Phenotyping of *Klf14* mouse white adipose tissue enabled by whole slide segmentation with deep neural networks

Ramón Casero[1], Henrik Westerberg[1], Neil R Horner[1], Marianne Yon[1], Alan Aberdeen[2,3], Vicente Grau[2], Roger D. Cox[1], Jens Rittscher[2], Ann-Marie Mallon[1]

(1) MRC Harwell Institute (United Kingdom)
(2) Institute of Biomedical Engineering, University of Oxford (United Kingdom)
(3) Ground Truth Labs Ltd (United Kingdom)

Abstract

White adipose tissue (WAT) plays a central role in metabolism, and multiple diseases and genetic mutations cause its remodeling, most notably obesity, which has reached pandemic levels. WAT is present in subcutaneous (SAT) and visceral (VAT) depots, and its main components are white adipocytes. Quantitative analysis of white adipocyte size and counts is of great interest to understand physiology and disease, due to intra- and inter-depot heterogeneity, as well as better prognosis for hypertrophy than hyperplasia, and for SAT expansion than VAT expansion. H&E histology of whole depot cross-sections provides excellent approximation of cell morphology and preserves spatial information. Previous studies have been limited to window subsampling of whole slides, and cell size analysis.
In this paper, we present a deep learning pipeline that can segment overlapping white adipocytes on whole slides and filter out other cells. We also propose a statistical framework based on linear models to study WAT phenotypes with three interconnected levels (body weight BW, depot weight DW and quartile cell area). Applying it to find *Klf14* phenotypes in mice using 147 whole slides of WAT H&E histology, we show sexual dimorphism, and different effects between depots, heterozygous parent of origin for the KO allele and genotype (WT vs. Het). In particular, whether variables are correlated (DW vs. BW and cell area vs. DW), and statistical differences between fitted linear models. We also find significant differences between hand-traced or window subsampling datasets and whole slide analysis. Finally, we provide heatmaps of cell size for all the slides, showing substantial spatial heterogeneity and local spatial correlations.

Introduction

White adipose tissue (WAT) provides the body's main long-term energy storage and plays a central role in metabolism. It is distributed in subcutaneous (SAT) and visceral (VAT) depots, with SAT depots located abdominally, gluteofemorally, intramuscularly and in the upper body

above and below the fascia, whereas VAT depots are omental, mesenteric, perirenal, retroperitoneal, gonadal and pericardial[1,2]. The main components of WAT are unilocular lipid-filled white adipocytes or fat cells, comprising 90% of WAT mass, but less than 20% to 25% of cell population count[3,4]. WAT also contains blood vessels, adipocyte precursor cells, lymph nodes and nerves of the sympathetic nervous system and immune cells[5], which interact with white adipocytes in adipose expansion and in metabolic disease[6–9].

Healthy white adipocytes may present inter-depot heterogeneity in size, with SAT cells generally larger than VAT[10–12]. There is also a sex effect, as females have significantly smaller adipocytes than males in VAT[11]. Females have larger adipocytes than males in SAT[11,13], but not significantly after adjusting for BMI, age and ancestry[11]. Obesity, or excessive WAT expansion occurs through two mechanisms: initial increase in cell size (hypertrophy) followed by increase in cell numbers (hyperplasia), the latter having poorer prognosis[14]. Obesity has reached pandemic levels and is strongly associated with a higher risk of type 2 diabetes, cardiovascular disease and cancer, as well as other chronic diseases such as fatty liver disease, hyperlipidaemia, hypertension, gout, restrictive lung disease, stroke, dementia, gallbladder disease, degenerative arthritis, and infertility[14–19]. Where the expansion occurs is also important, as VAT expansion (especially omental and mesenteric) is correlated with higher disease risk, whereas SAT expansion can have a protective metabolic effect[20–22]. Further, relative anthropometric measures of fat distribution such as waist-hip-ratio adjusted for BMI (WHRadjBMI) capture body fat percentage, and cardiovascular disease and type 2 diabetes risk, and may be better than BMI measures[17,23,24]. Adipocyte mean area correlates with obesity measured through BMI, more strongly in VAT than SAT[11]. The study of WAT remodeling is interesting beyond obesity, as WAT expansion or wasting can be caused by certain diseases such as hypogonadism, Cushing's syndrome, HIV, parasitic infection, or cancer-induced cachexia[2].

The laboratory mouse is the leading model organism for the study of human disease, due to 99% of its genes having human orthologs, a wide catalogue of inbred strains and mutant models, and ease of breeding[25]. Depending on the strain, 16 week mice on regular diets vary from 16.4% to 43.1% fat percentage in males and from 16.7% to 43.5% in females[26]. For the C57BL/6NTac strain that we use in this work, fat percentage varies from 29.0% in males to 23.2% in females on control diets, and from 43.3% in males on a 12 week high fat diet (HFD) to 31.6% in females on a 2 week HFD[27]. As in humans, mice present inter-depot heterogeneity, with visceral gonadal rather than SAT or visceral mesenteric expansion being associated with metabolic disorders[28]. Common mouse models to study obesity are numerous genetics models such as the ob/ob, db/db, POMC, MC4 knockout, ectopic agouti and AgRP, FABP4-Wnt10b, LXRβ−/−, Sfrp5−/− and Timp−/− models, surgical/chemical models and diet induced obesity[2,29]. Other models to study WAT phenotypes are R6/2 and CAG140 for Huntington's disease[30] and CRH for Cushing's syndrome[31]. In this paper, as an exemplar we focus on the C57BL/6NTac *Klf14* knockout model (*Klf14*$^{tm1(KOMP)Vlcg}$) previously developed by Small *et al.*[13]. The *KLF14* (Krüpple-Like family 14) transcription factor is associated with metabolic syndrome and regulates gene expression in adipose tissue. This single exon gene is imprinted and only expressed from the maternally inherited allele[32], and is expressed more highly in females than males[22]. Homozygous females with the *KLF14* risk allele had significantly larger SAT white adipocytes in the Oxford Biobank (BMI-matched)[13]

and GTEx (unmatched)[11] datasets, compared to non-risk homozygous females. The ENDOX dataset showed the same trend, but no statistical significance possibly due to the smaller data sample[11]. By contrast, VAT did not show significant adipocyte size differences in the GTEx[11] dataset.
Therefore, for human or animal model analysis, unbiased, high-throughput, global WAT quantitative analysis is of great interest. Adipocytes can be collagenase-isolated, counted and measured with a series of methods: microscope[33,34], hemocytometer[35], Coulter counter[36] and flow cytometry[37]. However, those approaches consume the tissue and discard its spatial information, whereas whole slide imaging using fluorescent or conventional brightfield microscopy preserves tissue architecture within their slices. Given a sufficiently large field of view, ideally surveying the entire depot area, 2D assessment provides an excellent approximation of cell morphology. Hence our work focuses on the analysis of digitised Hematoxylin and Eosin (H&E) stained tissue sections.

In the first part of this work, we tackle adipocyte segmentation. Manual segmentation of adipocytes is highly accurate[6], but also labour intensive and slow. Larger scale studies require semi- or automatic instance segmentation methods, that need to be robust against preanalytical variation in histopathology (e.g. staining variability, tissue deformations, tears) and imaging artefacts (e.g. out of focus regions, bubbles). Early approaches were either based on hand-tailored features combined in *ad hoc* pipelines, including colour conversion, median filters, mathematical morphological operators, thresholding and watershed algorithms[38–42]; or based on training a pixel classifier with feature vectors extracted by a set of predetermined general-use filters[43].

Advances in deep learning have revolutionised biomedical image analysis, for instance cell detection and segmentation[44,45]. In addition, these methods provide new ways for phenotyping specific cell types[46]. DeepCell[47,48] replaces predetermined filters[43] by deep convolutional neural networks (DCNNs) that learn optimal feature extraction for the target cell and microscopy modality, although its fully connected layers restrict input images to a pre-arranged size. To overcome this, Adipocyte U-Net[11] uses a Fully Convolutional Network (FCN)[49], so that images of variable size can be processed, up to the GPU memory limit. DCNNs successfully tackle stain variability and other colour variations using their generalisation capabilities, transfer learning, data augmentation (geometric transformations and colour) or a combination thereof. Even so, a DCNN-based whole H&E slide WAT segmentation pipeline needs to address several design decisions that we briefly review next.

Segmentation by pixel classification (typically as background, cell boundary or cell interior) is widely used[11,46,47], but the results may need to be regularised, for example with an active contour[47]. In addition, segmentation results tend to be worse where membranes touch, have less definition or are damaged. Segmentation results have been shown to improve by replacing pixel-classification by regression of the Euclidean distance transform (EDT)[50,51], i.e. distance of each pixel to the closest membrane point (see *Supplementary material* for insight in pixel-classification vs. EDT). Then, watershed seeds can be computed with peak detection[51] or with a DCNN trained as an object detector on the EDT[50], although neither can differentiate between adipocytes or other objects in the EDT. In this paper, we propose an

EDT DCNN followed by a contour detector DCNN and watershed for full segmentation of H&E images.

The aforementioned segmentation approaches do not tackle white adipocyte overlap (see Fig. OVERLAP). Cell overlap has been tackled with a number of options such as a Physics model of light attenuation through the cytoplasm[52] to ISOODL[53], which rotates the plane of each cell in 3D space so that they do not overlap, but those increase the problem complexity. A general solution is using a single Shot Multibox Detector followed by individual cell segmentation with a U-Net[46], but this approach does not suit our whole-tile segmentation. Instead, we propose a DCNN based on QualityNet1(Huang, Wu, and Meng 2016) that corrects each segmented object to account for cell overlap.

Efficient computation requires pre-segment tissue areas to avoid segmenting large areas of empty background space[54,55]; we tackle this problem with traditional image processing techniques. In addition, tissue areas are typically too large at full resolution to process in random-access or GPU memory, and need to be tiled and the results stitched together. Uniform titling is commonly used[11,55,56] as it is simple to implement, but the tile overlap needed to avoid dropping cells on the tile seams produces redundant computations; in this work, we propose an adaptive tiling algorithm to reduce that burden. Furthermore, it is necessary to differentiate between the cells of interest (mature white adipocytes) and other WAT components and image artefacts. For this, tiles can be accepted or rejected as a whole by an InceptionV3 network[11]. This, however, has poor granularity and favours areas away from tissue edges and where other components are less prevalent. For full granularity, one can first detect valid cells and then segment them[46], or as we do in this paper, first compute a whole-tile segmentation and then classify each object as a valid or invalid white adipocyte.

Building a training dataset for training/validating DCNNs can be done by cropping small histology windows, labelling them as valid/invalid for processing, and/or manually hand tracing the cells they contain. Previous multiple-cells-per-window approaches[11,47,50] need that all training pixels are labelled as either background / membrane / cell interior, because they all contribute to the network's loss function. But in practice, windows often contain ambiguous pixels, due to damaged, overlapping or unclear membranes. Furthermore, windows with non-adipocytes are precluded, or those objects need to be labelled as background or a new class, something laborious if they present intricate boundaries. Alternatively, one-cell-per-window approaches[46] can be trained granularly, as each training image contains a single cell. However, this also introduces redundant computation, as each training image must allocate space around the cell to provide spatial context. In this paper we search for a compromise, with a multiple-cells-per-window approach that can leave pixels unlabelled.

We address the challenges above to propose a whole slide white adipocyte segmentation pipeline called DeepCytometer. The challenges include: 1) whole slide processing of all tissue, ignoring the background; 2) tiling overlap compromise between segmenting all cells and reducing redundant computations; 3) colour variability in the slide; 4) cell segmentation considering that white adipocytes present as mostly background surrounded by a thin membrane, and that membranes can touch, overlap, be damaged or have poor definition; 5)

differentiate between adipocytes and other WAT components, as well as image artefacts; 6) choice of training scheme, e.g. one-window multiple-cells vs. one-window one-cell, full vs. partial segmentation. We validate the segmentation results with summary statistics (Dice coefficient and relative area error) as in previous literature, but also propose to examine segmentation errors as a function of cell size, to assess whether all subpopulations are equally well segmented. We integrate DeepCytometer with the open source web-based application AIDA (github.com/alanaberdeen/AIDA) to navigate whole slides with the segmentation results.

In the second part of this paper, we phenotype *Klf14*$^{tm1(KOMP)Vlcg}$C57BL/6NTac mice WAT, analysing DeepCytometer segmentations of 147 whole slides of H&E histology. We extend previous approaches that quantify median cell area from BMI-matched subjects[13] or mean area[11], and propose a phenotype framework with three interconnected linear model levels (body weight, depot weight and quartile cell area). Finally, we provide heat maps of cell area in whole slides, for qualitative assessment of spatial heterogeneity of subpopulations.

We provide all the code for the pipeline and experiments, and trained weights (github.com/MRC-Harwell/cytometer). We also provide the histology images, hand-traced and pipeline segmentations (TODO: upload to zenodo.org).

Results

In this section we present the DeepCytometer pipeline and its validation. Then, we present an analysis of WAT from *Klf14*$^{tm1(KOMP)Vlcg}$ C57BL/6NTac mice, using tissue samples and additional data generated as part of the Small *et al.* 2018 study[13]. This analysis is based on cell areas derived from automatic segmentations from our pipeline and body and depot weight. It should be noted that the single exon *Klf14* gene is imprinted and only expressed from the maternally inherited allele[32]. However, our current preliminary analysis of this dataset does *not* take into account this *Klf14* monoallelic expression, which we will include in our next round of analysis.

A main contribution of this paper is our DeepCytometer pipeline to segment white adipocytes from whole H&E histology slides. The pipeline (Fig. PIPE(a)) performs a coarse segmentation of the tissue, uses an adaptive tiling algorithm to select image blocks that fit in GPU memory, performs colour correction and feeds the blocks to a white adipocyte segmentation sub-pipeline (Fig. PIPE(b)) based on DCNNs. The sub-pipeline works by segmenting all objects in the image block, then classifying which ones are white adipocytes, and correcting their outlines to account for cell overlap. Pixels that belong to cropped cells on the edges are flagged to be processed in neighboring image blocks. A detailed description of slide preprocessing can be found in *Methods - Coarse tissue mask and adaptive tiling for full processing of whole slides*. The segmentation sub-pipeline is described in *Methods - Segmentation sub-pipeline*, with details of its constituent Deep CNNs, in *Methods - Deep CNN architectures*.

In this subsection we present three groups of experiments: “Adaptive tiling computational load reduction” quantifies the reduction in the number of pixels that need to be processed with our adaptive tiling algorithm compared to uniform tiling. “Tissue CNN validation” shows that the segmentation sub-pipeline correctly identifies white adipocytes and rejects other types of elements in the histology image with high sensitivity and specificity. “Segmentation validation” shows that the segmentation sub-pipeline outlines white adipocytes with low area errors, such that it provides a good representation of the adipocyte population in the slide.

Adaptive tiling computational load reduction

We measured the reduction of computational load provided by our adaptive tiling compared to uniform tiling of the tissue region with overlapping blocks, by comparing the number of pixels each approach needs to process. Uniform tiling was produced by splitting the image into (L_{max}, L_{max}) square blocks, where L_{max}=2,751 pixels is the maximum tile length allowed in our adaptive algorithm. Blocks overlapped by $2R_{\max} + ERF$ on each side, where $R_{\max} = 179$ pixels is the radius of the largest circular cell accepted by the pipeline, and $ERF = 131$ pixels is the maximum Effective Receptive Field of the CNNs. The sum $A_{\text{total,uniform}}$ of the areas of all the uniform blocks containing tissue in an image were compared to the sum $A_{\text{total,adaptive}}$ of the areas of the adaptive blocks (Fig. ADAPTBLOCK). The average ratio from the 147 whole slides used in the phenotyping experiments was $A_{\text{total,adaptive}}/\ A_{\text{total,uniform}}$ = 0.86 ± 0.13, corresponding to a reduction of 16.59% in the number of processed pixels (and correspondingly, time), from an average of $2.11 \cdot 10^{-9}$ pixel (uniform) to $1.81 \cdot 10^{-9}$ (adaptive) per slide.

Tissue CNN validation

The Tissue CNN was validated on 126 training images using 10-fold cross validation. We calculated the receiver operating characteristic (ROC) curve for the classification of white adipocytes vs. “other” objects, weighted by the number of pixels in each object (experiment details in *Methods - Segmentation sub-pipeline*, curve in Fig. CLASS_ROC(a)). The classifier performs very well, with area under the curve = 99.59%, and pixel-wise false positive rate (FPR) = 1.80% and true positive rate (TPR) = 97.71% for a white-adipocyte classification threshold $Z_{obj} \geq 0.5$. The low FPR means that cell population studies will contain a negligible number of false objects, and the high TPR indicates that the vast majority of white adipocytes will be detected in the slides. We also provide TPR and FPR values for other thresholds in Fig. CLASS_ROC(b).

Segmentation sub-pipeline validation

We validated the segmentation sub-pipeline on 55 hand-segmented images using 10-fold cross validation, both for the Auto and Corrected methods (experiment details in *Methods - Segmentation sub-pipeline*). We computed the Dice Coefficient (DC) between pairs of DeepCytometer and hand-traced (ht) cell contours, matched as described in *Methods*. For the Auto method we obtained DC_{Auto} = 0.89 (median), 0.85 (mean), 0.10 (std), and for the Corrected method, $DC_{Corrected}$ = 0.91 (median), 0.87 (mean), 0.10 (std). We also computed

the median relative errors, which were -10.19% (Auto) and 4.19% (Corrected) (Fig. SEG_VALIDATION(b)). Both DC and relative error suggest that DeepCytometer segments white adipocytes with an acceptable area error, and that the Corrected method performs better than the Auto method.

To evaluate whether DeepCytometer segmentations represent the training cell area population, we compared box-and-whisker plots of the hand traced, Auto and Corrected white adipocyte areas (Fig. SEG_VALIDATION(a)). The hand traced population had quartiles (Q1, Q2, Q3) = (1.5, 2.3, 3.9) 10^3 μm^2. The Auto method moderately underestimated cell areas, as expected due to the lack of cell overlap, (Q1, Q2, Q3) = (1.3, 2.1, 3.4) 10^3 μm^2. The Corrected method approximated the hand traced population better, with just a slight overestimation (Q1, Q2, Q3) = (1.5, 2.4, 4.0) 10^3 μm^2.

Furthermore, as discussed in *Methods*, we went beyond summary statistics commonly found in the current literature to evaluate whether segmentation errors remain constant across the cell population. For this, we plotted Auto and Corrected relative errors vs. hand traced cell area, the rolling median and interquartile range curves and the global median relative error (Fig. SEG_VALIDATION(c)-(f)). The curves suggest that relative errors remain constant for $Area_{ht}$ ≥ 780 μm^2, but shift towards more positive values for smaller cells. This would suggest less reliable phenotyping results for cells with $Area_{ht}$ < 780 μm^2, which comprise the bottom 15.9% of the population. This highlights the need to report segmentation errors by cell size in future literature.

Phenotype study of WAT using DeepCytometer segmentations

We broke down the study of *Klf14* phenotypes into three interconnected levels: the mouse level (body weight, BW), the depot level (depot weight, DW) and the cell level (quartile cell area). The mice were stratified by sex, and we tested separatedly for genotype (WT or Het) and “parent” (heterozygous parent of origin for the KO allele: father, PAT or mother, MAT) effects as exploratory analysis (the number of mice did not allow us to include both variables and their interactions in a model). Body and depot weight were measured in the laboratory, and cell areas were computed both for hand traced and DeepCytometer segmentations. (Methodology details are provided in *Methods, Phenotype framework for WAT*).

Mouse level

We studied the sex effect on BW, as well as genotype and parent. (We also checked that cull age had no significant effect, see *Suppl. Cull age effect on BW*).

Sex effect on BW

To assess the effect of sex on BW, we fitted a Robust Linear Model (RLM) (BW ~ sex) to the PAT mice (n_{female}=n_{male}=18). We used all PAT mice instead of only WT as controls as they do not display *Klf14* phenotypes. The RLM was preferred to an Ordinary Least Squares (OLS) model to moderate the leverage of a large male outlier. The model calculated a mean BW for

females of 25.06 g and 38.30 g for males (males 52.82% larger, β=13.24 g, p=6.00e-20). This sexual dimorphism is larger than genotype or parent effects that we found in the next sections at the BW, DW or cell level, so we stratify the data by sex in the rest of the study.

Genotype and parent effect on BW

Next we looked at the effect of parental inheritance on BW. We fitted separate OLS models (BW ~ genotype) and (BW ~ parent) to 76 mice stratified by sex (n_{female}=n_{male}=38). T-tests of the $\beta_{genotype}$(Het) coefficients do not show any significant genotype effect in females (p=0.22) or males (p=0.79) (Fig. BW(a)). On the other hand, there is a highly significant parent effect for females (β_{parent}=4.48 g, p=0.0061), where MATs are on average 4.48 g / 25.10 g = 17.86% larger than PATs (Fig. BW(c)).

Depot level

As discussed in the *Introduction*, SAT and VAT have different impacts on disease and phenotypes. In this section we study genotype or parent effects on DW of inguinal subcutaneous (for SAT) and gonadal (for VAT) depots adjusting for BW, as well as DW correlation with BW.

BW, genotype and parent effects on depot weight (DW)

We fitted OLS models (DW ~ genotype * BW/$\underline{BW}$) and (DW ~ parent * BW/$\underline{BW}$) to the same 76 mice stratified by sex and depot, where $\underline{BW}$=33.44 g is the mean BW of all animals used as a normalisation factor to lower the condition number of the linear model. We then used Likelihood ratio tests (LRTs) to compare those models to null-models (DW ~ BW/$\underline{BW}$) in order to test for genotype or parent effects. Even though the (DW ~ effect * BW/$\underline{BW}$) models already provide one fitted line for the control group (WT or PAT) and another for the effect group (Het or MAT), this assumes similar data variance and overlap in both groups. To be free from that requirement, in order to assess correlation we fitted new models (DW ~ BW/$\underline{BW}$) separately to the control and effect groups, and then used a t-test of their slopes to evaluate correlation between DW and BW. Said individual linear models are shown in Fig. BW(b) together with scatter plots of the original data (one point per mouse). Their intercept and slope values, and corresponding p-values are tabulated in Tables Table DW_BW_RLM_GENOTYPE and Table DW_BW_RLM_PARENT. The p-values of the slopes from the 8 genotype or parent models were jointly corrected using Benjamini-Krieger-Yekutieli[57].

Genotype effect: LRTs do not show a significant genotype effect in DW for any of the 4 sex/depot strata: female (gonadal LR=0.23, p=0.63, subcutaneous LR=0.03, p=0.87) or male (gonadal LR=2.13, p=0.14, subcutaneous LR=3.26, p=0.071). According to the individual models stratified by genotype (Table Table DW_BW_RLM_GENOTYPE), the p-values of the slopes suggest that DW is positively correlated with BW in female gonadal depots both for WTs and Hets ($\beta_{BW/\underline{BW}}$(WT)=1.52 g, p(WT)=0.015; $\beta_{BW/\underline{BW}}$(Het)=1.63 g, p(Het)=0.031), but uncorrelated in the other strata.

Parent effect: LRTs reveal significant parent effects in females (gonadal LR=5.23, p=0.022, subcutaneous LR=5.42, p=0.020). According to the individual models stratified by parent (Table Table DW_BW_RLM_PARENT), the p-values of the slopes suggest that BW is positively correlated with DW in female gonadal depots both for PATs and MATs ($\beta_{BW/\underline{BW}}$(PAT)=2.58, p(PAT)=0.0048; $\beta_{BW/\underline{BW}}$(MAT)=1.40, p(MAT)=0.010). In female subcutaneous depots, the trend is the same as in gonadal ones, but does not reach significance. In both depots, female MATs have lower DW for the same BW than PATs, and thus, lower fat percentages. In males, LRTs show very significant parent effects (gonadal LR=7.48, p=0.0063, subcutaneous LR=11.29, p=0.00078). However, what those LRTs are finding is a significant difference between a non-significant positive slope $\beta_{BW/\underline{BW}}$(PAT) and a non-significant negative slope $\beta_{BW/\underline{BW}}$(MAT). Thus, we consider that male DW is uncorrelated with BW under parent stratification, and that a parent effect in male DW is inconclusive.

To summarise, DW is positively correlated with BW in female gonadal depots, but not in female subcutaneous, or either male depot. In addition, there is no genotype effect on DW, but there is a parent effect in females, with MAT females having lower fat percentages than PAT ones.

Cell level

First, we calculate probability distribution functions (pdfs) in the hand traced data set and the DeepCytometer segmented whole slides to compare hand traced populations to DeepCytometer whole slide ones. Second, we study genotype or parent effects on cell area in the same depots as before, adjusting for DW, as well as cell area correlation with DW. Finally, we present heatmaps of cell areas computed from the DeepCytometer segmentations. This provides a clear picture of whether there are local correlations or a uniform spatial distribution of cell sizes across whole slides, as well as whether slides from the same stratum have similar cell population spatial distributions.

Area population distributions of hand traced cells

The hand traced dataset consists of 1,903 cells pooled from 60 subcutaneous windows and 20 mice (see *Data*). The cells measured between 66.0 μm^2 (321 pixel) and 19,058.2 μm^2 (92,544 pixel). To represent cell populations, we estimated probability density functions (pdfs) of the areas of hand traced cells (Fig. MANUAL_POPULATION_HISTOS(a)). We also calculated the cell area Harrell-Davis (HD) quartiles (Q1, Q2, Q3) with 95%-CIs (Fig. MANUAL_POPULATION_HISTOS(b)). Male cells were notably larger than female cells for each quartile. On the other hand, for each sex and quartile, PAT and MAT areas were similar, with overlapping 95%-CIs (although there is a trend for smaller MAT cells in each quartile).

Area population distributions of DeepCytometer segmented cells

We estimated pdfs from the areas of DeepCytometer segmented cells (with the Corrected method), one pdf per slide (Fig. SEG_POPULATION_HISTOS(a)-(b)). We segmented 75 inguinal subcutaneous and 72 gonadal whole histology slides, corresponding to 73 females

and 74 males, to produce 2,560,067 subcutaneous and 2,467,686 gonadal cells. We combined all the pdfs by computing pdf HD quantiles q={2.5%, Q1, Q2, Q3, 97.5%} (i.e. quantiles of density values instead of cell areas), and displayed them as shaded areas and solid curves in Fig. SEG_POPULATION_HISTOS(c)-(d): 2.5%-97.5% form the 95%-interval (light shaded area) and Q1-Q3 form the interquartile range (dark shaded area). The Q2 curve (solid blue) provides a median histogram for each stratum. In addition, we computed the cell area quartiles Q1, Q2, Q3 for each pdf, and their standard errors. We combined those estimates using the inverse-variance meta-analysis method to produce one combined $\hat{Q1}$, $\hat{Q2}$, $\hat{Q3}$ and their 95%-CIs per stratum. The combined $\hat{Q1}$, $\hat{Q2}$, $\hat{Q3}$ are displayed as vertical black lines in Fig. SEG_POPULATION_HISTOS(c)-(d), and their numerical values and 95%-CIs are provided in the tables of Fig. SEG_POPULATION_HISTOS(e)-(f).

.

Comparison of hand traced vs. DeepCytometer segmented cell populations

In section *Results - Segmentation sub-pipeline validation*, we showed that DeepCytometer automatic segmentation approximates hand tracing in training windows, and is faster. In this section, we test whether whole slide segmentation also adds valuable population information to training window segmentation. First, we compared hand tracing of the training windows sampled from 20 subcutaneous whole slides against the DeepCytometer segmentations of those same 20 whole slides. For the purpose of this experiment, it is enough to stratify by sex and parent, omitting genotype. We computed HD quartiles (Q1, Q2, Q3) from each mouse and combined them using the inverse-variance meta-analysis as in the previous section. Pdfs and quartiles are plotted in Fig. MANUAL_POPULATION_HISTOS(a) and (c), and quartile values and their 95%-CIs are tabulated in Fig. MANUAL_POPULATION_HISTOS(b) and (d). The area difference (%) between hand traced and DeepCytometer segmentations is tabulated in Fig. MANUAL_POPULATION_HISTOS(e). In all but one stratum, the DeepCytometer segmentation quartiles are larger than the hand traced ones. Whereas for males the area difference is between -8.72% and +20.42%, for female MATs it ranges between +57.65% and +65.15%. This suggests that our 1,903 hand traced cells from 60 windows and 20 mice, despite being a rather large training dataset, misrepresents the whole slide populations in the four strata, especially for female MATs. Namely, it undersamples the long tails on the right-hand side of the pdfs, i.e. the larger cells in the population. These errors are not systematic, and vary between strata. This could be partly due to hand tracing sampling a relatively small number of cells per mouse and pooling them. This highlights the need for whole slide analysis.

Furthermore, to test whether 20 whole slides are enough to represent cell populations, we computed pdfs from the other 55 subcutaneous DeepCytometer whole slide segmentations for a total of 75 subcutaneous pdfs (as well as from the 72 gonadal slides for completion) (Fig. SEG_POPULATION_HISTOS). The quartile area difference between the 20 and 75 whole subcutaneous slides is shown in Fig. SEG_POPULATION_HISTOS(g). Although the quartile areas are similar for females (from -4.10% to +9.00%), the subpopulation estimates change substantially for males (from -14.28% to +22.36%).

