate pathology reports from gigapixel whole slide images with HistoGPT.
18. Ash, J. T., Darnell, G., Munro, D. & Engelhardt, B. E. Joint analysis of expression levels and histological images identifies genes associated with tissue morphology. *Nat. Commun.* **12**, 1609 (2021).
19. Fu, Y. *et al.* Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat Cancer* **1**, 800–810 (2020).
20. Jones, A., Gundersen, G. W. & Engelhardt, B. E. Linking histology and molecular state across human tissues. *bioRxiv* 2022.06.10.495669 (2022) doi:10.1101/2022.06.10.495669.
21. Wagner, S. J. *et al.* Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. *Cancer Cell* **41**, 1650–1661.e4 (2023).
22. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).
23. Karras, T., Aila, T., Laine, S. & Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv [cs.NE]* (2017).
24. Karras, T., Laine, S. & Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. *arXiv [cs.NE]* (2018).
25. Casale, F. P., Bereket, M. D., Loomba, R. & Sanyal, A. Convolutional neural networks of H&E-stained biopsy images accurately quantify histologic features of non-alcoholic steatohepatitis. *J. At. Mol. Phys.*
26. Wang, X. *et al.* RetCCL: Clustering-guided contrastive learning for whole-slide image retrieval. *Med. Image Anal.* **83**, 102645 (2023).
27. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. Simclr: A simple framework for contrastive learning of visual representations. in *International Conference on Learning Representations* vol. 2 (2020).
28. Riasatian, A. *et al.* Fine-Tuning and training of densenet for histopathology image

representation using TCGA diagnostic slides. *Med. Image Anal.* **70**, 102032 (2021).

29. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visual-language foundation model for pathology image analysis using medical Twitter. *Nat. Med.* **29**, 2307–2316 (2023).
30. Mirza, M. & Osindero, S. Conditional Generative Adversarial Nets. *arXiv [cs.LG]* (2014).
31. Goodfellow, I. *et al.* Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **27**, (2014).
32. Wang, Q. *et al.* Genetic predisposition to papillary thyroid carcinoma is mediated by a long non-coding RNA TINCR enhancer polymorphism. *Int. Immunopharmacol.* **109**, 108796 (2022).
33. Kichaev, G. *et al.* Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am. J. Hum. Genet.* **104**, 65–75 (2019).
34. Open Targets Genetics. https://genetics.opentargets.org/Variant/9_97772921_C_G/associations.
35. Kurki, M. I. *et al.* Author Correction: FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **615**, E19 (2023).
36. Moran, C. *et al.* Genetic disorders of thyroid development, hormone biosynthesis and signalling. *Clin. Endocrinol.* **97**, 502–514 (2022).
37. Rurale, G., Marelli, F., Duminuco, P. & Persani, L. Glis3 as a Critical Regulator of Thyroid Primordium Specification. *Thyroid* **30**, 277–289 (2020).
38. Wang, F. *et al.* DUOX2 and DUOXA2 Variants Confer Susceptibility to Thyroid Dysgenesis and Gland-in-situ With Congenital Hypothyroidism. *Front. Endocrinol.* **11**, 237 (2020).
39. Lu, W., Xu, Y., Xu, J., Wang, Z. & Ye, G. Identification of differential expressed lncRNAs in human thyroid cancer by a genome-wide analyses. *Cancer Med.* **7**, 3935–3944 (2018).
40. Dong, K. *et al.* GREB1L overexpression is associated with good clinical outcomes in breast cancer. *Eur. J. Med. Res.* **28**, 510 (2023).
41. Ren, J., Zhang, F.-J., Wang, J.-H. & Tang, J.-D. LINC01315 promotes the aggressive

phenotypes of papillary thyroid cancer cells by sponging miR-497-5p. *Kaohsiung J. Med. Sci.* **37**, 459–467 (2021).

