Intraplaque haemorrhage quantification and molecular characterisation using attention based multiple instance learning

Authors

Francesco Cisternino[1]*, Yipei Song[2,3]*, Tim S. Peters[6]*, Roderick Westerman[6], Gert J. de Borst[7], Ernest Diez Benavente[6], Noortje A.M. van den Dungen[6], Petra Homoed-van der Kraak[8], Dominique P.V. de Kleijn[7], Joost Mekke[7], Michal Mokry[6], Gerard Pasterkamp[6], Hester M. den Ruijter[6,9], Evelyn Velema[9], Clint L. Miller[2,4,5]*, Craig A. Glastonbury[1,10]*, S.W. van der Laan[2,6]*.

* these authors contributed equally

Affiliations

[1] Human Technopole, Viale Rita Levi-Montalcini 1, 20157, Milan, Italy.
[2] Department of Genome Sciences, University of Virginia, Charlottesville, VA, USA.
[3] Department of Computer Engineering, University of Virginia, Charlottesville, VA, USA.
[4] Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA.
[5] Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA, USA.
[6] Central Diagnostic Laboratory, Division Laboratories, Pharmacy, and Biomedical genetics, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands.
[7] Vascular surgery, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands.
[8] Pathology, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands.
[9] Experimental Cardiology, Department Cardiology, Division Heart & Lungs, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands.
[10] Nuffield Department of Medicine, University of Oxford, Oxford, UK.

Corresponding author(s)

Dr. Sander W. van der Laan
Central Diagnostic Laboratory, Division Biomedical genetics, Laboratories, and Pharmacy
University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands
Department of Genome Sciences, University of Virginia, Charlottesville, VA, USA.
s.w.vanderlaan-2@umcutrecht.nl

Dr. Craig A. Glastonbury
Genomics Research Centre, Human Technopole, Italy
Nuffield Department of Medicine, University of Oxford, Oxford, UK
craig.glastonbury@fht.org

Dr. Clint L. Miller
Department of Genome Sciences, Department of Public Health Sciences, Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA, USA.
clintm@virginia.edu

Journal subject terms: Atherosclerosis, Intraplaque haemorrhage, cardiovascular disease, multiple instance learning, machine learning, genetics

Abstract

Intraplaque haemorrhage (IPH) represents a critical feature of plaque vulnerability as it is robustly associated with adverse cardiovascular events, including stroke and myocardial infarction. How IPH drives plaque instability is unknown. However, its identification and quantification in atherosclerotic plaques is currently performed manually, with high inter-observer variability, limiting its accurate assessment in large cohorts. Leveraging the Athero-Express biobank, an ongoing study comprising a comprehensive dataset of histological, transcriptional, and clinical information from 2,595 carotid endarterectomy patients, we developed an attention-based additive multiple instance learning (MIL) framework to automate the detection and quantification of IPH across whole-slide images of nine distinct histological stains. We demonstrate that routinely available Haematoxylin and Eosin (H&E) staining outperformed all other plaque relevant Immunohistochemistry (IHC) stains tested (AUROC = 0.86), underscoring its utility in quantifying IPH. When combining stains through ensemble models, we see that H&E + CD68 (a macrophage marker) as well as H&E + Verhoeff-Van Gieson elastic fibers staining (EVG) leads to a substantial improvement (AUROC = 0.92). Using our model, we could derive IPH area from the MIL-derived patch-level attention scores, enabling not only classification but precise localisation and quantification of IPH area in each plaque, facilitating downstream analyses of its association and cellular composition with clinical outcomes. By doing so, we demonstrate that IPH presence and area are the most significant predictors of both preoperative symptom presentation and major adverse cardiovascular events (MACE), outperforming manual scoring methods. Automating IPH detection also allowed us to characterise IPH on a molecular level at scale. Pairing IPH measurements with single-cell transcriptomic analyses revealed key molecular pathways involved in IPH, including TNF-α signalling, extracellular matrix remodelling and the presence of foam cells. This study represents the largest effort in the cardiovascular field to integrate digital pathology, machine learning, and molecular data to predict and characterize IPH which leads to better understanding how it drives symptoms and MACE. Our model provides a scalable, interpretable, and reproducible method for plaque phenotyping, enabling the derivation of plaque phenotypes for predictive modelling of MACE outcomes.

Introduction

Atherosclerosis is a chronic inflammatory process affecting arterial walls, which leads to the accumulation of lipids, inflammatory cells, and fibrous elements, resulting in the formation of atherosclerotic plaques[1,2]. These plaques are heterogeneous, with a multitude of endophenotypes present, such as lipid-rich necrotic cores, fibrous caps, calcification, and intraplaque haemorrhage (IPH), each offering unique insights into plaque stability and the likelihood of adverse cardiovascular events [3,4].

IPH marks the transition of a stable plaque into a vulnerable one, prone to rupture, and often associated with thrombus formation and acute cardiovascular events [5,6]. The process of IPH is multifaceted, involving the formation of leaky neo-vessels due to the lack of smooth muscle cells and endothelial gap junctions[7]. In the hypoxic environment of the plaque, IPH stimulates neutrophils to secrete angiogenic factors and lipid peroxidation products, while macrophages scavenge leaked erythrocytes, releasing haemoglobin and iron[8]. These processes underscore the complexity of IPH and its critical role in plaque vulnerability (**Figure 1A**).

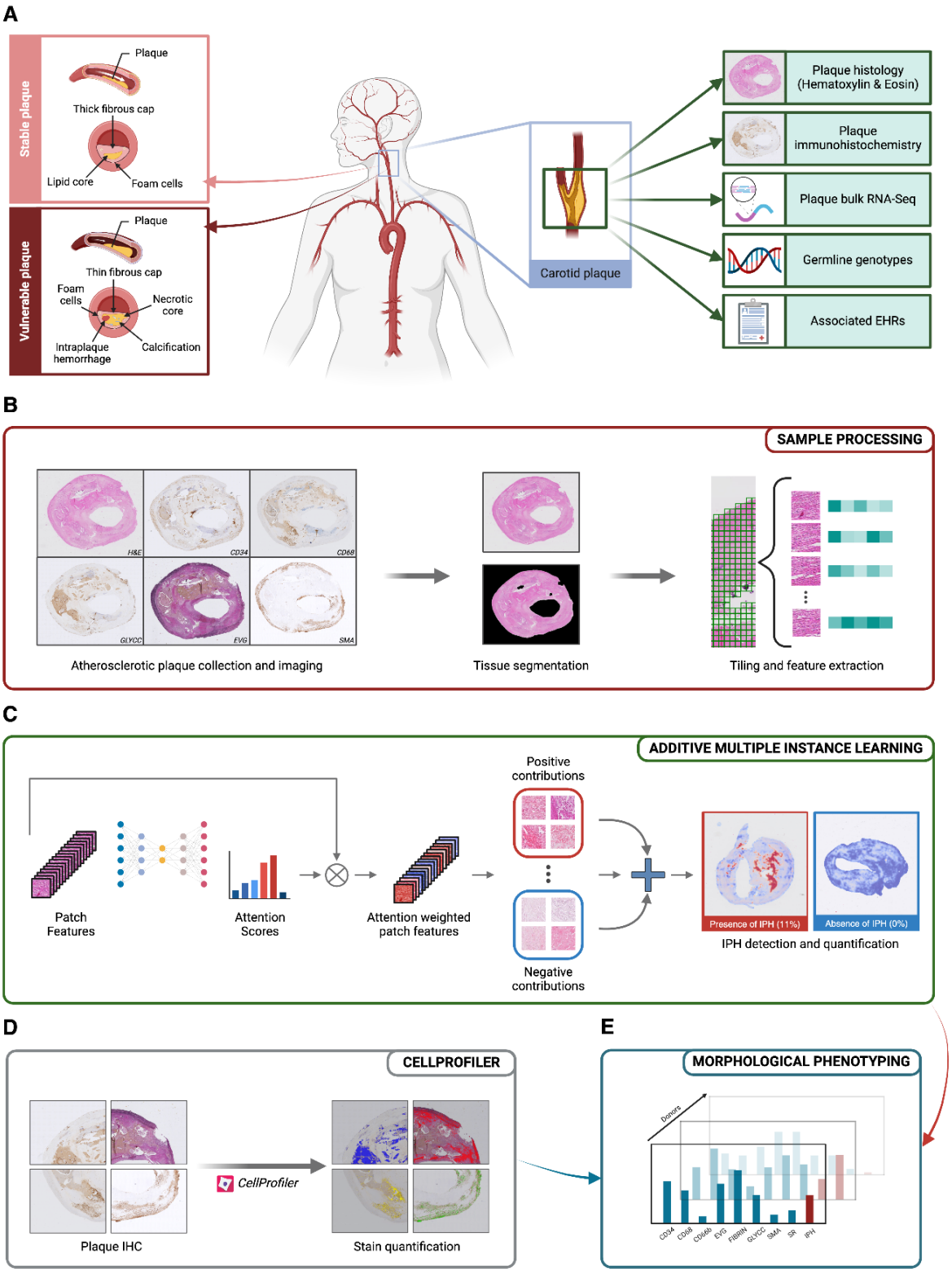


Figure 1: Schematic research overview. (**A**) Carotid plaque vulnerability shows up with specific biomarkers, like development of intraplaque haemorrhage (IPH) and calcification (Left). In this study carotid plaques were collected during endarterectomy. Samples were stained with Hematoxylin & Eosin and 8 additional histological and immuno-histochemical (IHC) markers, plaque bulk RNA-Seq were collected, together with germline genotypes; electronic health records are available for the whole cohort (Right). (**B**) After scanning the stained tissue samples, whole-slide images (WSI) were pre-processed by segmenting tissue regions from background, patching and performing feature extraction. (**C**)

Attention-based additive multiple instance learning was employed to predict, quantify and localize IPH in the tissue; coupling this image-derived biomarker with stain quantification from CellProfiler[9,10] (**D**) resulted in a complete morphological phenotyping (**E**) of carotid plaques.

To accurately assess IPH and other plaque phenotypes, histopathological examination traditionally rely on various stains, each tailored to reveal specific cellular or extracellular components[10]. Hematoxylin and Eosin (H&E) staining provides a general overview of plaque morphology, whilst Elastin von Giessen (EVG) specifically highlights elastic fibres, which are critical for assessing vascular remodelling in response to IPH[11]. Picrosirius Red (SR) identifies collagen fibres, offering insights into the structural impact of haemorrhage. CD34 highlights endothelial cells and is used to assess neovascularization[12], while CD68 identifies macrophages, which are strongly associated with IPH[13]. CD66b stains neutrophils, which are involved in atherosclerotic inflammatory processes[14], and glycophorin C (GLYCC) specifically stains red blood cells, directly correlating with haemorrhagic regions[10]. Fibrin staining identifies blood clots, and Smooth Muscle Actin (SMA) highlights smooth muscle cells, which contribute to fibrous cap integrity[15]. While the manual scoring of IPH is done using exclusively Hematoxylin and Eosin, all these stains together could provide a more comprehensive picture of the cellular and structural IPH landscape within atherosclerotic plaques. However, to date, no systematic comparison has been made to assess the utility or redundancy of these stains in capturing IPH.

Phenotyping plaques using several individual stains is labour intensive, subjective, and can lead to high inter-sample variability, necessitating the development of advanced computational techniques to enhance phenotyping. Machine learning (ML) has emerged as a powerful tool for analyzing complex biomedical images, offering the potential for objective, high-throughput, and quantitative phenotyping of atherosclerotic plaques[16–19]. For instance, machine learning models have been successfully employed for automated segmentation of plaque components such as lipid cores, fibrous caps, and calcifications in various imaging modalities[20–23]. ML has been applied to analyze inflammatory cell infiltration in tissue samples[24,25] and used to integrate imaging data with genomics and proteomics data to identify molecular pathways associated with several types of disease[26–28].

We present an attention-based additive multiple instance learning (MIL) model to address the challenges of IPH identification and quantification in 2,595 plaques using 9 distinct histological stains (**Figure 1B,C**). Estimating IPH area allows us to characterize whether IPH abundance relates to symptom severity and secondary outcomes. Besides IPH quantification, we systematically characterise the composition of all histological stains to capture the full diversity of cellular and extracellular plaque components (**Figure 1D**), enabling a more comprehensive and automated phenotyping of plaques (**Figure 1E**). Surprisingly, H&E histology alone is an excellent predictor of IPH. Rigorous validation against traditional histopathology demonstrates machine learning's potential to improve accuracy, efficiency, and reproducibility in atherosclerosis phenotyping. Further characterization of molecular properties of plaques with IPH was performed through differential expression analysis, highlighting macrophage-related pathways, TNF-α signaling, estrogen suppression, and the role of *HMOX1* in foam cells, which was validated by single-cell sequencing and follow-up IHC staining. This study presents a robust method for quantifying IPH plaque pathology and composition and investigating its molecular and cellular heterogeneity.

Results

Cohort and data description

This study uses part of the Athero-Express Biobank Study, an ongoing study conducted in two Dutch tertiary hospitals with the objective to investigate the etiological value of plaque characteristics for long-term clinical outcomes[29]. For this study, only patients undergoing carotid endarterectomy (CEA) were selected and involve 1,803 males and 792 females. Compared to males, females smoked more frequently (40.8% vs 32.8%) and used fewer anticoagulants (8.2% vs 12.5%). Females had a lower estimated glomerular filtration rate compared to males (71.44mL/min vs 74.84mL/min) and were operated on more restenotic lesions (7.6% vs 4.0%). Previous coronary artery disease was observed less frequently in females as compared to males (23.3% vs 32.8%), and the prevalence of IPH was significantly lower in females than in males (53.1% vs 64.9%). All clinical baseline characteristics are summarized in **Supplemental Table 1**. During the 3 year follow-up period, 452 (25.6%) males and 171 (22.1%) females reached a composite endpoint (see Methods).

Predicting Intraplaque haemorrhage from nine histological stains

Intraplaque haemorrhage is typically detected by identifying the presence of red blood cells (RBC) within the plaque, either through H&E staining or through specific markers of erythrocytes (e.g. glycophorin C). However, the manual detection and quantification of IPH in samples can be hard to achieve at scale. This motivated the development of an automated method to quantify IPH in carotid plaque histology. Nine stain-specific additive MIL models were trained using the manually labelled IPH as ground truth (**Supplemental Figure 1**). Our additive MIL framework allows us to leverage individual patch contributions to calculate the overall prediction for the sample. These patch contributions, which may be positive, negative, or irrelevant to the final outcome, provided clear interpretability of the precise morphological features associated with IPH. This approach allows both the detection of IPH and its quantification in terms of area of the plaque occupied by IPH, allowing us to understand the morphological and transcriptional changes associated with its variation.

Train, validation, and test set sizes alongside AUROC, accuracy, and F1-score metrics after 10-fold cross-validation are reported in **Supplemental Table 2**. The stain achieving the best performance for IPH prediction was H&E (area under the receiver operating characteristic curve (AUROC = 0.86 ± 0.017), followed by FIBRIN (AUROC = 0.85 ± 0.035) and EVG (AUROC = 0.84 ± 0.031). Importantly, whilst EVG had a comparable number of donors compared to H&E (N = 1,743 versus N = 1,715, respectively), there were 2.7x more H&E whole slide images as compared to FIBRIN (N = 628), suggesting FIBRIN is a highly specific stain to measure IPH on its own accurately. To assess whether any gain in performance can be achieved by leveraging multiple stains, we fit logistic regression ensemble models combining the probability of having IPH across multiple stains. For each logistic regression ensemble model, we combined the probabilities from H&E (the highest-scoring single stain model) with those from another stain. Train and test set sizes alongside AUROC, accuracy, and F1-score metrics after 10-fold cross-validation are reported in **Supplemental Table 3.** The combinations achieving the best AUROC were H&E+CD68 (AUROC = 0.92 ± 0.016) and H&E+EVG (AUROC = 0.92 ± 0.017), emphasising the importance of both macrophages and vascular remodelling in IPH.

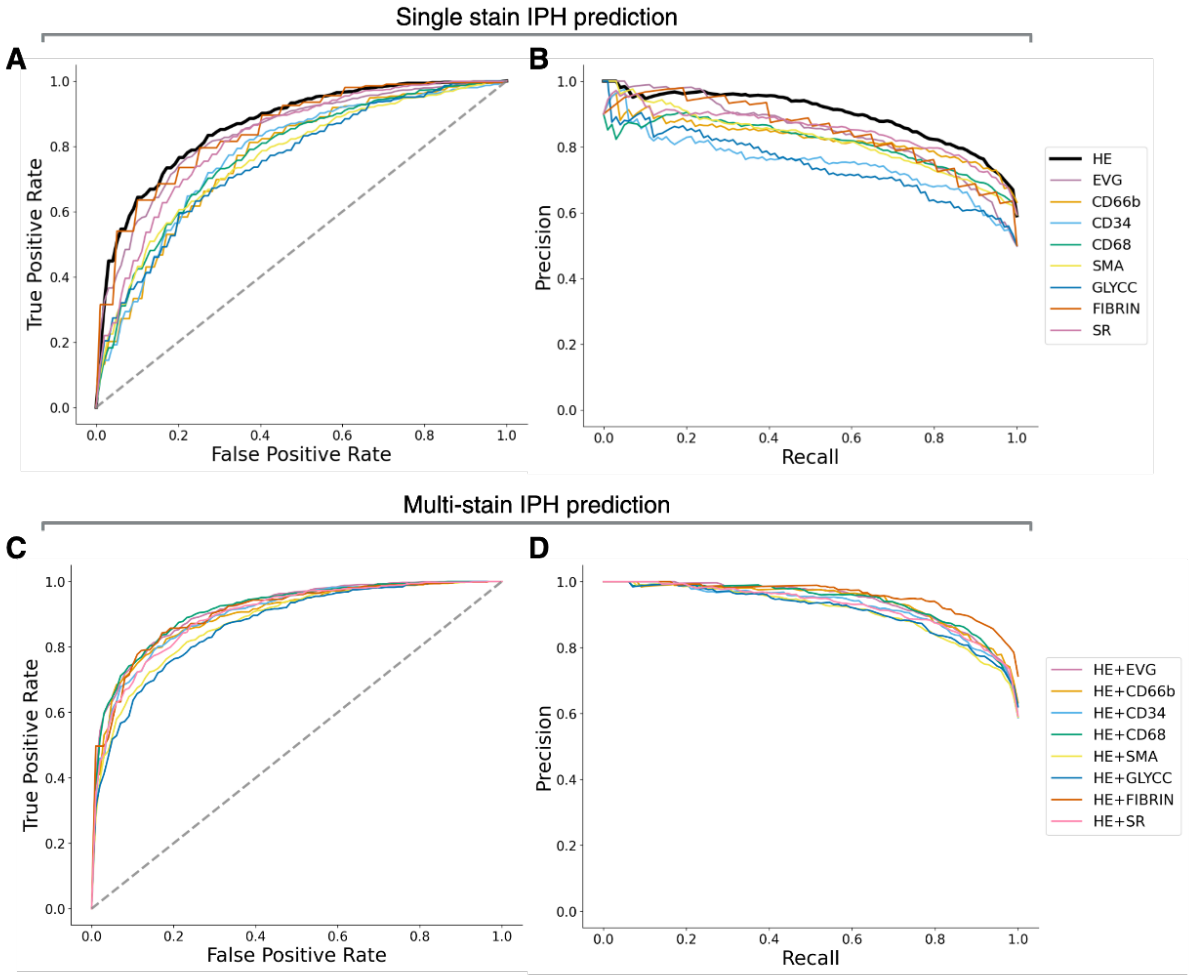


Figure 2: Additive MIL model performance. Receiver Operating Characteristic (ROC) curves (**A, C**) and Precision-Recall (PR) curves (**B, D**) for classification models trained using single histological stains (**A, B**) and multi-stain logistic regression ensemble models combining Hematoxylin & Eosin (H&E) with each of the other stains (**C, D**). The top panels show performance from models trained on individual stains: H&E, Elastin Verhoeff–Van Gieson (EVG), CD66b, CD34, CD68, Smooth Muscle Actin (SMA), glycophorin C (GLYCC), Fibrin, and Sirius Red (SR). The bottom panels present results from logistic regression ensemble combinations: H&E+EVG, H&E+CD66b, H&E+CD34, H&E+CD68, H&E+SMA, H&E+GLYCC, H&E+Fibrin, and H&E+SR. The diagonal dashed line in ROC curves represents the performance of a random classifier. H&E represent the best stain for predicting the presence of IPH. For the multi-stain comparison, H&E+CD68 achieved the highest AUROC score, while H&E+FIBRIN obtained the highest F1-score.

The use of an additive MIL model allows one to obtain instance (patch)-level contributions to the final probability score. These patch-level values can be used to localise and quantify the amount of haemorrhage across the plaque (see Methods); therefore, the proposed method can be applied to derive IPH area, even when the ground truth label is binary or categorical (**Figure 3**).

The single stain H&E model was used to derive the IPH area across all 2,220 donors. The percentage of plaque area occupied by regions of IPH varied substantially (N = 2,220, mean = 10%, SD = 1%, range: (0 - 63.6%) between subjects. The extent of IPH area is, on average, higher and more variable in men (mean = 11%, SD = 10%) than in women (mean = 7%, SD = 9%) (p-value = 6.64×10^{-25}), suggesting IPH area is sexually dimorphic. As reported by several studies[30–32], plaque rupture is more common in older women and men with high cholesterol

levels, whereas younger women tend to develop plaque erosion. This trend is confirmed in our cohort, where younger (age < 60, N = 125) women have a significantly lower mean IPH area than older (age ≥ 60, N = 550) women (mean = 5% vs mean = 8%, p-value = 1.01×10^{-5}). IPH is also more present in plaques with >10% of fat (mean = 14%, SD = 11%) than in plaques with a low fat percentage (mean = 10%, SD = 9%) (**Supplemental Figure 2)**. Finally, we see that IPH area significantly correlates positively with manually counted neutrophils (r-score = 0.41, p-value = 9.18×10^{-13}) and negatively with manually counted smooth muscle cells (r-score = -0.29, p-value = 1.92×10^{-31}). These findings collectively highlight the heterogeneity of atherosclerosis manifestations across age, sex, plaque fat content and cellularity.