Therefore, both whole slide analysis and the analysis of more mice significantly changed the cell population pdfs. Both increases are enabled by DeepCytometer's segmentation.

DW, genotype and parent effects on cell area quartiles

So far, our comparison of cell area quantiles has not taken into consideration confounders such as BW or DW. As we have previously studied DW as a function of BW, in this section we complete the three level phenotyping analysis with OLS models ($area_q$ ~ genotype ∗ DW) and ($area_q$ ~ parent ∗ DW), stratified by sex and depot, where $area_q$ are the area quartiles q={Q1, Q2, Q3} (Table AREAQ_DW_LRT). (We removed 2 gonadal and 1 subcutaneous slides from the analysis due to lack of BW and DW records of two mice.) We apply a similar approach as before, using LRTs to assess genotype and parent effects, and slopes to assess correlation. The p-values of the slopes $\beta_q=\beta_{DW}(q)$ were jointly corrected using Benjamini-Krieger-Yekutieli in the 24 models that correspond to 3 quartiles, 2 sexes, 2 genotypes/parents, and 2 depots.

Genotype effect: individual linear models are shown in Fig. AREAQ_DW_GENOTYPE_LINREG and coefficients provided in Table AREAQ_DW_GENOTYPE_LINREG. LRTs are shown in Table AREAQ_DW_LRT(a). For the gonadal depot, the individual linear models for WT and Het are visually very close, and LRTs comparisons show no significant difference between WT and Het for females or males. For the individual models, there is a statistically significant correlation between DW and cell area in female WTs (β_{Q2}=2360.7 μm^2/g, p_{Q2}=0.030; β_{Q3}=4420.8 μ^2/g, p_{Q3}=0.030), but not in female Hets. However, this could be due to slightly higher variance for Het values. Both visual assessment and LRT p-values suggest a trend of WT and Het female gonadal cell area increasing with DW. By contrast, in male gonadal depots, visual assessment and slopes and their p-values suggest constant cell area regardless of DW.

Subcutaneous depots visually show that cell area increases with DW in female and male WTs, and that cell areas are smaller in Hets, at least as a trend. However, in females, slopes β_q are not statistically significantly different from zero according to their t-test p-values. There is no evidence of a difference between female WT and Het models according to LRT p-values either. On the other hand, p-values for male WT slopes are very significant ($p_{Q1}=p_{Q2}=p_{Q3}$=0.0011) and indicate that cell area increases with DW (β_{Q1}=720.9 μm^2/g, β_{Q2}=1735.2 μm^2/g, β_{Q3}=2744.8 μm^2/g), whereas Het slopes are not significantly different from 0. The difference between WT and Het models is significant according to LRTs (p_{Q1}=0.0091, p_{Q2}=0.0058, p_{Q3}=0.014). Thus, there is evidence that cell area increases with DW for WTs, but there is no evidence of correlation with DW for Hets.

Parent effect: individual linear models are shown in Fig. AREAQ_DW_PARENT_LINREG and coefficients provided in Table AREAQ_DW_PARENT_LINREG. LRTs are shown in Table AREAQ_DW_LRT(b). The plots display positively correlated cell area to DW. However, the residuals reveal heteroscedasticity and autocorrelation, and t-tests of the β_q coefficients return non-significant p-values after multitesting correction due to the variance of area values. Thus, it is inconclusive whether DW and cell area are truly correlated. Nonetheless, the LRTs suggest

a very significant parent effect in female cells, both gonadal (p_{Q1}=0.01, p_{Q2}=0.0023, p_{Q3}=0.0022) and subcutaneous (p_{Q1}=0.00082, p_{Q2}=0.0015, p_{Q3}=0.0021). Visual inspection of the OLS plots suggests that this difference arises from larger MAT than PAT cell areas for a given DW. However, due to the aforementioned issues with the residuals, we qualify this phenotype as inconclusive.

For males, we have a case analogous to the depot level above, as in some cases the LRTs show a significant difference between PAT and MAT, but all the gonadal and subcutaneous slopes β_q are non-significant. Thus, we conclude that cell area and DW are uncorrelated, and there is effectively no parent effect.

Quantile colour map plots to assess cell size heterogeneity

In order to gain insight into the spatial distribution of adipocyte populations, we used the area-to-colour map described in *Methods - Quantile colour map plots to assess cell size heterogeneity* to visualise cell area distribution in all whole slides processed by DeepCytometer, both in AIDA to visually assess the segmentation, and to generate figures for this paper (Fig. COLORMAP_F_GWAT-Fig. COLORMAP_M_SCWAT). The colour map is linear with area quantile, rather than area, and we use separate colour maps for females and males, due to sexual dimorphism. The results clearly show local subpopulations or clusters of white adipocytes. These clusters are irregular in shape, and present high inter- and intra-slice variability, even within the same sex and depot stratum. They illustrate the challenges for statistical studies of cell populations performed on subsamples of whole slides, such as our hand traced dataset. Namely, cells within clusters are correlated observations, so although spatial analysis is without the scope of this paper, we conjecture that an apparently large number of cells (~2,000 in our hand traced dataset) may not properly represent the mixture of subpopulations in the original whole slides.

Summary of phenotype findings

Effect	**Model**	**Results**	
		Full body	
		Female	**Male**
Sex	**BW**	Male 52.8% heavier than female.	
Genotype		N.e.	N.e.
Parent		E: MAT 14.7% heavier than PAT.	N.e.

(a)

		Gonadal depot	
		Female	**Male**
Genotype WT → Het	**DW ~ BW**	N.e. C: β(WT), β(Het) > 0.	N.e. N.c.
	$area_q$ ~ DW	N.e. C: β(WT) > 0, β(Het) = 0.	N.e. N.c.
Parent: PAT → MAT	**DW ~ BW**	E: DW(MAT) < DW(PAT). C: β(PAT), β(MAT) > 0.	N.e. N.c.
	$area_q$ ~ DW	E: inconclusive. C: inconclusive.	N.e. N.c..

(b)

		Subcutaneous depot	
		Female	**Male**
Genotype WT → Het	**DW ~ BW**	N.e. N.c.	N.e. N.c.
	$area_q$ ~ DW	E: n.s. area(Het) < area(WT). N.c.	E: area(Het) < area(WT). C: β(WT) > 0, β(Het) = 0.
Parent: PAT → MAT	**DW ~ BW**	E: DW(MAT) < DW(PAT). C: n.s. β(PAT), β(MAT) > 0.	N.e. N.c.
	$area_q$ ~ DW	E: area(PAT) < area(MAT)? inconclusive. C: inconclusive.	N.e. N.c.

(c)

Table SUMMARY_FINDINGS. Summary of phenotyping analysis of three traits: (a) body weight (BW), (b) depot weight (DW), and (c) cell area. The main findings we report are genotype/parent effect (significant difference between WT and Het, or PAT and MAT, respectively) and correlation (β slope coefficient significantly different from zero) between trait and covariate (BW or DW). N.e.: no effect (genotype or parent, respectively). E: effect. N.c.: no correlation between continuous covariate and dependent variable. C: norrelation. n.s.: not significant.

Discussion

This paper tackled two parts: first, we presented DeepCytometer, a pipeline to segment white adipocytes from high resolution H&E histology, together with visualisation tools for the results. Second, we presented a phenotype framework for white adipose tissue that we applied to *Klf14* mouse data. Unlike previous methods that have been limited to processing small windows containing mostly white adipocytes with good image quality, DeepCytometer can process whole slides containing cell overlaps, other types of tissue, image artifacts and variations in image quality. In the first part of the paper, we addressed several problems that naturally arise in whole slide segmentation: **1) Coarse tissue mask.** By applying traditional image processing techniques to segment tissue areas on the 16× downsampled slide. This resolution was enough to obtain a tissue mask, avoided processing large empty background areas and its computation time was negligible compared to segmentation. **2) Adaptive tiling for whole slide processing.** We improved on previous tiling approaches by proposing a method that does not discard cells cropped by tile edges, chooses each tile's location and size to reduce overlap and redundant computations. We found a 16.59% reduction in the number of processed pixels on 147 histology images, compared to uniform tiling. As the processing time is linear with the number of pixels, this resulted in an overall speedup of whole slide segmentation. **3) White adipocyte segmentation.** We built on previous cell segmentation work, with a combination of FCNs and post-processing methods, in particular watershed and mathematical morphology (we call this our *Auto* method). As in previous work, we estimated an EDT from adipocyte membranes with a CNN, but followed it by a Contour CNN to find the EDT troughs, because we found that in practice, previously proposed simple post-processing methods did not work consistently in whole slides. In addition, we proposed the Correction CNN to correct Auto labels to account for cell overlap (*Corrected* method). Our median Dice coefficients were 0.89 (Auto) and 0.91 (Corrected), showing good agreement with the hand traced segmentation. The median relative area errors were -10.19% (Auto) and 4.19% (Corrected). Following previous practice in the literature, these measures would validate our segmentation method. However, in this paper we also proposed looking at segmentation errors over the population distribution. We found that the segmentation area error was roughly proportional to cell area for cells ≥ 780 μm^2, which is desirable, but shifts towards more positive values for smaller cells, the bottom 15.9% of the population. This illustrated how segmentation errors in subpopulations could go undetected using summary statistics, and highlighted the need for validating segmentation errors as a function of cell size. This did not affect our phenotyping evaluation, as we used the cell area population quartiles (25%, 50% and 75%) and not the smallest cells. Running time of the pipeline increased linearly with tissue area, with 5.5 h for a median size slide with 174.7 mm^2 (848.2 Mpixel) of tissue. Most of our slides contained two slices, so the median time to analyse a slice would be 2.75 h. Of the processing time, 43.9% corresponded to the Auto method, and 56.1% to the Corrected method. Thus, the Corrected method increases computation time by ×2.28 over using only the Auto method. Whether this trade off is acceptable depends on the available resources. In our quantitative experiments, we used Corrected results. **4) Tissue and object classifier.** To determine which segmented objects

are white adipocytes, first we classify each histology pixel, and then accept objects that contain ≥ 50% white adipocyte pixels. We weighed the object classifier's ROC by the number of pixels per object, to balance the contribution to classification errors of white adipocytes (small but numerous) with other objects (scarcer but large). The 0.97 area under the ROC indicated overall good performance of the classifier. With the 50% acceptance threshold, the false positive rate (FPR) = 15% and true positive rate (TPR) = 95%, which we considered acceptable for phenotyping. If a lower FPR was required in other work, the 50% threshold could be raised. Given the large cell counts in the slides, the resulting lower TPR may be an acceptable trade-off. **5) Ground truth / training data.** We created a ground truth / training data set for the EDT, Contour and Correction CNNs with 55 random windows from 20 mice, totalling 2,117 white adipocyte hand traced contours, for 10-fold cross validation. For the Tissue CNN, we added another 71 windows containing only non-white adipocyte regions. This was necessary to create a balanced set of $\approx 23.7 \cdot 10^6$ white adipocyte pixels and $\approx 45.1 \cdot 10^6$ non white adipocyte pixels. Non-white adipocyte areas were easier to manually segment because they are larger and do not need precise outlining. For the cell population studies, we added another 5 windows (for a total of 60) to two mice with undersampled populations, but removed 214 segmented objects (for a total of 1,903 cells left over) where we had doubts of being white adipocytes. We hope that the data set will be a useful resource in the field, and have made it available for download at XXXXXX. **6) Segmentation results visualisation.** Integration of our pipeline with AIDA allowed us to review whole slide segmentation results in real-time. AIDA was launched on the same GPU server as the pipeline, and enabled real-time review of the results from a desktop or laptop using a regular browser. The pipeline features saving each block as a separate layer or all in the same layer for display. AIDA also allows manual correction of labels. **7) Cell size heterogeneity visualisation.** To highlight spatial size heterogeneity, we proposed a colour scale proportional to the cell area's quantile. This scheme produced highly contrasted images readily showing cell area heterogeneity across tissue samples, with subpopulations of different sizes grouped in clusters (Fig. COLORMAP_F_GWAT-Fig. COLORMAP_M_SCWAT). **8) Suitability of hand traced dataset for cell population study.** The *Results - Comparison of hand traced vs. DeepCytometer segmented cell populations* experiment returned very significant differences between the distributions obtained from the hand traced data set and whole slide automatic segmentations used as ground truth. This suggests that even though the hand traced data set contained 1,903 cells from 60 random windows, it failed to represent the true distribution of white adipocyte areas. Although spatial analysis is without the scope of this work, we conjecture that intra-slice clustering introduces strong spatial local correlations in cell size, thus reducing the effective size of the training dataset (Fig. COLORMAP_F_GWAT-Fig. COLORMAP_M_SCWAT). Such difference between hand traced and DeepCytometer populations highlights the need for whole slide segmentation and analysis, with pipelines that can run on dozens or hundreds of slides. Furthermore, this begs the question for future work of whether a single whole slide provides an appropriate representation of a whole depot's cell population, or whether fat phenotype studies should contain multiple slices that cover each depot, or our work should be extended to 3D modalities such as fluorescent microscopy, considering the financial and computational cost increase.

In the second part of this paper, we performed an explanatory study of *Klf14* phenotypes in B6NTac mice at three nested levels –animal (body weight, BW), depot (depot weight, DW) and cell (quartile cell area)– in four strata defined by sex (female/male) and depot (gonadal/subcutaneous). Cell areas were obtained by DeepCytometer from 75 inguinal subcutaneous and 72 gonadal full histology slides. At the BW level, control mice presented marked sexual dimorphism, with males weighing 52.8% more than females on average. Thus, the rest of the study was stratified by sex.
At each level, we considered two phenotype assessments: 1) correlation between variables (e.g. BW vs DW) via t-tests of linear model slopes and 2) genotype (WT/Het) or parent (PAT/MAT) effects using LRTs. In addition, at the cell level we computed pdfs of cells areas.

At the **animal level** (BW), there was no genotype effect, but we found a parent effect, as MAT females were 14.7% heavier than PATs. This would suggest a phenotype where females with mothers that carry the KO allele are heavier, regardless of whether the daughter carries it herself. At the **depot level** (DW ~ BW), males had no genotype or parent effects, and their BW and DW were uncorrelated. Females showed no genotype effect either. Nonetheless, there was an interesting non-phenotypic difference between VAT and SAT: stratified by genotype, BW and DW were correlated in gonadal depots, but not in subcutaneous depots. Females showed a depot parent effect, with MAT depots being smaller than PAT ones. This is remarkable, because then, female MATs are both heavier but have smaller fat depots than their PAT counterparts, i.e. they are both larger and leaner mice. Furthermore, when stratifying by parent, we observed a similar non-phenotypic difference as above: BW and DW were correlated in gonadal depots, but the slope in subcutaneous depots is statistically non-significant. Thus, both at animal and depot weight, there are phenotypes for female MATs. At the **cell level** ($area_q$ ~ DW), it is noteworthy that males exhibited a phenotype, but only as a subcutaneous genotype effect, with cell area in Hets being smaller than in WTs. In that depot, male $area_q$ is correlated with DW in WTs but not in Hets. These two observations, together with the fact that there was no genotype effect in male DW itself, would suggest that male WTs and Hets have similar subcutaneous DW, with the distinction that in WTs, larger DW is achieved by white adipocyte enlargement, whereas in Hets, it is by cell multiplication. In female gonadal depots, there is no genotype effect according to the LRT. However, DW and $area_q$ are correlated in WTs but not in Hets; because in the depot level, DW and BW were correlated for both WTs and Hets, this would suggest that WT female gonadal white adipocytes grow with BW, but in Hets, cell growth is limited and instead DW is due to cell multiplication. The female gonadal parent effect analysis is inconclusive, due to the data variance and not suitability for a linear model. Finally, female subcutaneous depots show no correlation between DW and $area_q$ with genotype stratification, and only a weak trend of Het areas being smaller than WT ones. The parent effect is inconclusive as before, although there is a trend for MAT areas being larger than PAT ones.
In summary, our exploratory analysis reveals interesting phenotype leads by breaking down the study into animal, depot and cell level, and assessing parent and genotype effects. Of particular interest are strata where depot size can be explained in terms white adipocyte size vs. multiplicity, which would be directly connected to hypertrophy vs. hyperplasia. However, further investigation with more mice is necessary, to be able to examine genotype:parent interactions and provide explanatory mechanisms for the phenotypes.

In future work, the pipeline can also be redesigned with alternative deep learning approaches that find candidate objects and then segment them, as opposed to DeepCytometer, which first segments all objects and then classifies them, e.g. based on Faster R-CNN[58], Mask R-CNN[59] or TensorMask[60]. Although we have proved that DeepCytometer can analyse hundreds of slides with common resources, we would like to improve its architecture to scale it to studies with thousands of slides, without needing large cloud computing resources.

Data

Mouse description, tissue acquisition and imaging

To develop and evaluate our methods we analysed adipose tissue histological samples from mice carrying (Het) or not (WT) a *Klf14* gene knockout on either maternally (MAT) or paternally (PAT) inherited chromosomes, described previously[13]. The summary of the characteristics of the 20 *Klf14*-C57BL/6NTac (B6NTac) mice used for training and testing the DeepCytometer pipeline, as well as the hand traced population experiments, is shown in Table MICE. The mean ± standard deviation mouse weights were 26.6 ± 3.7 g (female PAT), 26.2 ± 2.9 g (female MAT), 37.6 ± 1.9 g (male PAT), 39.9 ± 3.8 g (male MAT).

The histopathology screen involved fixing, processing and embedding in wax, sectioning and staining with Hematoxylin and Eosin (H&E) both inguinal subcutaneous and gonadal adipose depots. For paraffin-embedded sections, all samples were fixed in 10% neutral buffered formalin (Surgipath) for at least 48 hours at RT and processed using an Excelsior™ AS Tissue Processor (Thermo Scientific). Samples were embedded in molten paraffin wax and 8 µm sections were cut through the respective depots using a Finesse™ ME+ microtome (Thermo Scientific). Sampling was conducted at 2sxns per slide, 3 slides per depot block onto simultaneous charged slides, stained with haematoxylin Gill 3 and eosin (Thermo scientific) and scanned using an NDP NanoZoomer Digital pathology scanner (RS C10730 Series; Hamamatsu).

Ground truth hand traced dataset for CNN training and cell population studies

We created two slightly different hand traced datasets. For DeepCytometer training and validation, we randomly sampled each of the 20 training histology slides to extract a total of 55 histology training windows with size 1001×1001 pixels and traced white adipocytes, other types of tissue (connective, vessels, muscle, brown adipocytes, etc.) and background areas. Another 71 windows were manually selected to add more examples of only other types of tissue to train the tissue classifier. This first dataset is summarised in Table MICE in black.

For cell population studies, we added 5 random windows from 2 mice whose cell population was undersampled, and we removed those white adipocyte objects with ambiguous interpretation, namely fully overlapped, suffering from image artifacts or comprising very small gaps between clear adipocytes. This second dataset is summarised in Table MICE in blue.

The total number of hand traced white adipocytes is 2,117 and 1,903, respectively, with a roughly balanced split between the four groups formed by the female/male and PAT/MAT partitions (Table CELL-NUM). The total number of "other tissue" objects is 232, and of "background", 24. (Note that cell objects tend to be much smaller than "other" or "background" objects). Each window contained one or more of these types of objects. Hand tracing was performed with the image editor Gimp (www.gimp.org) over a month. Automatic levels correction and manual curves adjustment was temporarily applied to each window to improve image contrast for the human operator. Contours were drawn as linear polygons on the outermost cell edge, accounting for the cell overlaps shown in Fig. OVERLAP(b), and exported with an *ad hoc* plugin as SVG files for further processing. For "other" or "background", representative linear polygons were drawn, avoiding complex boundaries. This approach produced partially labelled training windows.

ID	Sex	Genot.	BW (g)	Fold	Cells		Oth.	Back.	Total win.		Win. that contain cells	
16.2d	m	MAT Het	46.19	3	55	89	9	5	7	8	4	5
17.1c	f	MAT Het	22.07	2	165	86	15	3	5	5	1	1
17.2c	f	MAT Het	26.39	1	34	31	7	0	4	4	1	1
17.2f	m	MAT Het	40.87	9	150	135	14	0	7	7	3	3
18.1a	f	MAT Het	30.65	6	63	48	13	1	6	6	1	1
18.1e	m	MAT Het	40.02	8	25	128	9	0	7	11	2	6
18.2b	f	MAT Het	24.28	0	49	44	12	0	5	5	2	2
18.2d	f	MAT Het	27.72	0	190	148	12	0	5	5	2	2
18.2g	m	MAT Het	41.98	9	0	0	9	0	3	3	0	0
18.3b	m	MAT Het	34.52	8	83	73	11	3	8	8	4	4
18.3d	m	MAT Het	36.08	2	199	179	16	1	8	8	5	5
36.1a	f	PAT Het	31.42	5	96	82	5	0	5	5	2	2
36.1b	f	PAT WT	29.25	5	65	56	8	0	6	6	2	2
36.1c	f	PAT Het	27.18	7	187	159	8	1	5	5	2	2

36.1i	m	PAT Het	36.55	1	251	218	11	2	11	11	7	7
36.3d	m	PAT Het	40.77	6	111	100	21	1	8	8	5	5
37.1c	f	PAT WT	23.69	3	157	122	13	2	5	5	2	2
37.1d	f	PAT WT	21.20	4	17	13	15	0	5	5	2	2
37.2g	m	PAT Het	36.98	4	147	126	13	0	8	8	5	5
37.4a	m	PAT Het	36.11	7	73	66	11	5	8	8	3	3
Total					2,117	1,903	232	24	126	131	55	60

Table MICE. Description of mouse cohort used for CNN training (black), and hand traced cell population studies (blue). All slides acquired from subcutaneous tissue. Genot.: Parent and genotype (MAT/PAT: maternally/paternally inherited allele. WT: Wild type. Het: Heterozygous). BW: Body weight. Oth.: Other. Back.: Background. Cells/Oth./Back.: Number of hand traced white adipocytes/other tissue/background objects. Total win.: Number of 1,001×1,001 pixel windows extracted from the full slice at maximum resolution for training and testing. Win. that contain cells: Number of those windows that contain hand traced white adipocytes.

Sex → / Parent ↓	Female	Male	Total
PAT	522	582	1,104
MAT	501	512	1,013
Total	1,023	1,094	2,117

DeepCytometer training dataset

Sex → / Parent ↓	Female	Male	Total
PAT	432	510	942
MAT	357	604	961
Total	789	1,114	1,903

Cell population studies dataset

Table CELL-NUM. Number of hand traced white adipocytes stratified by sex and parent. These tables are summaries of Table MICE.

Methods

Coarse tissue mask and adaptive tiling for full processing of whole slides

Coarse tissue mask

Histology slides in Hamamatsu NDPI format can be read by blocks of arbitrary size at a fixed number of precomputed resolution levels with OpenSlide [61]. At the highest resolution level, our images have pixel size 0.454 μm. We read the whole histology slide at the precomputed ×16 downsampling level (7.26 μm pixel size), applied contrast enhancement and computed the mode Mo_{colour} and standard deviation σ_{colour} in each RGB channel of the image. We assume that the background colour is centered around Mo_{colour}, as background pixels are more numerous than tissue pixels. To segment the tissue, we thresholded the downsampled image for pixels that are darker than $Mo_{\text{colour}} - k_\sigma \sigma_{\text{colour}}$ in all RGB channels (see Fig. RMABb), where $k_\sigma = 0.25$. Then, we applied morphological closing with a 25×25 kernel at ×16 downsampling level, filled holes smaller than 8,000 pixels (421,759 μm^2), and removed connected components smaller than 50,000 pixels (2,635,994 μm^2).

Adaptive tiling

When using uniform tiling, to guarantee that any cell will be processed whole in at least one tile, adjacent tiles need to overlap by $RF + D_{\max}$, where RF is the receptive field or diameter of input pixels that affect each output pixel, and $D_{\max}$ is the diameter of the largest cell. This overlap introduces repeated processing of the same pixels and multiple segmentations of the same cells from adjacent tiles, but when it is ignored[11], cells cropped by tile edges need to be discarded.

To ameliorate this problem, we propose an iterative tiling algorithm that adapts the block size and overlap of each new tile according to the local cell size and tissue mask, such that the whole coarse tissue mask is covered, all cells are segmented, and redundant computations are reduced.

For this, first we flag all pixels in the coarse tissue mask as “to be processed”. To find the location of the first block, the mask is convolved with two small linear kernels. Pixels > 0 in both outputs are potential locations for the block’s top-left corner. Any of those locations guarantee that the block has at least one mask pixel on the top and left borders. The algorithm chooses the first of the candidate pixels, in row-column order. The bottom-right corner of the block is initially chosen to obtain a block with maximum size 2751×2751 pixels (maximum allowed by GPU memory). The block’s right side and bottom are then cropped to remove empty columns or rows, producing an adaptive size. Finally, the block is extended half the effective receptive field on each side, to prevent border effects. (This extra border is discarded after image processing operations.) If the block overflows the image, it is cropped accordingly. The image block is then segmented; pixels from cells cropped by the edges keep their “to be processed” flag, so they will be included in another block. The rest of the mask pixels are cleared, and the process is repeated iteratively choosing new adaptive blocks until the whole tissue mask is cleared. Pseudocode and details for the algorithm are provided in *Suppl. Pseudocode for adaptive tiling*, and an example of its behaviour is shown in Fig. RMAB:

Deep CNN architectures

In this section we describe the function of the four different DCNNs used in this work, with illustrative examples (Fig. DMAP-Fig. CORRECT), a summary of their architectures (Table CNN), and the calculation of their effective receptive fields (ERFs) (Table ERF). Training and

validation details are provided in *Supplementary material*. These networks are components of the *Segmentation sub-pipeline* that we describe in the next section. The networks are fully convolutional[49], so that tile size can be adjusted to available GPU memory and the needs of the adaptive tiling. We used a stride of 1 to avoid downsampling followed by deconvolution, and thus preserve high-resolution segmentation following [62]. We used atrous or dilated convolution[63–65] to enlarge the ERF following [47,62]. RGB 8-bit unsigned integer values in the image files were converted to 32-bit float type and scaled to [0, 1] or [-1, 1] depending on the specifications of the network they are being fed to.

Histology to EDT regression CNN (EDT CNN)

This network (see architecture in Table CNN(a)), based on previous work by [51,66] and similar to [50,67] —as discussed in the *Introduction*— takes an input histology image and estimates the Euclidean Distance Transform (EDT) as the Euclidean distance of each pixel to the closest label boundary point (see *EDT CNN and Contour CNN training dataset* below for details). This produces an image similar to an elevation map (Fig. DMAP), where each "hill" defines an object (e.g. a white adipocyte, an area of muscle tissue or a vessel cross-section). If the object is a white adipocyte, troughs represent the cell's membrane boundary or a compromise boundary between overlapping cells. Trough points are all critical points (extrema or saddle points) but have different values or "elevations". We found that troughs in our whole slides could not be segmented with simple segmentation methods like in [50,67]. Instead, we trained the following Contour network for that task.

EDT to Contour detection CNN (Contour CNN)

This network (see architecture in Table CNN(b)) classifies each EDT pixel as either belonging or not to a trough (Fig. CONT).

Pixel-wise tissue classifier CNN (Tissue CNN)

This network (see architecture in Table CNN(c)) classifies each pixel from the histology block as "other type of tissue" vs. "white adipocyte" or "background" (see Fig. CLASS). "White adipocyte" and "background" are combined in one class because a gap in the tissue and the inside of a white adipocyte have the same appearance in the histology. The output of this network is used to classify a segmented object as white adipocyte or not, according to the proportion of "white adipocyte" / "background" pixels it contains (see section *DeepCytometer pipeline* below for details).

Segmentation correction regression CNN (Correction CNN)

This network (see architecture in Table CNN(d)) takes the cropped and scaled histology of a single object, multiplied by a mask derived from its segmentation, and estimates which pixels underestimate (detection error) or overestimate (false positive) the segmentation. This output is then used to correct the object's segmentation (see section *DeepCytometer pipeline* below for details). Because each object's segmentation is corrected separately, corrected boundaries can overlap. To create an input for the network, the histology tile is

cropped using a square box with at least twice the size as the segmentation mask's bounding box, and scaled to a fixed size of 401×401 pixels to make the network blind to cell size. The scaling factor is $s = 401/L$, where $L \times L$ is the size of the cropping window. Then, the histology RGB values are multiplied by +1 within the segmentation mask, and by -1 without. By contrast, [68]'s QualityNet1 multiplied the histology RGB values by 0 without. Our approach preserves the information outside the segmentation, while still partitioning the histology image into two sets of inside/outside pixels. Moreover, instead of estimating one single quality measure for the whole input image as in QualityNet1, we estimate whether the segmentation is correct *per pixel*, computing a value between -1 (undetected pixel) to +1 (false positive pixel) through 0 (correctly segmented pixel).