42. Ling, Z., Long, X., Wu, Y., Li, J. & Feng, M. LMO3 promotes proliferation and metastasis of papillary thyroid carcinoma cells by regulating LIMK1-mediated cofilin and the β-catenin pathway. *Open Med.* **17**, 453–462 (2022).
43. Patel, K. N. *et al.* MUC1 plays a role in tumor maintenance in aggressive thyroid carcinomas. *Surgery* **138**, 994–1001; discussion 1001–2 (2005).
44. Xu, X., Lu, Y., Li, Y. & Prinz, R. A. Sonic Hedgehog Signaling in Thyroid Cancer. *Front. Endocrinol.* **8**, 284 (2017).
45. Lu, Y., Li, J. & Li, J. Estrogen and thyroid diseases: an update. *Minerva Med.* **107**, 239–244 (2016).
46. Pan, H. *et al.* Interferon-Induced Protein 44 Correlated With Immune Infiltration Serves as a Potential Prognostic Indicator in Head and Neck Squamous Cell Carcinoma. *Front. Oncol.* **10**, 557157 (2020).
47. Busse, D. C. *et al.* Interferon-Induced Protein 44 and Interferon-Induced Protein 44-Like Restrict Replication of Respiratory Syncytial Virus. *J. Virol.* **94**, (2020).
48. Wang, Y.-M. *et al.* Circular RNA hsa_circ_0003823 promotes the Tumor Progression, Metastasis and Apatinib Resistance of Esophageal Squamous Cell Carcinoma by miR-607/CRISP3 Axis. *Int. J. Biol. Sci.* **18**, 5787–5808 (2022).
49. Guo, K. & Li, K. Screening of four key genes in esophageal carcinoma based on TCGA and GEO data and verification of anti-proliferative effect of LAPTM4B knockdown in esophageal carcinoma cells invitro. *Arch. Biochem. Biophys.* **728**, 109352 (2022).
50. Ye, J. *et al.* MELK predicts poor prognosis and promotes metastasis in esophageal squamous cell carcinoma via activating the NF-κB pathway. *Int. J. Oncol.* **61**, (2022).
51. Rubio-Ramos, A., Labat-de-Hoz, L., Correas, I. & Alonso, M. A. The MAL Protein, an Integral Component of Specialized Membranes, in Normal Cells and Cancer. *Cells* **10**,

(2021).

52. Wu, N., Song, Y., Pang, L. & Chen, Z. CRCT1 regulated by microRNA-520 g inhibits proliferation and induces apoptosis in esophageal squamous cell cancer. *Tumour Biol.* **37**, 8271–8279 (2016).
53. Xie, D., Pei, Q., Li, J., Wan, X. & Ye, T. Emerging Role of E2F Family in Cancer Stem Cells. *Front. Oncol.* **11**, 723137 (2021).
54. Bretones, G., Delgado, M. D. & León, J. Myc and cell cycle control. *Biochim. Biophys. Acta* **1849**, 506–516 (2015).
55. Marei, H. E. *et al.* p53 signaling in cancer progression and therapy. *Cancer Cell Int.* **21**, 703 (2021).
56. Xie, N. *et al.* Glycolytic Reprogramming in Myofibroblast Differentiation and Lung Fibrosis. *Am. J. Respir. Crit. Care Med.* **192**, 1462–1474 (2015).
57. Lee, K. & Nelson, C. M. Chapter four - New Insights into the Regulation of Epithelial–Mesenchymal Transition and Tissue Fibrosis. in *International Review of Cell and Molecular Biology* (ed. Jeon, K. W.) vol. 294 171–221 (Academic Press, 2012).
58. Rognoni, E. *et al.* Role of distinct fibroblast lineages and immune cells in dermal repair following UV radiation-induced tissue damage. *Elife* **10**, (2021).
59. Steiner, B. M. & Berry, D. C. The Regulation of Adipose Tissue Health by Estrogens. *Front. Endocrinol.* **13**, 889923 (2022).
60. Glastonbury, C. A. *et al.* Machine Learning based histology phenotyping to investigate the epidemiologic and genetic basis of adipocyte morphology and cardiometabolic traits. *PLoS Comput. Biol.* **16**, e1008044 (2020).
61. Magdalon, J. & Festuccia, W. T. Regulation of adiposity by mTORC1. *Einstein* **15**, 507–511 (2017).
62. Afewerki, S. *et al.* In vitro high-content tissue models to address precision medicine challenges. *Mol. Aspects Med.* **91**, 101108 (2023).

63. Bradski, G. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer* **25**, 120–123 (2000).

64. GTEx Portal. https://gtexportal.org/home/downloads/adult-gtex/bulk_tissue_expression.

65. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

66. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

67. *pytorch_GAN_zoo: A Mix of GAN Implementations Including Progressive Growing*. (Github).

68. Brock, A., Donahue, J. & Simonyan, K. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv [cs.LG]* (2018).

69. Arjovsky, M., Chintala, S. & Bottou, L. Wasserstein Generative Adversarial Networks. *ICML* 214–223 (2017).

70. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv [cs.LG]* (2014).

71. Goodfellow, I. *et al.* Generative adversarial networks. *Commun. ACM* **63**, 139–144 (2020).

72. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).

73. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).

74. Lin, X. Variance component testing in generalised linear models with random effects. *Biometrika* **84**, 309–326 (1997).

75. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).

76. Moore, R. *et al.* A linear mixed-model approach to study multivariate gene–environment interactions. *Nat. Genet.* **51**, 180–186 (2018).

77. Hao, W., Song, M. & Storey, J. D. Probabilistic models of genetic variation in structured

populations applied to global human studies. *Bioinformatics* **32**, 713–721 (2016).

78. Engelmann, J. P., Palma, A., Tomczak, J. M., Theis, F. J. & Casale, F. P. Attention-based Multi-instance Mixed Models. *arXiv [cs.LG]* (2023).
79. Fang, Z., Liu, X. & Peltz, G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* **39**, (2023).
80. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425 (2015).
81. Ghoussaini, M. *et al.* Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.* **49**, D1311–D1320 (2021).