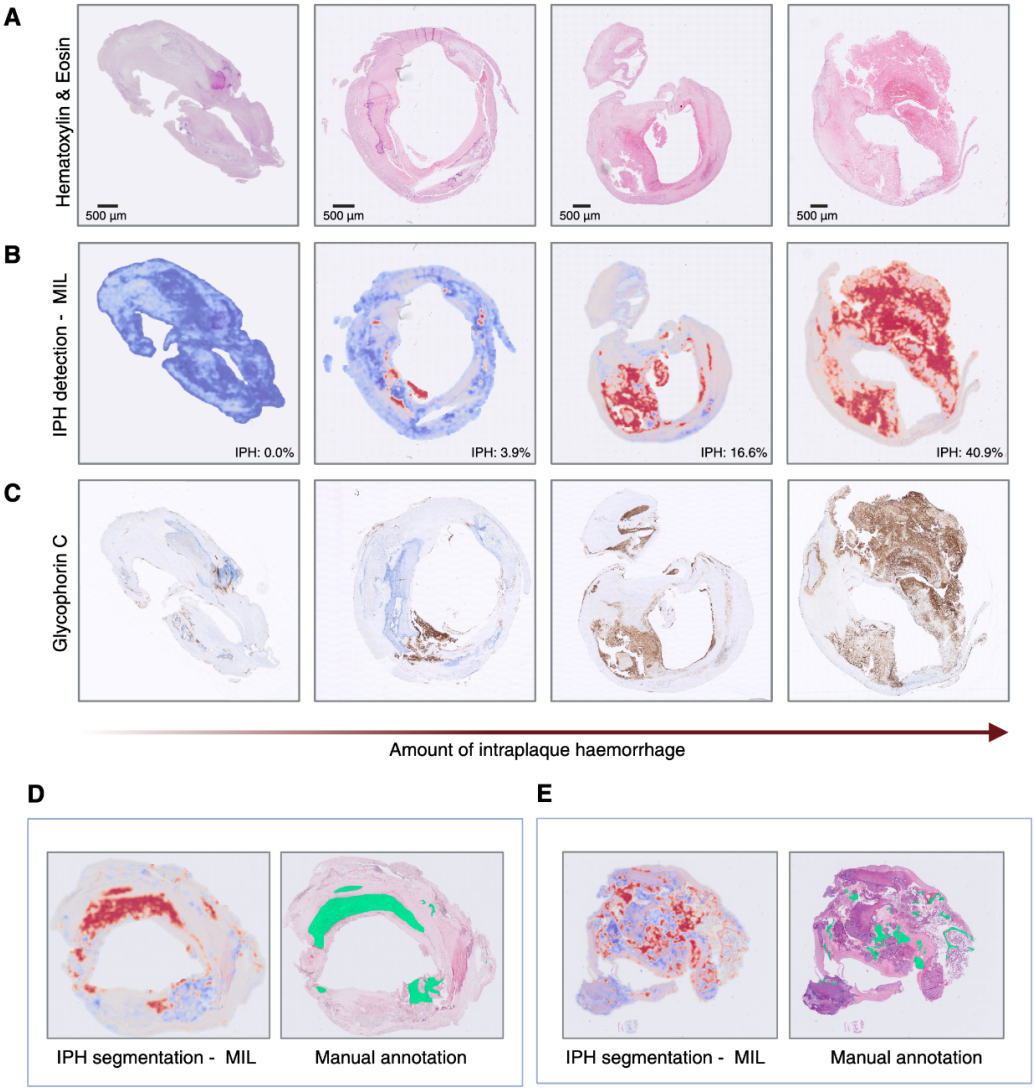


Figure 3: Additive MIL attention heatmaps. Intraplaque haemorrhage localization and quantification, expressed as a percentage of total plaque area (**B**) from H&E images (**A**) compared to glycophorin C staining (**C**); MIL-derived segmentation of IPH is validated with manual annotation (**D, E**).

Differential composition analysis reveals intricate plaque and IPH area composition changes

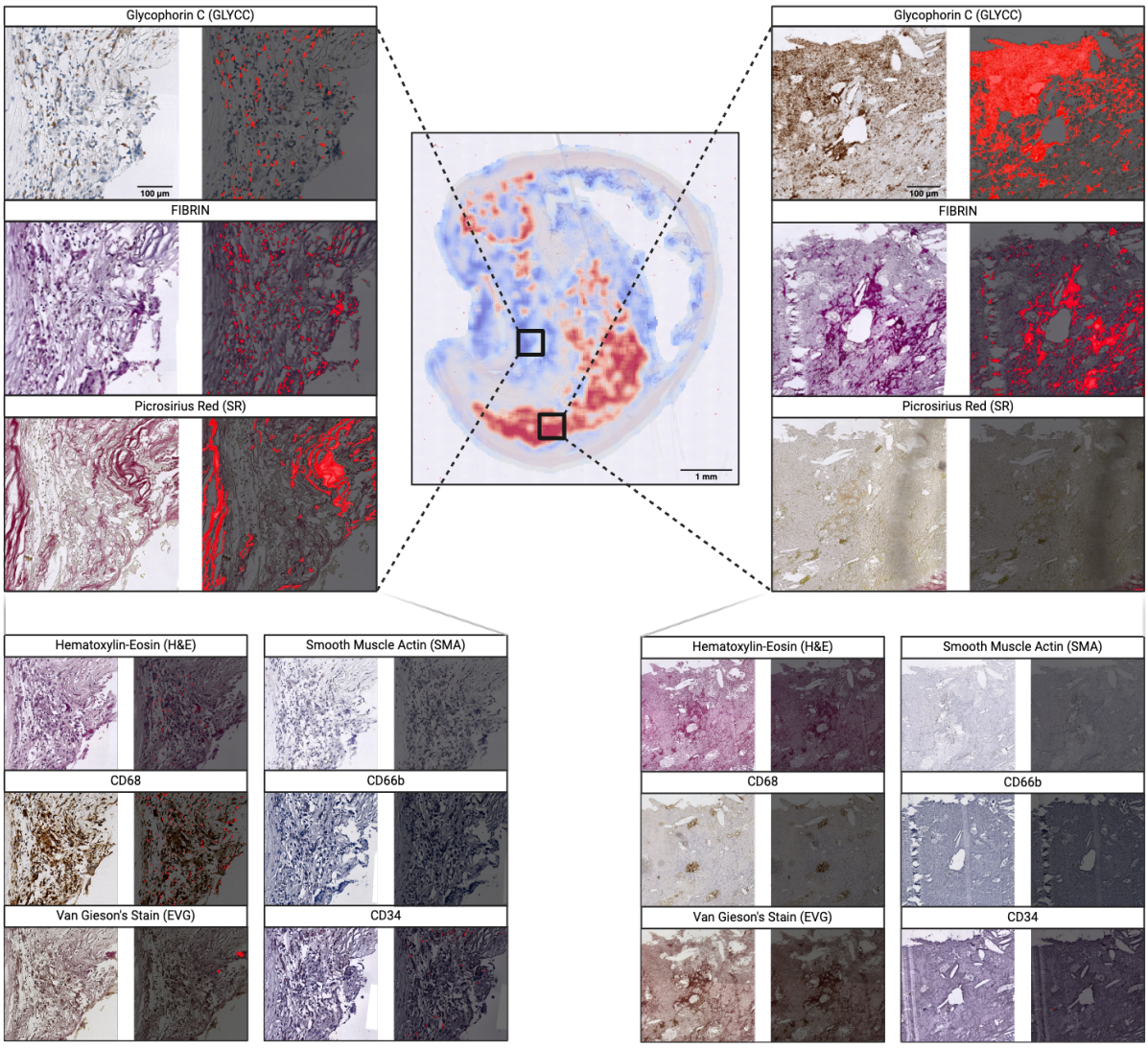


Figure 4: **CellProfiler differential composition results.** Plaque composition comparison between IPH positive vs. negative area. GLYCC, FIBRIN, and SR stains show the most significant difference. The red overlay (dark right-placed images) specifies the stained area as determined by slideToolKit[10,33] and CellProfiler[9] (completely dark indicates little to no staining). (**Left**) enlarged stain patches of IPH negative area. (**Right**) enlarged stain patches of IPH positive area. (Smaller bottom images are scaled-down versions of the larger top images, with the same scale bar applied proportionally.)

To investigate the compositional differences between plaques containing IPH and those that do not, we quantified the intensity of all 9 stains in Athero-Express using SlideToolKit[34] (see **Methods**). For each stain, we calculated the correlation between each plaque's stained ratio and the IPH area size (see Methods). In addition, we also directly compared each plaque's stained ratio between IPH-positive and IPH-negative regions (**Figure 4**). When looking at the correlation between each plaque's stained ratio and the IPH area size, multiple markers vary significantly: SMA (r-score = -0.30, p-value = 4.53×10^{-37}), GLYCC (r-score = 0.26, p-value = 2.02×10^{-21}), FIBRIN (r-score = 0.30, p-value = 1.80×10^{-15}), EVG (r-score = -0.18, p-value = 1.13×10^{-11}) and SR (r-score = -0.08, p-value = 4.01×10^{-3}) (**Figure 5A-I, Supplemental Table 4**). We see that SMA, EVG and SR have significantly less staining in plaques with an increasing IPH area, while GLYCC and FIBRIN markers positively correlate with IPH area.

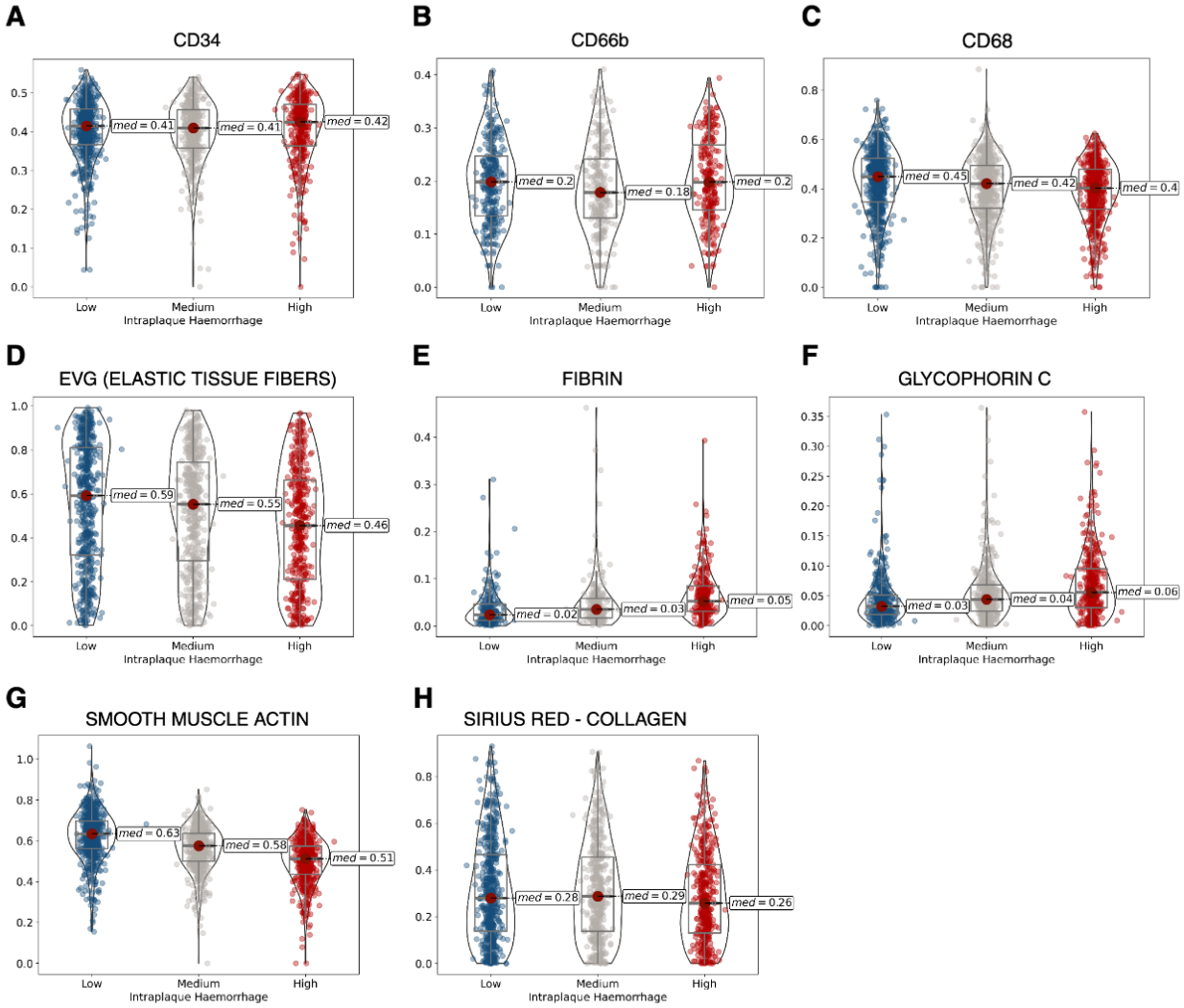


Figure 5: Histological stain quantification across IPH area. Violin plots of SlideToolKit[10] derived biomarkers across different levels of intraplaque haemorrhage (low, medium and high): CD34 (**A**), CD66b (**B**), CD68 (**C**), elastic tissue fibres (**D**), fibrin (**E**), glycophorin C (**F**), smooth muscle actin (**G**) and sirius red (**H**).

When direcly comparing IPH-positive and IPH-negative regions, there are substantial compositional differences, including: less SMA (r-score = -0.79, p-value = 1.69×10^{-234}), H&E (r-score = -0.49, p-value = 1.09×10^{-92}), SR (r-score = -0.52, p-value = 2.22×10^{-87}), CD34 (r-score = -0.46, p-value = 2.30×10^{-63}), EVG (r-score = -0.33, p-value = 1.87×10^{-36}), CD68 (r-score = -0.25, p-value = 9.93×10^{-25}), and CD66b (r-score = -0.15, p-value = 6.39×10^{-6}), whereas there is significantly more GLYCC staining (r-score = 0.56, p-value = 4.20×10^{-91}) and FIBRIN (r-score = 0.63, p-value = 5.02×10^{-77}) (**Supplemental Table 5**).
Collectively, these findings reflect a complex interplay of extracellular matrix (ECM) degradation, increased coagulation and haemorrhage, decrease in smooth muscle cells and neovascularization, and a modest decrease of neutrophil infiltration in areas of IPH.

Intraplaque haemorrhage is associated with preoperative symptoms and Major Adverse Coronary Events (MACE)

We sought to assess whether IPH is predictive of both preoperative symptom presentation as well as postoperative major adverse cardiovascular events (MACE, see Methods). The analysis of plaque morphology in the context of reported symptoms (i.e. the reason why patients had been admitted for carotid endarterectomy) provides the possibility to identify which plaque characteristics predispose to certain clinical manifestations, and to which extent. We trained a CatBoost model to predict two primary symptoms: Transient Ischemic Attack (TIA) and stroke (see Methods). The model is able to predict TIA symptoms with 64% accuracy whilst a slightly higher accuracy of 66% is observed for stroke. For both types of symptom presentation, we see that the most predictive features as measured via feature importance scores, are IPH area, age, BMI and Glomerular Filtration Rate (GFR) (**Supplemental Figure 3**).

Due to missingness, we imputed missing values (see Methods) (**Supplemental Figure 4**), with an average certainty of the imputation of 0.903 (with a more stratified certainty per variable in **Supplemental Table 6**) After imputation, we ran a Cox regression using 883 patients (626 males, 257 females). We found a significant association for the manual IPH score (HR=1.38, 95% CI, 1.00-1.90, p-value = 0.05) and the model's predicted probability score (HR=1.70, 95% CI, 1.05-2.77, p-value = 0.03) (**Figure 6A**). As previously demonstrated[35], when stratified by sex (**Figure 6B,C**), there was only a significant association identified in males. The manual IPH scoring had a hazard ratio of 1.52 (95% CI, 1.02-2.26, p-value = 0.04), with our predicted IPH classification having a comparable hazard ratio to that of the manual scoring: 1.46 (95% CI, 1.01-2.10, p-value = 0.04). Finally, the IPH probability measure had a hazard ratio of 1.94 (95% CI, 1.09-3.46, p-value = 0.02), highlighting greater sensitivity for recovering patients with IPH. For IPH area quantification, we did not see any significant association (**Figure 6, Supplemental Figure 5**).

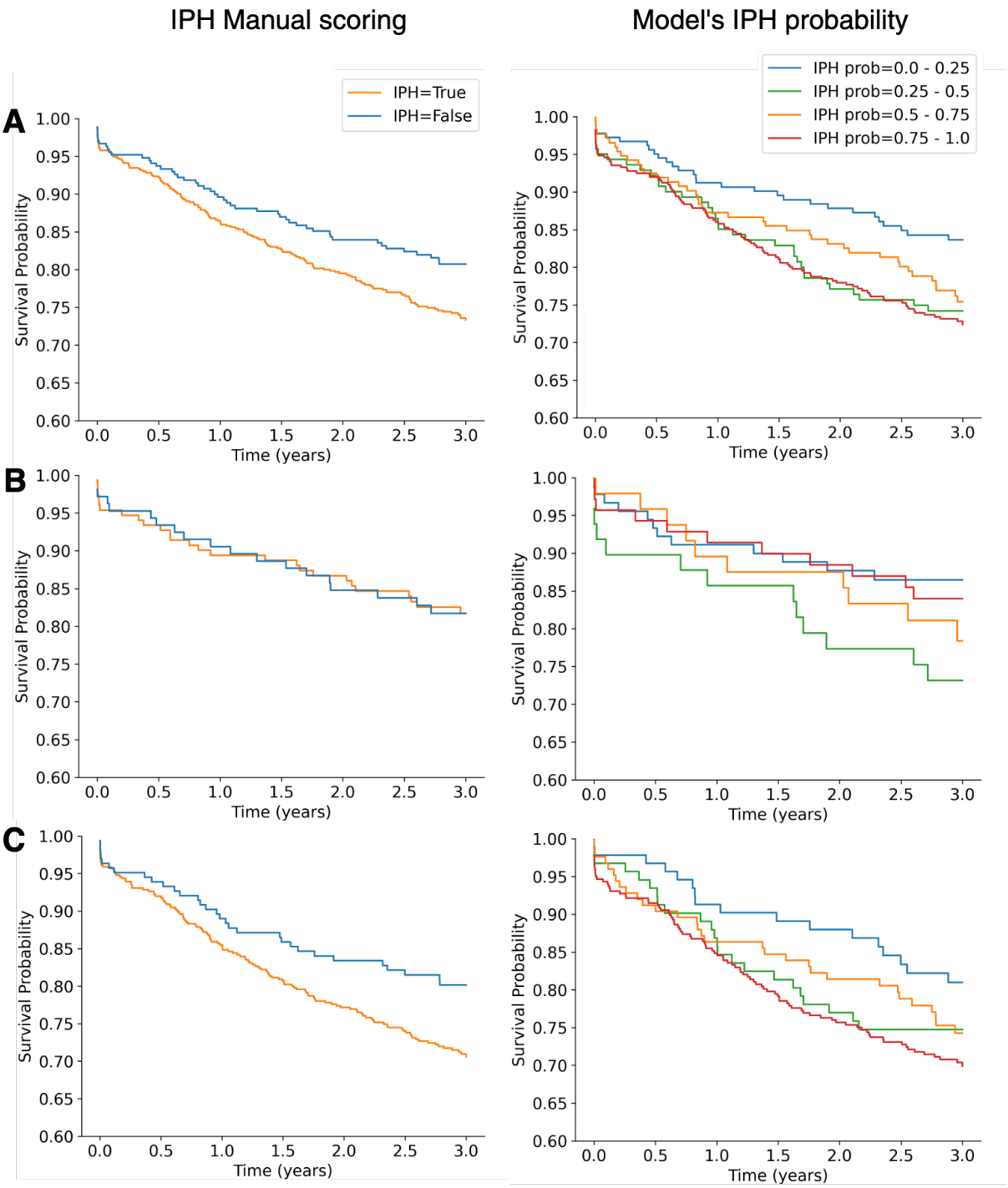


Figure 6: **Survival analysis Kaplan-Meier curves.** Cox regression Kaplan-Meier curve of IPH manual score and predicted IPH probability score for all patients (**A**), only females (**B**), and only males (**C**). Showing how IPH scoring affects the event-free survival probability.