Effective receptive field (ERF)

The theoretical receptive field (span of input pixels that contribute to an output pixel) can be computed considering the properties of convolutions, downsampling and pooling. However, the weight of an input pixel's contribution decreases quickly towards the edges of the span, and thus the *effective receptive field* (ERF) is much smaller than the theoretical one[69]. To estimate the ERF, we used gradient backpropagation[69], but replaced ReLU activations by Linear activations and Max Pooling by Average Pooling to avoid numerical instabilities. The ERF was around 131×131 pixels (Table ERF), or 37.1% of maximum cell diameter (160 µm or 353 pixels), causing the EDT CNN to clip the estimated distance to the membrane for distances larger than the ERF. However, the segmentation validation showed no performance drop for large cells. This is because distant points do not contribute essential information to the Contour CNN. This was convenient, as increasing the ERF is computationally expensive, generally requiring deeper networks.

Segmentation sub-pipeline

The segmentation sub-pipeline combines the EDT, Contour, Tissue and Correction CNNs with traditional image processing methods (Fig. PIPE(b)) to segment an input histology image tile, producing one label per white adipocyte.

Histology colour correction

To estimate the typical background colour of the training data set, we computed density histograms with 51 bins between 0-255 for each RGB channel of the 126 training images (Table MICE). The 50% HD quantile was computed for each bin, producing a median histogram for each channel. The modes in each median channel were taken as the typical background colour, R_{target} = 232.5, G_{target} = 217.5, B_{target} = 217.5. Colour correction was applied for inference but not for training, to reduce overfitting. To apply colour correction to a histology slide, we estimated the mode intensity of each channel R_{mode}, G_{mode}, B_{mode}. Then, each colour channel was corrected as $\hat{I}_R = I_R - R_{\text{mode}} + R_{\text{target}}$, $\hat{I}_G = I_G - G_{\text{mode}} + G_{\text{target}}$ and $\hat{I}_B = I_B - B_{\text{mode}} + B_{\text{target}}$.

White adipocyte label segmentation without overlap (Auto)

The colour-corrected histology image was used as input to the EDT CNN, and its output to the Contour CNN. To conservatively detect pixels inside objects, we thresholded the resulting image with a zero threshold (pixels on or near EDT troughs have values > 0). This produced one connected component inside each cell or object. We filled holes with fewer than 10,000 pixels, and each connected component was given a different label. Components with fewer than 400 pixels were removed. The colour-corrected histology image was also fed to the Tissue CNN, and pixels with score > 0.5 were labelled as white adipocyte pixels. All non white adipocyte pixels that do not already belong to a seed were labelled as a single new seed. Seeds were expanded using a watershed algorithm on the negative of the EDT surface (the negative sign turned hills into basins). Each watershed basin corresponded to a candidate object. Objects were rejected if they: 1) were smaller than 1,500 pixel (308.9 μm^2), 2) did not overlap at least 80% with the coarse tissue mask, 3) did not contain at least 50% white adipocyte pixels, 4) touched an edge, 5) were larger than 200,000 pixels (41,187.4 μm^2) - the largest training cell was 92,544 pixel (19,058.2 μm^2). Objects that were inside another object were merged into the surrounding object. Each surviving label was considered to segment one white adipocyte, but without overlap.

Segmentation label correction with overlap (Corrected)

The colour-corrected histology image was cropped and resized around each valid white adipocyte label, and passed through the Correction CNN as described above. Output pixels with scores $\geq$ 0.5 were added to the label, and pixels with scores $\leq$ -0.5 were removed. Label holes were filled, and the connected component that had the best overlap with the input label was kept as the corrected label. Finally, the corrected label was smoothed using a closing operator with an 11×11 pixel square kernel. (See Fig. CORRECTc).

Segmentation output to AIDA user interface

Each corrected label was converted to a linear polygon with vertices $c = \{(x_0, y_0), \ldots, (x_{P-1}, y_{P-1})\}$ using Marching Squares [70]. Point coordinates were converted from the processing window to the whole histology image using $(x'_i, y'_i) = (x_i/s, y_i/s) + (\delta_x, \delta_y)$, where s is the scaling factor and (δ_x, δ_y) are the coordinates of the processing window's top-left pixel within the whole histology image. The contours were then written to a JSON file that can be read by the browser interface AIDA. The results are illustrated in Fig. AIDA.

Tissue CNN validation

The Tissue CNN was applied to each of the 126 training images (see Table MICE) according to the 10-fold split, and classification scores > 0.5 were labelled as white adipocyte pixels. Then, for each hand traced contour we computed the white adipocyte score $Z_{obj} = \#\mathrm{WA} \;/\; (\#\mathrm{WA} + \#\mathrm{NWA}) \in [0, 1]$, where #WA stands for the number of white

adipocyte pixels within the object, and #NWA for the number of non white adipocyte pixels. This Z_{obj} was compared to the ground truth score $Z_{gt} = \{0, 1\}$ to compute the receiver operating characteristic (ROC) curve, weighting each object by its number of pixels. This way, the ROC takes into account the fact that white adipocyte objects tend to be much smaller and more numerous than other objects. Namely, this allows us to interpret the classification error in terms of the more balanced tissue area classification error (our training contours contain ≈ 23.7 · 10^6 white adipocyte pixels and ≈ 45.1 · 10^6 non white adipocyte pixels).

Segmentation sub-pipeline validation

We applied the segmentation sub-pipeline to each of the 55 colour-corrected images with hand traced cells (see Table MICE), using 10-fold cross validation. This produced Auto and Corrected contours for each image that we compared to the hand traced ones for validation.

The literature relies on summary statistics for segmentation validation. For instance, per-image average diameter[40], mean Dice coefficient (DC) and mean Jaccard Index[47], median cell area/volume per subject[13], IoU and F1-score [71] and mean cell area[11]. As summary statistics, we used the DC and relative area error between each pipeline-produced contour and the hand traced (ht) contours in the image. The $DC = 2a_{ht \cap pipeline}/(a_{ht} + a_{pipeline})$, where a are areas computed with polygon operations. The highest DC was considered the best match. DC ≤ 0.5 were considered no match, as they are usually contours that segment an adjacent object. To compare cell populations, cell area distributions were computed using box-and-whisker plots with median notches for the hand traced, Auto and Corrected contours with a match (Fig. SEG_VALIDATION). The lower whisker was set at the lowest datum above Q1 - 1.5 (Q3 - Q1), and the upper whisker, at the highest datum below Q3 + 1.5 (Q3 - Q1). Data outside the whiskers was displayed as circles. The relative area error was computed as $\epsilon_{Auto} = a_{Auto}/a_{ht} - 1$ and $\epsilon_{Corrected} = a_{Corrected}/a_{ht} - 1$, respectively.

Although they are the standard, summary statistics hide important subpopulation information. Ideally, we would like a constant relative error for all cell sizes, so that segmentation errors do not create spurious changes in cell subpopulations. To assess this for our algorithm, we plotted Auto and Corrected relative errors vs. hand traced cell area. We then sorted the errors by hand traced cell area and computed rolling Harrell-Davis (HD) estimates for quartiles={Q1, Q2, Q3} of the relative errors using a rolling window of 100 points that shrinks as it overflows at the extremes down to a minimum size of 20 points. The rolling median and interquartile range Q1-Q3 were plotted as a solid curve and red shaded area, respectively (Fig. SEG_VALIDATION(c)-(f)). Finally, we computed the global HD estimate for Q2 and plotted it as a horizontal green line.

Phenotype framework for WAT

We use two main tools to study WAT: cell area probability density functions (pdfs) to represent cell populations, and linear models to phenotype body and depot weight, and cell area.

To estimate cell area pdfs, we applied Kernel Density Estimation with a Gaussian kernel and bandwidth = 100 μm to preserve sharp peaks in the distributions. Quartiles (Q1, Q2, Q3) or other quantiles were computed with Harrell-Davis (HD) quantile estimation(Harrell and Davis 1982). For a weighted average of quantiles from multiple mice, for each p-quantile for the i-th mouse $q_i(p)$ we computed the corresponding HD standard error $se_i(p)$, using jackknife and applied the meta-analysis inverse-variance method [72,73]. The combined quantile estimate is

$$\hat{q}(p) = \frac{\sum_i q_i(p)/se_i^2(p)}{\sum_i 1/se_i^2(p)}$$

and the combined standard error is

$$\hat{se}(p) = \sqrt{\frac{1}{\sum_i 1/se_i^2(p)}}$$

The 95%-CI for the combined quantile estimate is

95%-CI(p) = $(\hat{q}(p) - 1.96\ \hat{se}(p),\ \hat{q}(p) + 1.96\ \hat{se}(p))$

Previous *KLF14* phenotype studies stratified datasets by sex[11,13] and compared two groups (sex or risk allele vs. non-risk allele) with summary statistics, namely median adipocyte area with a Wilcoxon signed-rank test[13], or mean area with inverse variance fixed effects meta-analysis[11]. Neither adjusted for BW (the Wilcoxon signed-rank test compared pairs of BMI-matched subjects and the meta-analysis study selected subjects within the normal BMI range), although BW is a known general phenotype confounder[74,75], in particular of mouse adipocyte diameter and DW[28]. In addition, summary statistics could miss changes in cell subpopulations that become apparent when comparing population quantiles[76].

To tackle equivalent issues in the mouse model, we propose a *Klf14* phenotype framework with three interconnected levels: the mouse level (body weight, BW), the depot level (depot weight, DW) and the cell level (quantile cell area), to remove confounding effects. Also, we study cell sizes at different quantiles; for the sake of simplicity, we use the three quartiles in this work, but the model could be applied to any other quantiles. In each phenotype level, we used linear models to quantify the trait (e.g. cell area) vs. categorical effect (e.g. genotype) and continuous covariate (e.g. depot weight). For this, we built upon the mixed model proposed by Karp et al.[74,75] (trait ~ genotype*sex + body_weight + (1|batch)), where “batch” groups animals processed the same day, and adipocyte size or DW vs. BW non-linear models by van Beek et al. [28]. Our approach is different in six ways: 1) as we have a smaller number of animals per stratum, we did not consider “batch”, and thus, replaced mixed models for simpler Ordinary Least Squares (OLS) models or Robust Linear Models (RLMs). In larger studies, batch and other random effects (litter, mother) could be considered. 2) Due to sexual dimorphism that would dominate other effects, we stratified the study by sex, rather than include sex as a covariate. 3) We considered two effects, genotype and parent, instead of just genotype. Due to limited data (~18 mice per sex/depot/effect stratum), we performed separate exploratory analysis for each effect rather than combining both in the same model. 4) We considered two covariates to adjust for, body weight (BW) and depot weight (DW),

instead of just BW. 5) We added an interaction, “effect * covariate = effect + covariate + effect:covariate”, to allow for different slopes in the trait vs. covariate relationship. 6) We use adipocyte area instead of adipocyte diameter[28] to linearise the relationship with DW.

Accordingly, the three-level linear models we propose are: (BW ~ effect), (DW ~ effect * BW/BW) and ($area_q$ ~ effect $*$ DW), where BW=33.44 g is the mean BW of all animals, effect $\in$ {genotype, parent} and q={Q1, Q2, Q3} are quartiles. BW, DW and $area_q$ are continuous variables, and effect (genotype or parent) are categorical variables. In addition, an intercept is included in all models, but omitted in the formula for simplicity.

We extract two results from the models. First, the two-tailed t-test of the continuous variable’s slope β tells us whether trait and variable are correlated (due to the equivalence between slope and Pearson’s coefficient tests); for a significant correlation with p-value ≤ 0.05, β also provides e.g. the increase rate of DW with BW/BW. Second, a Likelihood Ratio Test (LRT) evaluates whether the categorical variable has a significant effect in the trait. The LRT compares the null-model without the effect to the model with the effect, e.g. ($area_q$ ~ DW) vs. ($area_q$ ~ genotype $*$ DW), both applied to the same data. (Details are provided in *Suppl. Likelihood Ratio Test*).

Body and depot weight were measured with Satorius BAL7000 scales. For cell area quantification, we applied DeepCytometer to 75 inguinal subcutaneous and 72 gonadal full histology slides (with one or two tissue slices each), including the 20 slides sampled for the hand traced data set. Each slide belongs to an animal and depot (gonadal or subcutaneous), stratified by sex —female (f) or male (m)—, heterozygous parent —father (PAT) or mother (MAT)— and genotype —wild type (WT) or heterozygous (Het)—. For segmentation of training slides, we used the pipeline instance that did not see any part of it in training. Slides not used for training were randomly assigned to one of the 10 pipeline instances.

When fitting multiple linear models, e.g. 8 models for 2 sexes, 2 depots and 2 effect categories, the p-values of all slopes undergoing t-tests where corrected using the FDR 2-stage Benjamini-Krieger-Yekutieli method [57] for a significance level α=0.05.

Quantile colour map plots to assess cell size heterogeneity

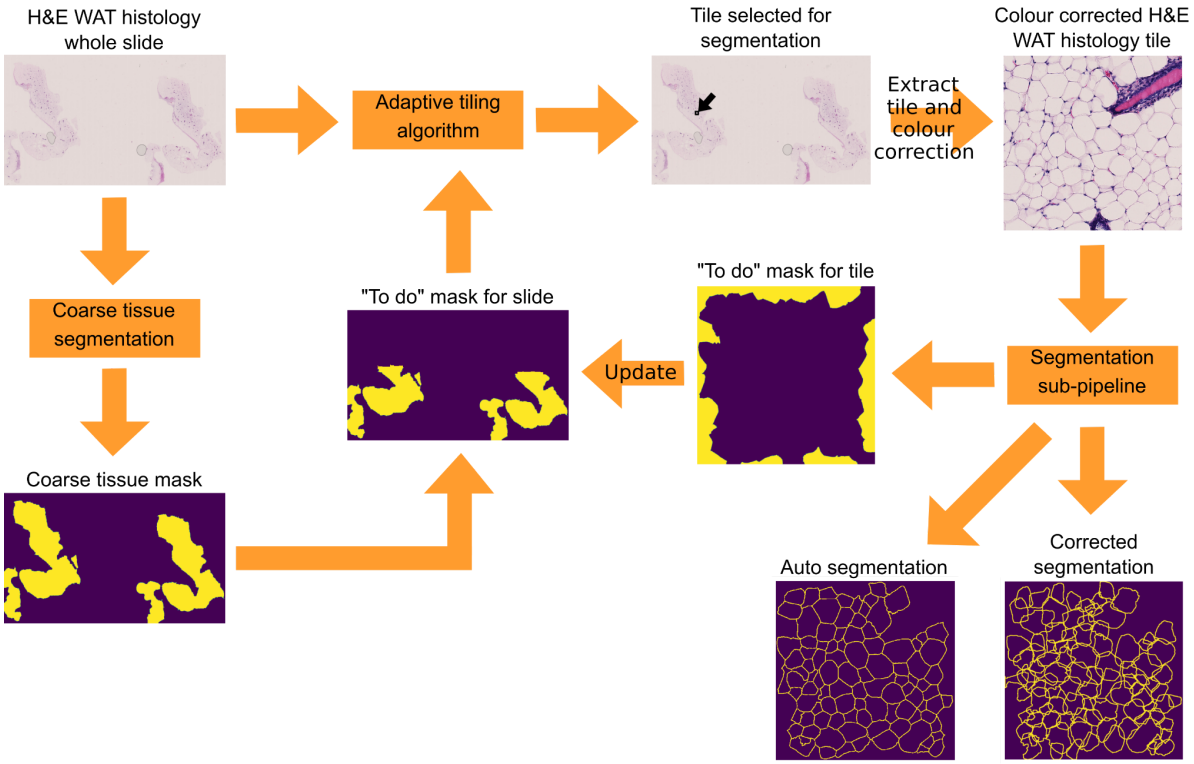
Because of skewness and long tails in the histogram (Fig. COLORMAP_F(a) and Fig. COLORMAP_M(a)), overlaying a colour map proportional to cell area or radius on the tissue sample produces a low contrast image that offers very little visual information about spatial patterns. In this section, we propose a colour map that provides a much clearer picture, by making the colour scale proportional to the ECDF of cell area.

We applied the pipeline to whole slides, to produce the scatter map $(x_i, y_i) \mapsto a_i$, where (x_i, y_i) are the coordinates of the i-th white adipocyte centroid and a_i is the white adipocyte’s area. We created a mask with all the pixels that belong to at least one white adipocyte. We used Delaunay triangulation and linear interpolation to rasterise the scatter map to pixel coordinates within the mask. Finally, area values were mapped to quantiles,

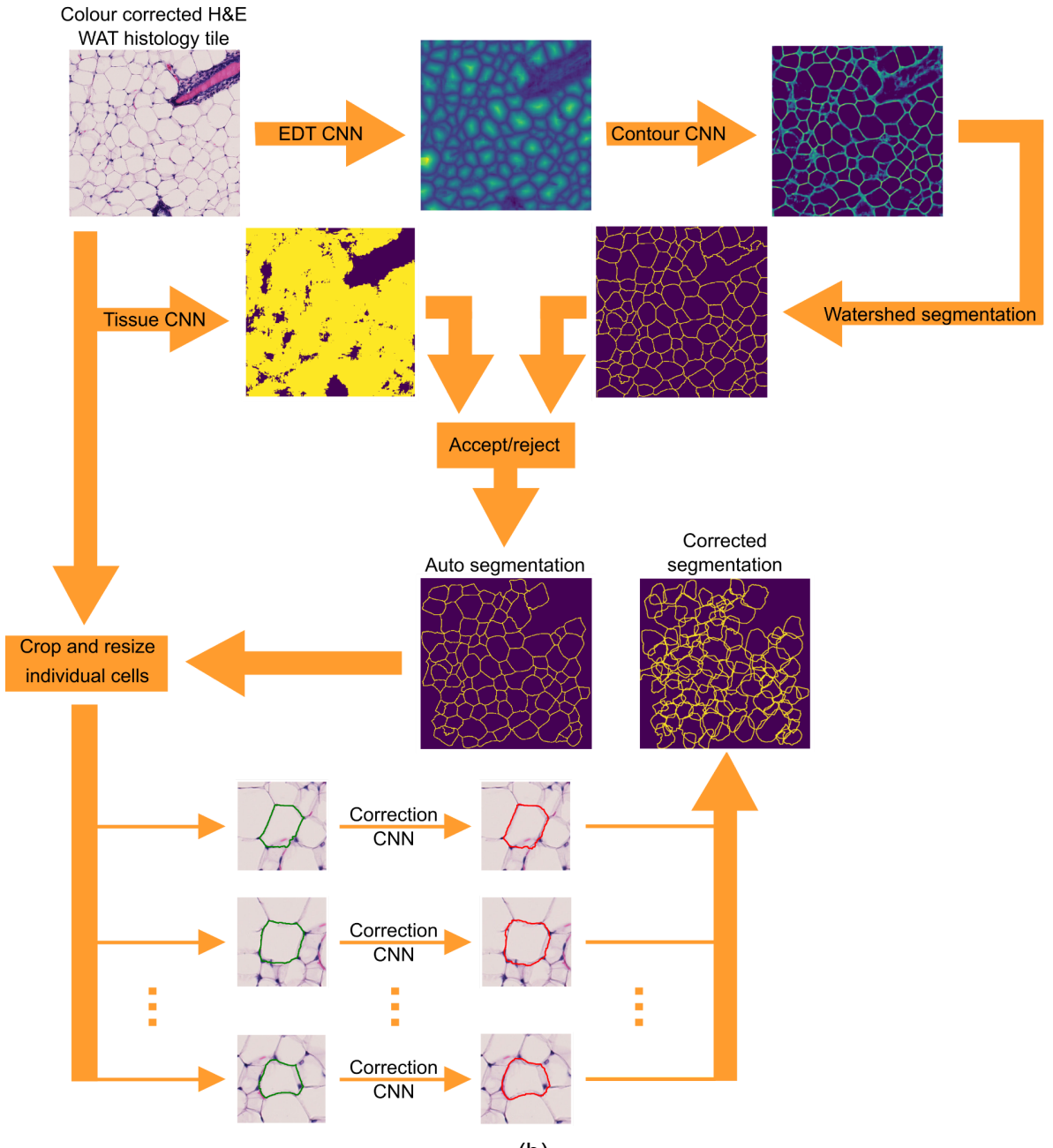
$q_i = f^{-1}_{\mathrm{HD,manual}}(a_i)$, where $f_{\mathrm{HD,manual}}$ is the HD quantile estimate computed on the hand traced training data set. Because of the significant sex effect in cell area (Fig. SEX_EFFECT), we stratified the training data set into females and males and computed separate ECDFs and colour maps for each group. Areas smaller or larger than in the hand traced data set were clipped to $q_i = 0.0$ or 1.0, respectively. The colour map was created in Hue/Saturation/Lightness (HSL) mode, by setting $S = 0.44,\ L = 0.69$ and $H_i = \frac{315}{360} q_i$. To sum up, each cell area was mapped to a colour using the relationship

$$a_i \mapsto q_i \in [0,\ 1] \mapsto (H,\ S,\ L) = \left(\frac{315}{360} q_i,\ 0.44,\ 0.69\right) \quad \text{(Eq_COLORMAP)}$$

Figures



(a)



(b)

Fig. PIPE. Pipeline diagrams. (a) DeepCytometer pipeline. (b) White adipocyte segmentation sub-pipeline.

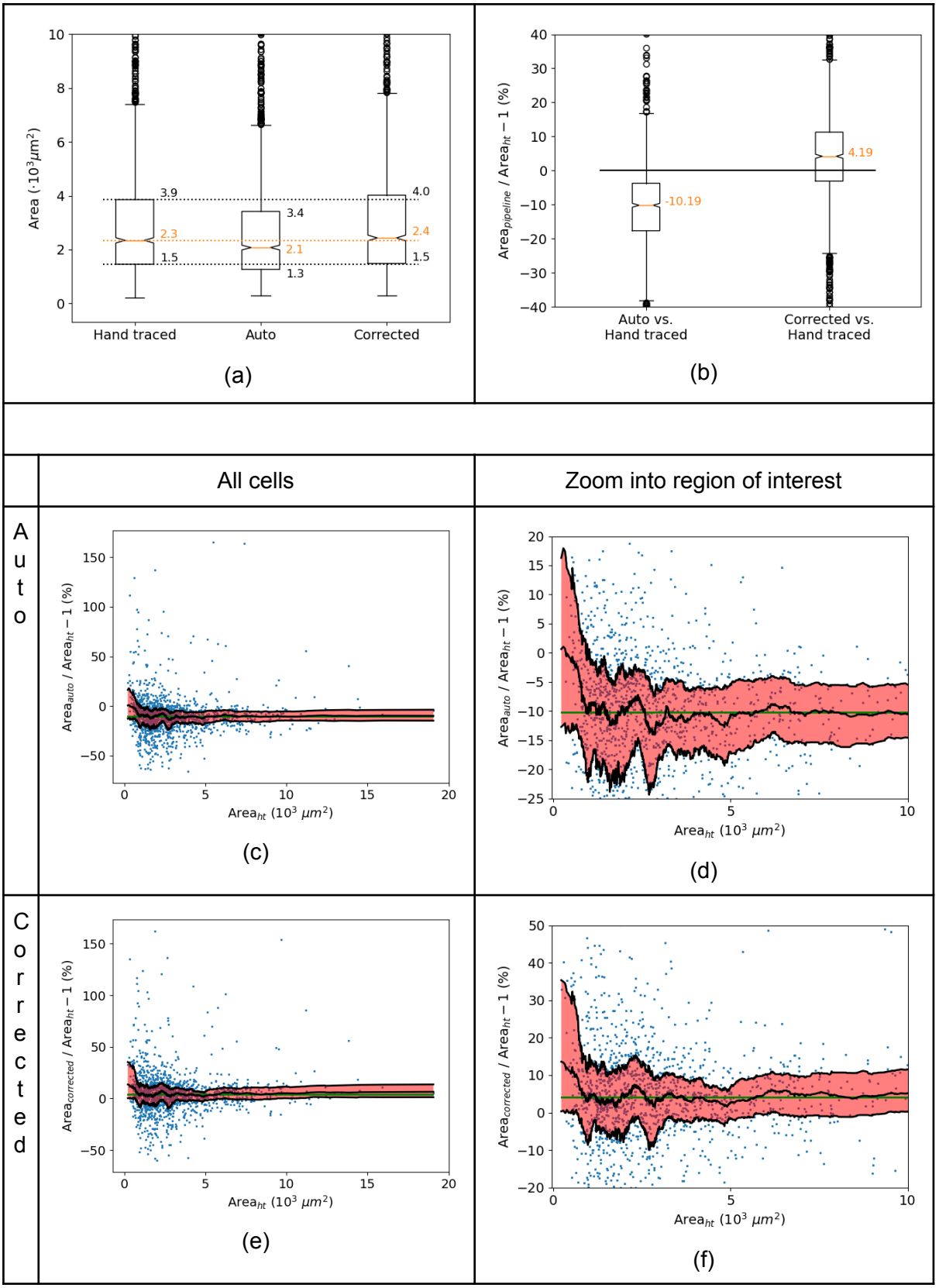
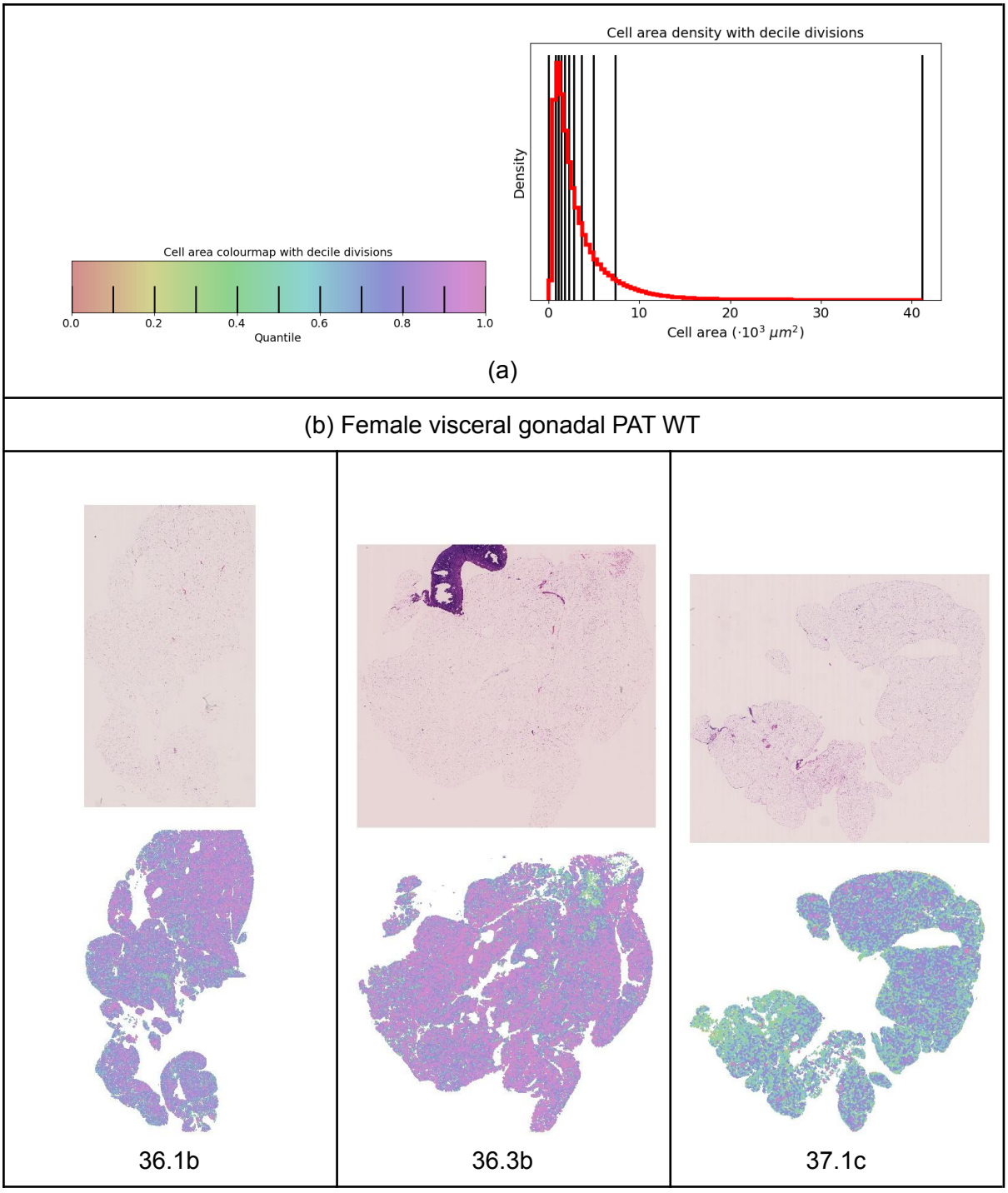
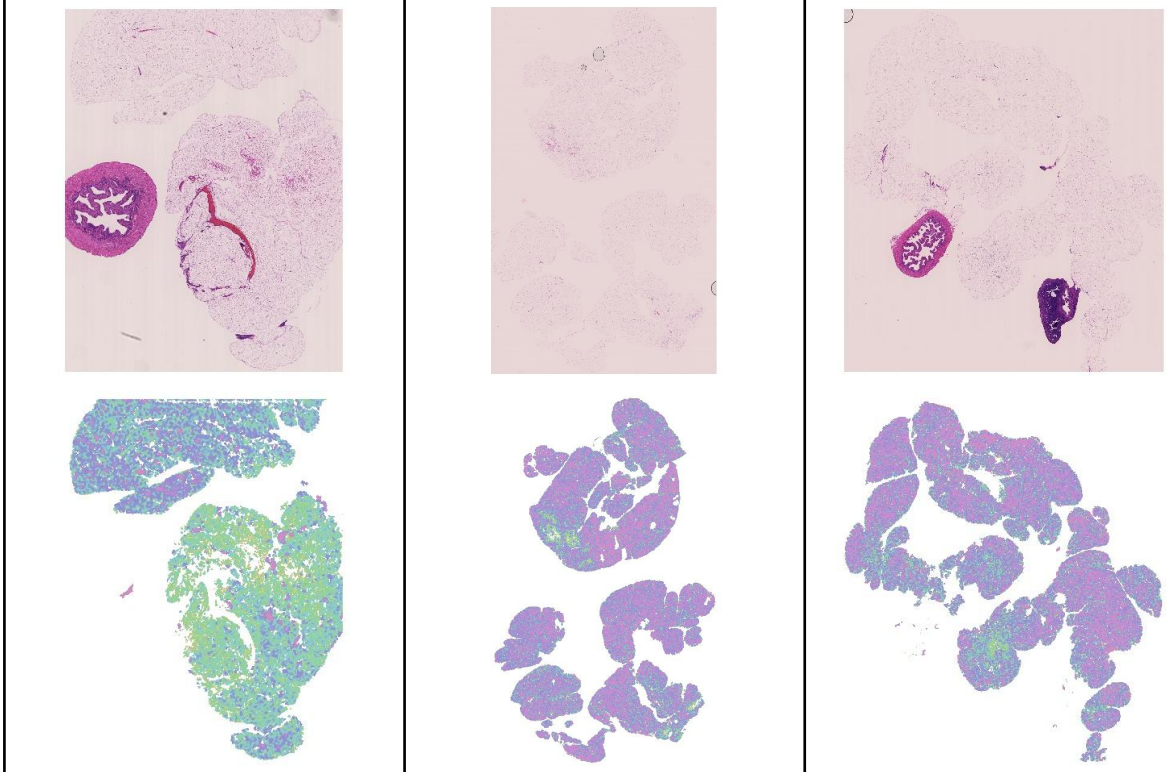