Genetic and transcriptomic characterisation of IPH variability reveals a macrophage & foam cell signature

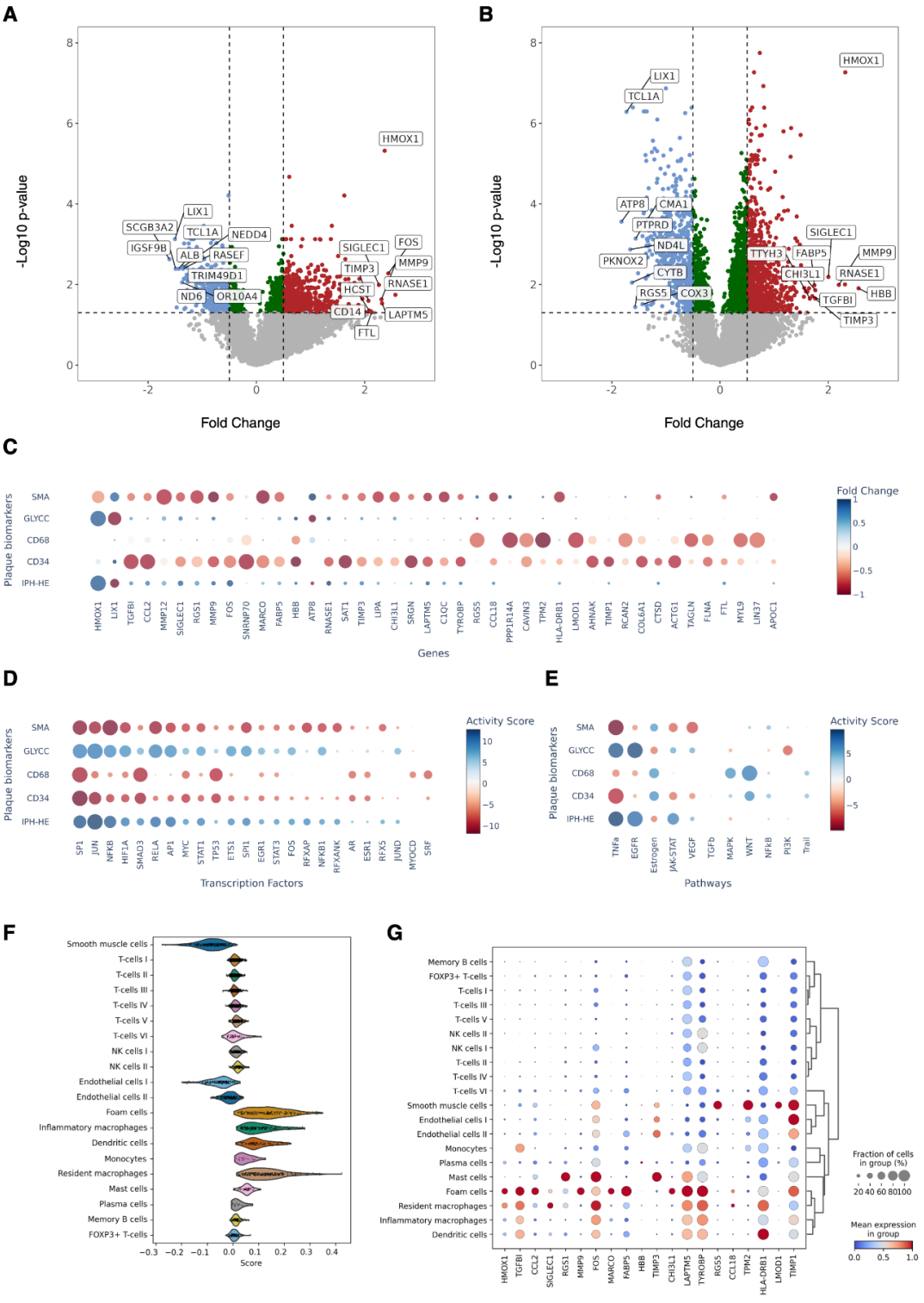


Figure 7: Differential gene expression and downstream analysis. Volcano plots of the differential gene expression using the quantified IPH area (**A**) and the amount of glycophorin C (**B**) as continuous phenotypes. (**C**) Top 10 differentially expressed genes per phenotype, ordered by total absolute fold change; the color indicates the fold change, while the dot size correlates with its significance. (**D**) Transcription factor and (**E**) pathway activity score per phenotype; in both cases, the colour of the dot

represents the activity score, whereas its size indicates the significance of the association. (**F**) Violin plot of the set of IPH area-associated genes scored across the cell types of a single-cell reference; the plot clearly shows a positive enrichment of foam cells, inflammatory and resident macrophages, and a negative enrichment of smooth muscle cells. (**G**) Mean expression of the top 20 genes reported in panel (C) per cell type; the colour of the dot depicts the mean expression value and its size denotes the fraction of cells expressing the gene.

To investigate the genetic basis of IPH presence, we conducted a genome-wide association study (GWAS) using both the manually labelled IPH phenotype and the predicted IPH phenotype from our model (**Supplemental Figure 6**). Whilst both GWAS resulted in no significant hits, we observed a strong correlation between the beta values of the two models, and a broader range of variants (p-value = 5×10^{-6}) for the image-derived IPH GWAS, indicating that our model recapitulates the same associations as the manual scoring.
To further identify genes associated with IPH and its variability, we profiled the transcriptome of n=1,026 plaques and performed differential gene expression (DGE) analyses on both IPH abundance and the amount of plaque biomarkers derived from 8 concurrent Immunohistochemistry (IHC) stains (see **Methods**). The DGE analysis of the IPH area identified a total of 1,412 significant (FDR 5%, p-value < 0.05) differentially expressed genes (854 highly expressed, 558 low expressed).
For instance, we were able to identify several genes associated with amount of GLYCC (n=686), CD34 cells (n=728), CD68 cells (n=892) and SMA (n=896) inside plaques, while CD66b (n=426), FIBRIN (n=362), picro-sirius red (n=748) and EVG (n=764) revealed no significant association with gene expression. (see **Supplemental Data**). Highly expressed and low expressed genes show high concordance between the differential gene expression analyses of derived IPH area and GLYCC (**Figure 7A, B**). Several genes (e.g. *HMOX1*, *LIX1, FOS*) are significantly differentially expressed across multiple phenotypes; again, the fold change and significance of genes almost overlap across IPH and GLYCC, meaning that the model captures regions of the plaque containing abundant erythrocytes (e.g. haemorrhage) (**Figure 7C**). The gene with the highest expression was *HMOX1*. We validated the association of *HMOX1* with the IPH area by demonstrating that areas of IPH have more *HMOX1* staining as determined by IHC (**Figure 8**). Others[36] previously showed that HO1 (the protein encoded by *HMOX1*) was upregulated in vulnerable lesions with lipid and macrophage accumulation and low collagen and vascular SMC content. Besides, high expression of HO1 was correlated with the presence of thrombus in lesions.

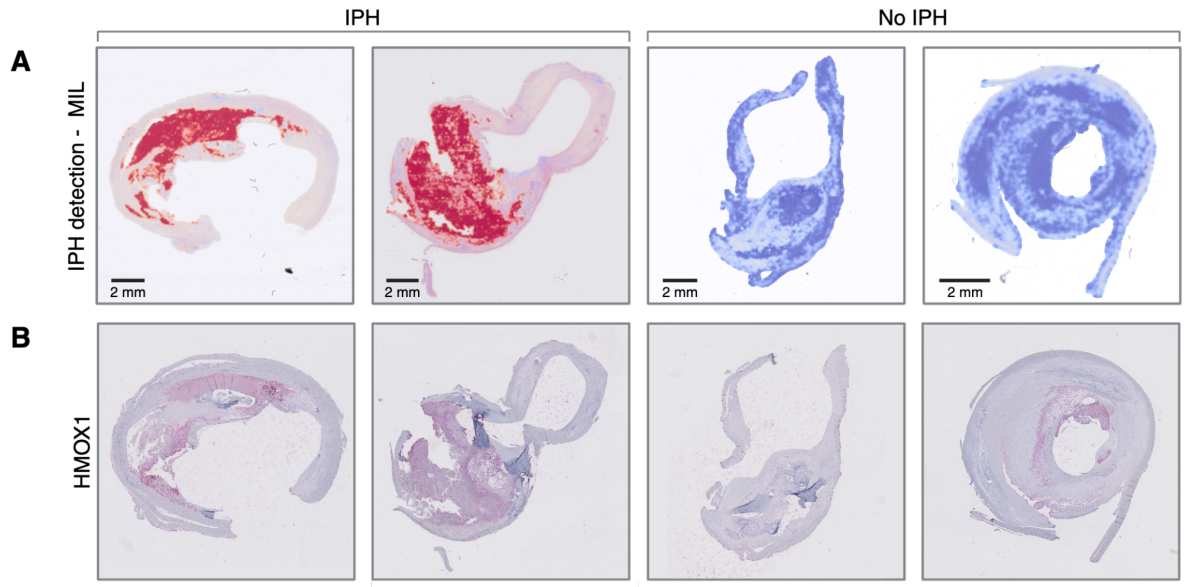


Figure 8: HMOX1 staining correlates with IPH area. Intraplaque haemorrhage localization (**A**) from H&E images compared to HMOX1 IHC staining (**B**), showing that areas of IPH express more HMOX1 staining.

We looked for associations between differentially expressed genes and atherosclerosis-related outcomes in the literature. *TIMP3* and *MMP9* are associated with increased IPH area. It is known that the balance between *TIMP3* (an inhibitor of matrix metalloproteinases) and MMPs plays a critical role in the stability of the ECM, whose disruption can lead to adverse cardiovascular events[37–39]. In Open Targets[40], *MMP9* is reported to be associated with increased triglycerides and decreased HDL cholesterol levels [41]. Several of the identified differentially expressed genes (*ATP2B1*, *CACNB2*, *LPA*, *LSP1* and *TRIB1*) have also been previously prioritised as the likely targets in a GWAS for cardiovascular disease[42] (**Supplemental Table 7**).

We investigated the derived expression signature of intraplaque haemorrhage through functional enrichment with *decoupler*. The inference of transcription factor (TF) activity identified *JUN* (p-value = 1.05×10^{-37}), *SP1* (p-value = 2.65×10^{-32}), *NFkB* (p-value = 1.05×10^{-30}) as the most active transcription factors for IPH; similar results were obtained for GLYCC. On the contrary, these TFs are mostly inactive in CD34, CD68 and SMA (**Figure 7D**). *JUN*, *SP1* and *NFkB* all play critical roles in regulating genes involved in inflammation (cytokine response, leukocyte recruitment, foam cell formation and activation), vascular cell adhesion, ECM remodelling and oxidative stress, processes that are all critical for sustaining a local inflammatory response and contribute to the destabilisation of the plaque microenvironment. These enrichments were further corroborated by pathway activity inference analyses, demonstrating a large positive enrichment for TNF-α (p-value = 1.83×10-12) and EGFR (p-value = 1.83×10^{-12}), whilst a negative enrichment for estrogen (p-value = 1.0×10^{-6}) and PI3K signaling (p-value = 1.80×10^{-5}). For CD34 and SMA, instead, a significant negative enrichment of TNF-α is observed (**Figure 7E**). It is known that TNF-α is a critical master regulator of inflammation and leukocyte recruitment into the plaque [43], whilst estrogen has been shown to have protective vascular effects in women[44]. Functional enrichment of biological terms with MSigDB revealed that many of our differentially expressed (DE) genes have a significant overlap with several gene sets, like heart macrophages markers (FDR p-value =

1.70×10^{-38}), innate immune response (FDR p-value = 2.46×10^{-14}), regulation of leukocyte adhesion to arterial endothelial cells (FDR p-value = 2.03×10^{-2}) and eosinophil activation and chemokine binding (FDR p-value = 6.28×10^{-5}).
Using single-cell RNA sequencing data, we scored the IPH-related differentially expressed genes in order to identify the cell types that are mostly associated with this phenotype: foam cells, inflammatory and resident macrophages had high positive scores, whereas smooth muscle cells were lowly scored (**Figure 7F**), confirming our previously observed negative correlation between IPH and SMA and increased performance in detecting IPH when adding CD68 and EVG stains to the model.

In summary, we derived an expression signature of IPH and other image-derived plaque biomarkers through differential expression, transcription factor and pathway activity analyses and gene scoring in a single cell reference, demonstrating the importance of ECM stability and the important role of foam cells.

Discussion

This study represents the largest digital pathology analysis of intraplaque haemorrhage (IPH) to date, leveraging attention-based additive multiple instance learning (MIL) models and a comprehensive dataset from the Athero-Express Biobank Study. This biobank includes 2,595 plaques from patients undergoing carotid endarterectomy, with extensive histological, transcriptional, and clinical data. The MIL framework we introduce here, enabled automated detection and quantification of IPH across nine histological stains, with Hematoxylin and Eosin (H&E) achieving the highest AUROC (0.86) with multi-stain ensemble logistic regression models for H&E+CD68 and H&E+EVG reaching an AUROC of 0.92. This methodology addresses the limitations of manual histopathological evaluation, including labour intensity, subjectivity, and variability, offering a scalable approach to characterize IPH and its related morphological features. By providing instance-level contributions from each whole slide image (WSI) patch, the additive MIL model facilitates the localization and quantification of IPH, laying the groundwork for downstream analyses linking plaque composition to clinical outcomes and molecular mechanisms. Integrating histological findings with clinical health records and transcriptional data, this study highlights IPH as a critical feature of plaque vulnerability, with significant implications for risk stratification and personalized cardiovascular care.

Consistent with previous studies[35,45], IPH emerges as a key phenotype linked to plaque instability and adverse clinical outcomes, including stroke and major adverse cardiovascular events (MACE). In this study, we demonstrate the association between IPH probability and composite endpoints, providing further evidence that IPH is a significant predictor of cardiovascular events. Our machine learning-derived IPH probability metrics showed stronger predictive power compared to manual annotations, particularly in male patients, consistent with earlier studies highlighting sex-based differences in IPH outcomes[35]. Beyond IPH, other plaque characteristics, such as smooth muscle actin (SMA), fibrin, and endothelial density (CD34), contribute to symptom development, reinforcing the multifactorial nature of atherosclerosis. Integrating these phenotypes with clinical data using machine learning classifiers achieved symptom prediction accuracies of 64% for transient ischemic attack (TIA) and 66% for stroke, demonstrating the utility of extracted phenotypes in understanding the role of plaque composition in clinical presentations.

The attention-based MIL pipeline delivered robust results for IPH detection, with H&E, EVG, and fibrin stains performing best in plaque characterization. Structural stains like Picrosirius Red (SR) and EVG excelled due to their ability to highlight extracellular matrix components critical to plaque integrity. The weaker performance of Glycophorin C (GLYCC) may be attributed to its high specificity for red blood cells, which limits its ability to capture broader haemorrhage related features. This may also explain why other cell-specific stains like CD68, CD34, CD66b, or SMA underperformed compared to stains that highlight structural or compositional changes. Variability in stain quality and sample size also influenced model performance, with stains like SMA and CD68 showing inconsistent results due to uneven biomarker distributions and the high variability of inflammatory cell density across samples. Despite these challenges, the pipeline's interpretability, enabled by patch-level contributions, adds a layer of transparency to the predictions, distinguishing it from traditional black-box models.

Further downstream analysis using differential gene expression (DGE) analysis revealed transcriptional signatures associated with IPH and other plaque features, shedding light on the biological mechanisms underpinning plaque vulnerability. Key genes such as *TIMP3* and

MMP9 were upregulated in plaques with increased IPH, reflecting the critical balance between extracellular matrix stability and degradation[37–39]. Genes linked to inflammation, including *CXCL1*, *CXCL4*, and *SIGLEC1*, further highlight the inflammatory pathways driving plaque instability. Enrichment analyses identified TNF-α activation and estrogen suppression as central pathways, aligning with the structural and inflammatory features of plaques[43,44]. These molecular insights corroborate the strong predictive performance of structural stains like SR and EVG, linking histological findings to transcriptional activity.

While this study advances IPH phenotyping, certain limitations must be acknowledged. Symptom prediction, while informative, remains challenging due to the multifactorial nature of atherosclerosis, non-biological influences (e.g., prior surgeries), and missing data. Sex-based differences in outcomes were noted, with IPH showing stronger predictive associations in males, but this study did not stratify analyses by sex comprehensively.

Additionally, the absence of image registration between serial stained plaque sections limited the ability to fully integrate stain-specific information, potentially affecting composition analyses. Image registration, which ensures that corresponding structures across serial plaque sections are spatially aligned, is crucial for accurately comparing stained regions. Without this, misalignments or variations in tissue orientation can introduce discrepancies, making it more difficult to reliably correlate features across different stains. Future studies should address these limitations by incorporating sex-stratified analyses and refining data integration methods.

This study sets the stage for future investigations by integrating spatial transcriptomics with histological and clinical data to deepen our understanding of plaque biology. Spatially resolved gene expression data could further elucidate the molecular mechanisms linking IPH to plaque vulnerability, complementing the DGE results presented here. Expanding this framework to include additional datasets will enhance the generalizability and clinical relevance of the findings. Additionally, exploring advanced multimodal learning approaches to leverage multi-stain data may unlock new opportunities for improving predictive accuracy and phenotyping capabilities. These advancements will ultimately contribute to personalized cardiovascular risk assessment and targeted therapeutic strategies. In summary, we present several novel contributions: (1) quantitative measurement of intraplaque haemorrhage by training an attention-based MIL model on a binary label, (2) extensive phenotyping of plaque morphology across thousands of donors, (3) an in-depth characterisation between derived biomarkers, symptoms and MACE and, finally, (4) an integration of imaging with transcriptomics to characterize the molecular basis of plaque vulnerability. These contributions collectively provide a robust framework for understanding the pathophysiology of atherosclerotic plaques and how aspects of plaque biology relate to symptom presentation and MACE risk.

Methods

Athero Express Biobank

Study design

Carotid atherosclerotic plaques were obtained from patients undergoing arterial endarterectomy and included in the Athero-Express Biobank Study (AE), an ongoing biobank study at the University Medical Centre Utrecht (Utrecht, The Netherlands) and the St. Antonius Hospital (Nieuwegein, The Netherlands)[29]. The medical ethical committees of the respective hospitals approved the AE which was registered under number 22/088. This study complies with the Declaration of Helsinki, and all participants provided written informed consent.

Sample collection

Blood samples were obtained prior to surgery and stored at -80℃. Carotid plaque specimens were removed during surgery and immediately processed in the laboratory following standardized protocols. Specimens were cut transversely into segments of 5 mm[29]. The culprit lesion (the region with the most severe stenosis) was identified, fixed in 4% formaldehyde, embedded in paraffin, and processed for histological examination[29]. The remaining segments were stored at -80℃.

Clinical characteristics

Various clinical characteristics are described throughout the paper, with additional details provided in **Supplemental Table 8.** The composite endpoint mentioned includes any death of presumed vascular origin (fatal stroke, fatal myocardial infarction, sudden death, other vascular death), nonfatal stroke, nonfatal myocardial infarction, and any arterial vascular intervention that had not already been planned at the time of inclusion (eg, carotid surgery or angioplasty/stenting, coronary bypass, percutaneous coronary intervention, peripheral vascular surgery or angioplasty/stenting).

Histological staining

We described the standardized (immuno)histochemical analysis protocols used in the AE before[3,29]. In short, 10-micron cross-sections of the paraffin-embedded segments were cut using a microtome, stained using standardized protocols, and examined under a bright-light microscope. The following stainings are performed to characterize the plaque: Picro Sirius red (SR, collagen and fat using polarized light), CD68 (macrophages), smooth muscle cell alpha-actin (SMA, smooth muscle cells), hematoxylin and eosin (H&E, thrombus and calcifications and elastin), Verhoeff's Van Gieson (EVG, elastic fibres), FIBRIN (fibrin deposits and blood clots), glycophorin C (GLYCC, red blood cells), CD34 (endothelial cells), and CD66b (neutrophils). The immunohistochemical staining for CD68 and SMA stains are performed fully automated (Ventana Medical Systems, Tucson, Arizona) (**Supplemental Table 9**).
Intraplaque haemorrhage (IPH) was scored as no/yes using a hematoxylin and eosin staining (H&E) as described before[10]. All histological observations were performed by the same dedicated technician and interobserver analyses have been reported previously[46].

Whole-slide scanning

We set up ExpressScan to obtain whole-slide images (WSIs) by scanning stained slides at 40x using a Roche Ventana iScan HT or Hamamatsu C12000-22 Digital slide scanner. WSIs were stored digitally as z-stacked .TIF (Roche) at 0.25 micron/pixel or .ndpi (Hamamatsu) at 0.23 micron/pixel brightfield microscopy images[10].

Bulk RNA isolation and sequencing

A total of 1,093 plaque segments were selected from patients included in the Ather-Express Biobank Study between 2002 and 2017 in two subsequent experiments, Athero-Express RNA Study 1 and 2 (AERNAS1: n=622, 2002-2016; AERNAS2: n=471, 2002-2017)[47]. The atherosclerotic plaques were grounded while frozen with liquid nitrogen, and, after that, TriPure (Roche, Cat# 11667165001) was added, and the plaque pieces were further disrupted by a Precellys 25 homogenizer (Bertin Instruments). The sample was incubated at room temperature for 5 minutes and centrifuged at 20,000g for 1 minute at 4°C. The supernatant was mixed with chloroform and incubated at room temperature for 15 minutes. The sample was centrifuged at 12,000g for 5 minutes at 4°C, and the upper phase was used for RNA isolation. Then, isopropanol and GlycoBlue (Invitrogen, Cat# 10301575) were added to the aqueous phase to precipitate the RNA and centrifuged at 12,000g for 10 minutes at 4°C. The pellet was washed with 75% ethanol and resuspended in RNase-free water.

The RNA isolated from the archived advanced atherosclerotic lesion is fragmented. They have, therefore, employed the CEL-seq2 library preparation protocol[48]. CEL-seq2 yielded the highest mappability reads to the annotated genes compared to other library preparation protocols. The methodology captures the 3′ end of polyadenylated RNA species and includes unique molecular identifiers (UMIs), which allow direct counting of unique RNA molecules in each sample. Then, 50 ng of total RNA was precipitated using isopropanol and washed with 75% ethanol. After removing ethanol and air-drying the pellet, a primer mix containing 5 ng of primer per reaction was added, initiating primer annealing at 65°C for 5 minutes. Subsequent reverse transcription reaction was performed: first-strand reaction for 1 hour at 42°C; heat-inactivated for 10 minutes at 70°C; second-strand reaction for 2 hours at 16°C; and then put on ice until proceeding to sample pooling. The primer used for this initial reverse transcription reaction was designed as follows: an anchored polyT, a unique 6-bp barcode, a UMI of 6 bp, the 5′ Illumina adapter and a T7 promoter, as described. Each sample now contained its own unique barcode due to the primer used in the RNA amplification, making it possible to pool together complementary DNA (cDNA) samples at seven samples per pool. cDNA was cleaned using AMPure XP beads (Beckman Coulter, Cat# A63882), washed with 80% ethanol and resuspended in water before proceeding to the in vitro transcription (IVT) reaction (Thermo Fisher Scientific, Cat# AM1334) and incubated at 37°C for 13 hours. Next, primers were removed by treating with Exo-SAP (Affymetrix, Thermo Fisher Scientific, Cat# 78201.1.ML), and amplified RNA (aRNA) was fragmented and then cleaned with RNAClean XP (Beckman-Coulter, Cat# A63987), washed with 70% ethanol, air-dried and resuspended in water. After removing the beads using a magnetic stand, RNA yield and quality in the suspension were checked by Bioanalyzer (Agilent Technologies).

cDNA library construction was then initiated by performing a reverse transcription reaction using SuperScript II reverse transcriptase (Invitrogen/Thermo Fisher Scientific, Cat# 18064022) according to the manufacturer's protocol, adding randomhexRT primer as a random primer. Next, polymerase chain reaction (PCR) amplification was done with Phusion High-Fidelity PCR Master Mix with HF buffer (New England Biolabs, F531L) and a unique indexed RNA PCR primer (Illumina) per reaction, for a total of 11–15 cycles, depending on aRNA concentration, with 30 seconds of elongation time. PCR products were cleaned twice with AMPure XP beads (Beckman Coulter, Cat# A63882). Library cDNA yield and quality were checked by Qubit fluorometric quantification (Thermo Fisher Scientific, Cat# Q32851) and

Bioanalyzer (Agilent Technologies), respectively. Libraries were sequenced on the Illumina NextSeq 500 platform, paired-end, 2 x 75 bp.

Single-cell RNA isolation and sequencing

In total 46 atherosclerotic plaque segments were used for scRNA sequencing and isolation described in detail by Depuydt *et al.* [49] and Slenders *et al.*[50]. In summary, all samples were collected from patients undergoing a carotid endarterectomy procedure. To examine the transcriptome, the plaque samples were enzymatically dissociated, filtered, and cryopreserved for subsequent fluorescence-activated cell sorting (FACS) and single-cell RNA sequencing (scRNA-seq). Viable cells were identified using Calcein AM and Hoechst staining before being sorted on a Beckman Coulter MoFlo Astrios EQ. Individual cells were dispensed into 384-well plates, lysed, and processed following the SORT-seq protocol, incorporating CEL-Seq2[48] library preparation and sequencing on an Illumina NextSeq 500. scRNA-seq data processing involved stringent quality control, normalization using Seurat[51], and clustering via canonical correlation analysis (CCA) with subsequent tSNE visualization. Cell type annotation was performed using differential gene expression analysis against BLUEPRINT reference datasets[52], and pathway analysis was conducted using EGSEA[53] with hallmark gene sets from GSEA. Clustering of cell populations provided a high resolution of the cellular landscape in atherosclerotic lesions, further distinguishing macrophages, lymphocytes, mast cells, and endothelial cells.

DNA isolation and genotyping

The Athero-Express genotyping has previously been described in detail[12,54]. In short, DNA was extracted from 2,652 EDTA blood or (when no blood was available) plaque samples of 2,266 consecutive patients and genotyped in 3 batches irrespective of artery type. Note that 1 patient can have multiple samples, thus we work with a unique patient identifier (UPID) which should not be confused with nor is it linked to a hospital ID, and a sample identifier (studynumber). In the first Athero-Express Genomic Study (AEGS1), 822 patients between 2002 and 2007 were included and genotyped (440,763 markers) using the Affymetrix Genome-Wide Human SNP Array 5.0 (SNP5) chip (Affymetrix Inc., Santa Clara, CA, USA). In the second Athero-Express Genomic Study (AEGS2), 1,022 patients between 2002 and 2013 were included (not overlapping with AEGS1) and genotyped (587,351 markers) using the Affymetrix Axiom® GW CEU 1 Array (AxM). For the Athero-Express Genomics Study 3 (AEGS3), 658 patients between 2002 and 2016, were included and genotyped (693,931 markers) using the Illumina GSA MD version 1 BeadArray (GSA) at the Human Genomics Facility (HUGE-F in Rotterdam, The Netherlands).

After genotyping, we followed community standard quality control and assurance (QC) procedures to clean the whole-genome data obtained in AEGS1, AEGS2, and AEGS3[55]. In the first step, samples that failed genotyping (due to hybridization) and with low average genotype calling (based on manufacturer recommendations) were excluded. Samples with other issues (patient withdrew from the study), sex discrepancies based on available clinical data, relatedness (pi-hat > 0.20) where none should exist, and inbreeding and contamination (average heterozygosity rate ± 3.0 s.d.) were excluded. Samples that turned out to be within (same study number and UPID) or between (different study number but same UPID) related (multiple) pairs were retained. After these initial sample QC steps, variants were filtered out

based on positional issues compared to the reference (1000G phase 3, version 5, b37)[56], duplicate variants IDs, duplicate variant positions, multi-allelic variants, and Hardy-Weinberg Equilibrium (HWE p-value < 1.0×10^{-3}). After sample and SNP QC, 780 samples (535 males, 245 females) and 409,023 SNPs in AEGS1, 900 samples (605 males, 295 females) and 529,193 SNPs in AEGS2, and 587 samples (407 males, 180 females) and 661,053 SNPs in AEGS3 remained and were included for imputation.
These 2,267 samples (1,547 males, 720 females) and 1,416,258 unique variants across the three datasets were further pre-processed and taken forward for imputation against TOPMed b38 (R3) using TOPMed Imputation Server (https://imputation.biodatacatalyst.nhlbi.nih.gov/) which includes 445,600,184 variants across chromosome 1-22 and X from 133,597 people[57]. After imputation we performed principal component analysis to assess the population stratification against 1000G phase 3 data (version 5, including, b38)[56]; 88 samples project to non-EUR populations, and 20 samples project to non-NL populations with EUR. Finally, we randomly picked 2nd and 1st-degree relatives, replicates, or duplicates and post-imputation QC 2,054 samples (1,398 males, 656 females) remain (1,989 NL/EUR, 59 non-EUR, 6 non-NL) covering 73,603,379 unique variants across chromosome 1-22 and X, 1,436 Y-variants and 193 MT-variants (**Supplemental Figure 7**).

Multiple instance learning

WSI preprocessing

The main preprocessing steps applied to the WSIs are: tissue segmentation, tiling and feature extraction. By segmentation, the tissue is separated from the white background of the scanned image: to do this, a U-Net pre-trained on H&E stained samples from multiple tissue types was used. We subsequently retrieve from the tissue area squared tiles (512×512 pixels, corresponding to about 118×118 microns). Finally, features are extracted from these tiles using a ResNet50 convolutional neural network pre-trained on ImageNet.

IPH prediction

The prediction of intra plaque haemorrhage (IPH) from a stained whole slide image (WSI) relies on the additive multiple instance learning paradigm[58].

Multiple Instance Learning (MIL) is a common framework in digital histopathology: tissue patches are grouped into bags, representing the full sample, and each bag is associated with a label. A bag is labelled positive if it contains at least one positive instance, and negative otherwise. Attention-based MIL is an extension that incorporates an attention mechanism aimed at assigning different levels of importance to instances within a bag. During training, the model learns attention weights, which enable it to focus on the most relevant instances while ignoring less significant ones. This mechanism helps in cases where the predictive information is localized in specific regions, providing also an attention map, which is useful for interpretability [58].

The Additive MIL model builds further upon attention-based MIL through a reformulation in the model's calculations, by replacing attention weights with instance-wise contributions to the logits. The additive structure, beyond making the training more stable and robust, provides a clearer and direct interpretability of the tissue patches that actually contribute to the final output score.

All these improvements combined make the additive MIL model easier to validate, which helps to build trust in the model. Through the visualization of the attentive regions per class, a user can verify that the model's predictions are based upon features that are supported by scientific evidence and similar to those identified by human experts [58].

The additive MIL model has been trained to predict IPH from the stained images based on the manual scoring (presence/absence); therefore, it performs binary classification, but it also provides patch-wise scores that can be considered to compute a proxy of IPH area.

IPH quantification

For each sample, the additive MIL model provides both the probability of the sample having IPH and the patch-wise scores that are used to compute that probability.
We first multiplied these scores by the sample size (to make them comparable across samples with different numbers of patches) and then applied the sigmoid function, to bring scores in the range [0,1]. Positive contributions (patches with a score > 0.75) are hypothesised to be predictive of the IPH area within the plaque, and their percentage with respect to the sample size (total number of patches) is the model-derived IPH amount.

Manual Annotation and Model-Based Intraplaque Hemorrhage Quantification

To establish a reference standard for IPH detection, a clinical expert at UMCU manually annotated 95 WSIs, ensuring that all available histological stains were consistently labelled. The annotation process was performed using QuPath[59], an open-source bioimage analysis tool (https://qupath.github.io/), allowing precise delineation of IPH regions. To prevent bias, the annotator was blinded to the model-generated results. These manually annotated regions were then used for evaluating the performance of our computational model in predicting IPH presence and spatial distribution.

For model-based IPH quantification, a deep learning pipeline was developed to generate patch-wise probability scores that indicate the likelihood of IPH in each region of a WSI. The model was applied to extracted patch features stored in `.h5` files, along with the corresponding spatial coordinates. It returned logits representing the confidence of IPH presence, which were converted into probability scores using a sigmoid activation function. These scores were then scaled by the total number of extracted patches (bag size) to normalize predictions across different WSIs. A pseudo-mask was generated by selecting high-confidence patches, defined as those with a probability score exceeding 0.75, ensuring that only regions with a strong likelihood of IPH were retained.

To visually interpret model predictions, the identified positively contributing patches were overlaid onto the original slide, creating a greenish pseudo-mask representing the predicted IPH areas. Additionally, a spatial heatmap was generated by mapping patch-wise scores back to their corresponding tissue locations. The probability values were visualized using a 'coolwarm' colourmap, effectively highlighting regions with varying degrees of IPH presence. This heatmap was then superimposed onto a downscaled version of the WSI, allowing for a direct comparison between model predictions and manual annotations.

SlideToolkit

The Athero-Express Biobank stores a large amount of scanned histological slides at high magnification. With the growing amount of whole slide images (WSIs), the demand for fast,

accurate, and reproducible histological quantification increases. Since manual quantification is time-consuming and subject to observer variability, we developed slideToolKit[33] to provide a toolset for automated feature analysis and histological quantification. It consists of a collection of open-source libraries and scripts that assist in processing a WSI at each step of the analysis procedure: acquisition, preparation, tiling, and analysis. For the acquisition phase, slideToolKit contains scripts to rename slides, convert slides into other TIFF formats when needed, or extract metadata (e.g. scan time, magnification, image compression). In the preparation step, directories for the downstream analysis are created and segmentation masks are retrieved, that have previously been made in the multiple instance learning (MIL) model's WSI preprocessing step, to hide any artefacts in the WSI during the analysis step. The tiling step splits the WSI into smaller tiles used in the analysis step. The analysis step utilizes CellProfiler[60] (version 4.2.6) to quantitatively measure histological features in the tiles.
SlideToolkit analyses a WSI patch by patch using a predetermined size. Because the Additive MIL models are trained using 512×512 patches, this resolution was also used in the SlideToolkit analysis. This also allowed us to extract the composition for the patches deemed to be in the IPH area. For each stain, we divided either the stained area or cell count by the total tissue size to accurately compare stain density between samples. For EVG, FIBRIN. GLYCC, and SR we used the staining area. We used the cell counts for CD34, CD68, CD66b, SMA, and HE.
We analysed the composition in two ways, first, we related the total composition between plaques with IPH versus plaques without IPH. Second, we compared the composition within plaques between IPH-positive and IPH-negative areas (determined from the attention heatmaps). Because each stain is from a serial section, the images for each stain do not always align with each other. Therefore, without image registration, we cannot use the H&E attention heatmaps to determine the IPH area for each stained image. Instead, we determine this IPH area for each stain using its own trained MIL model, which we use to determine the IPH area composition. As some trained models perform worse than H&E this might impact the composition results.

Logistic regression ensemble model

Each histological stain binds to a specific biomarker, illuminating various cellular and extracellular components within tissue samples. For instance, EVG highlights elastin, Picrosirius Red emphasizes collagen, CD68 targets macrophages, and CD34 is specific for endothelial cells. Training models on individual stains allows for specialized feature extraction, enhancing the ability to identify features related to intraplaque haemorrhage (IPH) by focusing on the unique details each stain reveals.
To see what stain adds the most additional value, we used logistic regression ensemble models to combine H&E (the best-performing single stain model) with other stains to predict IPH presence. We combined the single stain model's IPH probabilities of H&E and the other stain to create a set of 2 probabilities per sample. We used this to train a logistic regression model (Python 3.12.0, scikit-learn 1.4.1) to predict IPH presence. We trained using 10-fold cross-validation to validate the reliability and robustness of the logistic regression model performance.

Prediction of reported symptoms

The prediction of symptoms from the integration of plaque phenotypes with electronic health records (35 features in total) was performed using a machine learning classifier (CatBoost).

We fitted two models, one for stroke vs asymptomatic patients and a second for TIA vs asymptomatic patients. The two cohorts are composed of 1,080 (479 stroke, 601 asymptomatic) and 1,288 (687 TIA, 601 asymptomatic) donors respectively. CatBoost classifiers have been fitted for each task, using 10-fold cross-validation to measure performance; feature importances have been computed for each fold and then averaged.

Differential gene expression analysis

Bulk RNA-Seq counts were transformed using variance stabilization transformation. We fitted linear models to perform differential gene expression on both continuous (amount of IPH, IHC markers quantification) and binary (presence of IPH) phenotypes. Specifically, we modelled gene expression as the sum of the phenotype of interest with other covariates (age, sex and sequencing batch id). The resulting p-values have been corrected using False Discovery Rate. Since the bulk RNA-Seq data contained both gene-level and transcript-level counts, we performed DGE at transcript resolution (without aggregation), where possible.

Functional enrichment analysis and single-cell mapping

In order to identify biological processes, molecular functions, or pathways significantly associated with the set of differentially expressed genes, we performed functional enrichment analysis using decoupleR. We estimated transcription factor activity through the univariate linear model (providing *collectri* as network) and performed pathway activity inference with the multivariate linear model and the weights from *PROGENy*. In both cases, genes have been ordered by the absolute fold change from the DGE analysis. The Over Representation Analysis was run to infer enrichment scores of biological terms from the Human Molecular Signatures Database (MSigDB).
Lastly, as we had access to a single cell reference of atherosclerotic plaque and we wanted to identify the cell types involved in plaque vulnerability regulation, we used the *score_genes* function available in *Scanpy*, providing the significant (p-value < 0.05) DE genes as a gene list.

Genome-wide association study

Genome-wide association study (GWAS) was performed using regenie (https://rgcgithub.github.io/regenie/)[61], which consists of 2 steps. In the first step, a whole genome regression model is fitted using a subset of the genetic markers. This regression model captures a fraction of phenotype variance attributable to genetic effects. In the second step, the predictions from step 1 are used as covariates when each imputed SNP is tested.

It is important to filter out low-quality genetic data to improve the robustness of the analysis. First, using PLINK2 (https://www.cog-genomics.org/plink/2.0/)[62], we converted the VCF files to PGEN (PLINK 2 binary) files to use in the regenie pipeline. We use hard-called genotypes in regenie step 1 and 2. Before processing with step 1 in regenie, the following quality control (QC) steps are taken and executed using PLINK2:

- Call Rate filtering: Removing variants with a call rate of less than 10%, as these may introduce noise and uncertainty.
- HWE filtering: Exclude variants with an HWE p-value higher than $1.0x10^{-3}$, as these might indicate genotyping errors or biases, making them unreliable.
- Minor Allele Frequency (MAF) filtering: Filter out variants with an MAF of less than 10%, as these could lead to spurious associations due to insufficient power.

- Pruning for independent variants: Remove variants with low linkage disequilibrium (LD) with an r^2 threshold of 0.2. By pruning we select variants that provide unique information, improving the efficiency of the analysis.
- Exclude long-range LD blocks: Long-range LD can distort associations by creating artificial correlations. (**Supplemental Table 10**)
- Remove non-autosomal variants: Autosomal variants are generally present in both sexes and across different populations, making findings more universally applicable.
- Exclusion of A/T and C/G variants: Removing A/T and C/G variants helps minimize the impact of genotyping errors.

In both steps 1 and 2 of the regenie pipeline we used age, sex, principal component 1 (PC1), PC2, and year of surgery as covariates. Both regenie steps also used a blocksize (number of variants processed together (bsize)) of 1000.
Post-GWAS was done using Python (version 3.9.18) and GWASLab[63] (version 3.4.44; https://cloufield.github.io/gwaslab/). During post-GWAS processing we perform GWASlab's basic check function which makes sure that 1) SNPIDs are of the form chr:bp; 2) orders the data; 3 all alleles are capitalized; 4) does sanity checks on data. After this we remove duplicate and multi-allelic variants and harmonize the data with Human Genome Build 38 to 1) make sure alleles are oriented according to the reference; 2) assign rsIDs; 3) flip alleles (and effect size) when necessary. Next, we check for palindromic variants and indistinguishable indels and remove these. Last, we filter based on the Derived Allele Frequency (DAF) (>-0.12 and <0.12) to exclude variants where the derived allele is relatively common in the population (above 12%), focusing on rare or less frequent variants.

Imputation

For the survival analysis using a Cox Proportional Hazards Survival Regression model, missing values in the clinical data were imputed using the MICE[64] package in R (version 4.4.1). This package helps to impute missing values with plausible data which are drawn from a distribution specifically designed for each missing data point.

Missing data can be classified into three categories: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR means that the probability of missing a data point is the same for all samples. This implies that the cause of the missing data is unrelated to the data. MAR is broader than MCAR. It means that the probability of missing a data point is the same within groups defined by the observed data. MNAR means that the probability of missing a data point depends on unknown reasons. For example, some patient measurements are not taken if the patient's age is below 40. MNAR is the most complex imputation case and requires varying strategies[65].

In our models features like CRP and HDL, used by Hellings et. al, are missing in more than 40% of the patients (**Supplemental Figure 4)**. For all features with missing values, we assume a case of MCAR or MAR. For all missing variables, we used the predictive mean matching (PMM) imputation method. It is an easy-to-use, versatile, and robust method. PMM fills in a missing entry by randomly picking the value from a set of 5 candidate donors. The set of candidate donors is determined by looking at which regression-predicted values are closest to the regression-predicted value for the missing value from the simulated regression model. This method makes sure that all values are plausible and fall within the observed data range.

Using multiple imputation we create multiple versions of a complete dataset. Each analysis starts with the incomplete dataset and replaces the missing data with plausible data values, based on a determined imputation method. After creating multiple versions of the dataset, they are pooled together into a single final estimate. The differences between imputed values across these datasets reflect the uncertainty of the imputation. We ran the imputation for 10 iterations with 100 imputed datasets per iteration. The average certainty of the imputation is 0.903, with a more stratified certainty per variable in **Supplemental Table 6**. For continuous variables, the certainty measure is based on the standard deviation ratio between imputed and observed values. For categorical variables, the certainty is based on the mode proportion of imputed values compared to observed values. Mode proportion determines how well the dominant category in imputed values matches the observed values. This certainty measure ranges from 0 to 1, where < 0.5 is a very low certainty (Imputed values differ significantly from the original distribution) and > 0.9 is a high certainty (Imputed values closely match the original distribution)

Cox Proportional Hazards Survival Regression

To validate the model's results, we replicated results from studies by Hellings *et al*.[45] and Vrijenhoek *et al.*[35]. The study by Hellings *et al.* showed that intraplaque haemorrhage, scored as Yes/No, was significantly correlated with the event rate after surgery (HR=1.7, 95% CI, 1.2-2.6). The subsequent study by Vrijenhoek *et al.* stratified their patients by sex and found that IPH was only significantly associated with the event rate in men (HR=1.9, 95% CI, 1.2-2.8). In women, there was no significant correlation (HR=1.0, 95% CI, 0.6-1.7).

To reproduce these results we created a Cox proportional hazards survival regression in Python (version 3.12.0) using the package “lifelines”. We used the same covariates as used by Hellings *et al.* While they used the same Athero-Express dataset, exact variable names were not provided. We chose all variables based on the description given in the paper (**Supplemental Table 8**). The study by Hellings *et al.* only included patients up until March 11, 2008, which we therefore also used as a cutoff.

Data and code availability

Model: https://github.com/CirculatoryHealth/PHENOMICL
Downstream analyses: https://github.com/CirculatoryHealth/PHENOMICL_downstream
GWASToolKit: https://github.com/swvanderlaan/GWASToolKit
slideToolKit: https://github.com/swvanderlaan/slideToolKit
GWAS summary statistics: GWAS Catalog accession identifiers will be shared as soon as they are available.
DGEA results: https://doi.org/10.34894/ZODL42.

Links

Roche Ventana iScan HT:
https://diagnostics.roche.com/global/en/products/instruments/ventana-iscan-ht.html
Hamamatsu C12000-22 Digital slide scanner:
https://www.hamamatsu.com/eu/en/product/life-science-and-medical-systems/digital-slide-scanner/index.html
1000G high coverage phase 3 data b38:
https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20220422_3202_phased_SNV_INDEL_SV/

Acknowledgements

We are thankful for the support of the Netherlands CardioVascular Research Initiative of the Netherlands Heart Foundation (CVON 2011/B019 and CVON 2017-20: Generating the best evidence-based pharmaceutical targets for atherosclerosis [GENIUS I&II]), the ERA-CVD program ‘druggable-MI-targets’ (grant number: 01KL1802), the Leducq Foundation ‘PlaqOmics’, ‘AtheroGen’ and ‘COMET’, and the Chan Zuckerberg Initiative ‘MetaPlaq’. The research for this contribution was made possible by the AI for Health working group of the EWUU alliance (https://aiforhealth.ewuu.nl/). The collaborative project ‘Getting the Perfect Image’ was co-financed through use of PPP Allowance awarded by Health~Holland, Top Sector Life Sciences & Health, to stimulate public-private partnerships.