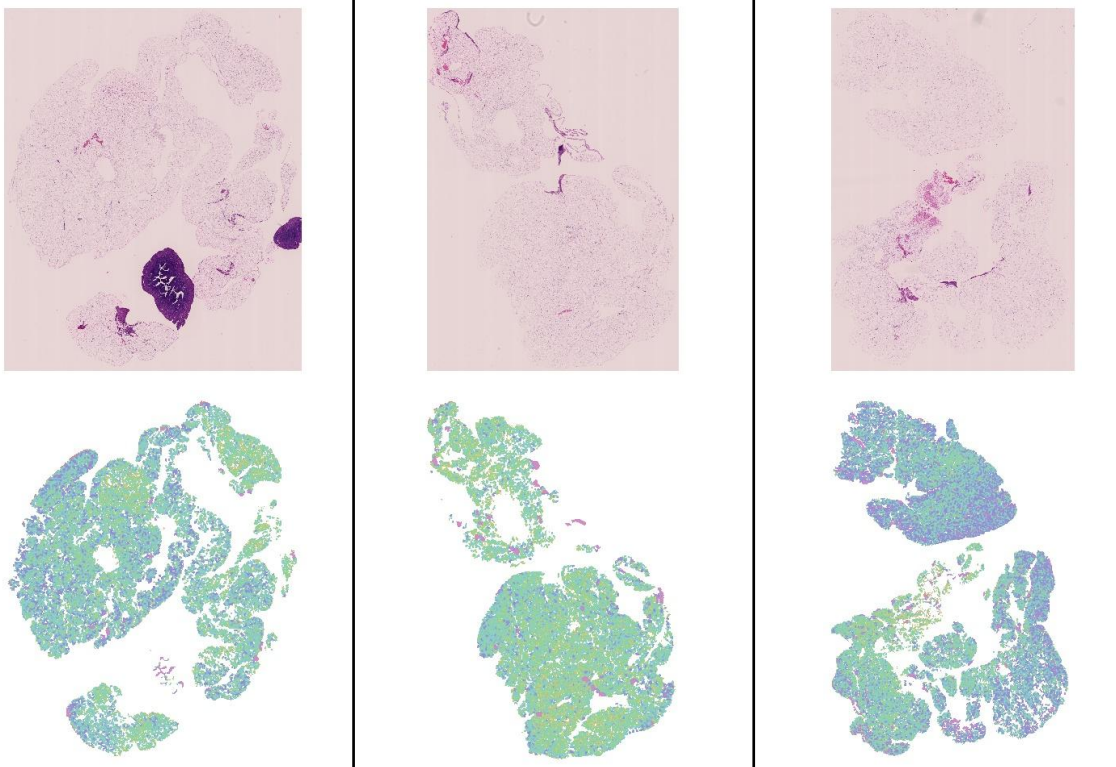
Fig. SEG_VALIDATION. Validation of the segmentation algorithm on the training dataset. (a) Comparison of cell area distribution between the hand traced dataset, the automatically segmented labels with the watershed algorithm (Auto) and segmentation correction (Corrected). Only matches with DC_{Auto} > 0.5 were considered valid. (b) Relative area

segmentation error with respect to hand traced cells. (c)-(f) Segmentation error as a function of hand traced cell area. For the whole cell population with Auto method (c) and Corrected method (e). Zoom into regions of interest for Auto (d) and Corrected (f). Blue dots correspond to individual cells in the training dataset. Black solid curves represent the HD quartiles (Q1, Q2, Q3) on points sorted by Area_{ht}, with space between the curves highlighted as a shaded red area. Green horizontal line represents the overall HD Q2 for all cells.



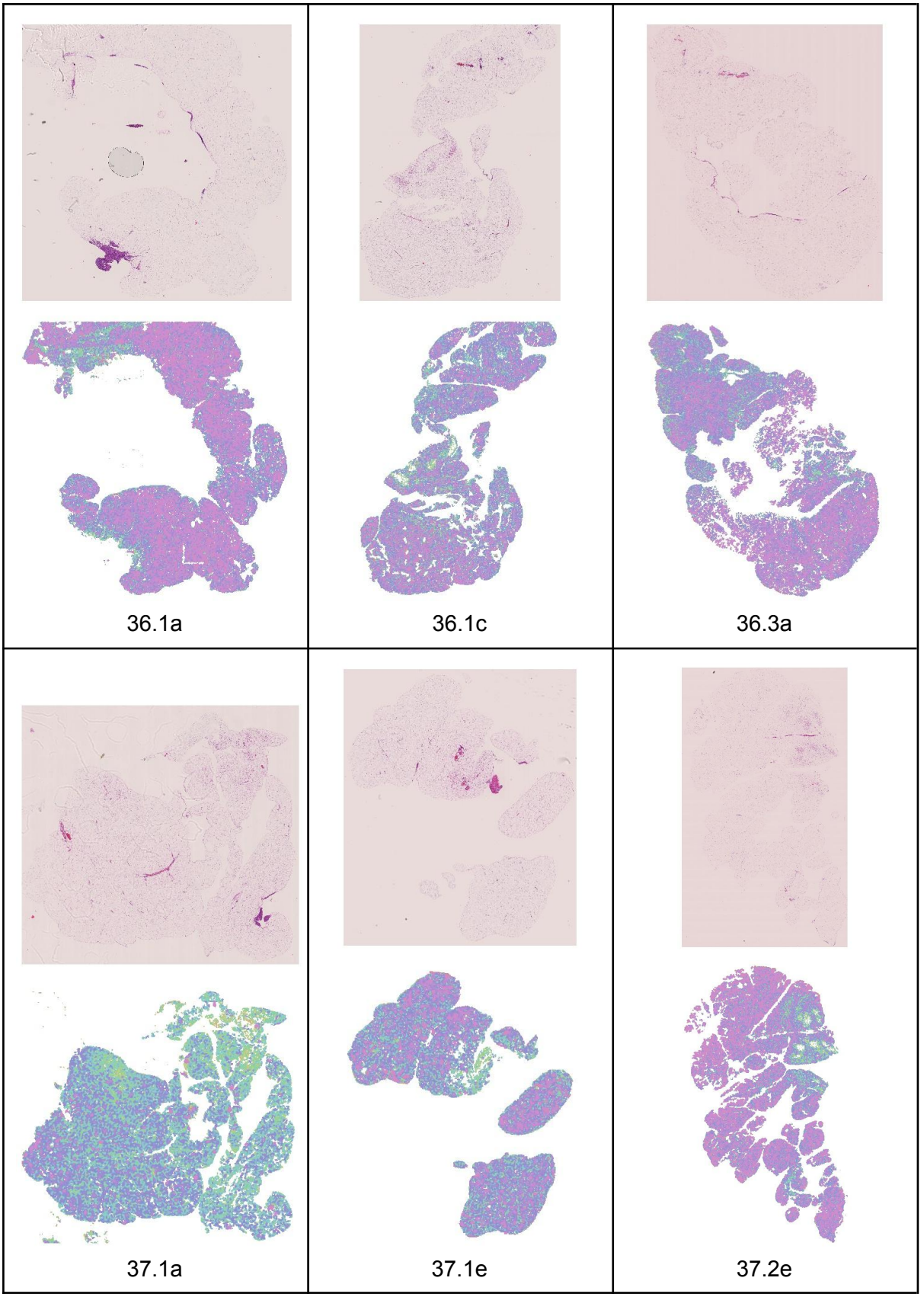


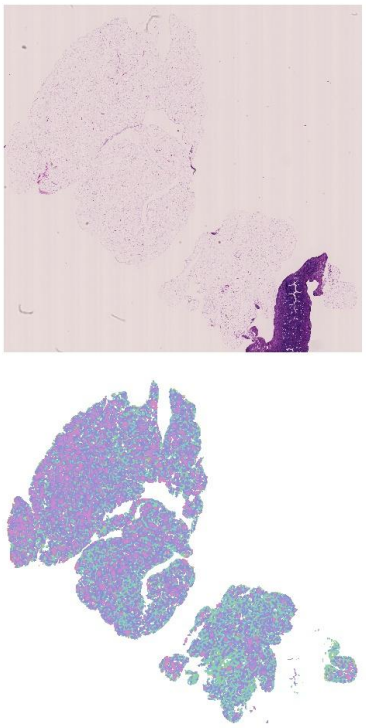
37.1d | 37.3a | 37.3c



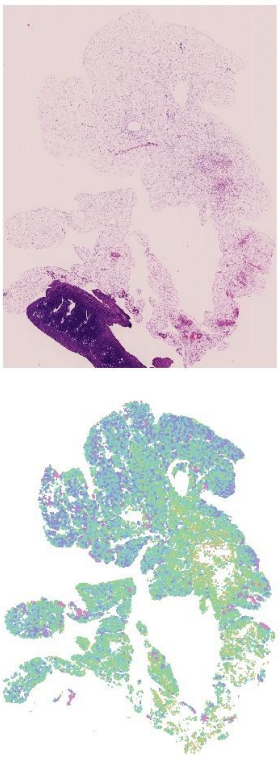
37.2b | 37.2d | 39.1h

(c) Female visceral gonadal PAT Het



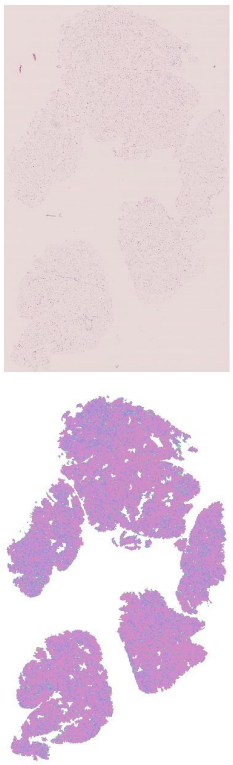


37.1b

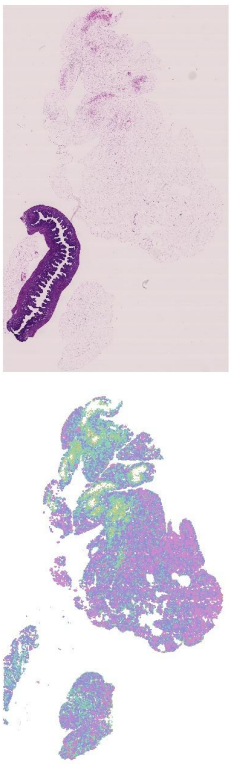


37.2c

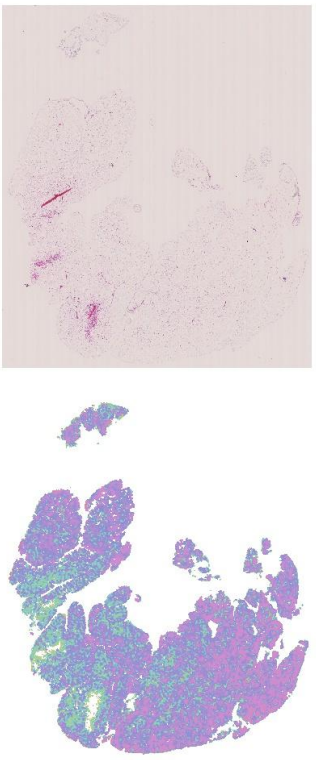
(d) Female visceral gonadal MAT WT



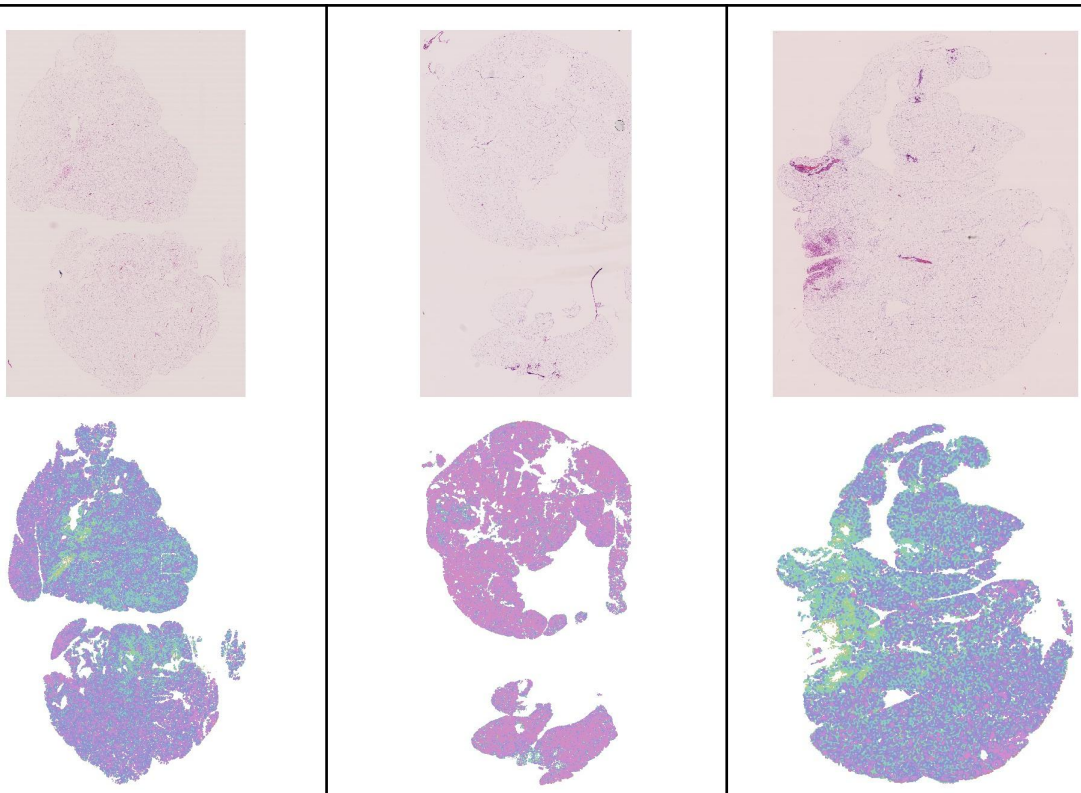
16.2a



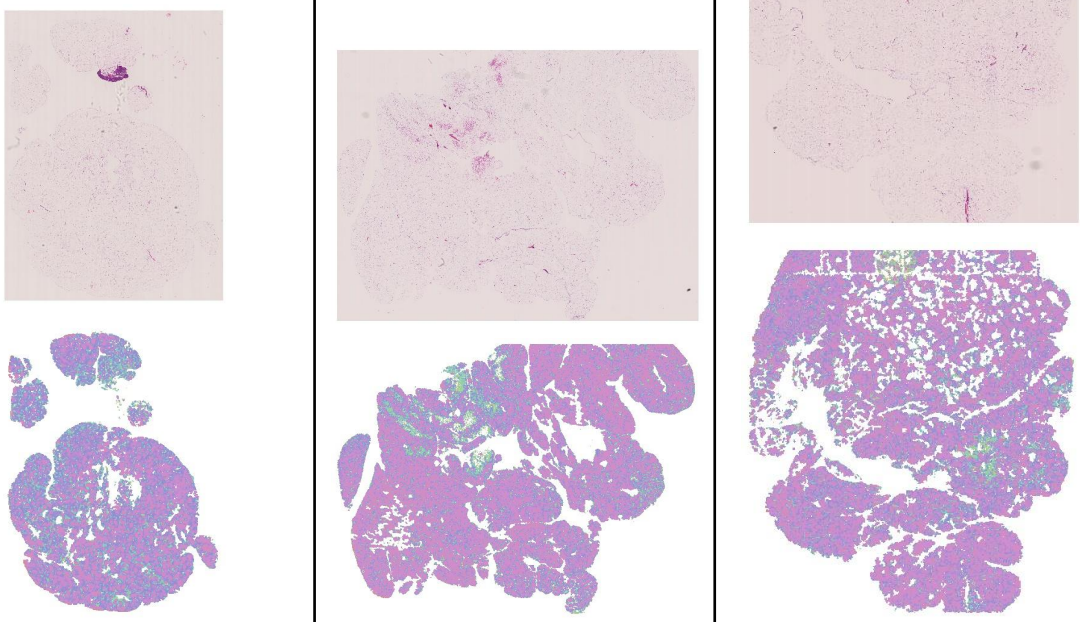
17.1d



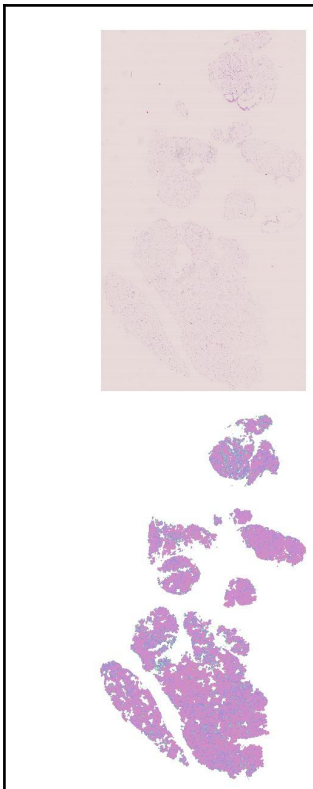
17.2b



17.2d | 18.1c | 19.2b

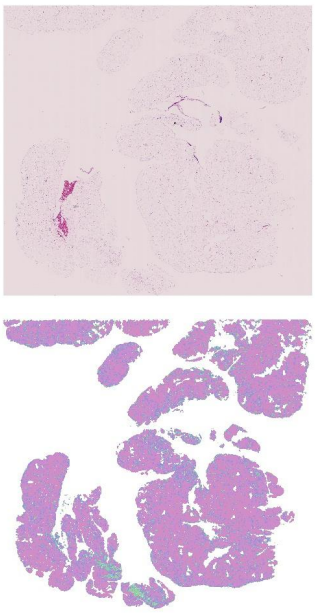


17.2a | 18.2c | 19.2c

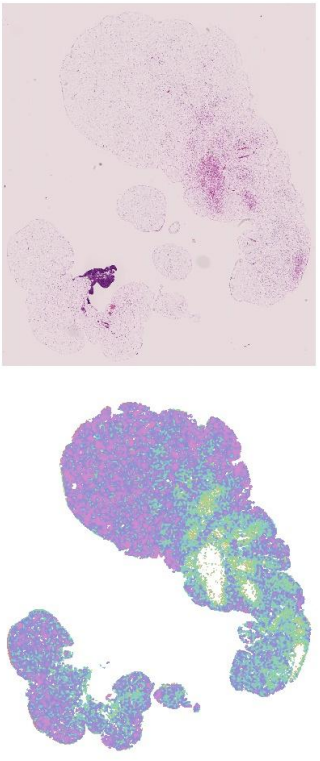


19.2e

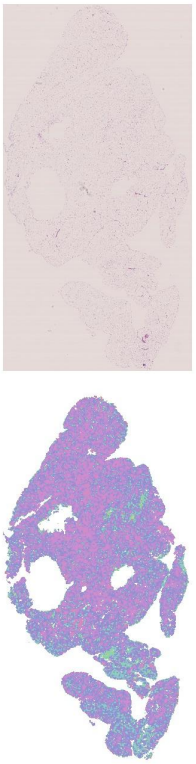
(e) Female visceral gonadal MAT Het



17.1b



17.1c



17.2c

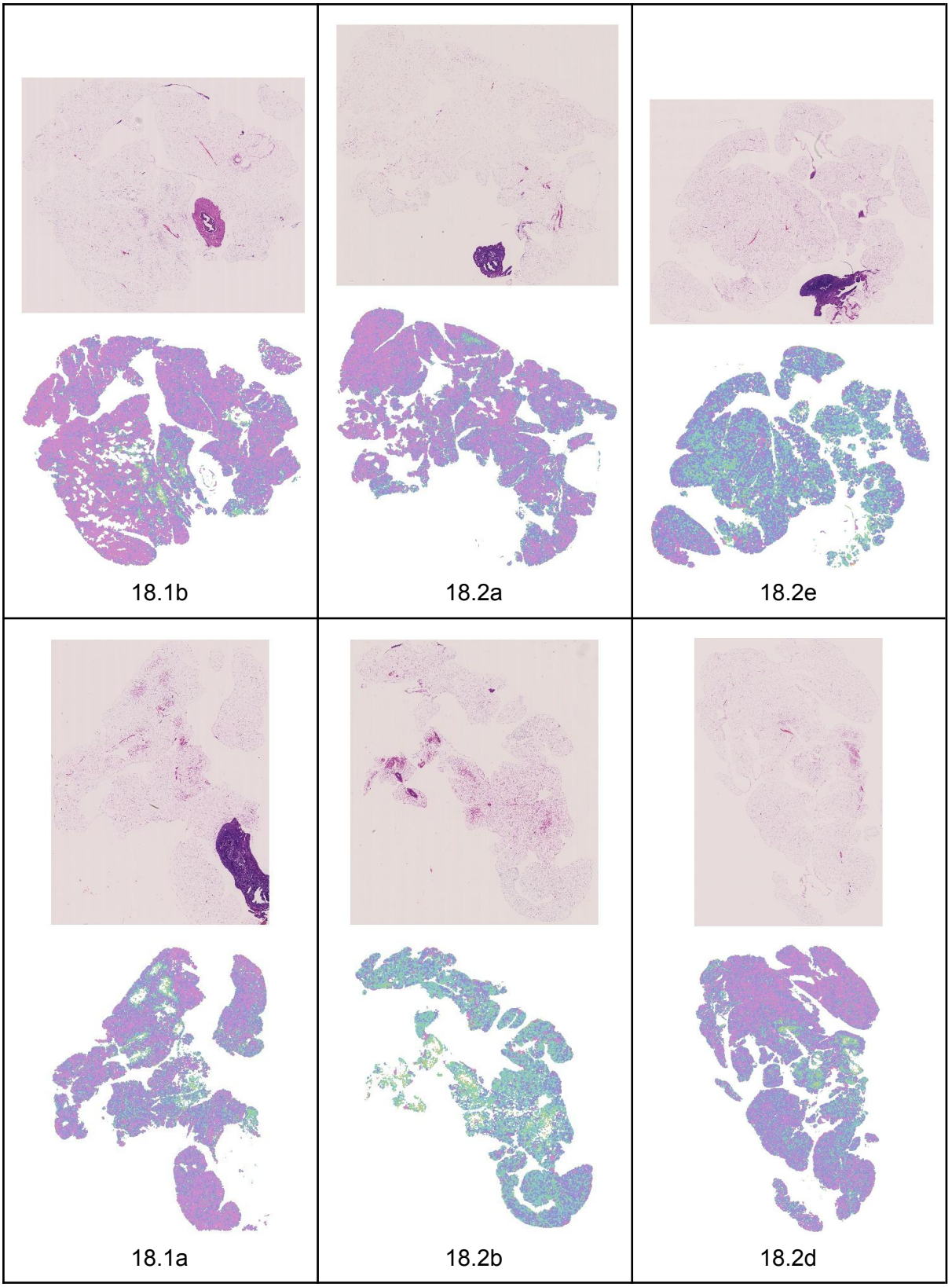
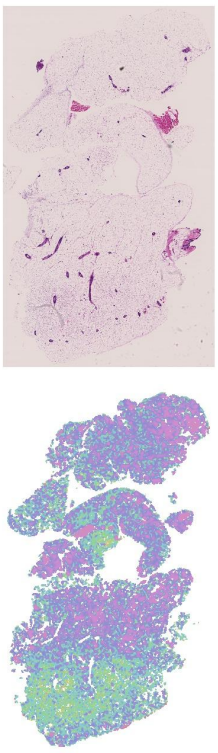
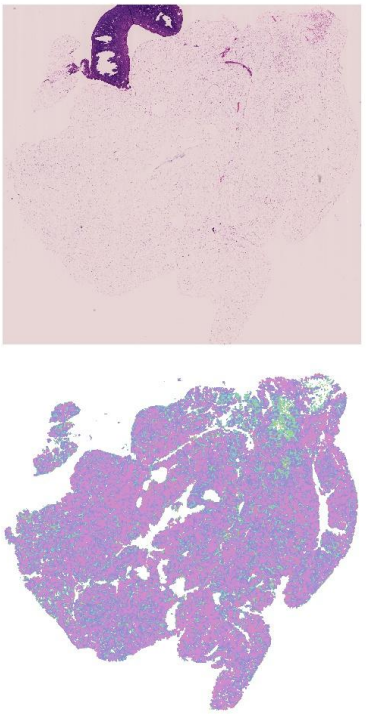


Fig. COLORMAP_F_GWAT. White adipocyte tissue histology and area quantile heatmaps for female visceral gonadal depot. (a): Quantile colour map and cell area density for female Corrected segmentation with deciles plotted for reference (vertical black lines). (b): PAT WT. (c): PAT Het. (d): MAT WT. (e): MAT Het.

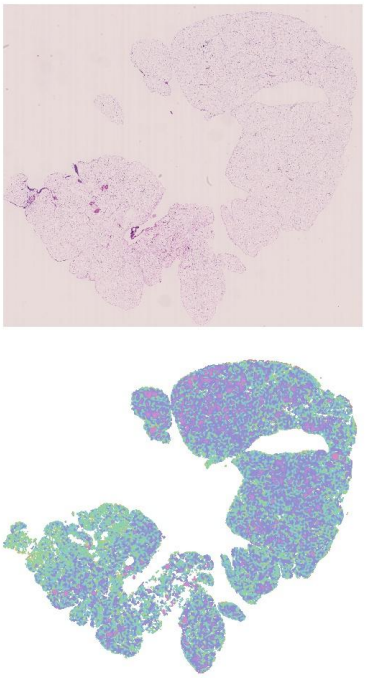
(a) Female inguinal subcutaneous PAT WT



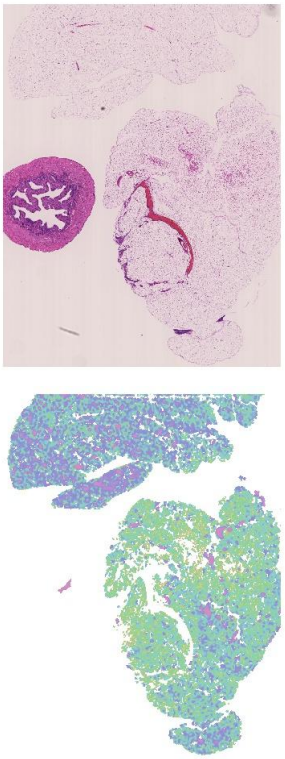
36.1b



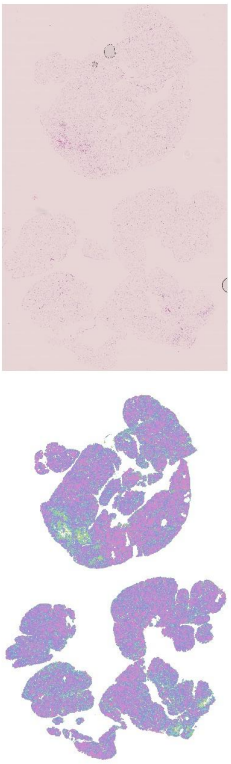
36.3b



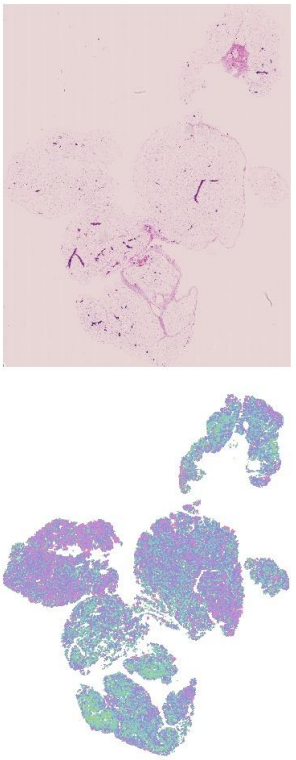
37.1c



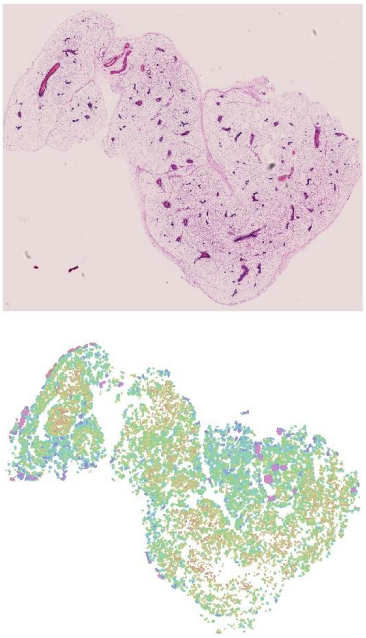
37.1d



37.3a

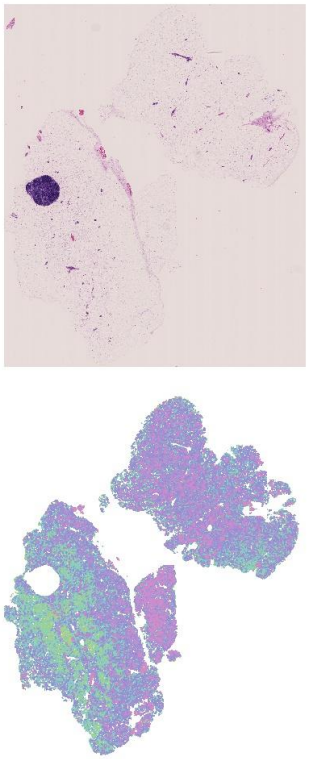


37.3c

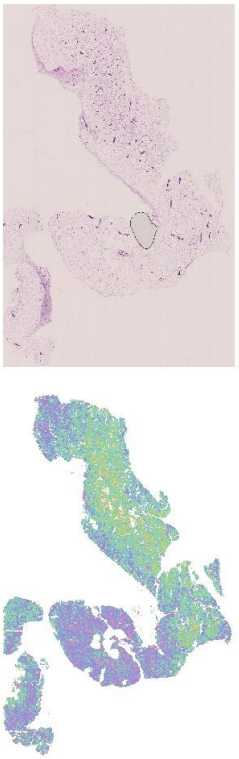


37.2b

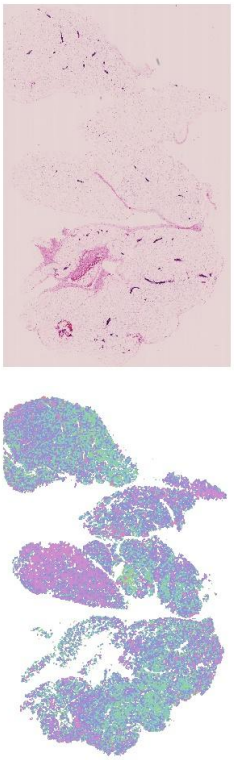
Female inguinal subcutaneous PAT Het



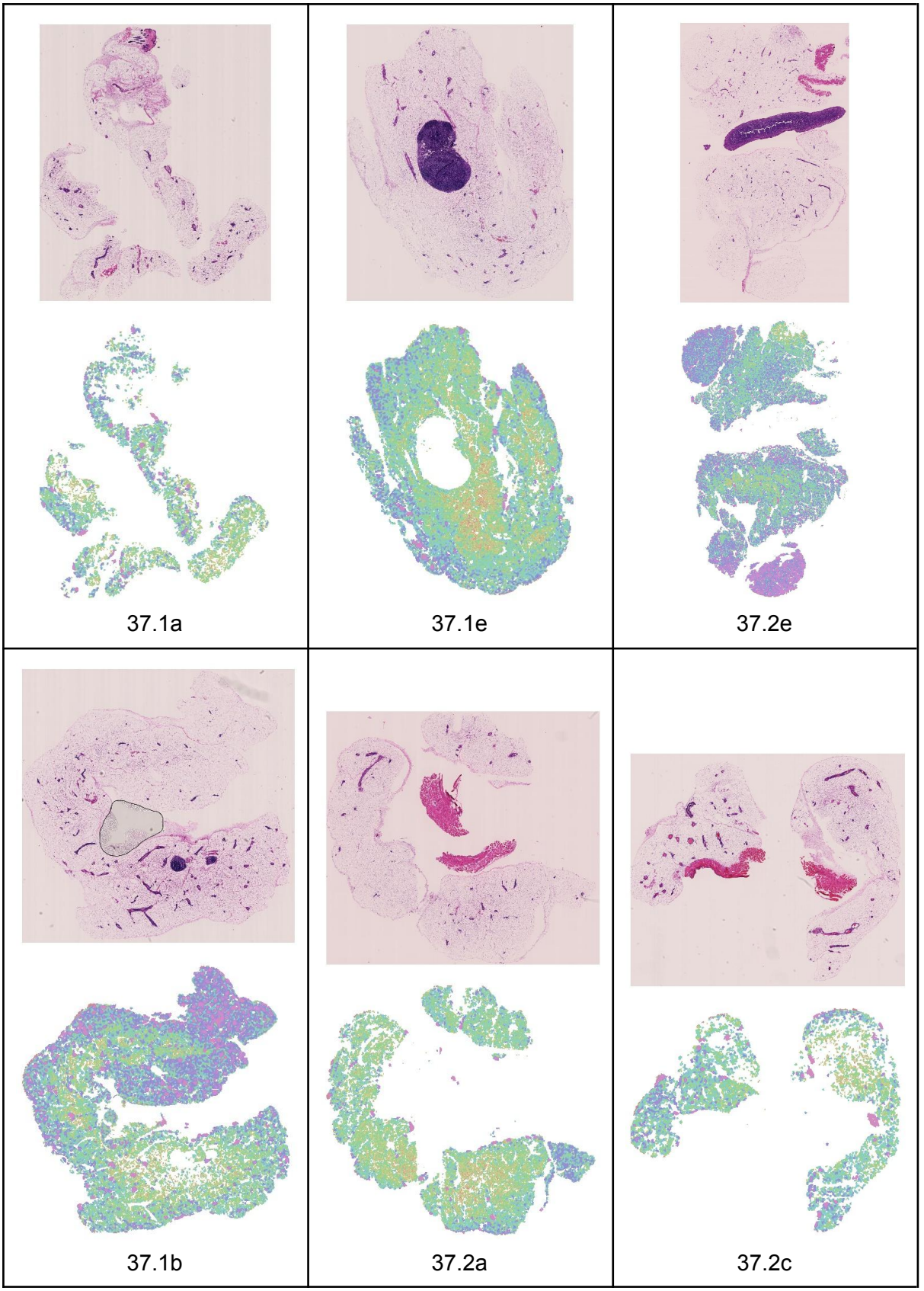
36.1a

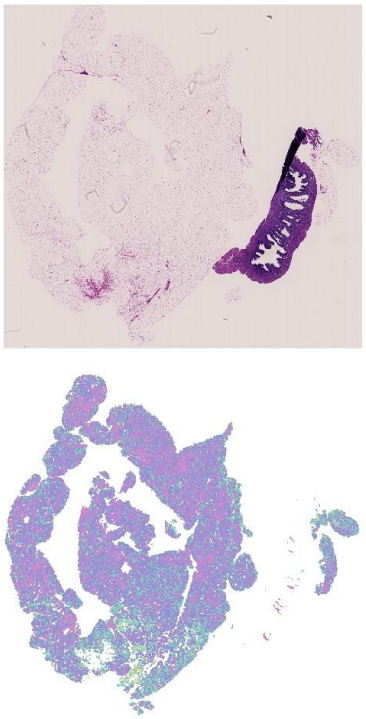


36.1c



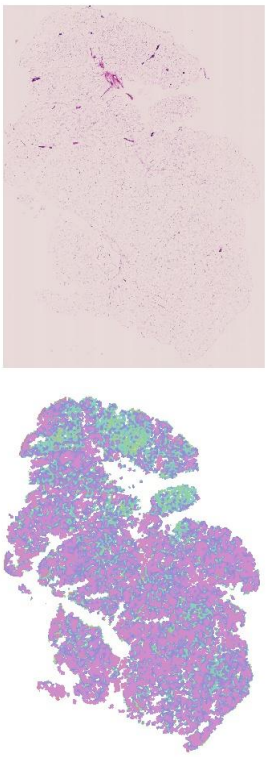
36.3a



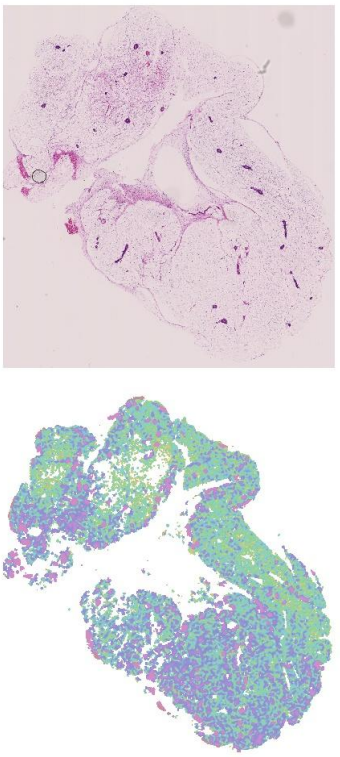


38.1a

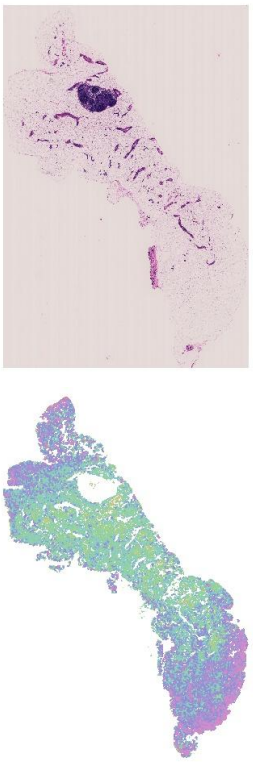
Female inguinal subcutaneous MAT WT



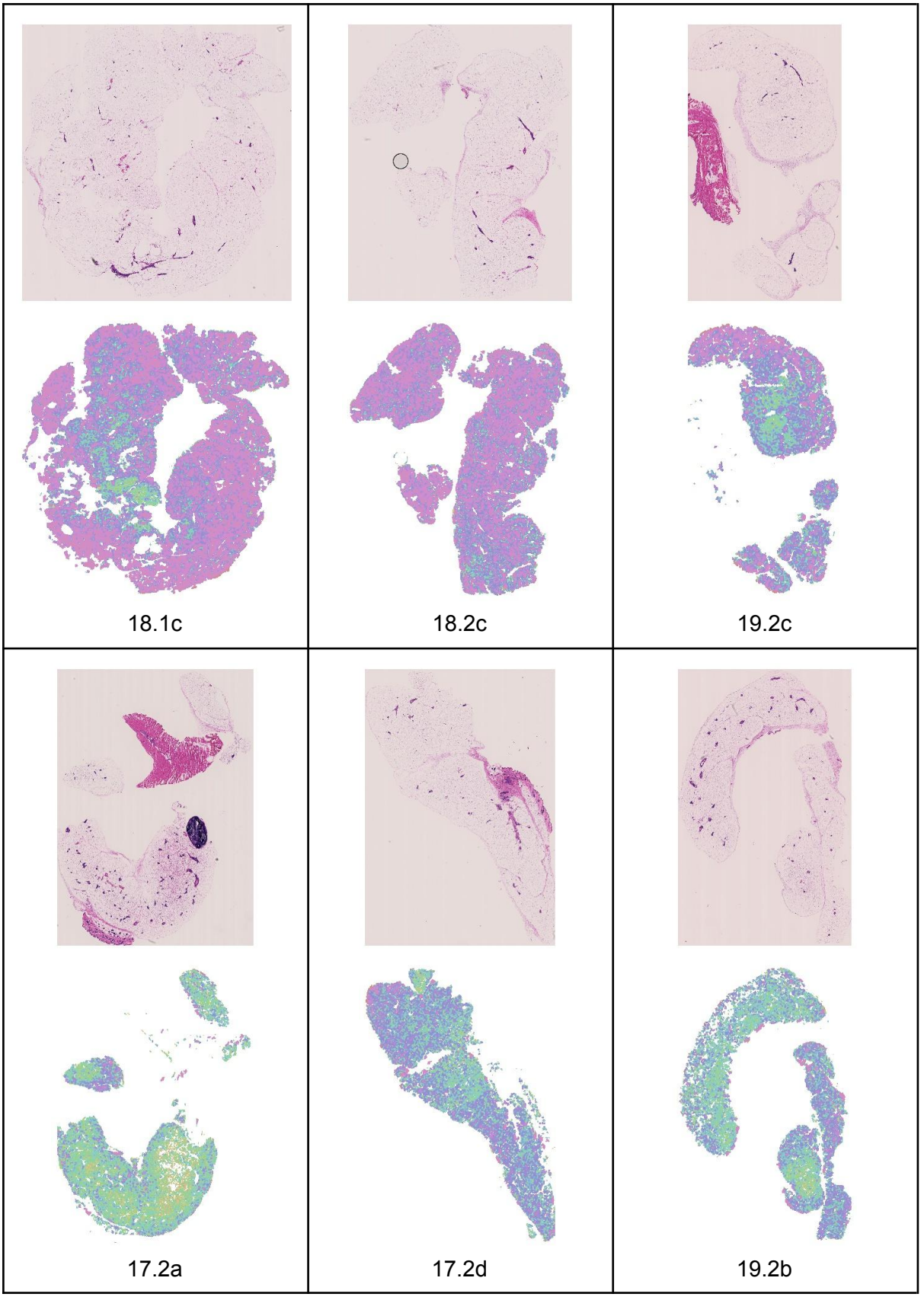
16.2a

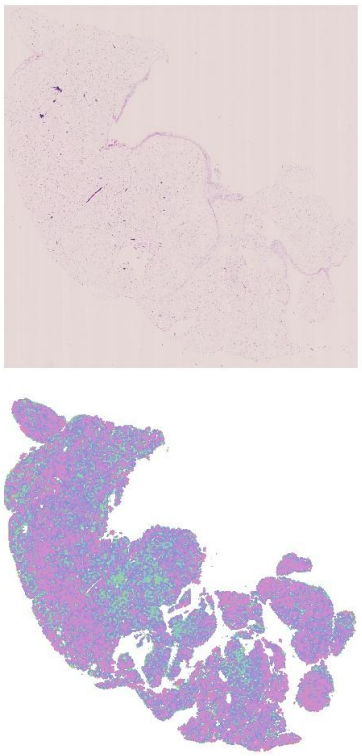


17.1d



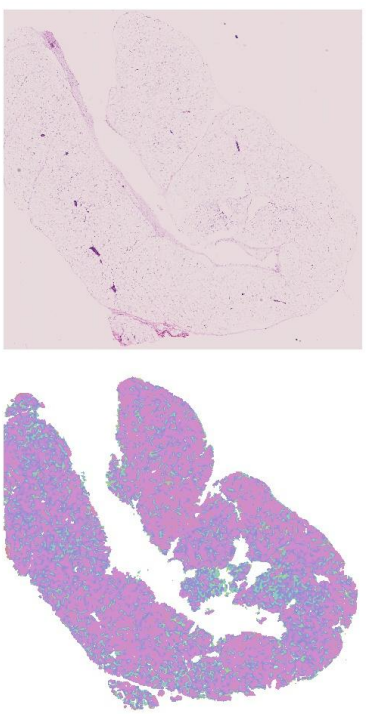
17.2b



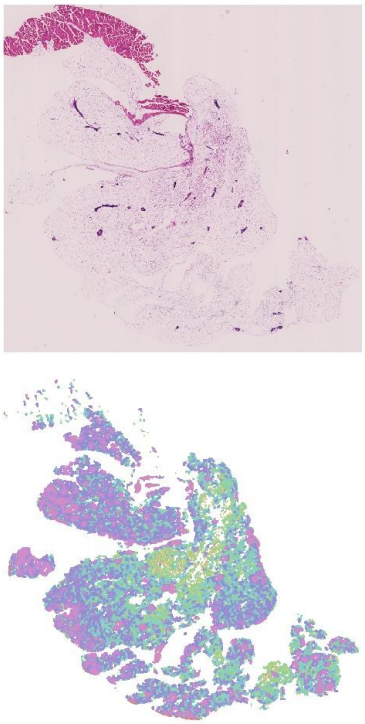


19.2e

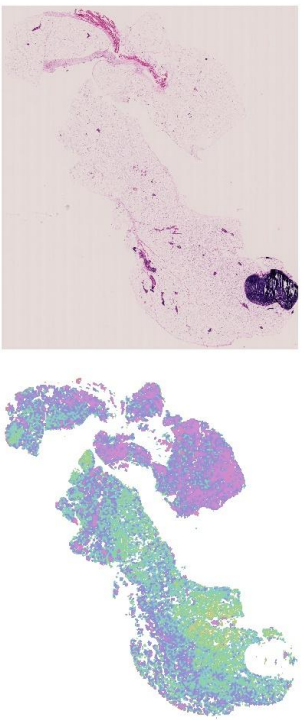
Female inguinal subcutaneous MAT Het



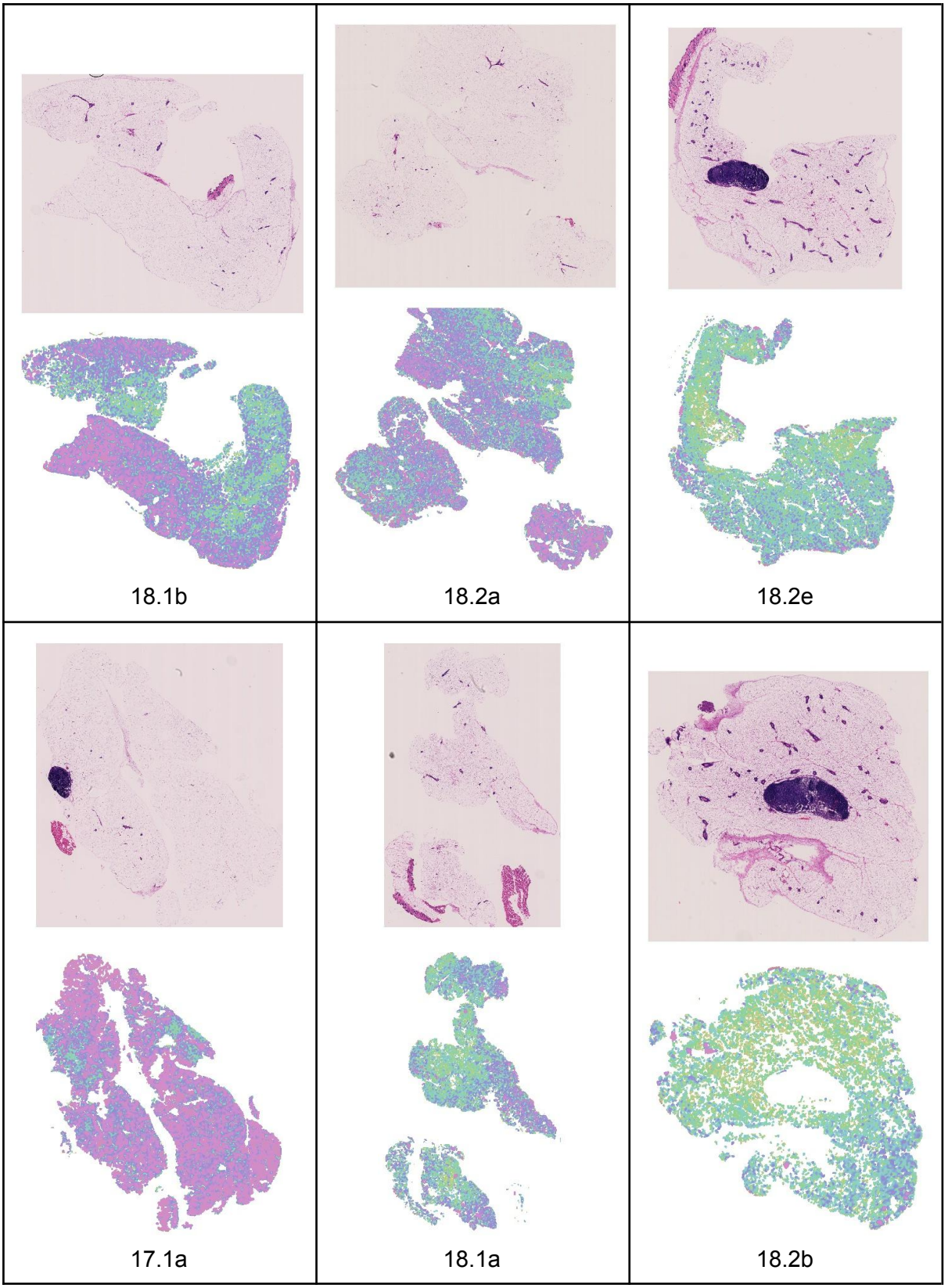
17.1b



17.1c



17.2c



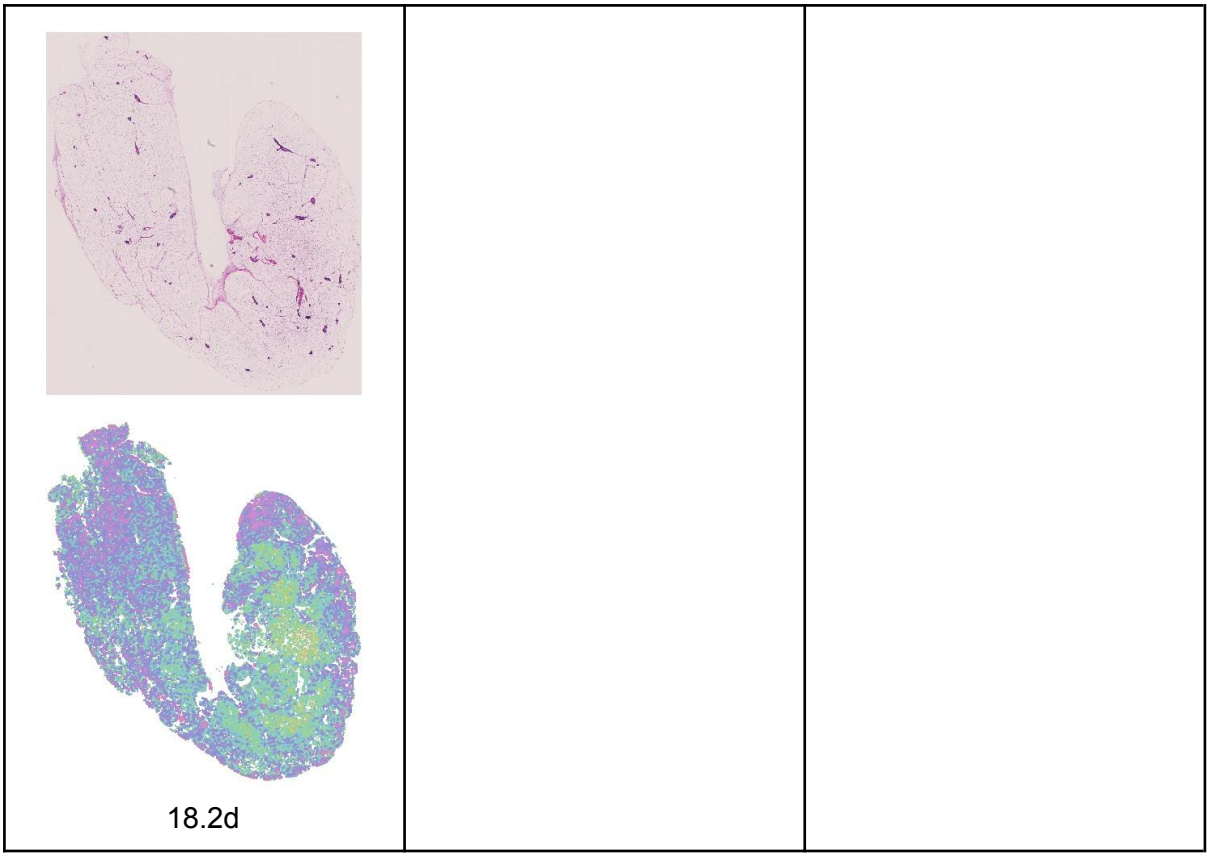
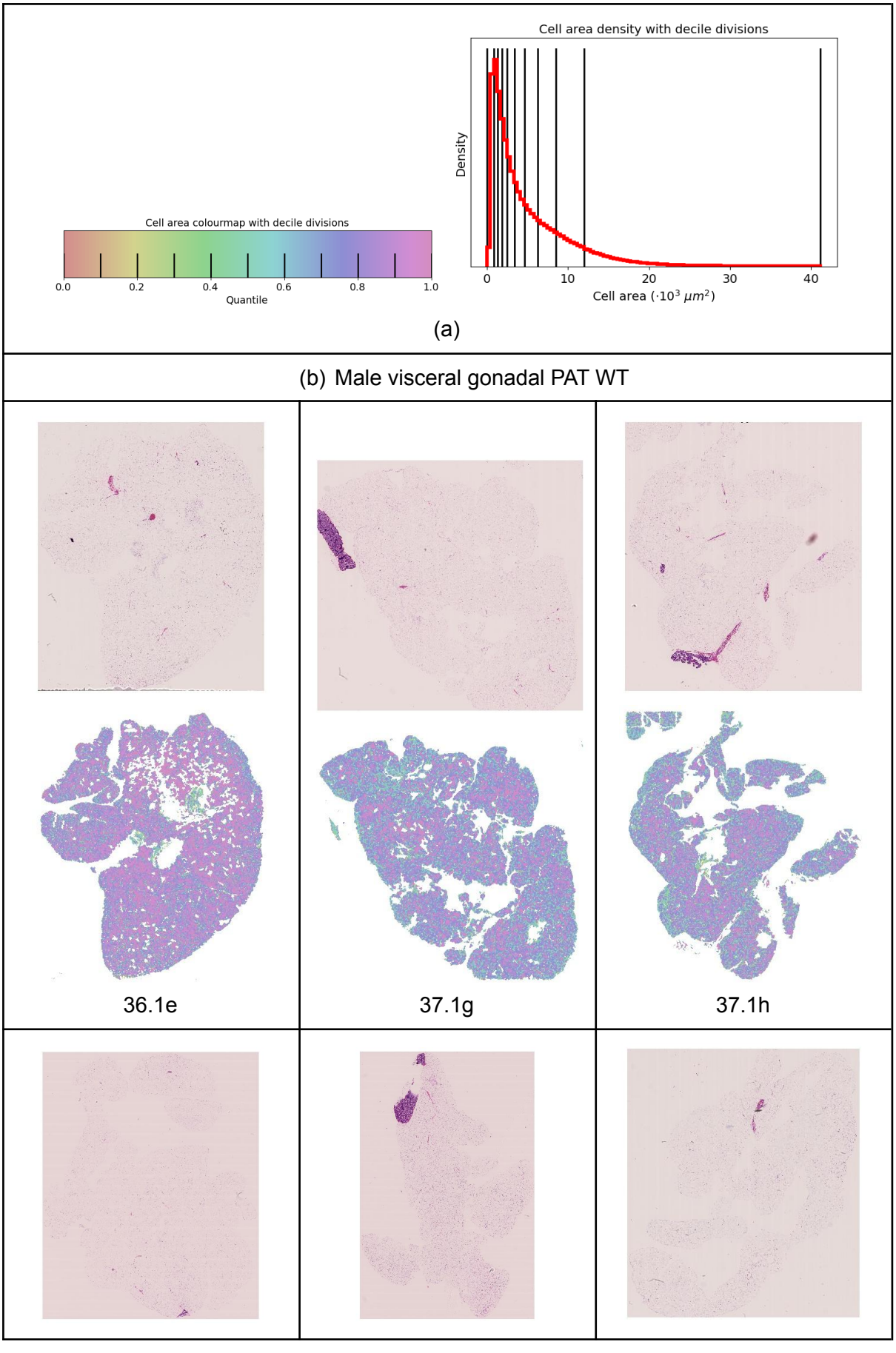
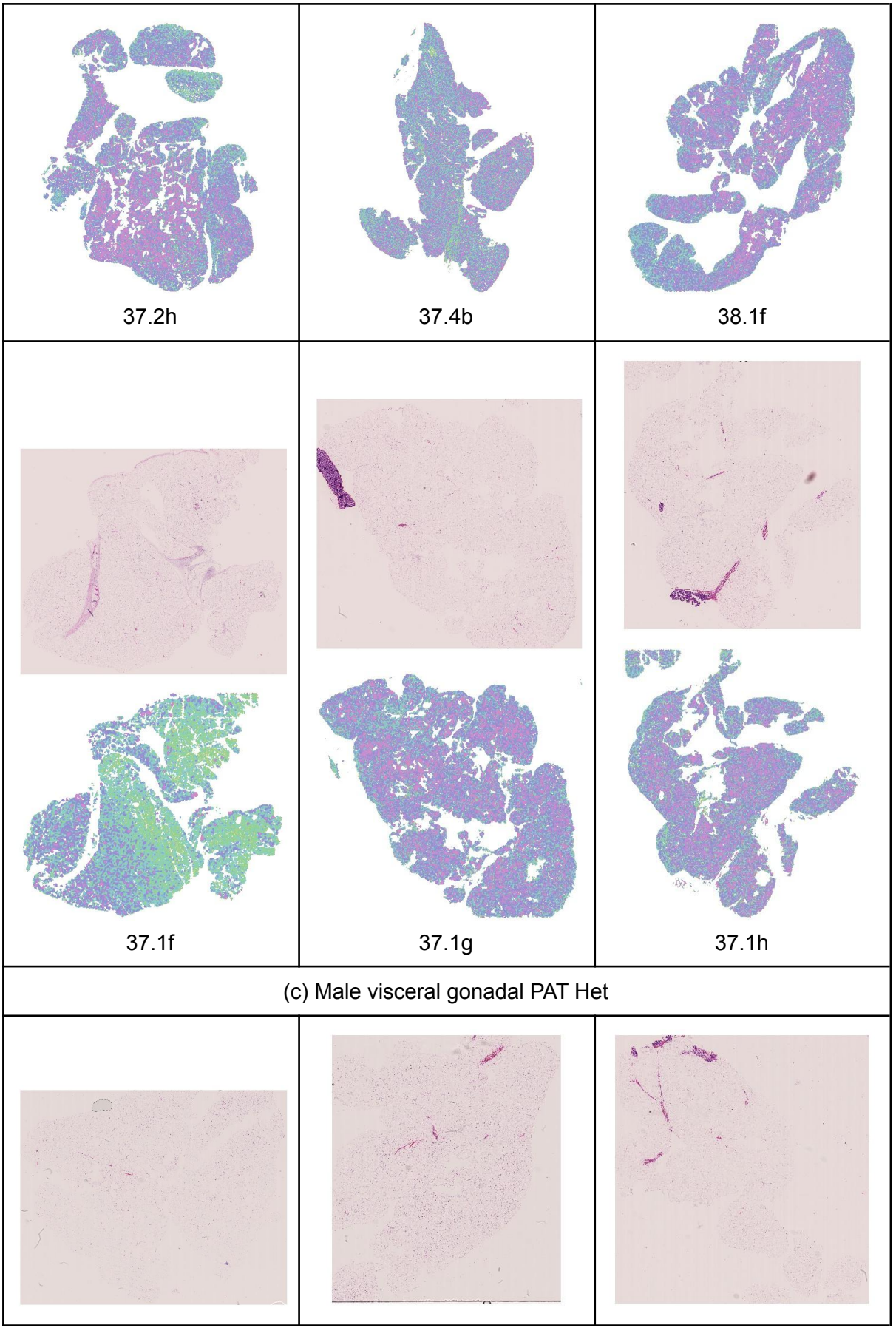
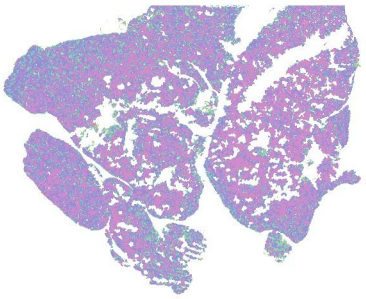