We would like to thank all the (former) employees involved in the Athero-Express Biobank Study of the Departments of Surgery of the St. Antonius Hospital Nieuwegein and University Medical Center Utrecht for their continuing work. In particular we would like to thank (in no particular order) Marijke Linschoten, Arjan Samani, Tim Bezemer, Tim van de Kerkhof, Joyce Vrijenhoek, Ben van Middelaar, Sander Reukema, Sander M. van de Weg, Arjan H. Schoneveld, Robin Reijers, Joëlle van Bennekom, and Bas Nelissen for their work and support on ExpressScan and slideToolKit. Lastly, we would like to thank all participants of the Athero-Express Biobank Study; without you these kinds of studies would not be possible.

Funding sources

Dr. Sander W. van der Laan is funded through EU H2020 TO_AITION (grant number: 848146), and EU HORIZON MIRACLE (grant number: 101115381). Tim S. Peters and Dr. Sander W. van der Laan are funded through HealthHolland PPP Allowance ‘Getting the Perfect Image’. Dr. Clint L. Miller is funded by National Institutes of Health (NIH) grants (R01HL148239, R01HL164577 and U01DK142283), Leducq Foundation network grant ‘COMET’ (24CVD02), Chan Zuckerberg Initiative Data Insights grant ‘MetaPlaq’, and American Heart Association

Transformational Project Award (24TPA1300556). Dr. Clint L. Miller and Dr. Sander W. van der Laan are funded through EU HORIZON NextGen grant (101136962).

Disclosures

Dr. Sander W. van der Laan and Prof. Dr. Gerard Pasterkamp received Roche funding for unrelated work. Dr. Craig A. Glastonbury is an advisor to Population Health Partners (PHP) and their portfolio companies. Dr. Clint L. Miller has received funding from AstraZeneca for unrelated work. Roche, PhP, and AstraZeneca have had no part in this study, neither in the conception, design and execution of this study, nor in the preparation and contents of this manuscript. PHP is unrelated to the topics of advisory work performed to date.

References

1. Ross, R. Atherosclerosis--an inflammatory disease. *N Engl J Med* **340**, 115–126 (1999).
2. Lusis, A. J. Atherosclerosis. *Nature* **407**, 233–241 (2000).
3. van Lammeren, G. W. *et al.* Atherosclerotic Plaque Vulnerability as an Explanation for the Increased Risk of Stroke in Elderly Undergoing Carotid Artery Stenting. *Stroke* **42**, 2550–2555 (2011).
4. Khachigian, L. M. *High-Risk Atherosclerotic Plaques: Mechanisms, Imaging, Models, and Therapy*. (CRC Press, 2004).
5. Sakamoto, A. *et al.* Significance of Intra-plaque Hemorrhage for the Development of High-Risk Vulnerable Plaque: Current Understanding from Basic to Clinical Points of View. *Int J Mol Sci* **24**, (2023).
6. Mekke, J. M. *et al.* Plasma Lipoprotein Lipase Is Associated with Risk of Future Major Adverse Cardiovascular Events in Patients Following Carotid Endarterectomy. *Eur J Vasc Endovasc Surg* **65**, 700–709 (2023).
7. Bentzon, J. F., Otsuka, F., Virmani, R. & Falk, E. Mechanisms of plaque formation and rupture. *Circ Res* **114**, 1852–1866 (2014).
8. Mura, M. *et al.* Carotid intraplaque haemorrhage: pathogenesis, histological classification, imaging methods and clinical value. *Ann Transl Med* **8**, 1273 (2020).
9. Carpenter, A. E. *et al.* CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006).
10. Mekke, J. M. *et al.* The accumulation of erythrocytes quantified and visualized by Glycophorin C in carotid atherosclerotic plaque reflects intraplaque hemorrhage and pre-procedural neurological symptoms. *Sci. Rep.* **13**, 17104 (2023).
11. Van der Donckt, C. *et al.* Elastin fragmentation in atherosclerotic mice leads to intraplaque neovascularization, plaque rupture, myocardial infarction, stroke, and sudden death. *Eur. Heart J.* **36**, 1049–1058 (2015).
12. van der Laan, S. W. *et al.* Variants in ALOX5, ALOX5AP and LTA4H are not associated

with atherosclerotic plaque phenotypes: the Athero-Express Genomics Study. *Atherosclerosis* **239**, 528–538 (2015).

13. Balmos, I. A. *et al.* Intraplaque Neovascularization, CD68+ and iNOS2+ Macrophage Infiltrate Intensity Are Associated with Atherothrombosis and Intraplaque Hemorrhage in Severe Carotid Atherosclerosis. *Biomedicines* **11**, (2023).
14. Ionita, M. G. *et al.* High neutrophil numbers in human carotid atherosclerotic plaques are associated with characteristics of rupture-prone lesions. *Arterioscler. Thromb. Vasc. Biol.* **30**, 1842–1848 (2010).
15. Bennett, M. R., Sinha, S. & Owens, G. K. Vascular smooth muscle cells in atherosclerosis. *Circ. Res.* **118**, 692–702 (2016).
16. Cisternino, F. *et al.* Self-supervised learning for characterising histomorphological diversity and spatial RNA expression prediction across 23 human tissue types. *Nat Commun* **15**, 5906 (2024).
17. 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).
18. Gupta, A. *et al.* SubCell: Vision foundation models for microscopy capture single-cell biology. *bioRxiv* (2024) doi:10.1101/2024.12.06.627299.
19. Sakkers, T. R. *et al.* Atherosclerotic fibrous plaques in females are characterized by endothelial-to-mesenchymal transition and linked to smoking. *medRxiv* (2024) doi:10.1101/2024.10.01.24314739.
20. Lee, J. *et al.* Deep learning segmentation of fibrous cap in intravascular optical coherence tomography images. *Scientific Reports* **14**, 1–13 (2024).
21. Intravascular ultrasound-based deep learning for plaque characterization in coronary artery disease. *Atherosclerosis* **324**, 69–75 (2021).
22. Lv, R. *et al.* Combining IVUS + OCT Data, Biomechanical Models and Machine Learning Method for Accurate Coronary Plaque Morphology Quantification and Cap Thickness and Stress/Strain Index Predictions. *Journal of Functional Biomaterials* **14**, 41 (2023).

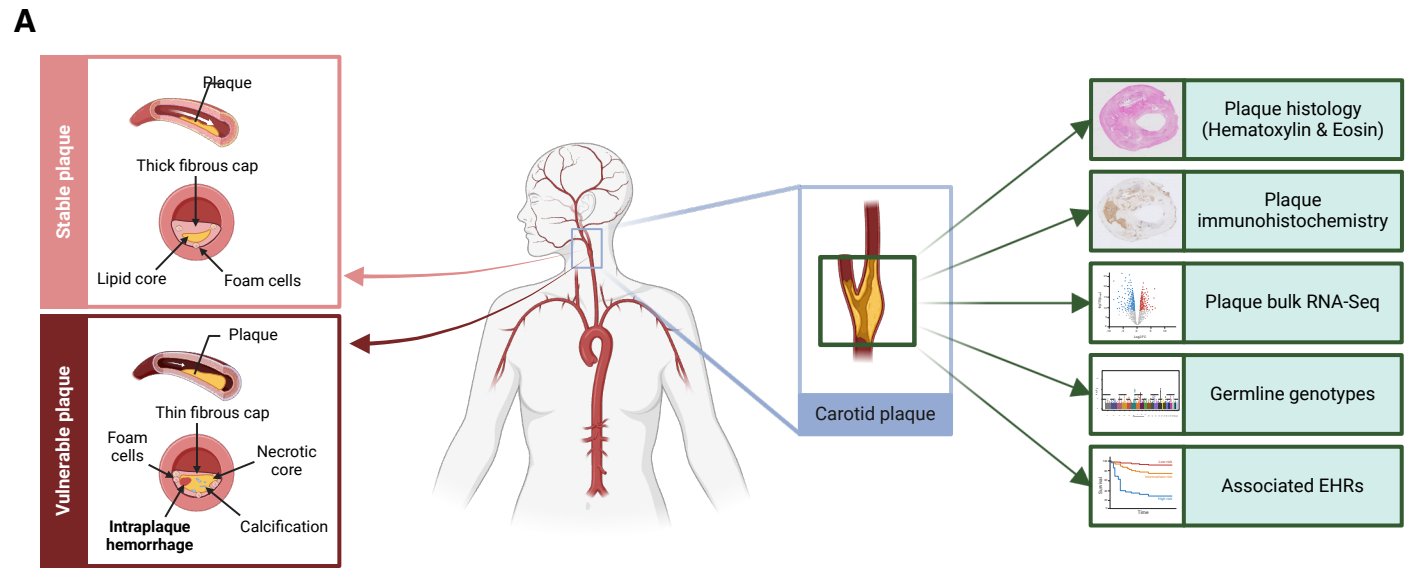
23. Guo, X. *et al.* Predicting plaque vulnerability change using intravascular ultrasound + optical coherence tomography image-based fluid–structure interaction models and machine learning methods with patient follow-up data: a feasibility study. *BioMedical Engineering OnLine* **20**, 1–18 (2021).
24. Wu, L.-D. *et al.* Analysis of potential genetic biomarkers using machine learning methods and immune infiltration regulatory mechanisms underlying atrial fibrillation. *BMC Medical Genomics* **15**, 64 (2022).
25. Deep Learning Analysis of Histologic Images from Intestinal Specimen Reveals Adipocyte Shrinkage and Mast Cell Infiltration to Predict Postoperative Crohn Disease. *The American Journal of Pathology* **192**, 904–916 (2022).
26. Wei, L. *et al.* Artificial intelligence (AI) and machine learning (ML) in precision oncology: a review on enhancing discoverability through multiomics integration. *Br J Radiol* **96**, 20230211 (2023).
27. Koh, D.-M. *et al.* Artificial intelligence and machine learning in cancer imaging. *Communications Medicine* **2**, 1–14 (2022).
28. Ozaki, Y., Broughton, P., Abdollahi, H., Valafar, H. & Blenda, A. V. Integrating Omics Data and AI for Cancer Diagnosis and Prognosis. *Cancers (Basel)* **16**, (2024).
29. Verhoeven, B. A. N. *et al.* Athero-express: Differential atherosclerotic plaque expression of mRNA and protein in relation to cardiovascular events and patient characteristics. Rationale and design. *Eur. J. Epidemiol.* **19**, 1127–1133 (2004).
30. Prasad, K. *et al.* Gender-based in vivo comparison of culprit plaque characteristics and plaque microstructures using optical coherence tomography in acute coronary syndrome. *J. Cardiovasc. Thorac. Res.* **13**, 277–284 (2021).
31. Seegers, L. M. *et al.* Sex differences in culprit plaque characteristics among different age groups in patients with acute coronary syndromes. *Circ. Cardiovasc. Interv.* **15**, e011612 (2022).
32. Konishi, T. *et al.* Plaque histological characteristics in individuals with sudden coronary death. *Vascul. Pharmacol.* **153**, 107240 (2023).

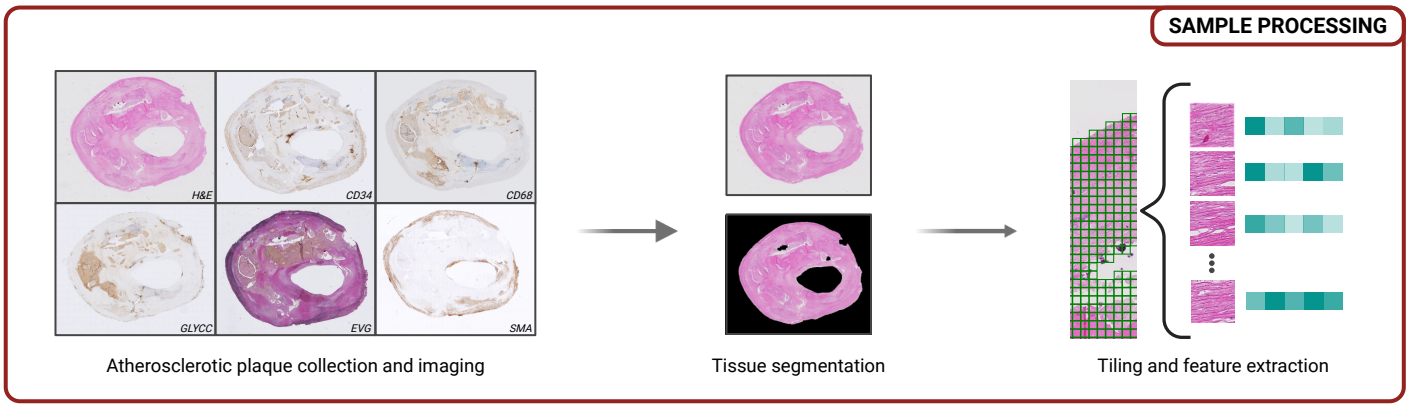
33. Nelissen, B. G. L., van Herwaarden, J. A., Moll, F. L., van Diest, P. J. & Pasterkamp, G. SlideToolkit: an assistive toolset for the histological quantification of whole slide images. *PLoS One* **9**, e110289 (2014).
34. Vrijenhoek, J. E. P. *et al.* High reproducibility of histological characterization by whole virtual slide quantification; an example using carotid plaque specimens. *PLoS One* **9**, e115907 (2014).
35. Vrijenhoek, J. E. P. *et al.* Sex is associated with the presence of atherosclerotic plaque hemorrhage and modifies the relation between plaque hemorrhage and cardiovascular outcome. *Stroke* **44**, 3318–3323 (2013).
36. Cheng, C. *et al.* Heme oxygenase 1 determines atherosclerotic lesion progression into a vulnerable plaque. *Circulation* **119**, 3017–3027 (2009).
37. Cabral-Pacheco, G. A. *et al.* The roles of matrix metalloproteinases and their inhibitors in human diseases. *Int. J. Mol. Sci.* **21**, 9739 (2020).
38. Peeters, W. *et al.* Collagenase matrix metalloproteinase-8 expressed in atherosclerotic carotid plaques is associated with systemic cardiovascular outcome. *Eur. Heart J.* **32**, 2314–2325 (2011).
39. Sluijter, J. P. G. *et al.* Matrix metalloproteinase 2 is associated with stable and matrix metalloproteinases 8 and 9 with vulnerable carotid atherosclerotic lesions: a study in human endarterectomy specimen pointing to a role for different extracellular matrix metalloproteinase inducer glycosylation forms: A study in human endarterectomy specimen pointing to a role for different extracellular matrix metalloproteinase inducer glycosylation forms. *Stroke* **37**, 235–239 (2006).
40. Buniello, A. *et al.* Open Targets Platform: facilitating therapeutic hypotheses building in drug discovery. *Nucleic Acids Res.* **53**, D1467–D1475 (2025).
41. Barton, A. R., Sherman, M. A., Mukamel, R. E. & Loh, P.-R. Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. *Nat. Genet.* **53**, 1260–1269 (2021).
42. Dönertaş, H. M., Fabian, D. K., Valenzuela, M. F., Partridge, L. & Thornton, J. M.

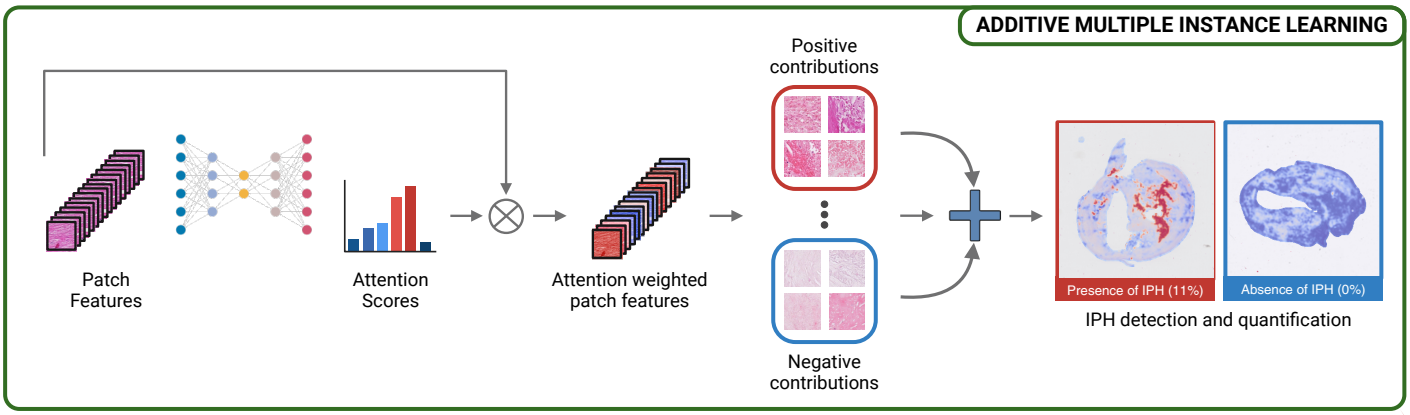
Common genetic associations between age-related diseases. *Nat. Aging* **1**, 400–412 (2021).

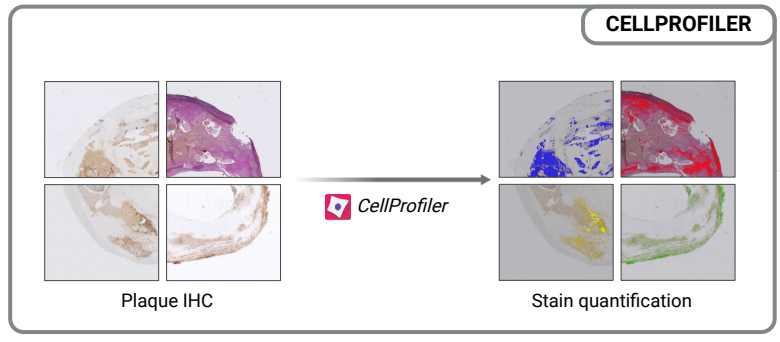
43. Gusev, E. & Sarapultsev, A. Atherosclerosis and inflammation: Insights from the theory of general pathological processes. *Int. J. Mol. Sci.* **24**, (2023).
44. Mendelsohn, M. E. & Karas, R. H. The protective effects of estrogen on the cardiovascular system. *N. Engl. J. Med.* **340**, 1801–1811 (1999).
45. Hellings, W. E. *et al.* Composition of carotid atherosclerotic plaque is associated with cardiovascular outcome: a prognostic study: A prognostic study. *Circulation* **121**, 1941–1950 (2010).
46. Hellings, W. E. *et al.* Intraobserver and interobserver variability and spatial differences in histologic examination of carotid endarterectomy specimens. *J. Vasc. Surg.* **46**, 1147–1154 (2007).
47. Mokry, M. *et al.* Transcriptomic-based clustering of human atherosclerotic plaques identifies subgroups with different underlying biology and clinical presentation. *Nat. Cardiovasc. Res.* **1**, 1140–1155 (2022).
48. Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).
49. Depuydt, M. A. C. *et al.* Microanatomy of the Human Atherosclerotic Plaque by Single-Cell Transcriptomics. *Circ. Res.* **127**, 1437–1455 (2020).
50. Slenders, L. *et al.* Intersecting single-cell transcriptomics and genome-wide association studies identifies crucial cell populations and candidate genes for atherosclerosis. *Eur Heart J Open* (2021) doi:10.1093/ehjopen/oeab043.
51. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
52. Martens, J. H. A. & Stunnenberg, H. G. BLUEPRINT: mapping human blood cell epigenomes. *Haematologica* **98**, 1487–1489 (2013).
53. Alhamdoosh, M. *et al.* Combining multiple tools outperforms individual methods in gene set enrichment analyses. *Bioinformatics* **33**, 414–424 (2017).

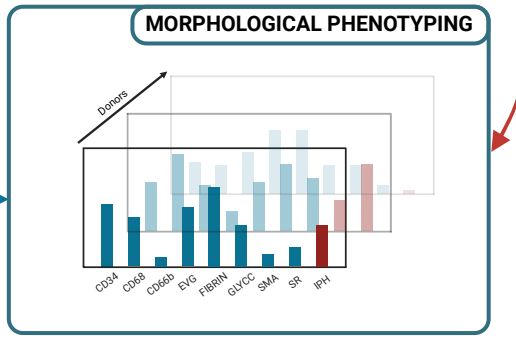
54. van der Laan, S. W. *et al.* Genetic Susceptibility Loci for Cardiovascular Disease and Their Impact on Atherosclerotic Plaques. *Circ Genom Precis Med* **11**, e002115 (2018).
55. Laurie, C. C. *et al.* Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol.* **34**, 591–602 (2010).
56. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440.e19 (2022).
57. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
58. Javed, S. A. *et al.* Additive MIL: Intrinsically Interpretable Multiple Instance Learning for Pathology. *arXiv [cs.CV]* (2022).
59. Bankhead, P. *et al.* QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* **7**, 16878 (2017).
60. Bray, M.-A., Vokes, M. S. & Carpenter, A. E. Using CellProfiler for automatic identification and measurement of biological objects in images. *Curr. Protoc. Mol. Biol.* **109**, 14.17.1–14.17.13 (2015).
61. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
62. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
63. He, Y., Koido, M., Shimmori, Y. & Kamatani, Y. Jxiv https://doi.org/10.51094/jxiv.370 (2023).
64. van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations inR. *J. Stat. Softw.* **45**, 1–67 (2011).
65. van Buuren, S. *Flexible Imputation of Missing Data, Second Edition*. (Taylor & Francis, London, England, 2021). doi:10.1201/9780429492259.