Fig. COLORMAP_F_SCWAT. White adipocyte tissue histology and area quantile heatmaps for female inguinal subcutaneous depot. (a): PAT WT. (b): PAT Het. (c): MAT WT. (d): MAT Het.



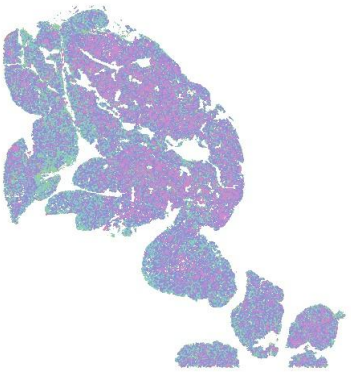


(c) Male visceral gonadal PAT Het

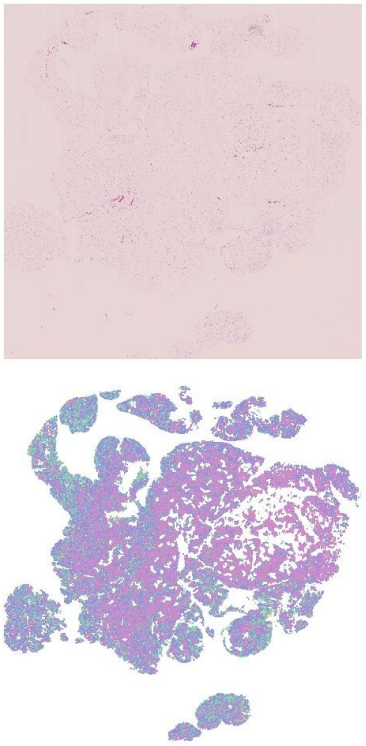


36.1d

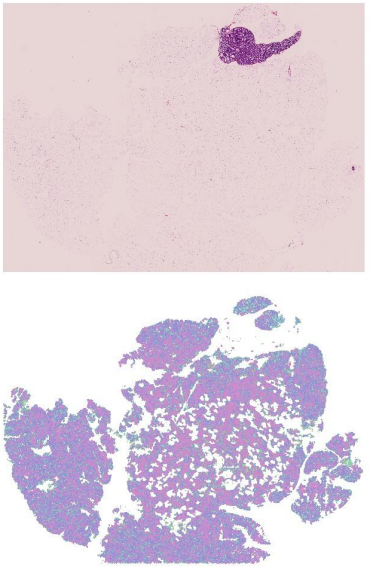
36.1f



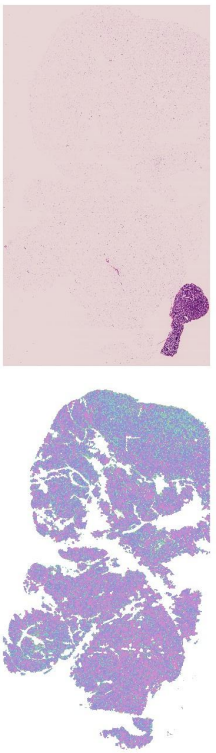
36.1j



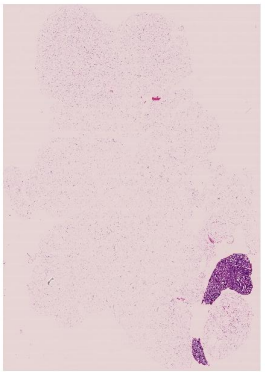
37.2g

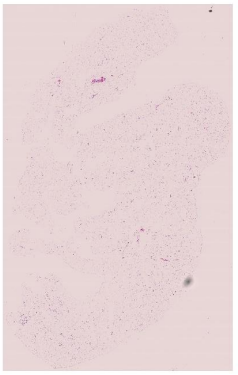


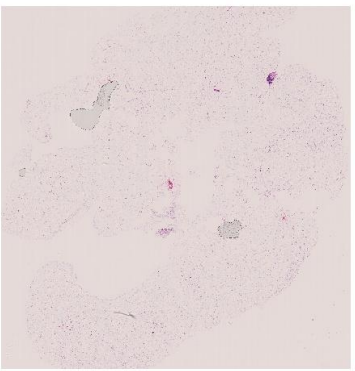
37.4a

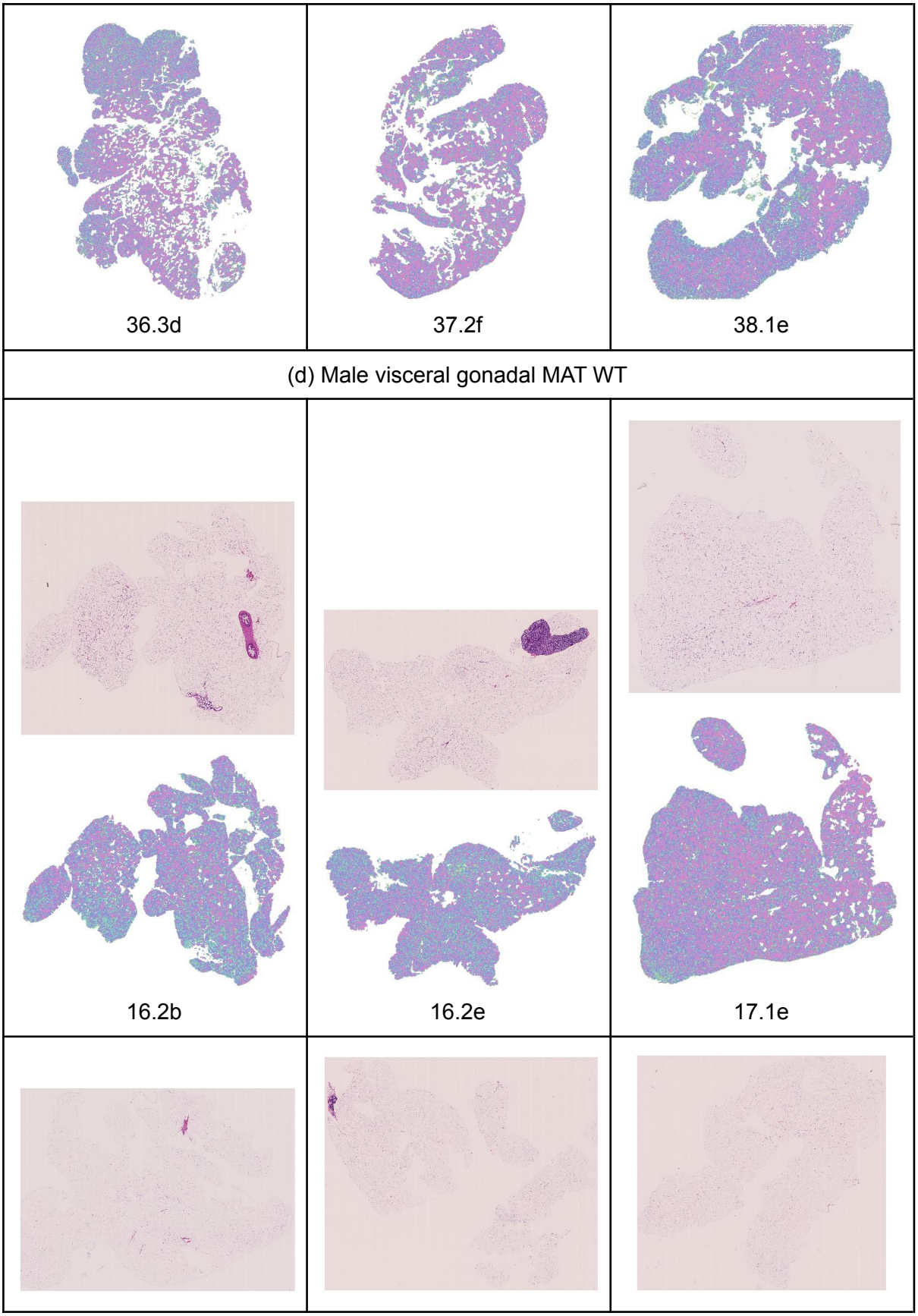


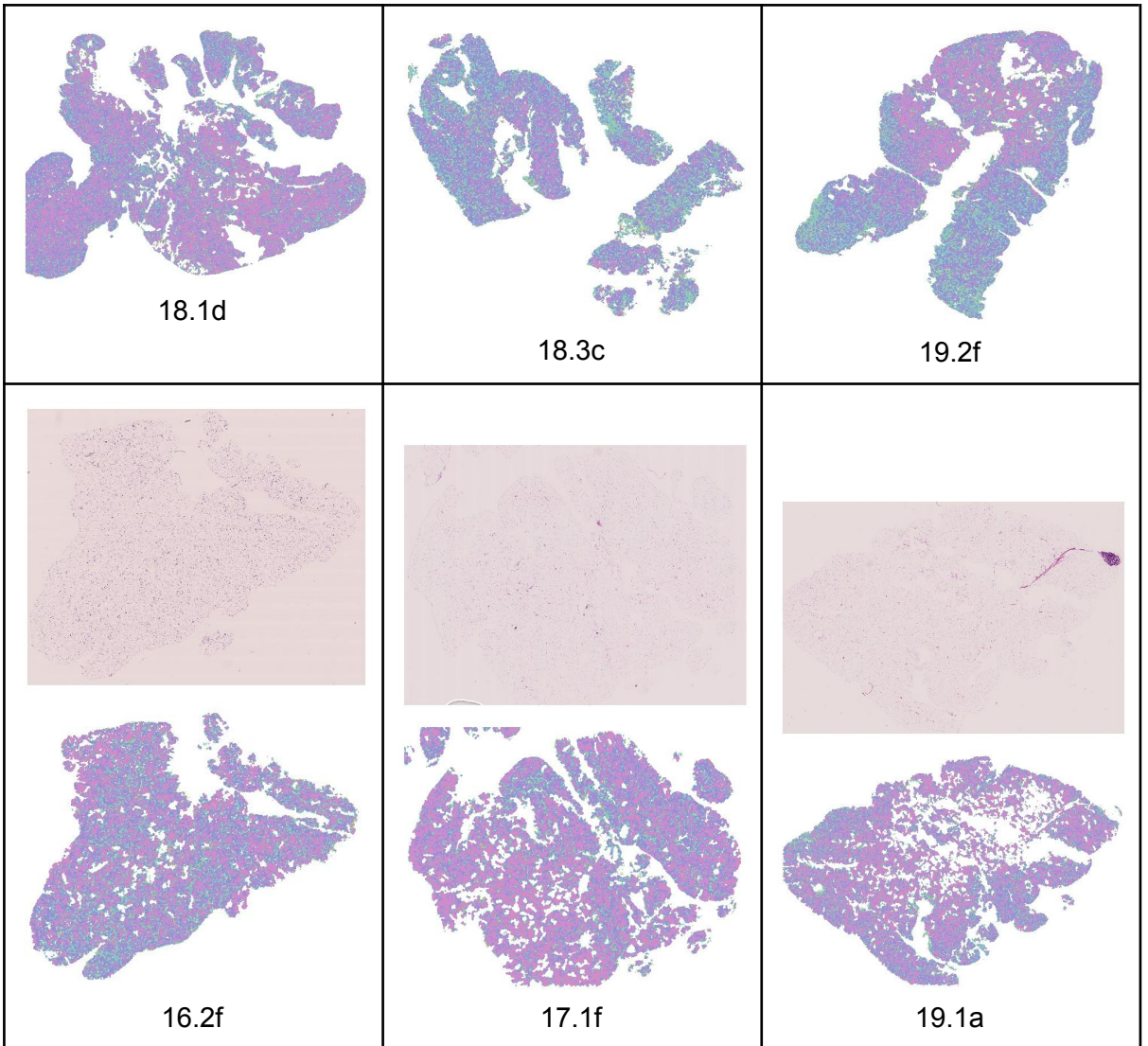
39.2d



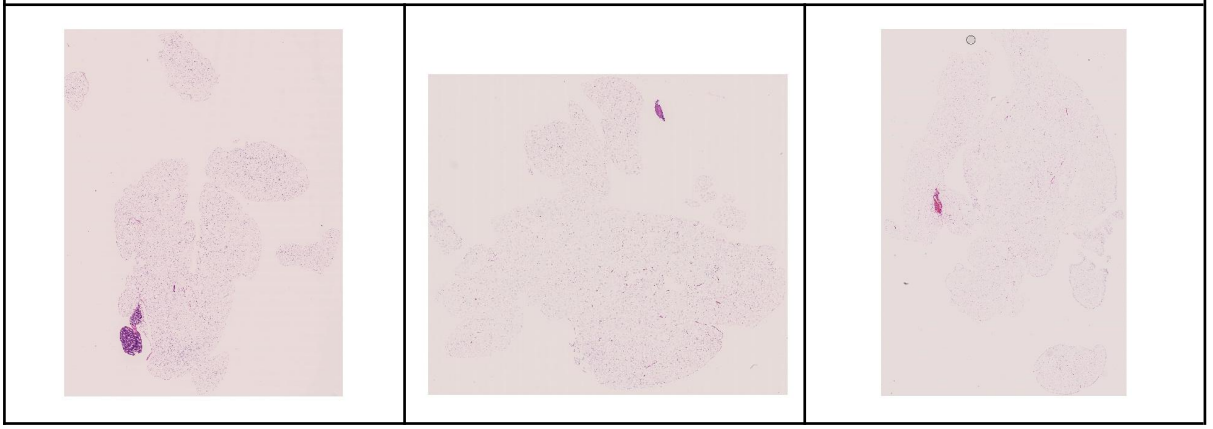


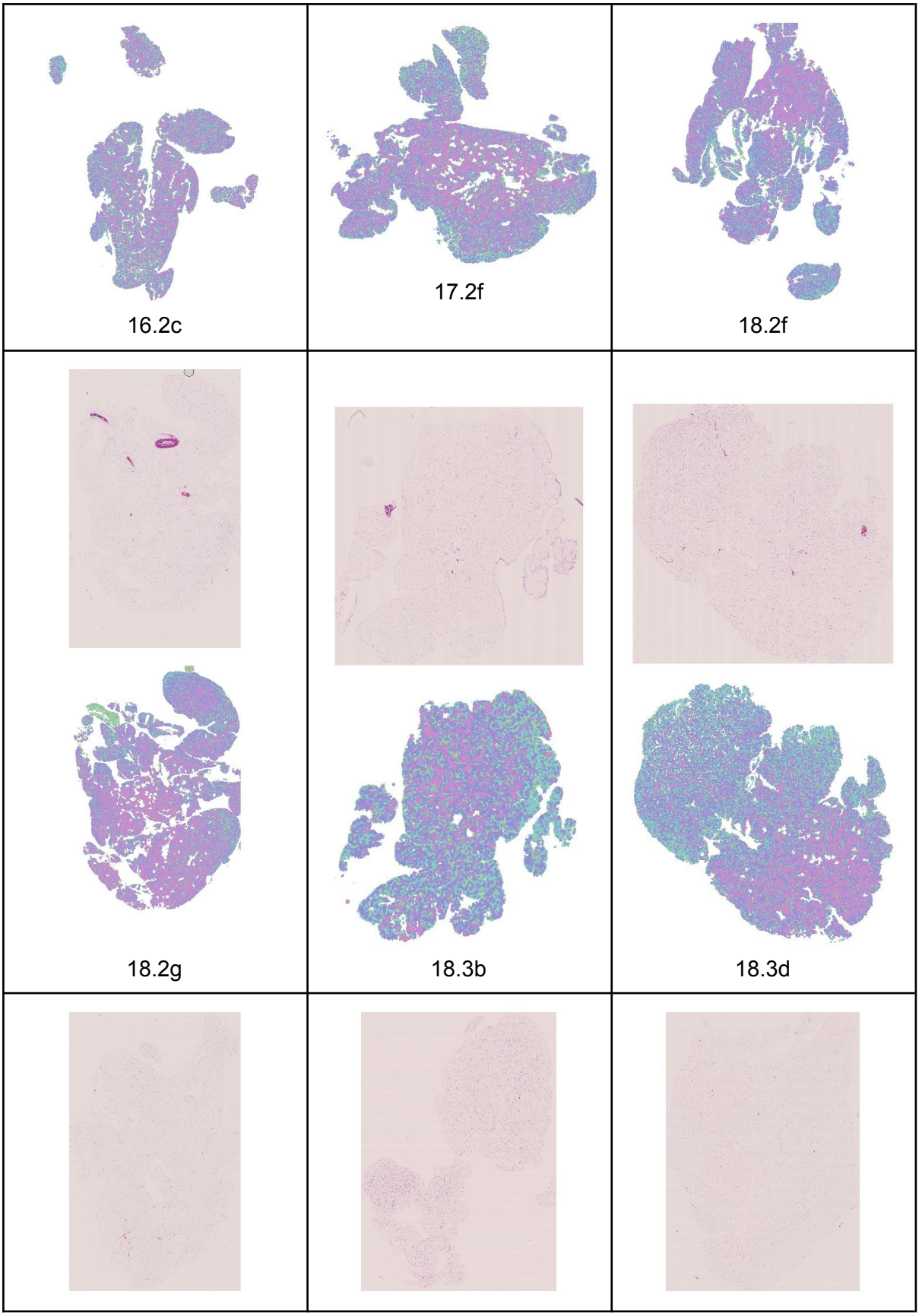






(e) Male visceral gonadal MAT Het





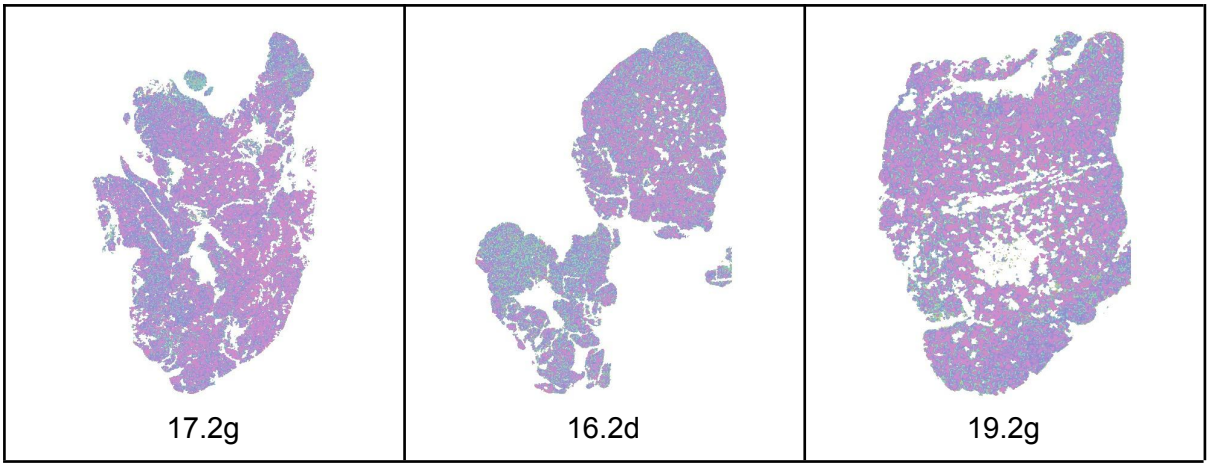
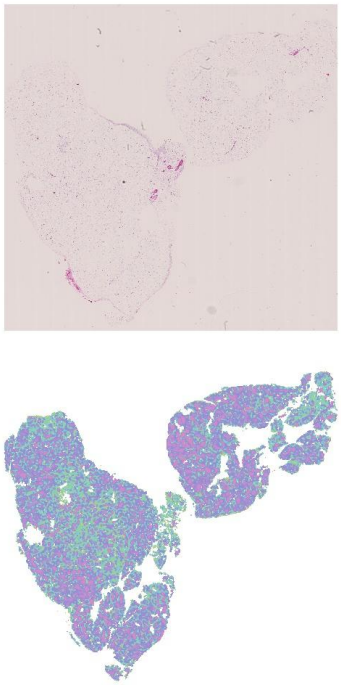
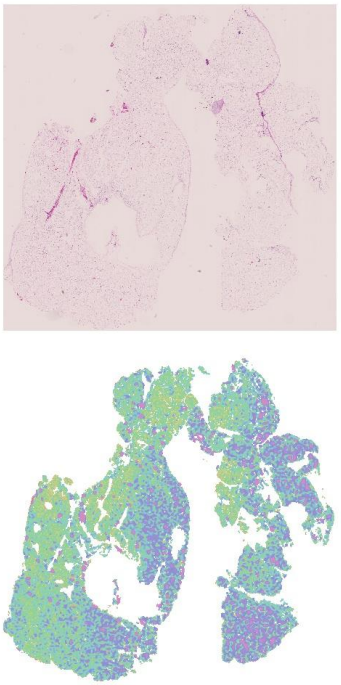


Fig. COLORMAP_M_GWAT. White adipocyte tissue histology and area quantile heatmaps for male visceral gonadal depot. (a): Quantile colour map and cell area density for male Corrected segmentation with deciles plotted for reference (vertical black lines). (b): PAT WT. (c): PAT Het. (d): MAT WT. (e): MAT Het.

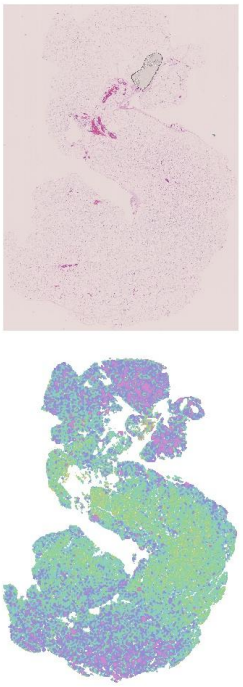
Male inguinal subcutaneous PAT WT



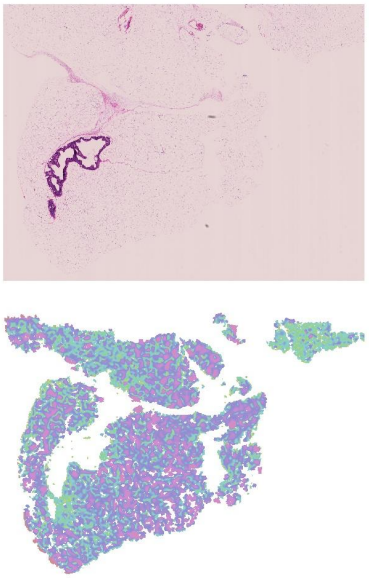
36.1e



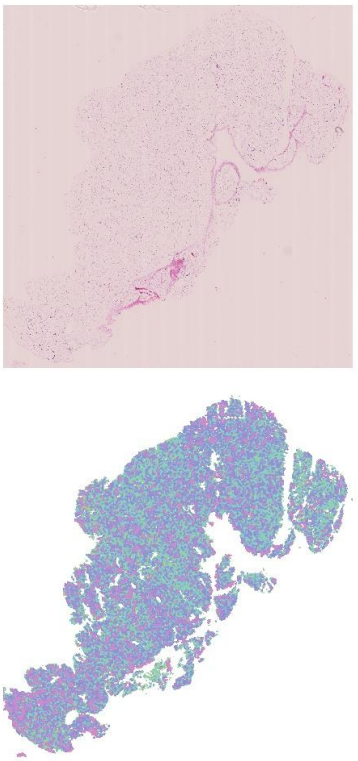
37.1g



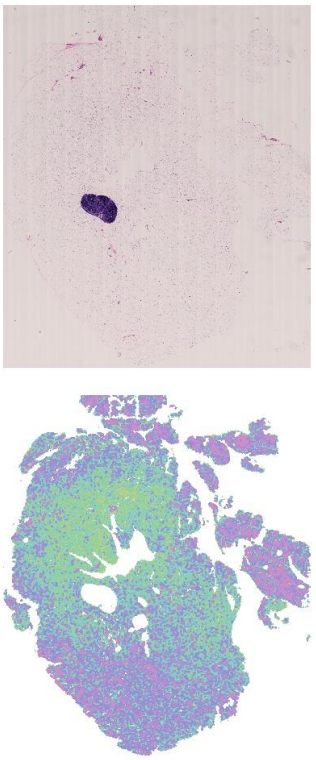
37.1h



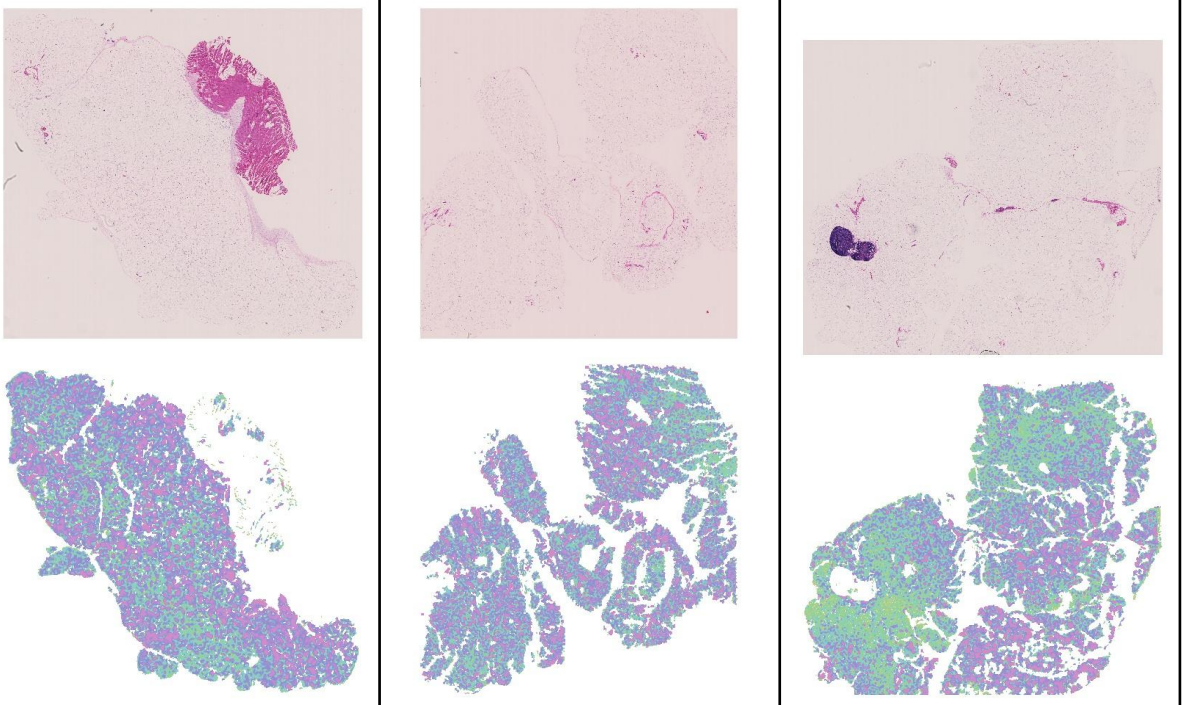
37.2h



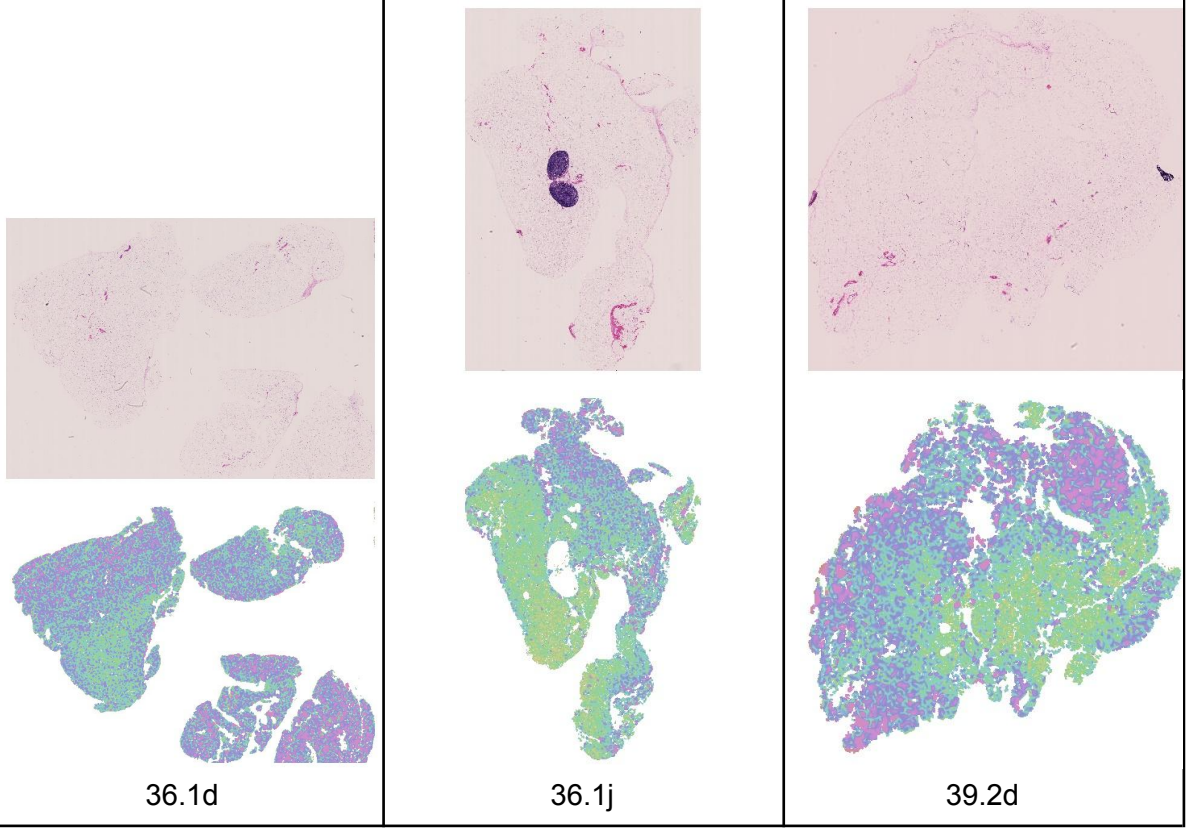
37.4b

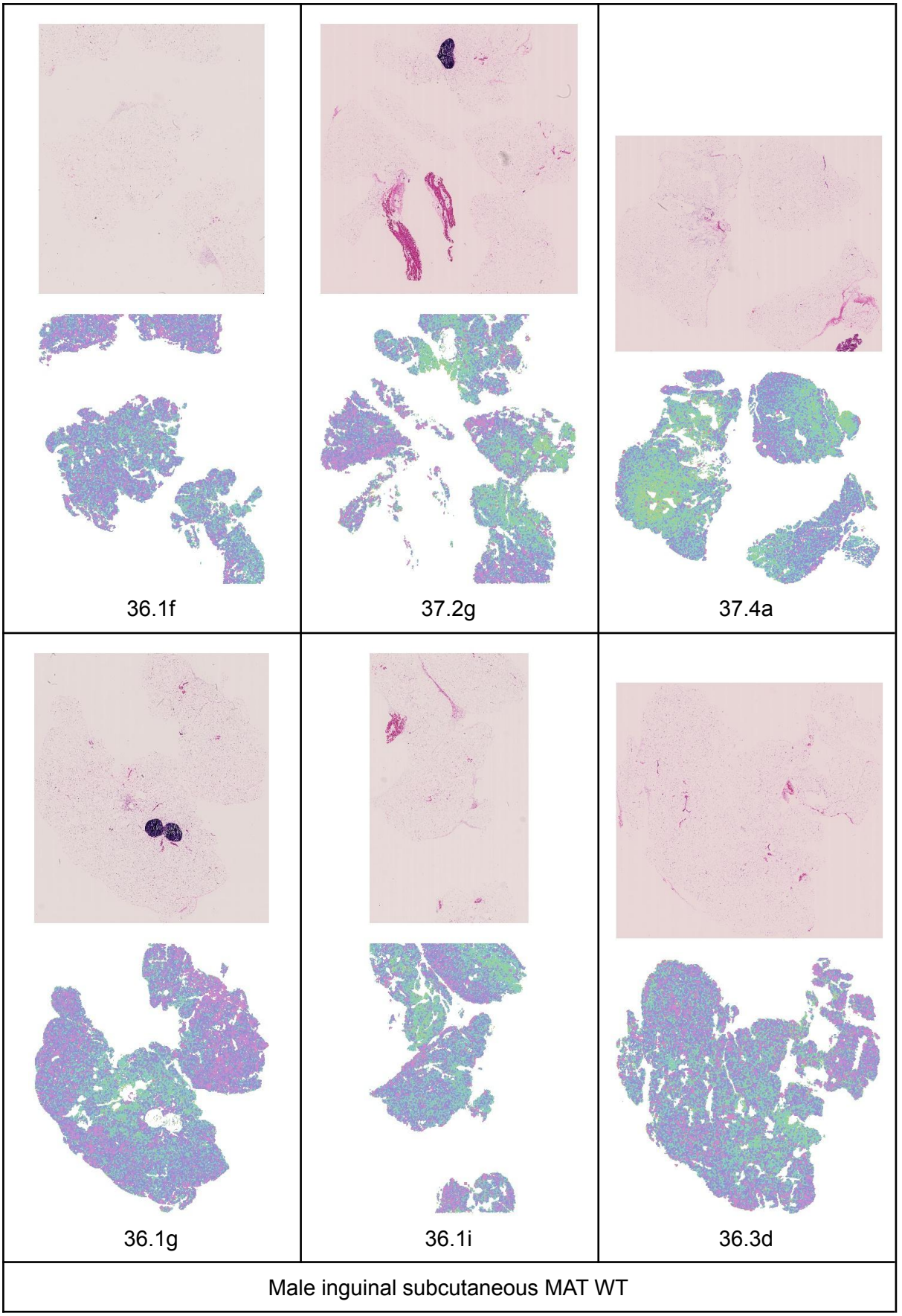


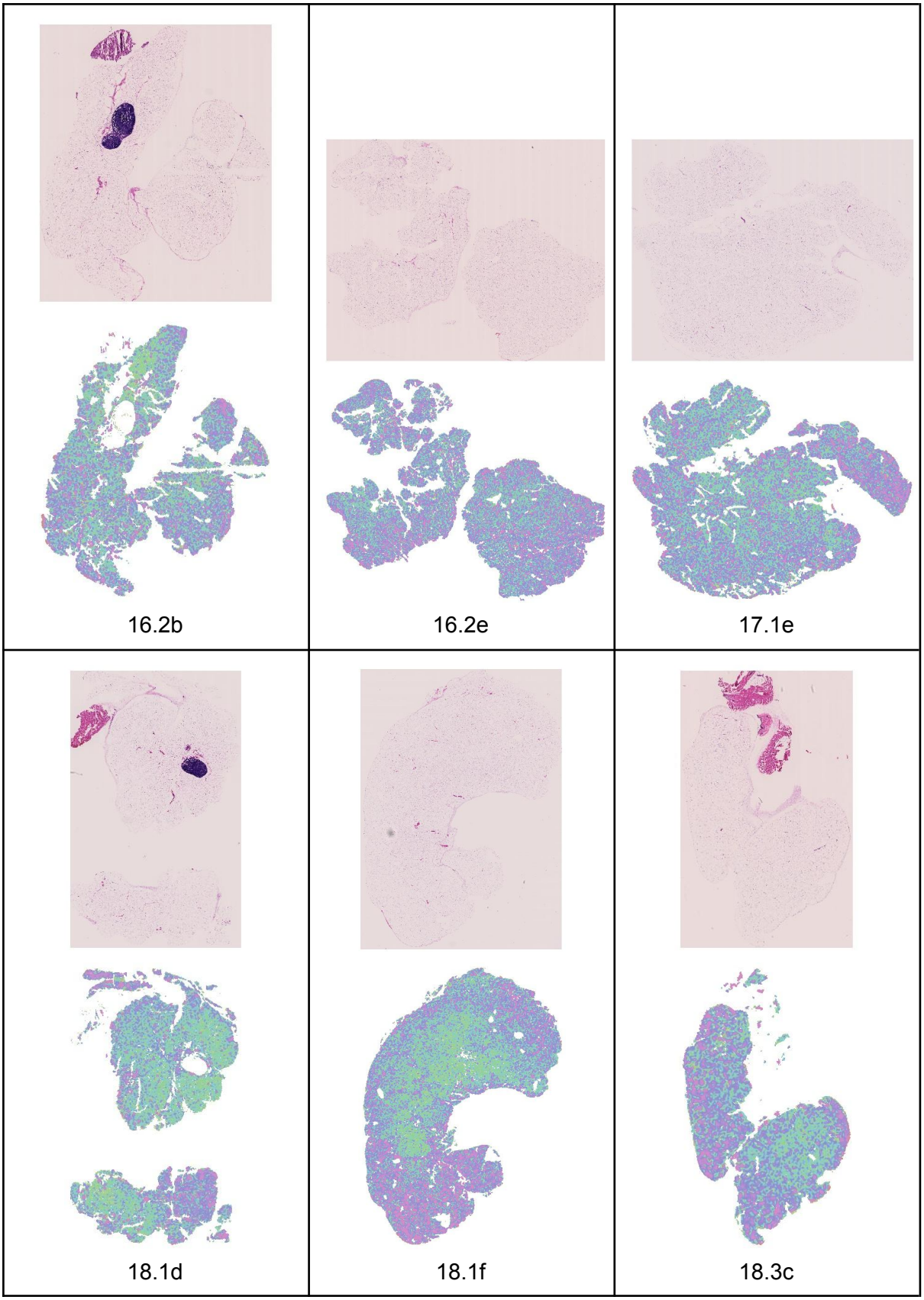
38.1f

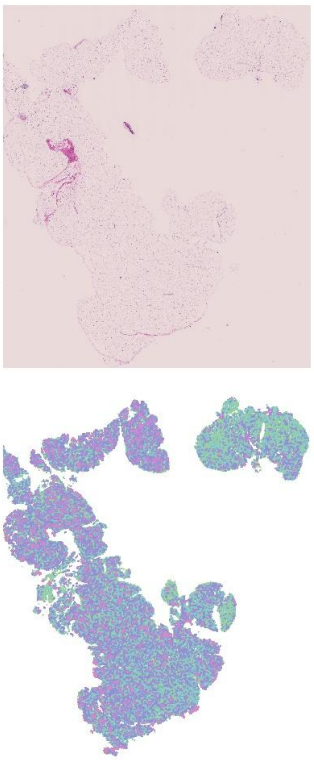


Male inguinal subcutaneous PAT Het

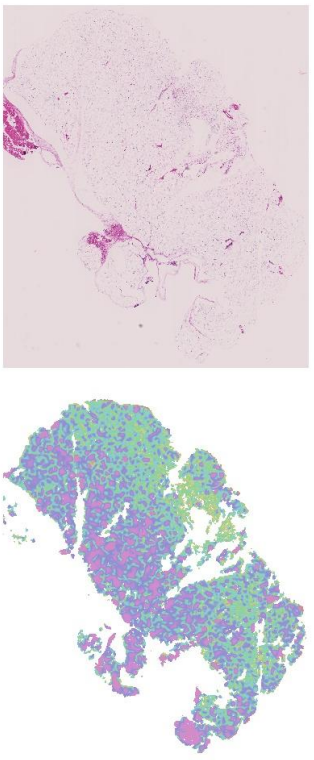




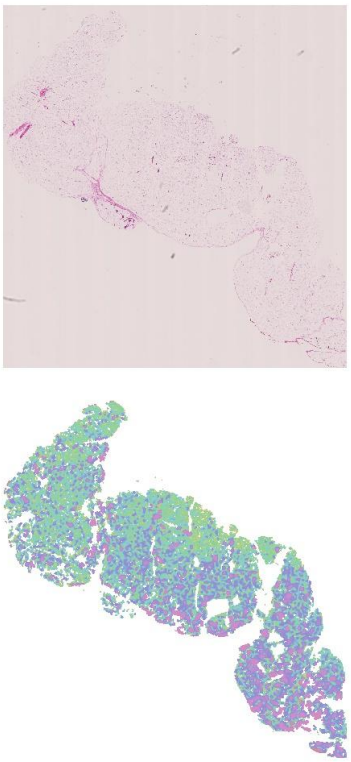




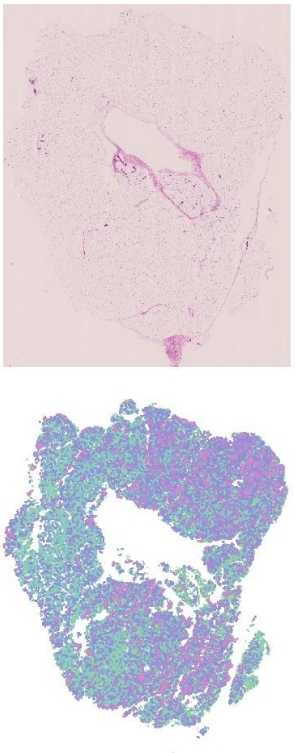
16.2f



17.1f

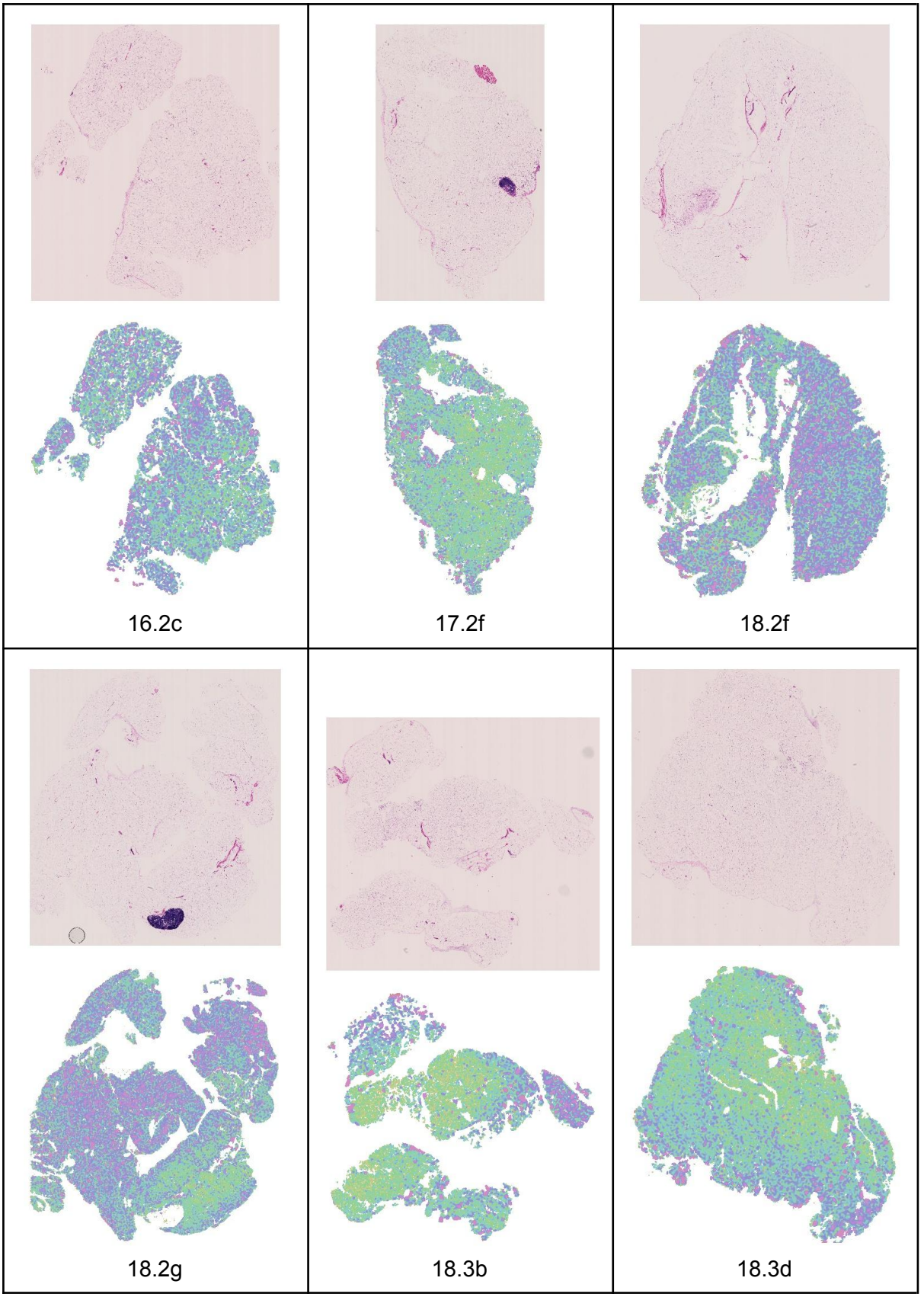


19.1a



19.2f

Male inguinal subcutaneous MAT Het



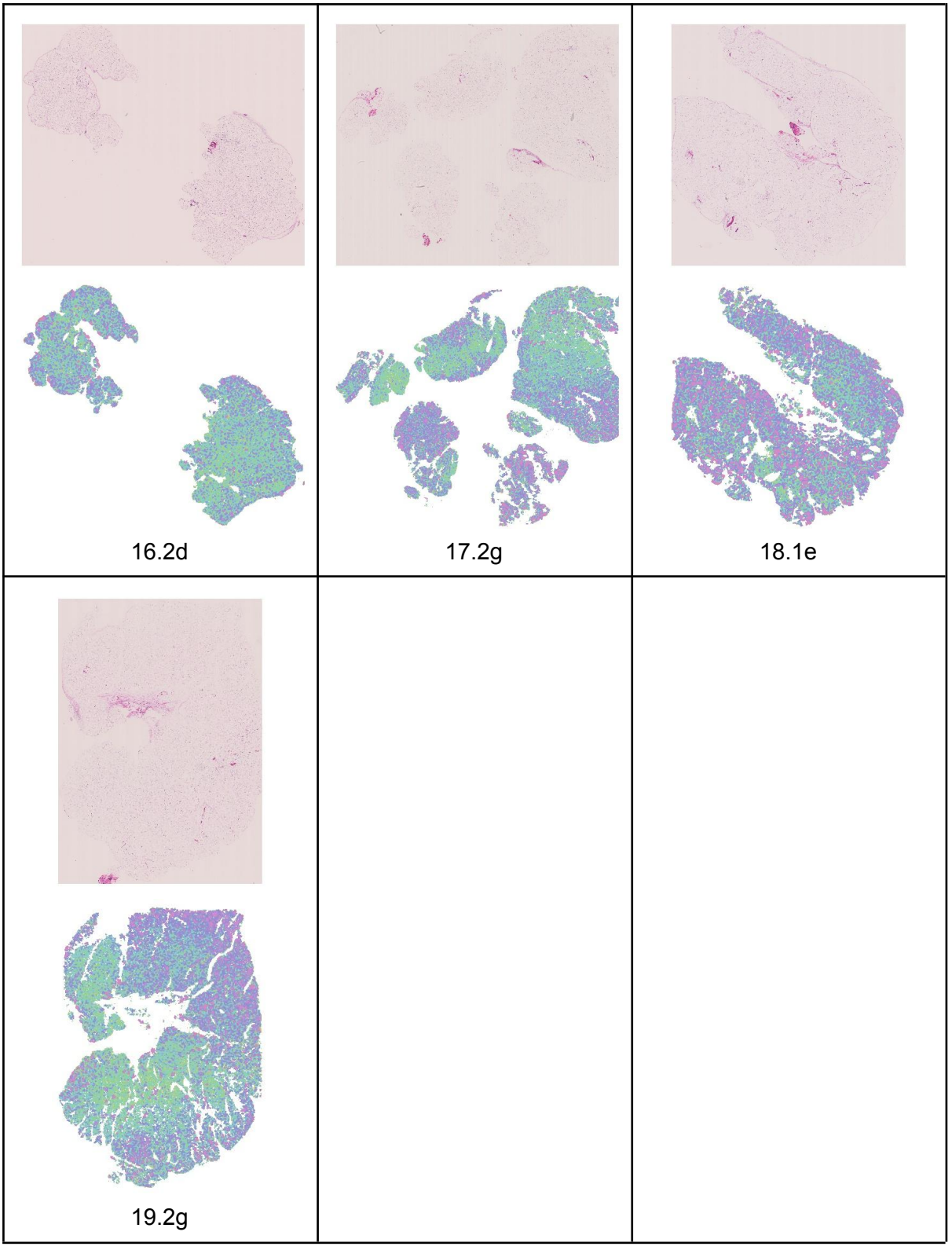


Fig. COLORMAP_M_SCWAT. White adipocyte tissue histology and area quantile heatmaps for male inguinal subcutaneous depot. (a): PAT WT. (b): PAT Het. (c): MAT WT. (d): MAT Het.

Supplementary material

Relation between pixel-classification and EDT regression.

We note that the pixel-classification[47] and EDT[50,51,66] approaches reviewed in the *Introduction* are related, as pixel-classification can be seen as a simplification of signed EDT regression by applying the sign function

$$\operatorname{sign}(\mathrm{EDT}_{\mathrm{signed}}) = \begin{cases} +1, & \text{cell interior} \\ 0, & \text{boundary} \\ -1, & \text{background} \end{cases}$$

Thus, the EDT representation is richer, because it not only estimates whether a pixel is inside a cell, but how far from the membrane it is. This could account for the better segmentation results found by Wang et al.[50].

Effective receptive field (ERF) of CNNs

The ERF of each CNN and fold was computed and is shown in Table ERF. The maximum cell size in our manual dataset is just under 20,000 μm^2. This corresponds to a diameter of 160 μm for a circular cell, or 353 pixels for pixel size 0.454 μm. An ERF=131 pixels covers around 37.1% of the largest cell diameter.

Effective receptive field of:
EDT CNN

Fold	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Height	128	131	131	131	131	131	131	130	131	124
Width	131	131	131	131	129	131	127	131	131	123

Contour CNN

Fold	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Height	131	131	131	131	131	131	131	131	131	131
Width	131	131	131	131	131	131	131	131	131	131

Tissue CNN

Fold	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*

Height	123	131	130	131	131	131	131	131	131	131
Width	131	131	131	131	131	131	131	131	131	131

Correction CNN

Fold	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Height	131	131	131	131	131	131	131	131	131	131
Width	131	131	131	131	131	131	131	131	131	131

Table ERF. Effective Receptive Field (ERF) of the DeepCytometer pipeline CNNs. All sizes in pixels.

Pseudocode for adaptive tiling

***adaptive_block_algorithm()*:**

1. Let
    a. *histology_filename* be the filename of the full resolution histology slide.
    b. *mask_todo_lores* := coarse tissue mask computed in previous section.
    c. *downsample_factor* := 8
    d. *max_window_size* := [2751, 2751] # due to GPU memory limit
    e. *border* := [65, 65] # Overlap with other windows to account for receptive field
2. Open file pointer to full resolution histology without reading it into memory
   im := OpenSlide(*histology_filename*)
3. Loop until *mask_todo_lores* is empty.
    a. Compute coordinates of next processing block, both in full resolution and downsampled image coordinates
    [*box_coords_hires*, *box_coords_lores*] :=
    get_next_roi_to_process(*mask_todo_lores*, *downsample_factor*, *max_window_size*, *border*)
    b. Load block to process from full resolution image
    im_box := OpenSlide.read_region(*im*, *box_coords_hires*)
    c. Extract low resolution mask for the block and upsample to full resolution
    mask_box_lores := *mask_todo_lores*[*box_coords_lores*]
    mask_box_hires := resize(*mask_box_lores*, *downsample_factor*, 'nearest neighbour')
    d. Segment histology to obtain one label per cell, and mask of cells on the edge
    labels_hires, *mask_edge_hires* :=
    DeepCytometer_pipeline(*im_box*, *CNN_models*, *segmentation_parameters*)
    e. If no cells found, wipe out current box from coarse tissue mask to avoid infinite loops
    mask_todo_lores[*box_coords_lores*] := 0
    go to next iteration in 3.

f. Downsample mask of edge objects so that the coarse tissue mask can be updated
mask_edge_lores :=
resize(*mask_edge_hires*, size(*box_coords_lores*), 'nearest neighbour')
g. Convert labels to contours for AIDA display
h. Update current block of coarse tissue mask with edge mask
mask_todo_lores[*mask_box_lores*] := *mask_edge_lores*

Function get_next_roi_to_process() convolves the coarse tissue mask with a vertical and horizontal line kernel and combines the outputs to find the location of the next processing block.

***get_next_roi_to_process()*:**

1. Compute convolution kernel size
L := int((*max_window_size* - 2 *border*) / *downsample_factor*)
2. Let *kh* and *kv* be convolution kernels with size *L*×*L* pixels. The kernels are all zeros except for a horizontal or vertical line of ones through the middle, respectively (Fig. RMABc).
3. Compute Fast Fourier Transform (FFT) convolution of coarse tissue mask, with output cropped to mask size
zh := *mask_todo_lores* ⊛ *kh*
zv := *mask_todo_lores* ⊛ *kv*
4. Compute hits where the processing block would both have a mask pixel on the top and left edges as the Hadamard or pointwise product (Fig. RMABd)
hits_lores := *zh* ∘ *zv*
5. Choose first found hit (*x0*, *y0*) in *hits_lores* as top-left corner of block (Fig. RMABe).
6. Choose bottom-right corner of block
xend := *x0* + *max_window_size* - 2 *border*
yend := *y0* + *max_window_size* - 2 *border*
7. Reduce block size if there are empty mask rows/columns at the bottom/right
xend := min(*xend*, last column in block with mask pixel)
yend := min(*yend*, last row in block with mask pixel)
8. Add a border around the block to account for the effective receptive field. Crop the border if it overflows the image edges, where the downsampled image has size (Rd, Cd) pixels
x0 := max(*x0* - *border*, 0)
y0 := max(*y0* - *border*, 0)
xend := min(*xend* + *border*, Cd-1)
yend := min(*yend* + *border*, Rd-1)
9. Upsample block coordinates for full resolution image
x0_hires := round(*x0* * *downsample_factor*)
y0_hires := round(*y0* * *downsample_factor*)
xend_hires := round(*xend* * *downsample_factor*)
yend_hires := round(*yend* * *downsample_factor*)

Deep CNNs training methodology

Training for 10-fold cross validation

All CNNs were trained with the same 10-fold cross validation described in Table MICE: 20 mice randomly partitioned in 10 sets of 18 mice for training and 2 for validation.