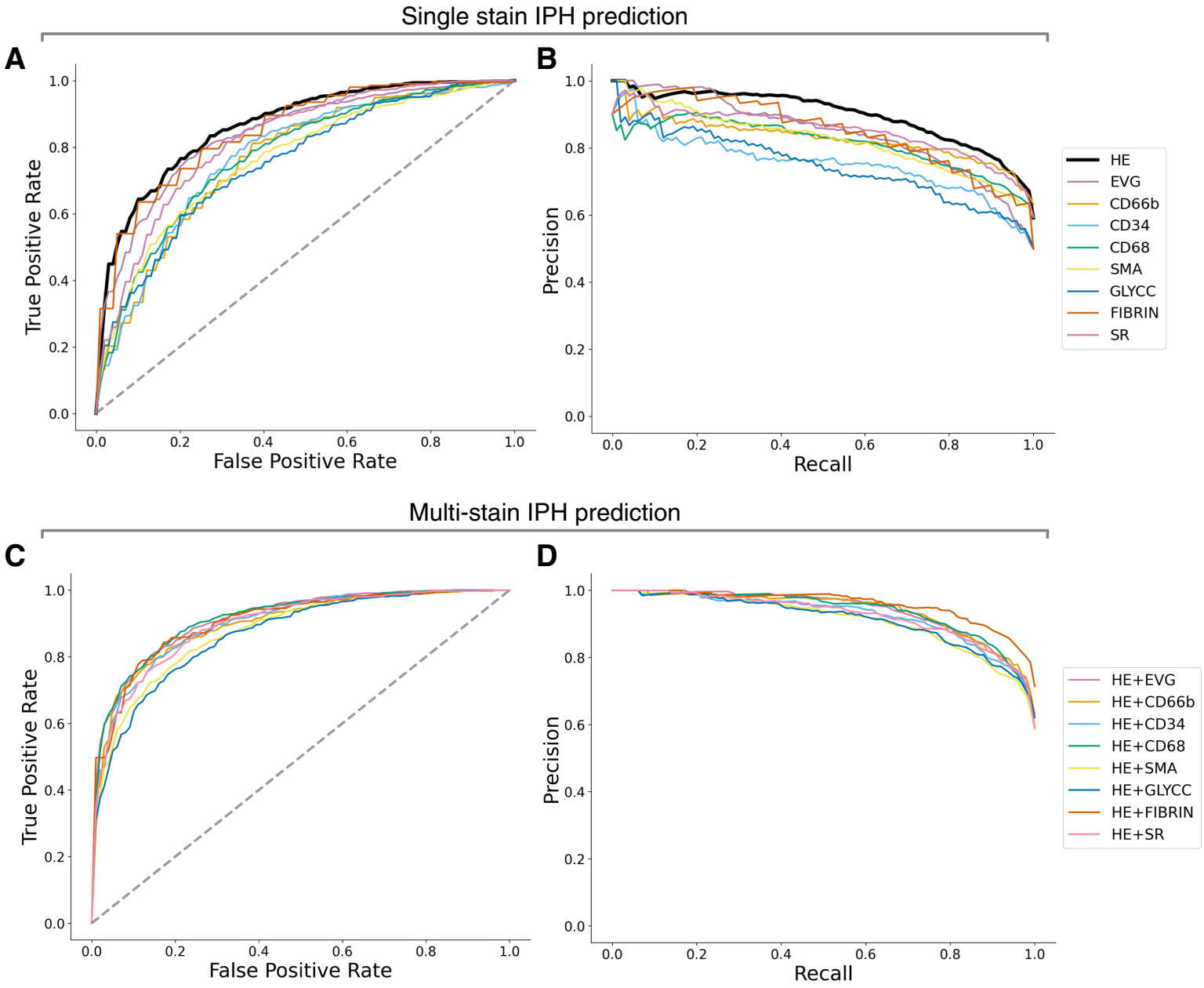


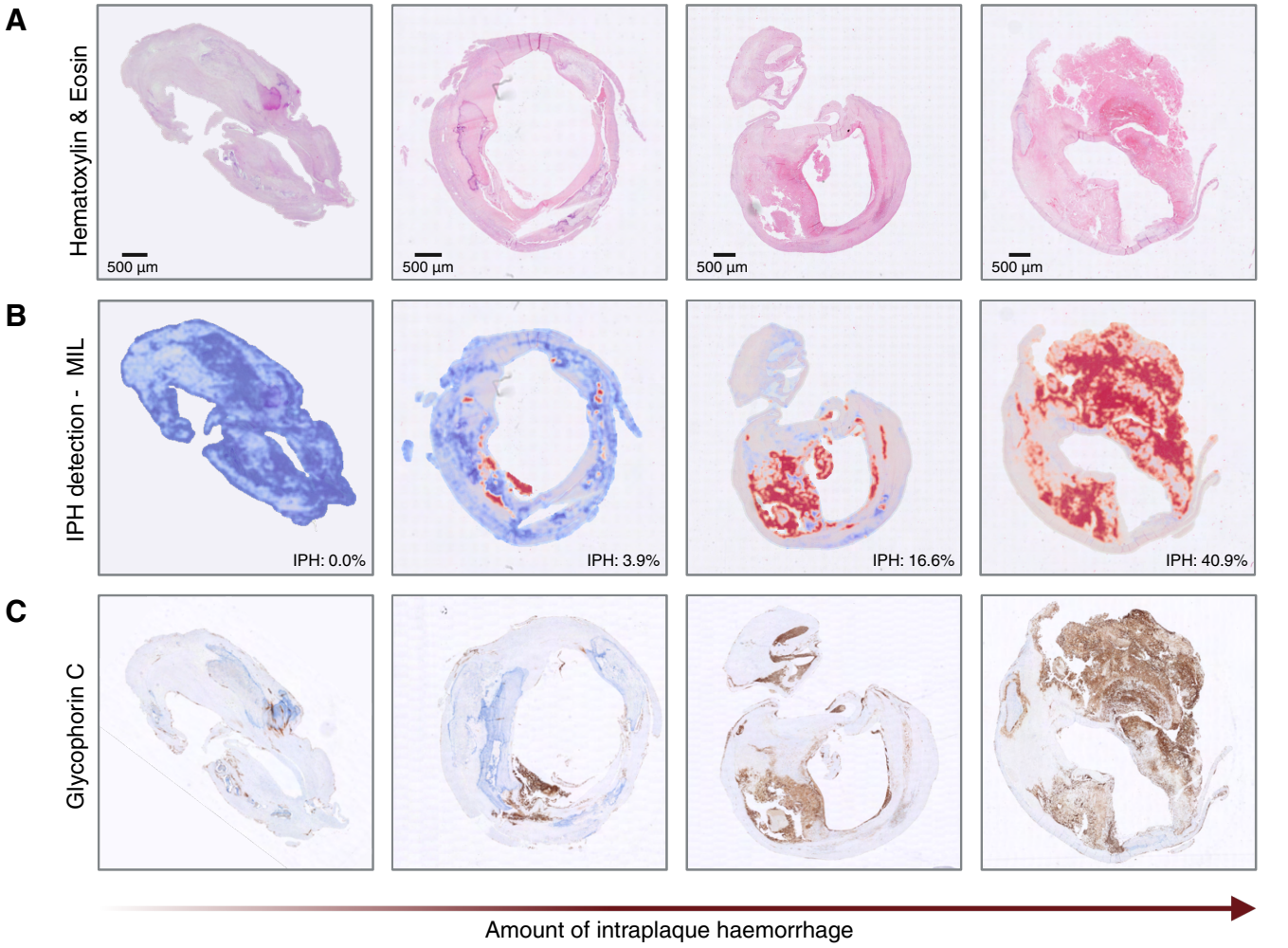


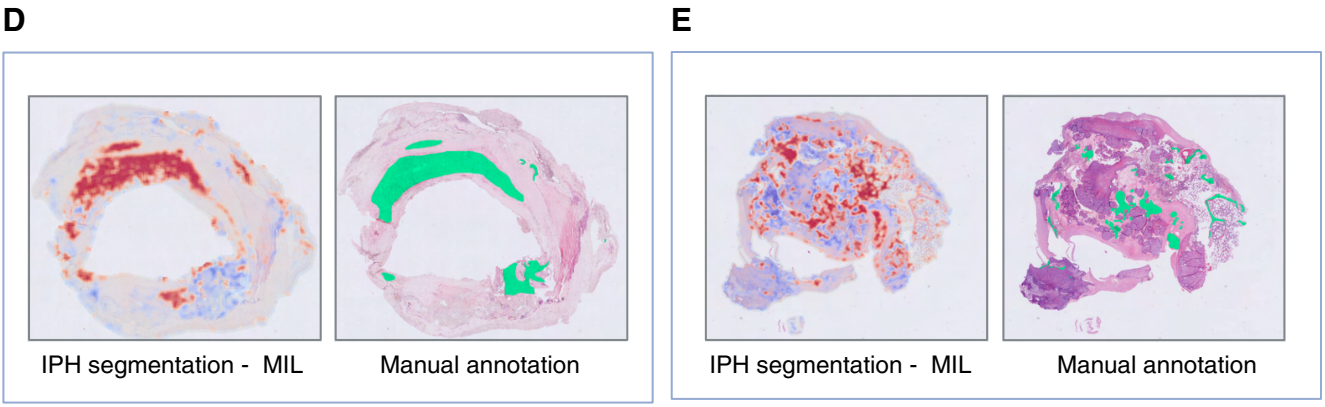


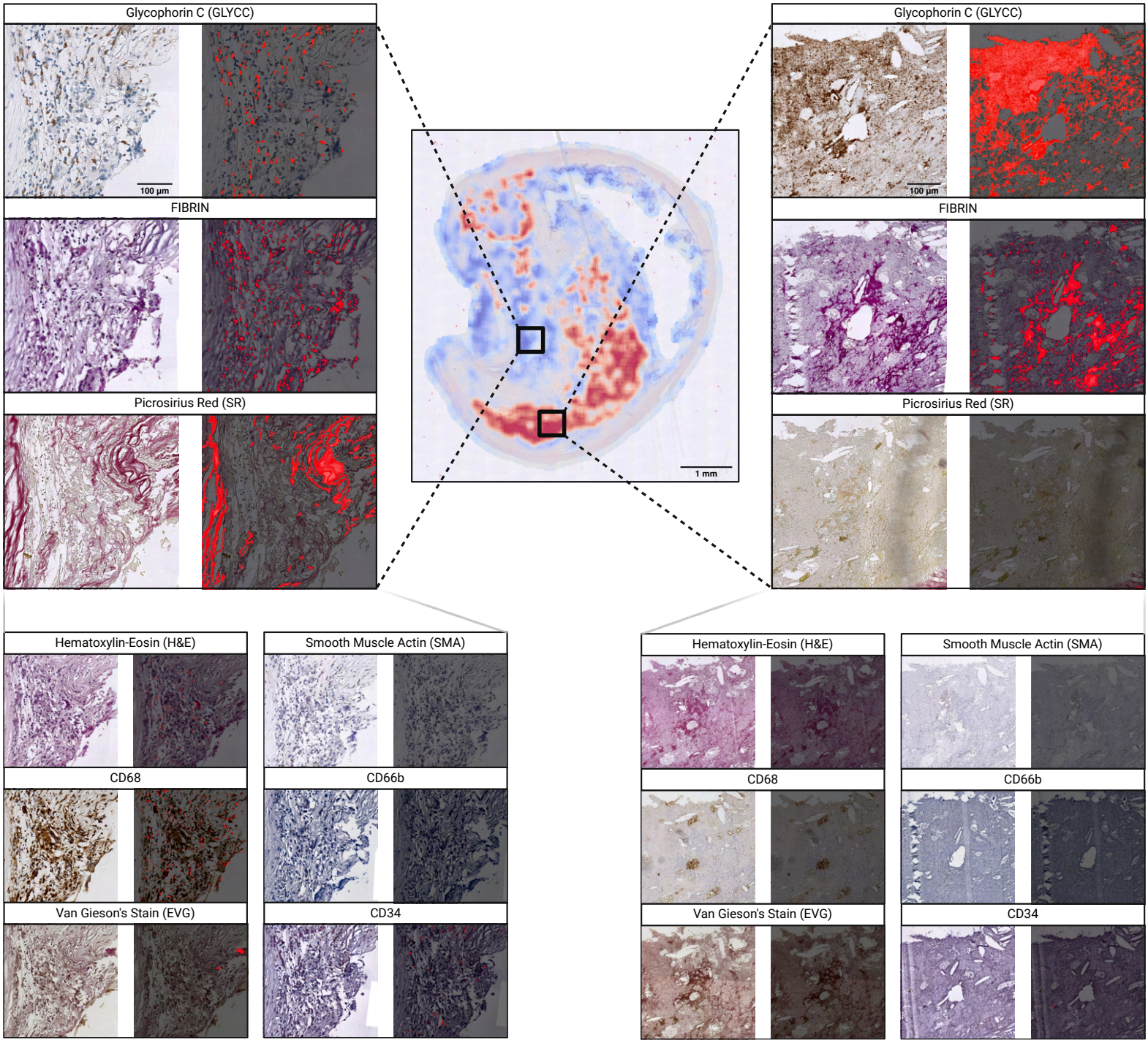


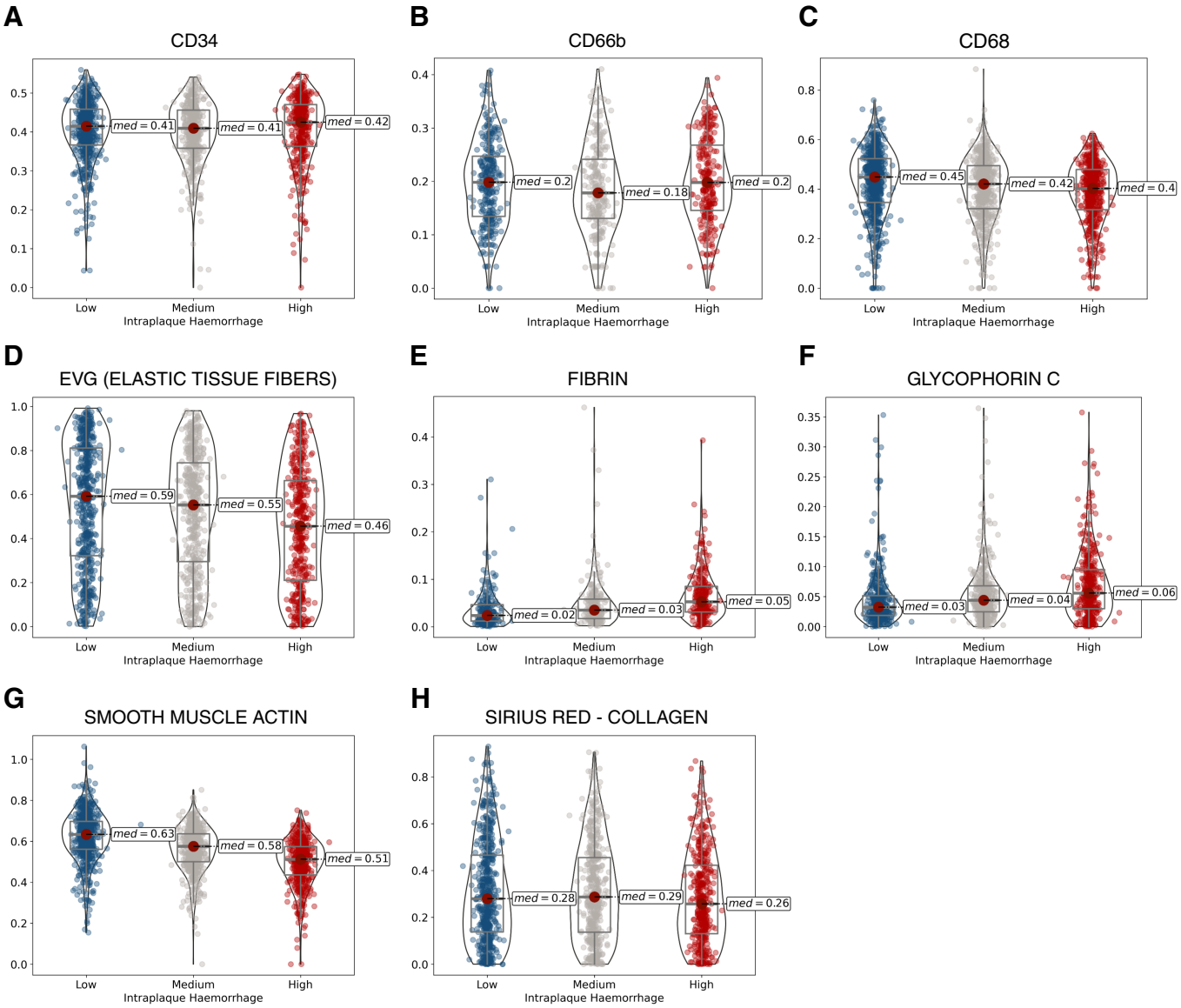


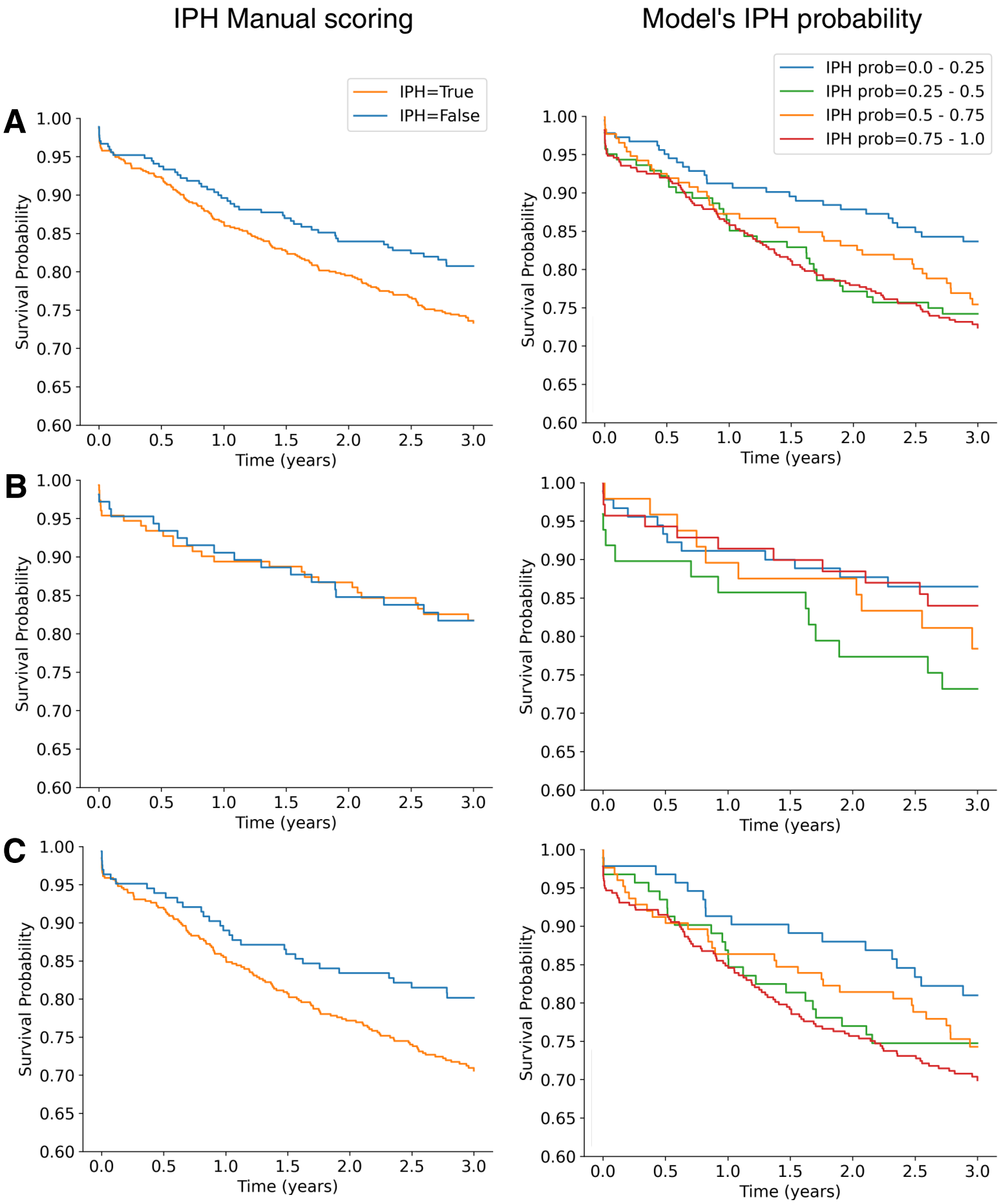


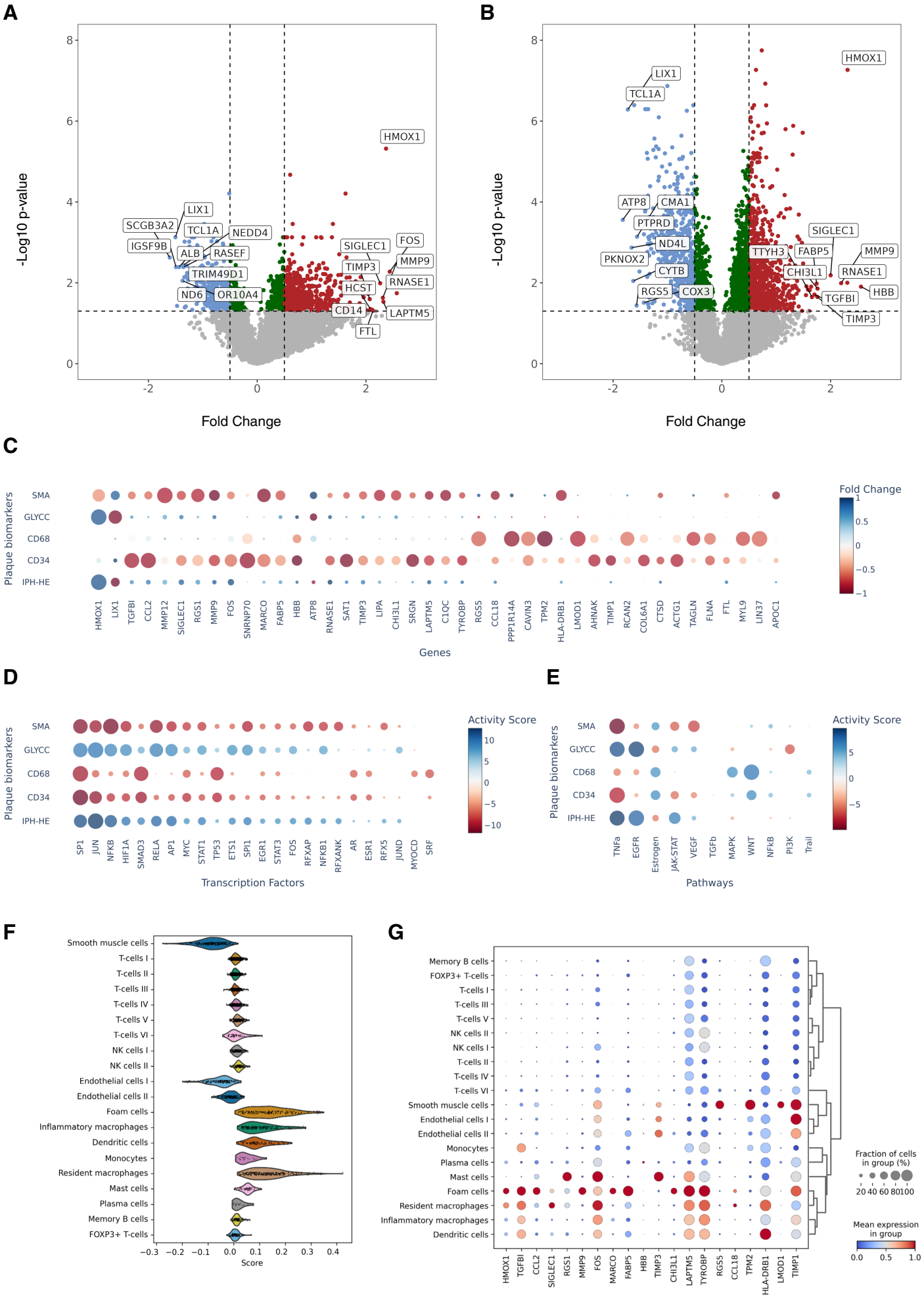


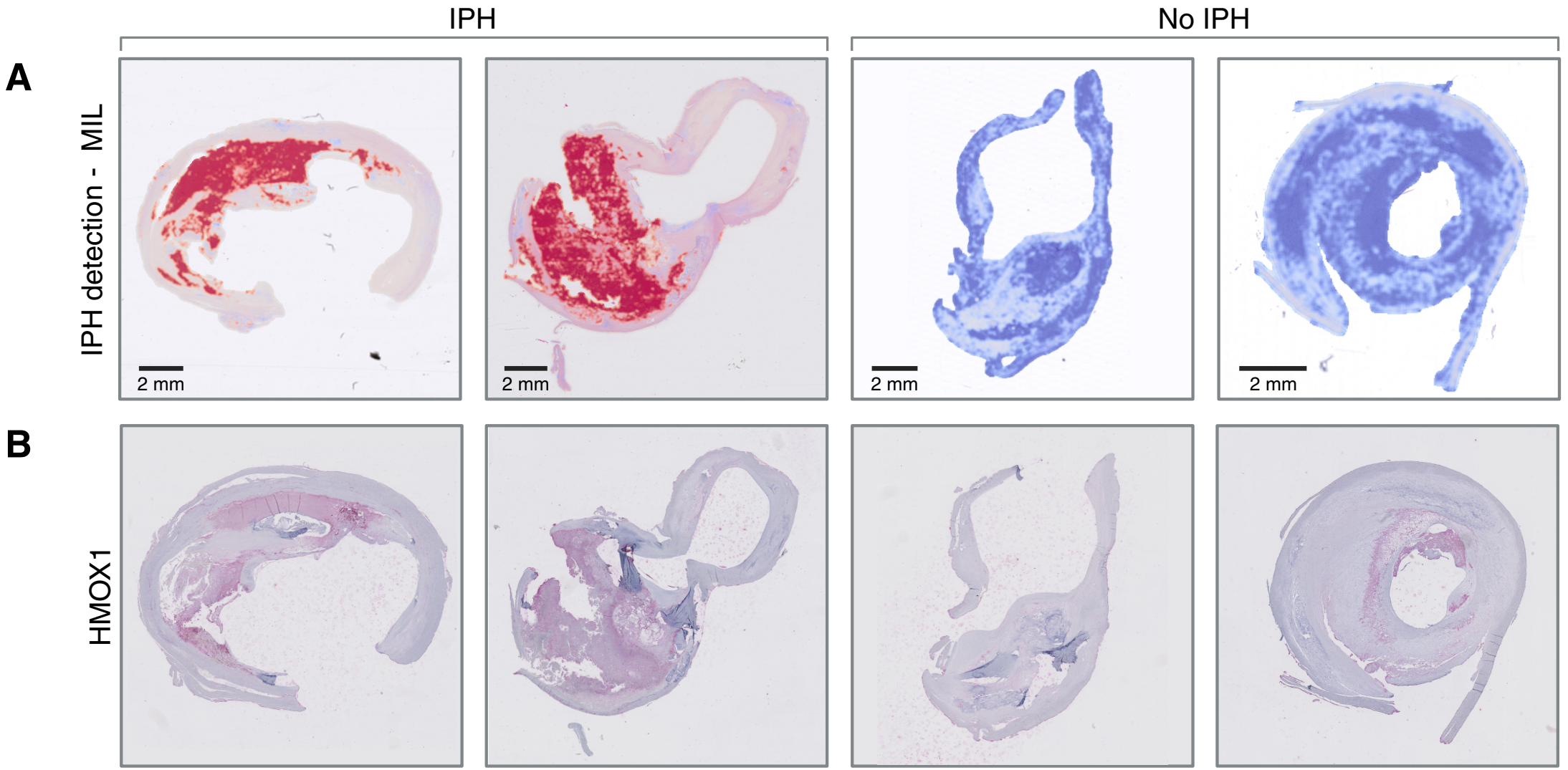


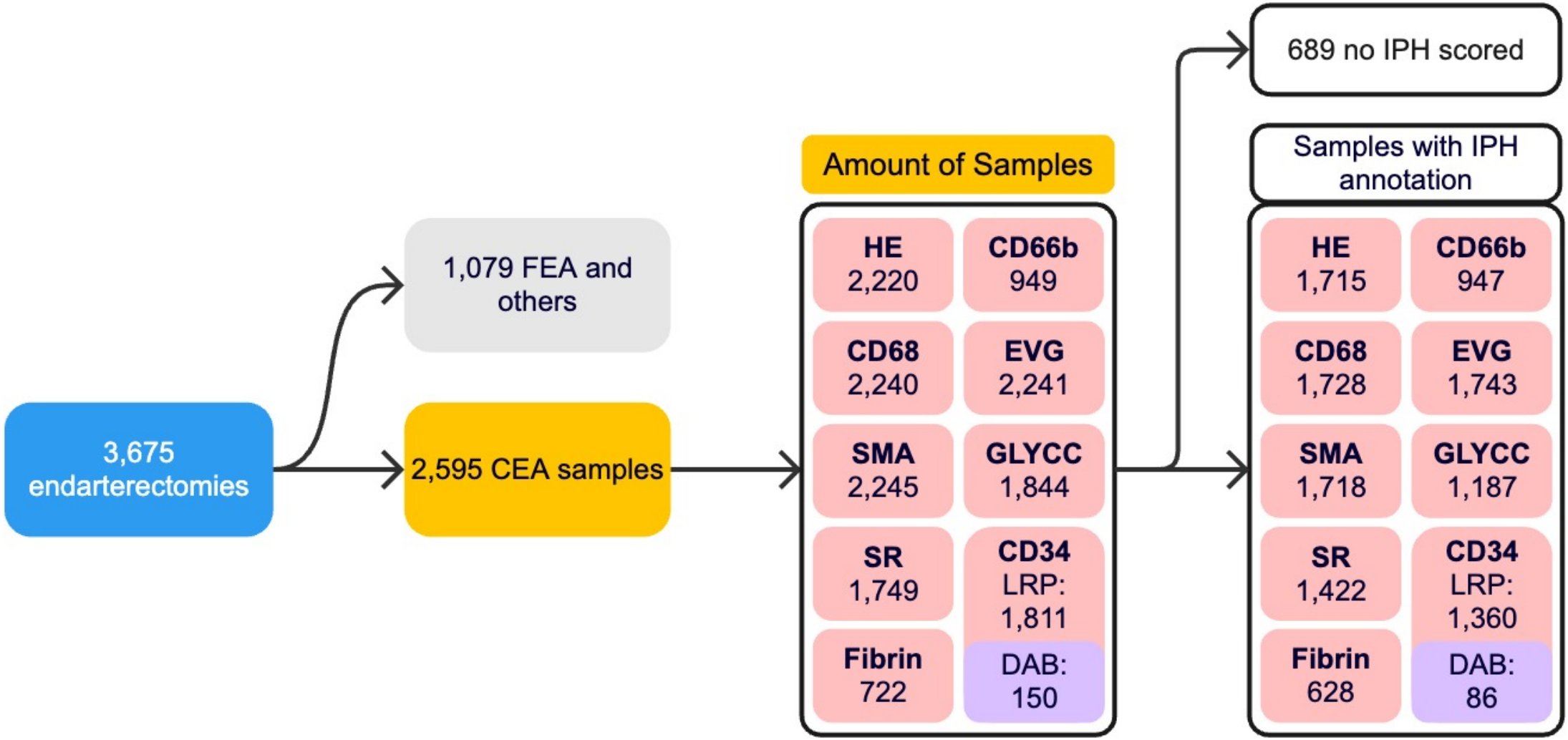


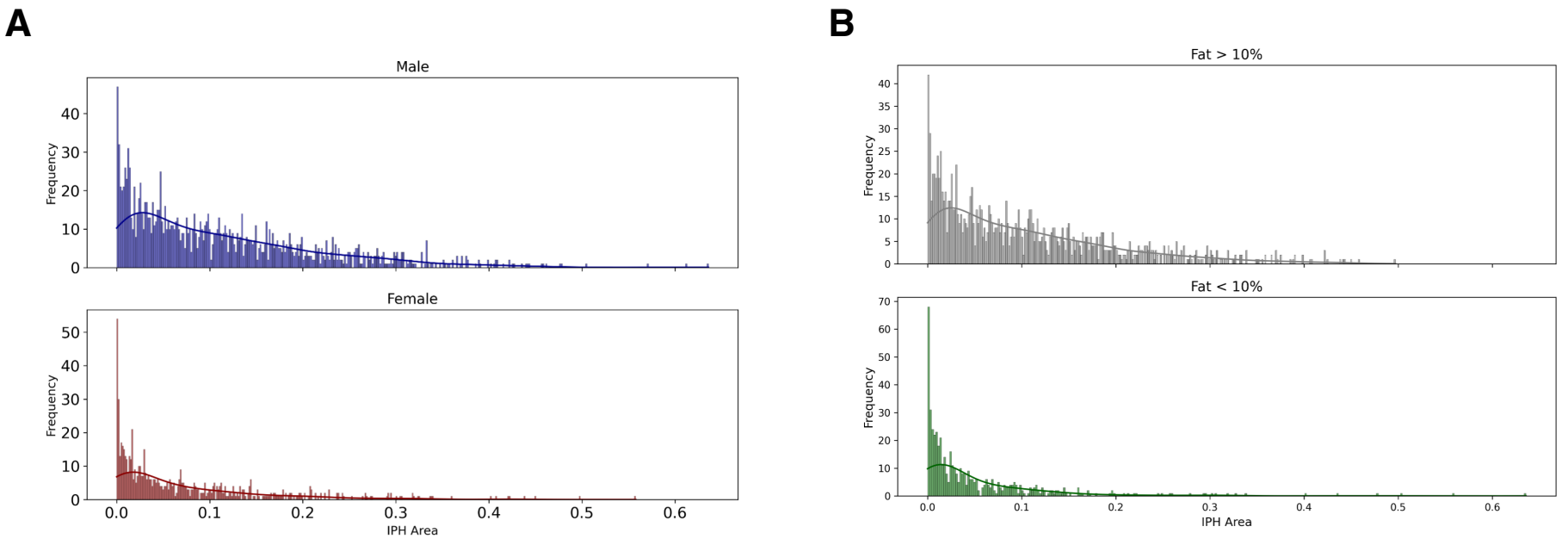




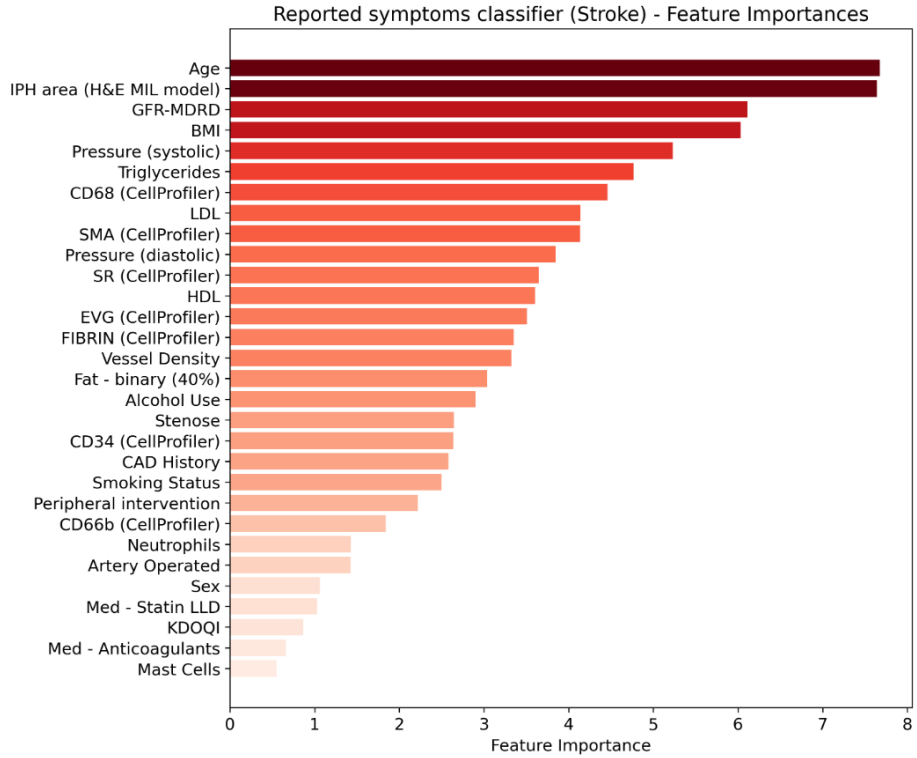



IPH
No IPH
A
IPH detection - MIL
2 mm
2 mm
2 mm
2 mm
B
HMOX1

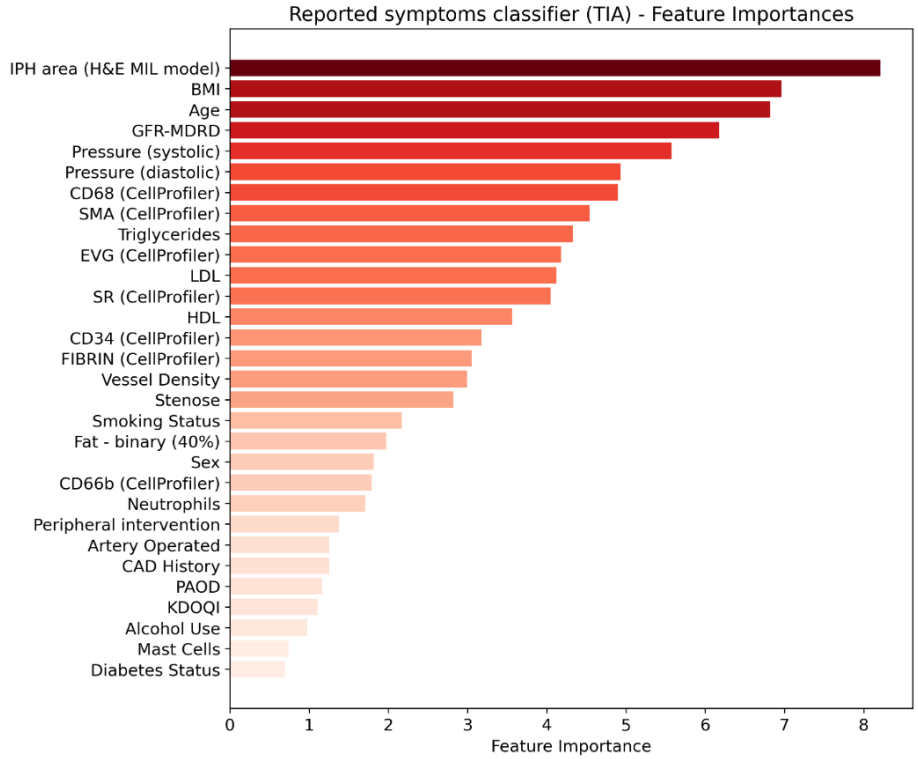

3,675 endarterectomies
1,079 FEA and others
2,595 CEA samples
Amount of Samples
HE 2,220
CD66b 949
CD68 2,240
EVG 2,241
SMA 2,245
GLYCC 1,844
SR 1,749
CD34 LRP: 1,811
Fibrin 722
DAB: 150
689 no IPH scored
Samples with IPH annotation
HE 1,715
CD66b 947
CD68 1,728
EVG 1,743
SMA 1,718
GLYCC 1,187
SR 1,422
CD34 LRP: 1,360
Fibrin 628
DAB: 86


A
Male
Frequency
0
10
20
30
40
Female
Frequency
0
10
20
30
40
50
0.0
0.1
0.2
0.3
0.4
0.5
0.6
IPH Area
B
Fat > 10%
Frequency
0
5
10
15
20
25
30
35
40
Fat < 10%
Frequency
0
10
20
30
40
50
60
70
0.0
0.1
0.2
0.3
0.4
0.5
0.6
IPH Area

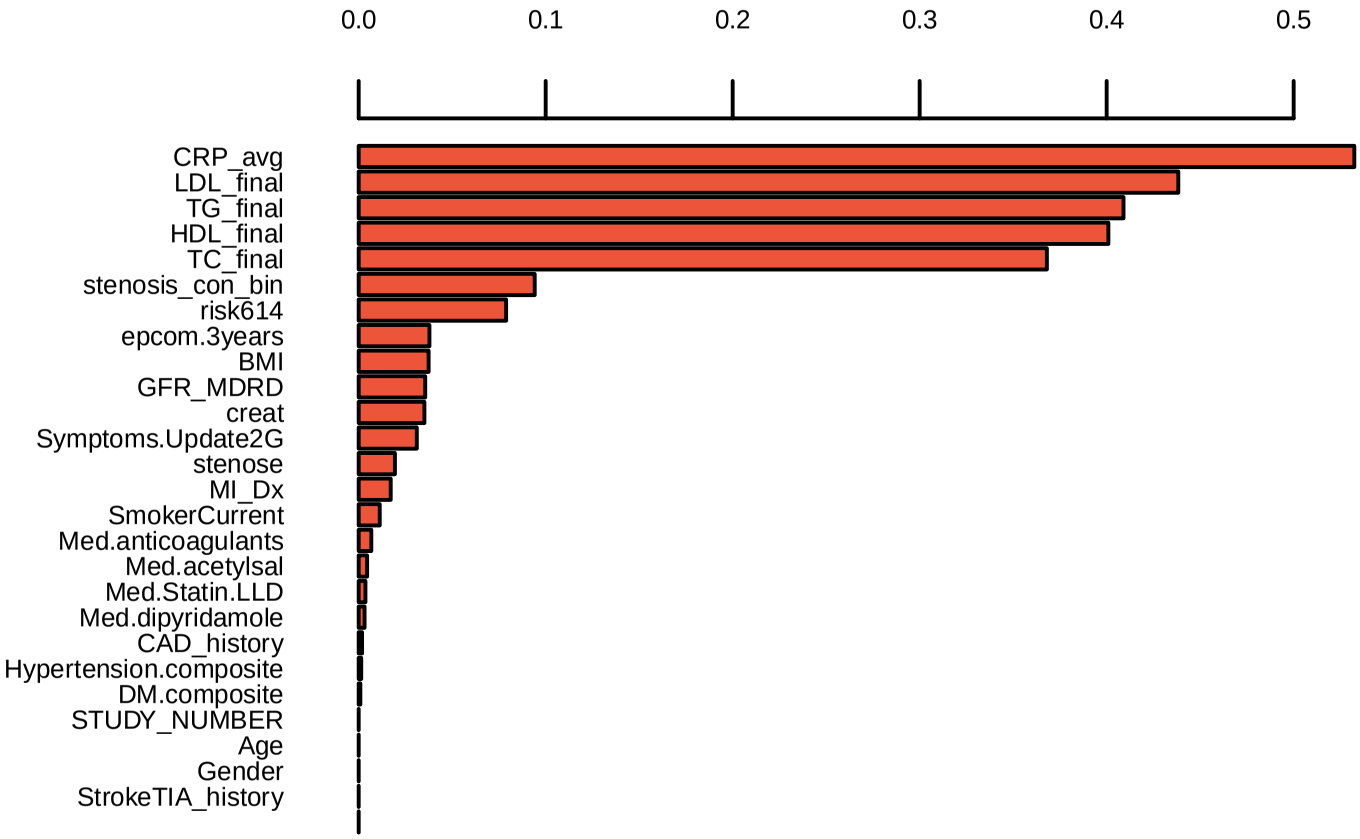
A


Reported symptoms classifier (Stroke) - Feature Importances
Age
IPH area (H&E MIL model)
GFR-MDRD
BMI
Pressure (systolic)
Triglycerides
CD68 (CellProfiler)
LDL
SMA (CellProfiler)
Pressure (diastolic)
SR (CellProfiler)
HDL
EVG (CellProfiler)
FIBRIN (CellProfiler)
Vessel Density
Fat - binary (40%)
Alcohol Use
Stenose
CD34 (CellProfiler)
CAD History
Smoking Status
Peripheral intervention
CD66b (CellProfiler)
Neutrophils
Artery Operated
Sex
Med - Statin LLD
KDOQI
Med - Anticoagulants
Mast Cells
0
1
2
3
4
5
6
7
8
Feature Importance

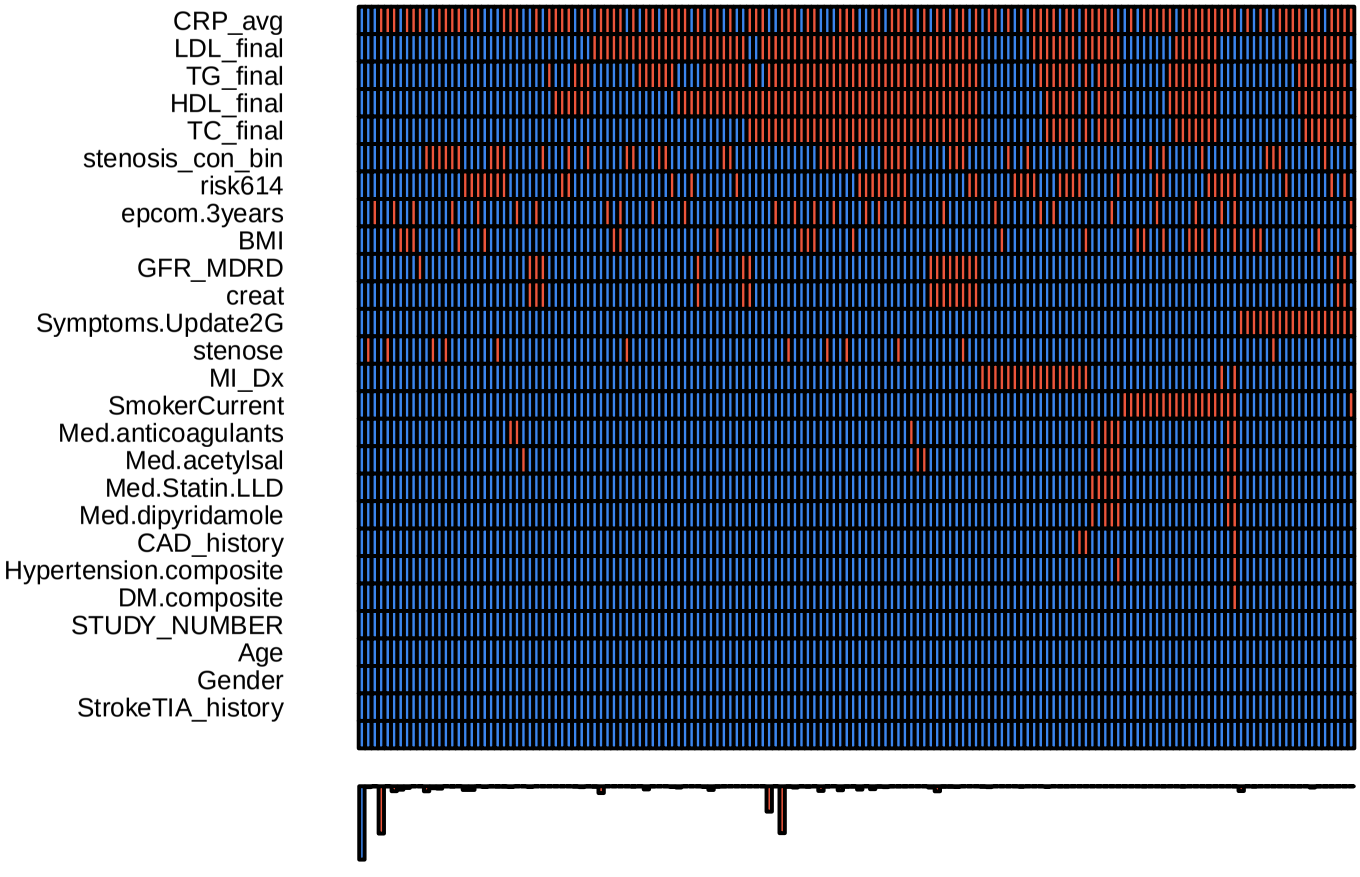
B


Reported symptoms classifier (TIA) - Feature Importances
IPH area (H&E MIL model)
BMI
Age
GFR-MDRD
Pressure (systolic)
Pressure (diastolic)
CD68 (CellProfiler)
SMA (CellProfiler)
Triglycerides
EVG (CellProfiler)
LDL
SR (CellProfiler)
HDL
CD34 (CellProfiler)
FIBRIN (CellProfiler)
Vessel Density
Stenose
Smoking Status
Fat - binary (40%)
Sex
CD66b (CellProfiler)
Neutrophils
Peripheral intervention
Artery Operated
CAD History
PAOD
KDOQI
Alcohol Use
Mast Cells
Diabetes Status
0
1
2
3
4
5
6
7
8
Feature Importance

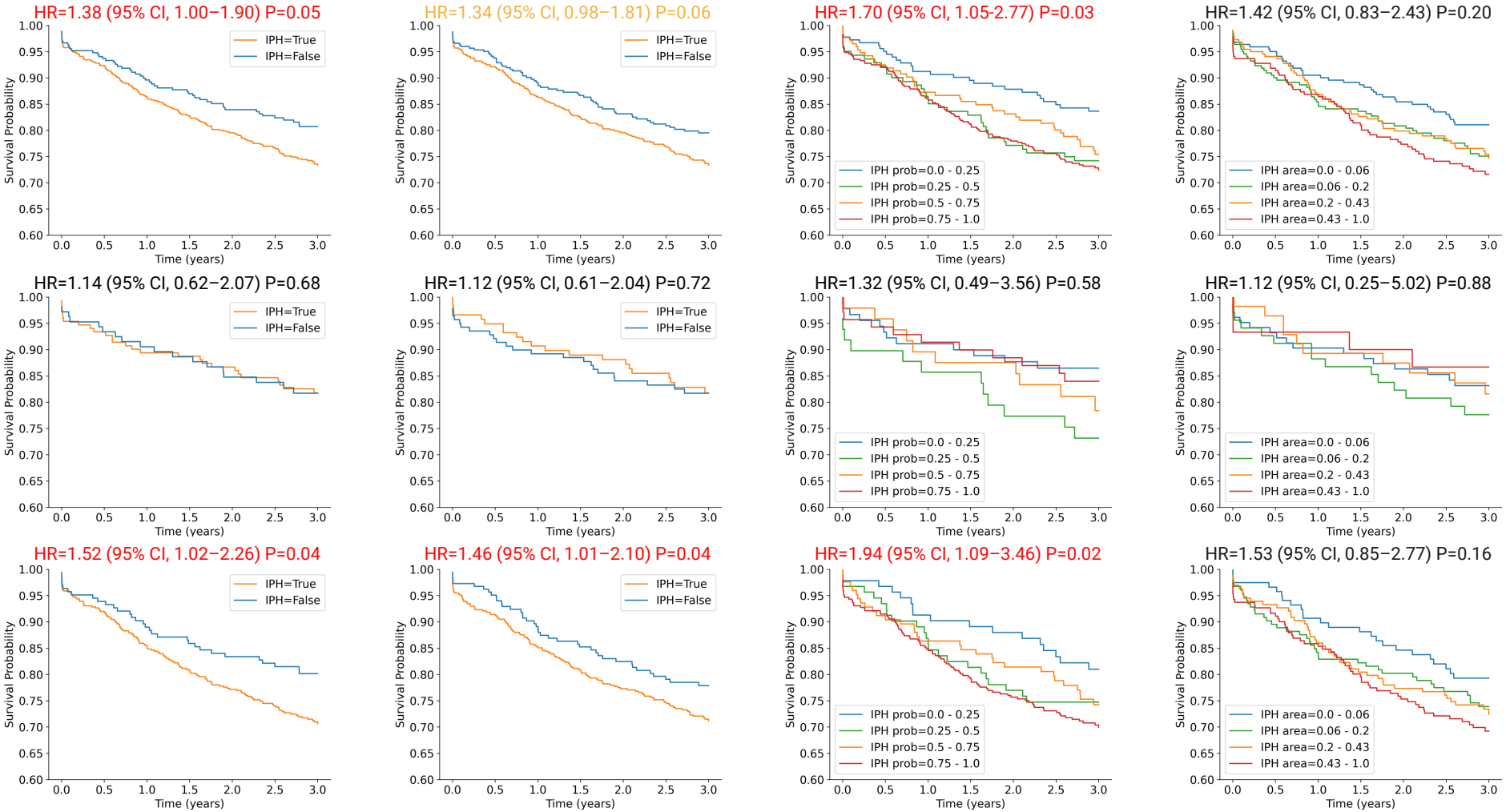
Proportion of missings



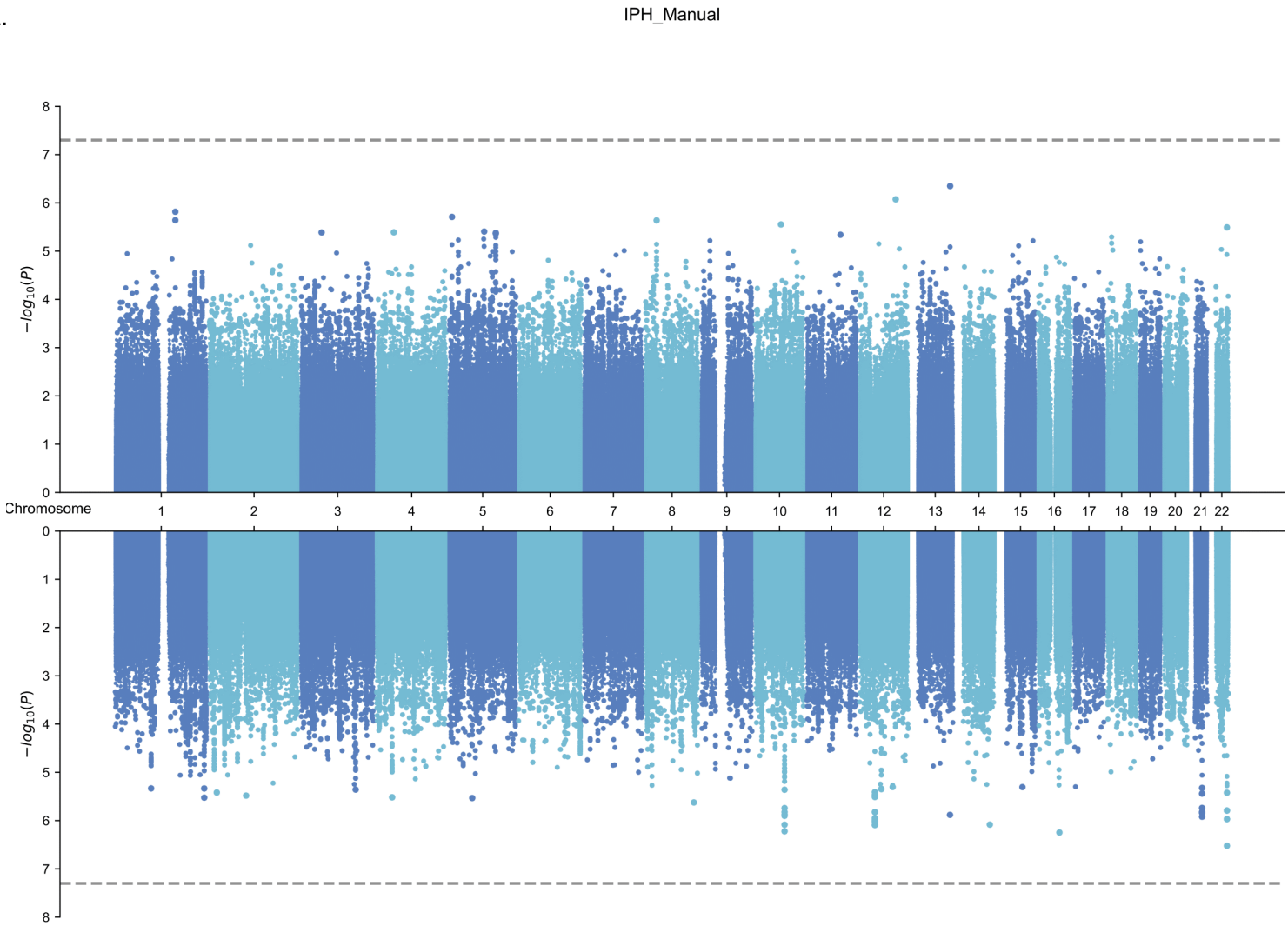
Combinations



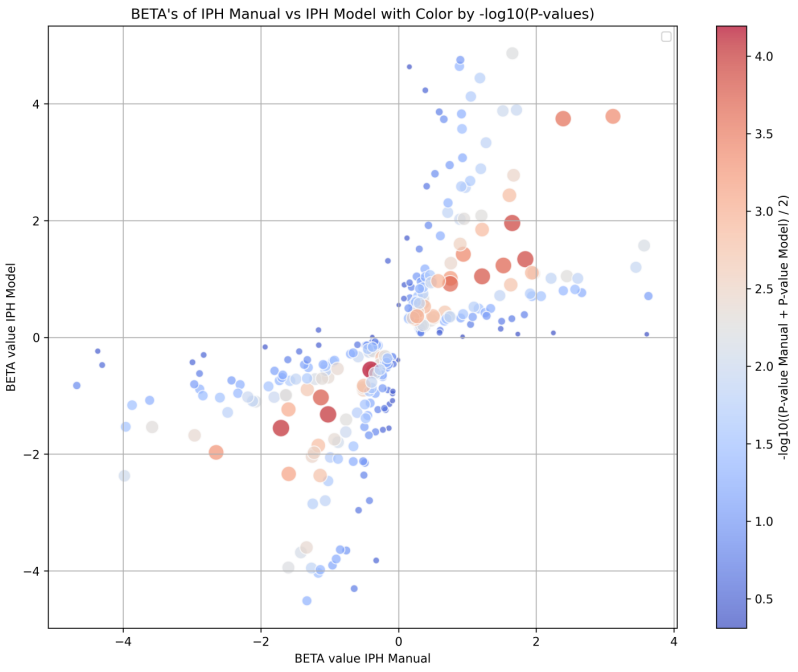
	IPH Manual scoring	Model's IPH Yes/No	Model's IPH probability	Model's IPH area
Overall	HR=1.38 (95% CI, 1.00–1.90) P=0.05	HR=1.34 (95% CI, 0.98–1.81) P=0.06	HR=1.70 (95% CI, 1.05-2.77) P=0.03	HR=1.42 (95% CI, 0.83–2.43) P=0.20
Female	HR=1.14 (95% CI, 0.62–2.07) P=0.68	HR=1.12 (95% CI, 0.61–2.04) P=0.72	HR=1.32 (95% CI, 0.49–3.56) P=0.58	HR=1.12 (95% CI, 0.25–5.02) P=0.88
Male	HR=1.52 (95% CI, 1.02–2.26) P=0.04	HR=1.46 (95% CI, 1.01–2.10) P=0.04	HR=1.94 (95% CI, 1.09–3.46) P=0.02	HR=1.53 (95% CI, 0.85–2.77) P=0.16



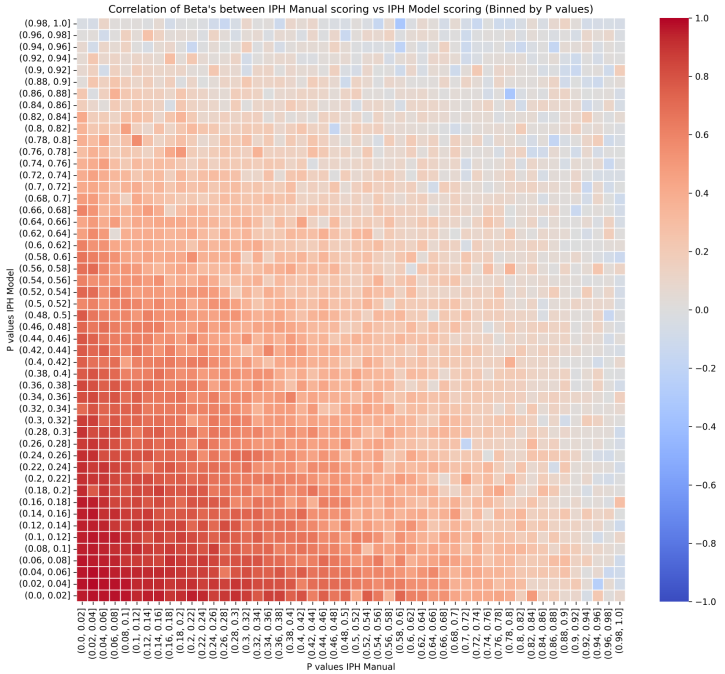
A.

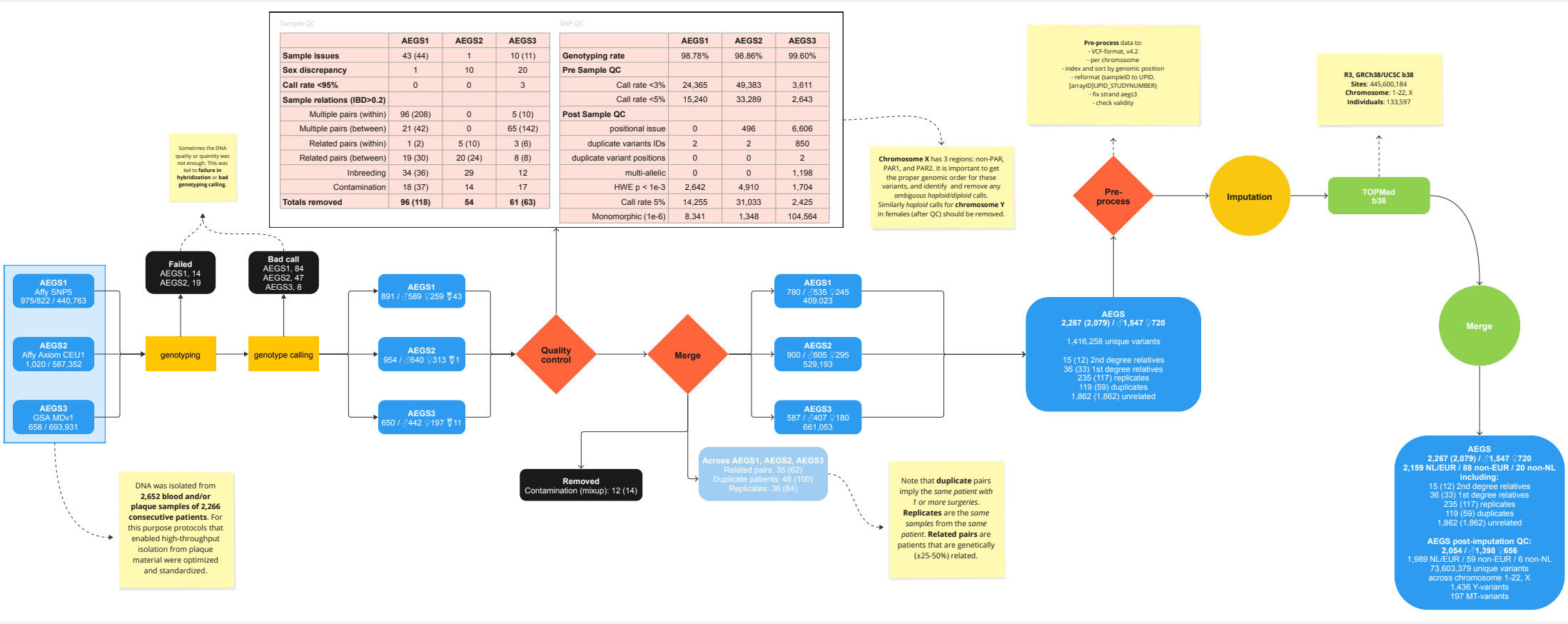


B.



C.





	AEGS1	AEGS2	AEGS3
Sample issues	43 (44)	1	10 (11)
Sex discrepancy	1	10	20
Call rate <95%	0	0	3
Sample relations (IBD>0.2)			
Multiple pairs (within)	96 (208)	0	5 (10)
Multiple pairs (between)	21 (42)	0	65 (142)
Related pairs (within)	1 (2)	5 (10)	3 (6)
Related pairs (between)	19 (30)	20 (24)	8 (8)
Inbreeding	34 (36)	29	12
Contamination	18 (37)	14	17
Totals removed	**96 (118)**	**54**	**61 (63)**

	AEGS1	AEGS2	AEGS3
Genotyping rate	98.78%	98.86%	99.60%
Pre Sample QC			
Call rate <3%	24,365	49,383	3,611
Call rate <5%	15,240	33,289	2,643
Post Sample QC			
positional issue	0	496	6,606
duplicate variants IDs	2	2	850
duplicate variant positions	0	0	2
multi-allelic	0	0	1,198
HWE p < 1e-3	2,642	4,910	1,704
Call rate 5%	14,255	31,033	2,425
Monomorphic (1e-6)	8,341	1,348	104,564