Training with a combination of labelled and unlabelled pixels

As advanced in the Introduction, it is convenient to use training images with a combination of labelled ('White adipocyte', 'Background', 'Other tissue') and unlabelled pixels. Instead of becoming a new class ('Void'), unlabelled pixels should not contribute to the training process, as if they were not part of the training dataset at all. This functionality is not available in Keras 2.2. Thus, we implemented an extension[1] that enables element-wise weighting of pixel-wise scores. Let z be the score matrix of size (R, C), where each element $z_{i,j}, i = 0, \ldots, R-1, j = 0, \ldots, C-1$ is the contribution of an output pixel to the loss. Let w be a weighting matrix such that the loss is

$$f_{\text{loss}} = \frac{1}{K} \sum_{i,j} w_{i,j} z_{i,j}$$

where $K \leq RC$ is the number of output pixels where $w_{i,j} \neq 0$. If

$$w_{i,j} \begin{cases} 0, & \text{unlabelled pixel} \\ 1, & \text{labelled pixel} \end{cases}$$

then f_{loss} is the average score of labelled pixels.

EDT CNN and Contour CNN training dataset

We used the 55 histology images with 2,117 ground truth white adipocyte (WA) hand traced contours from Table MICE. The corresponding SVG files containing the description of the hand traced WA contours (as described in *Ground truth hand traced dataset for CNN training*) were read. Each contour was rasterised as a closed polygon. Pixels that belonged to a single polygon were labelled as seeds. Then a watershed algorithm expanded the seeds over areas where polygons overlap. This effectively found a compromise boundary between overlapping cells. Boundaries were computed as pixels between two labels or between label and background. Ground truth EDTs were computed with respect to those boundaries (Fig. DMAP(c)). Boundaries were then dilated with a 3×3 kernel because we found that the Contour CNN training was unsatisfactory on 1-pixel thick boundaries. Masks of labelled pixels were computed for the loss function from the union of all polygons, and then were dilated with a 3×3 kernel. The histology windows, EDTs, boundaries and masks dataset was then 10× augmented with random rotations up to ±90º, a scaling factor in $[0.9, 1.1]$, and

[1] https://github.com/rcasero/keras

horizontal and vertical flips. The augmented images were split into 4 blocks due to GPU memory restrictions. Blocks where the mask had fewer than 1,900 pixels were discarded.

EDT CNN training

We trained the 10-fold CNNs using the augmented histology windows (input), ground truth EDTs (output), masks of labelled pixels (loss function masks), Adadelta optimisation [77] with He uniform variance scaling initialisation [78], mean absolute error (MAE) loss, MAE and mean squared error (MSE) metrics for validation with the left-out data, a batch size of 10 and 350 epochs until the loss and metrics converged.

Contour CNN training

We trained the 10-fold CNNs using the EDTs estimated by the previous CNN (input), ground truth dilated boundaries (output), masks of labelled pixels (loss function masks), Adadelta optimisation [77] with He uniform variance scaling initialisation [78], binary cross entropy loss, accuracy metric for validation with the left-out data, a batch size of 10 and 500 epochs until the loss and metric converged.

Tissue CNN training dataset

We used all 126 histology images with hand segmentations in Table MICE, containing a total of 2,117 “white adipocyte” (WA) objects and 232 “other” (NWA) objects. All WA and NWA contours were read. Each contour was rasterised as a polygon. Pixels within each polygon were labelled as “1” for WA and background, and “0” for other types of tissue to create the classifier ground truth. Pixels in overlap areas between a WA and NWA were considered NWA. The histology windows, classifier ground truth and masks dataset were 10× augmented with random rotations up to ±90º, a scaling factor in $[0.9, 1.1]$, horizontal and vertical flips and shear angle in $[-15^{\circ}, 15^{\circ}]$ to an output shape of 1,416×1,416 to avoid cropping out training pixels. The augmented images were split into 4 blocks due to GPU memory restrictions.

Tissue CNN training

We trained the 10-fold CNNs using the augmented histology windows (input), classifier ground truth (output), masks of labelled pixels (loss function masks), Adadelta optimisation [77] with He uniform variance scaling initialisation [78], binary focal loss [79] with $\gamma = 2,\ \alpha = 0.4$, accuracy metric for validation with the left-out data, a batch size of 8 and 37 epochs until the loss and metric converged. We used a cyclical learning rate[80] with a triangular cycle that scales initial amplitude by half each cycle, initial learning rate 10^{-7}, upper boundary 10^{-2} and number of training iterations per half cycle equal to 8× training iterations in epoch.

Correction CNN training dataset

We used the same 55 histology images, hand traced contours and rasterised polygons as for the EDT and Contour CNNs. Let a_h be the hand traced contour polygon area, and $r = \sqrt{a_h/\pi}$ the radius of a circle with the same area. We generated a series of incorrect segmentations by eroding and dilating the ground truth segmentation using an $l_{kernel} \times l_{kernel}$ square kernel with length $l_{kernel} = \lceil 2r|k| + 1 \rceil$, where the parameter $k \in \{\pm 0.03,\ \pm 0.07,\ \pm 0.10,\ \pm 0.15,\ \pm 0.20\}$ ($k < 0$ for erosion and $k > 0$ for dilation). Thus, we established a correspondence between a histology cropped image, the eroded/dilated segmentation mask m_k and the segmentation error $m_k - m_h$, where m_h is the hand traced contour polygon. Note that $m_k - m_h$ is -1 for pixels that underestimate the segmentation, 0 for correctly segmented pixels and +1 for pixels that overestimate the segmentation. Next, we multiplied the histology by +1 within m_k and by -1 without. We also computed a mask for the loss function by dilating $m_k \cup m_h$ by a factor $k' = 0.30$. Finally, the resulting image, the segmentation error $m_k - m_h$ and the loss function mask were cropped and resized according to the bounding box of m_k, as described above in the Correction CNN architecture section.

Correction CNN training

We trained the 10-fold CNNs using the cropped and resized -1/+1 masked histology (input), segmentation error $m_k - m_h$ (output) and loss function mask, the same cyclical learning rate as in the Tissue CNN, Adadelta optimisation [77] with He uniform variance scaling initialisation [78], mean squared error (MSE) loss, mean absolute error (MAE) and MSE metrics for validation with the left-out data, a batch size of 12 and 100 epochs until the loss and metrics converged.

DeepCytometer running times

We computed running times on a random sample of 95 automatically segmented whole slides. We measured tissue area as the area covered by the coarse tissue mask. Ordinary Least Squares model (time ~ tissue area) showed a linear relationship with intercept β_0=297.1 ± 2,495.8 s, slope β_1=111.9 ± 10.6 s / mm^2. Tissue areas were between 41.5 - 612.3 mm^2 (201.5 Mpixel - 2973.2 Mpixel), corresponding to a computation time of 4,939 - 6,8784 s (1.4 - 19.1 h) in the linear model. The area HD quartiles were (Q1, Q2, Q3) = (102.4, 174.7, 276.2) mm^2 corresponding to (3.3, 5.5, 8.7) h. It should be noted that each slide contained two tissue slices.

Segmentation times were calculated applying the Corrected method to the 60 training images for population studies. The Auto part of the pipeline took 43.9% ± 1.6% of the total time, and overlap correction took the other 56.1% ± 1.6%. Thus, overlap correction increased Auto computation time by a factor ×(2.28 ± 0.08).

Likelihood Ratio Test

To assess whether an independent variable X (parent, genotype) produces a significant effect on a dependent measure Y (body weight, depot weight) in a linear model, we use the Likelihood Ratio Test (LRT). With the LRT, we compare the fitting of the data to a null model (without X) with the fitting to an alternative model (with X). The LRT statistic is

$$\lambda_{LR} = 2\,(\ln(L1) - \ln(L0))$$

where ln(L0), ln(L1) are the log-likelihoods of the null and alternative models, respectively. The λ_{LR} statistic follows a χ^2 distribution with 1 degree of freedom. The test's null hypothesis (H0) is that the data is fully specified under the null model ($\lambda_{LR}=0$). The alternative hypothesis (H1) is that the alternative model, with variable X, is significantly better ($\lambda_{LR}>0$). The test produces a p-value, $p=\chi^2(\lambda_{LR}, 1)$, to reject H0 for H1. Using the LRT is equivalent to using the Akaike Information Criterion (AIC). The AIC is a measure that combines the goodness of fit of the model and its parsimony

$$AIC = 2\,k - 2\,\ln(L)$$

where k is the number of parameters of the model and ln(L) is the log-likelihood as above, so a lower AIC corresponds to a better model. Comparing the null model to the alternative model yields

$$AIC_1 - AIC_0 = 2 - 2(\ln(L1) - \ln(L0))$$

where by convention X produces a worthy improvement if $AIC_1 - AIC_0 < -2 \Rightarrow \ln(L1) - \ln(L0) > 2$. Note that under a χ^2 distribution with 1 degree of freedom, this corresponds to $p=\chi^2(4, 1)=0.046$. Thus, the LRT with significance threshold $\alpha=0.050$ is just slightly more lenient than the AIC with $AIC_1 - AIC_0 < -2$.

Cull age effect on BW

Mice were culled between 133 and 146 days. OLS models (BW ~ cull_age) stratified by sex suggest a mildly significant cull age effect in females (β=0.6355 g/day, p=0.041) and males (β=0.4356 g/day, p=0.046). Thus, we investigated whether cull age needs to be considered in the BW models. OLS models (cull_age ~ genotype) show no statistically significant difference between WTs and Hets' cull age in females (β=-0.47 days, p=0.60) or males (β=-0.34 days, p=0.74). OLS models (cull_age ~ parent) show no statistically significant difference between PATs and MATs' cull age in males (β=0.77 days, p=0.47). There is a statistically significant difference in females (β=2.32 days, p=0.0070), which amounts to MATs being older by 2.32 days / 137.28 days=1.69% on average.

Supplementary figures

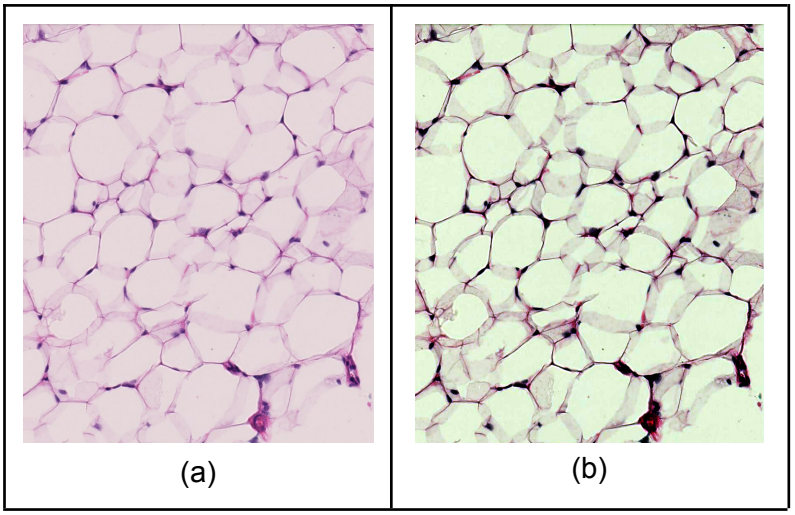
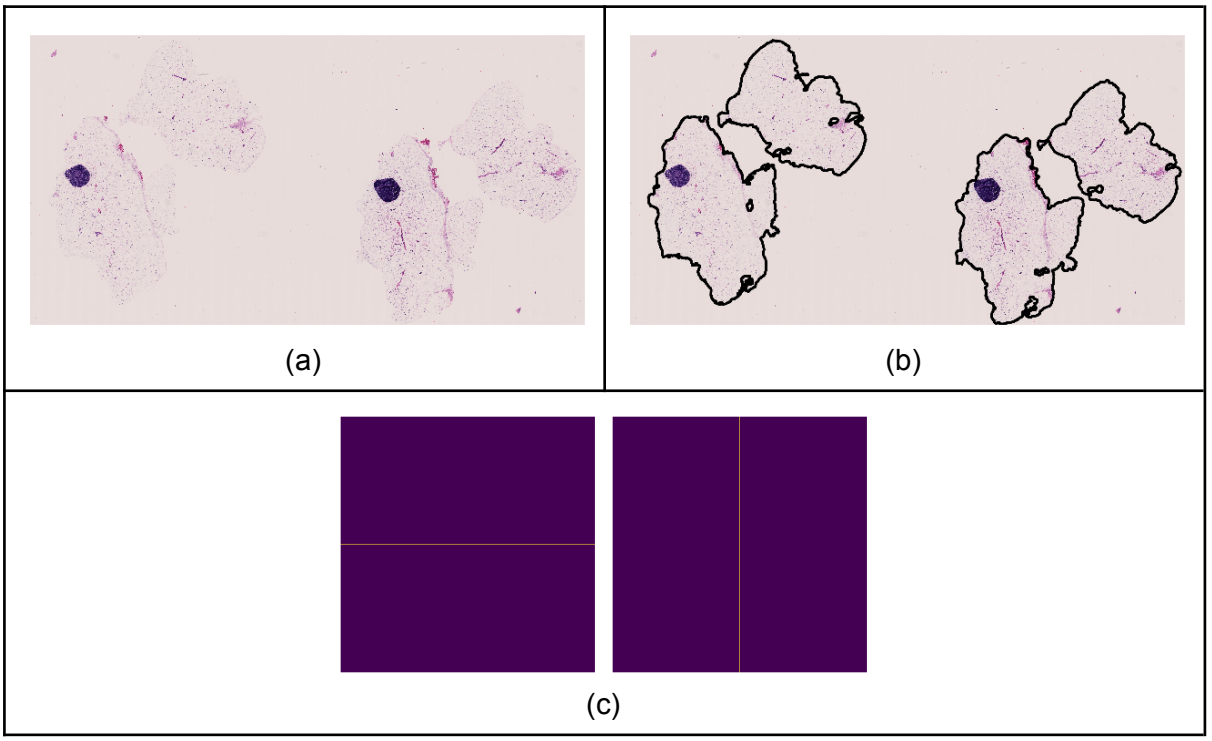


Fig. OVERLAP. H&E histology section of white adipose tissue (WAT) displaying clear cytoplasm overlap between adipocytes. Cut thickness is 8 µm, but we have observed similar overlaps at 4 µm, 6 µm and 10 µm. These overlaps are effectively the 2D projection of adjacent cells slightly mounting each other in 3D. (a) Original microscope image. (b) Enhanced contrast by automatic levels rescaling and manual curves adjustment for better visualisation of overlaps.



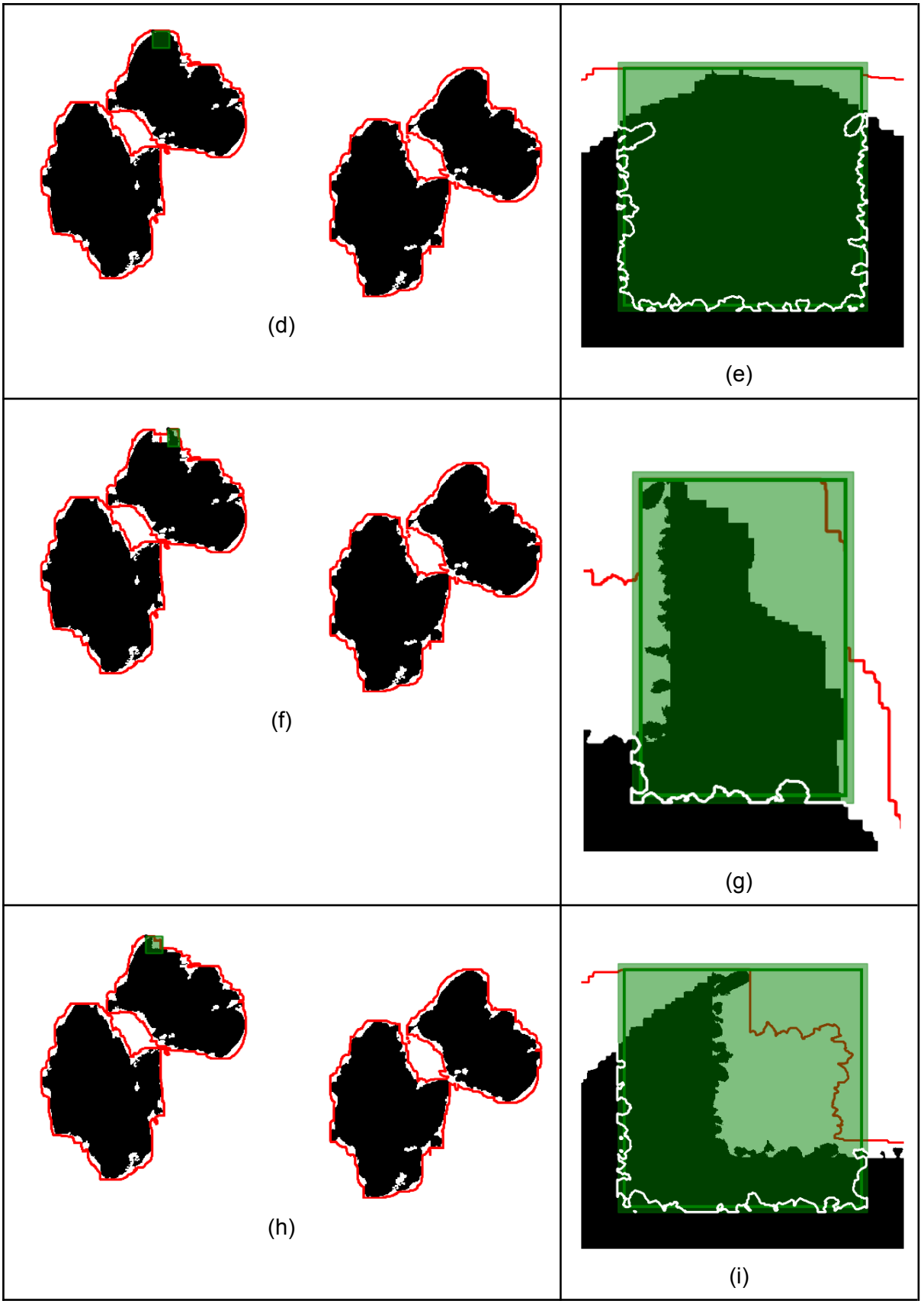


Fig. RMAB. Coarse tissue segmentation and adaptive tiling. (a) Whole histology slide, female MAT. (b) Coarse tissue mask. (c) Horizontal and vertical line convolution kernels.

(d)-(i) Black mask: Coarse tissue mask at three consecutive iterations. Red contour: Boundary of hits or mask convolutions with horizontal and vertical line kernels. Green square/rectangle: Block chosen for histology segmentation at each iteration. Green solid line: Separates inner part of the block from added border to account for effective receptive field. White contour: Update to coarse tissue mask. Note that segmented objects on the edges are not removed from the mask, for further processing in later iterations. (d), (f), (h): Whole slide. (e), (g), (i): Detail around processing block.

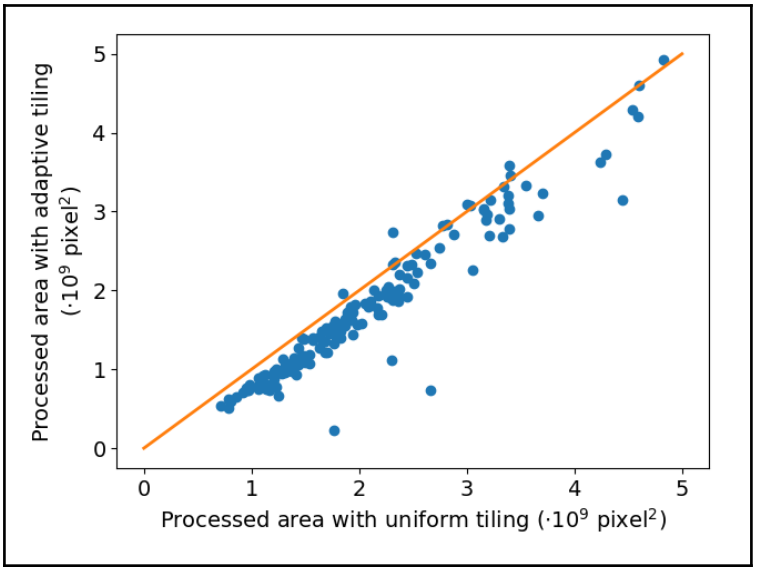


Fig. ADAPTBLOCK. Comparison of total area processed by uniform tiling of histology images vs. our adaptive tiling. Each point corresponds to one histology image. The orange line is the identity line.

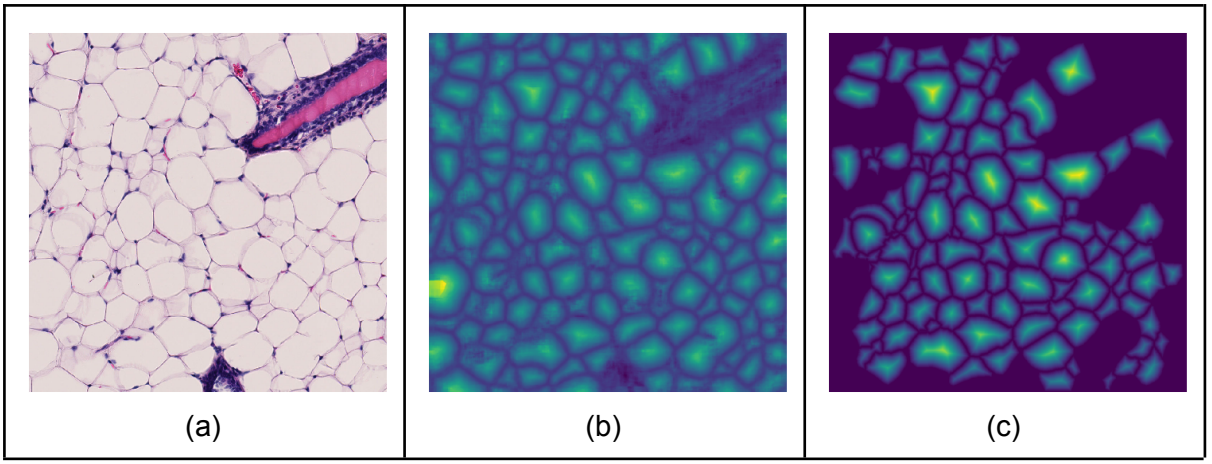
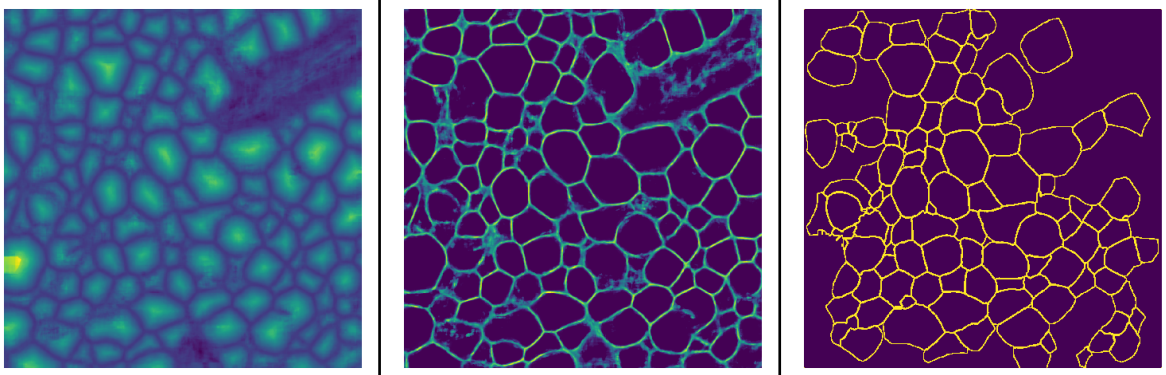
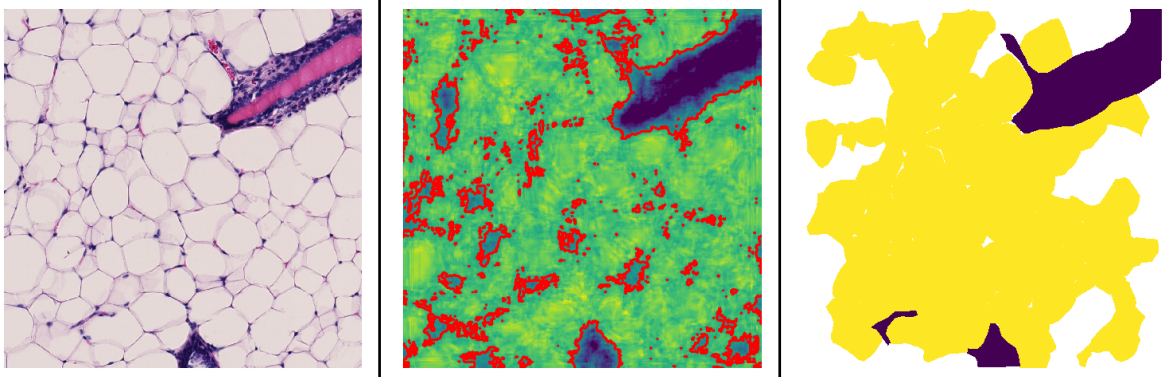


Fig. DMAP. Illustration of EDT CNN. (a) Input histology image. (b) EDT computed by the network, trained without the test data. (c) Ground truth EDT computed from the contours in Fig. CONTb.



(a) (b) (c)

Fig. CONT. Illustration of Contour CNN. (a) The Input to the network is the output from Fig. DMAPc. (b) Contours computed by the network, trained without the test data. (c) Ground truth contours derived from hand segmentations with overlaps removed (the hand segmentation leaves unlabelled pixels, as discussed in “Introduction - Deep Learning segmentation”).



(a) (b) (c)

Fig. CLASS. Illustration of Tissue CNN. (a) Input histology image. (b) Classification computed by the network from 0 (dark blue) to 1 (yellow). Red contours correspond to classification threshold 0.5. (c) Ground truth for classification. In white, unlabelled pixels.

(a) (b) (c)

Fig. CORRECT. Illustration of Correction CNN. (a) Cropped and scaled histology region around Auto segmentation (red contour). Size 401×401 pixel. Intensity values multiplied by

+1 inside segmentation, and by -1 outside segmentation. (b) Network estimation of whether pixels are correctly segmented: Oversegmented (yellow), correct (green), undersegmented (blue). (c) Correction overlayed with manual ground truth (white contour) and input segmentation (red contour).

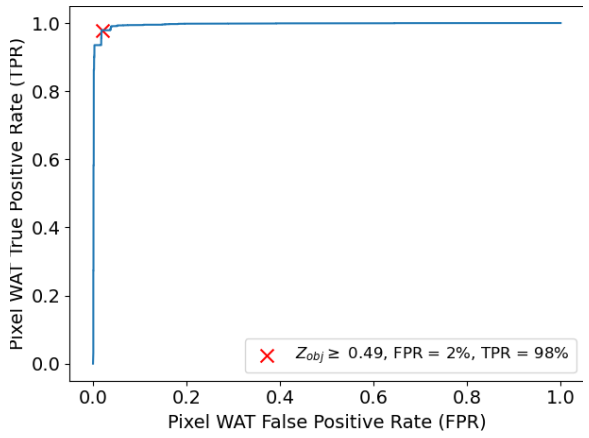
Nomenclature	**Description**
Input(F)	Input, any (row, col) input size, F features
Conv(K, F, D)	2D convolution, K×K kernel, F features, D dilation rate
MaxPool(P)	2D max pooling, P×P pool size
BN	Batch Normalization (Ioffe and Szegedy, 2015)
ReLU	Rectified Linear Unit Activation

Table NOM. Nomenclature for layers of pipeline CNNs. All Conv and MaxPool layers have stride 1, and zero padding so that their output has the same (row, column) size as their input.

EDT CNN	**Contour CNN**
Input(F=3)	Input(F=1)
Conv(K=5, F=32, D=1) + ReLU + MaxPool(P=3)	Conv(K=5, F=32, D=1) + ReLU + MaxPool(P=3) + BN
Conv(K=5, F=48, D=2) + ReLU + MaxPool(P=5)	Conv(K=5, F=48, D=2) + ReLU + MaxPool(P=5) + BN
Conv(K=3, F=64, D=4) + ReLU + MaxPool(P=9)	Conv(K=3, F=64, D=4) + ReLU + MaxPool(P=9) + BN
Conv(K=3, F=98, D=8) + ReLU + MaxPool(P=17)	Conv(K=3, F=98, D=8) + ReLU + MaxPool(P=17) + BN
Conv(K=3, F=256, D=16) + ReLU	Conv(K=3, F=256, D=16) + ReLU + BN
Conv(K=1, F=1, D=1)	Conv(K=1, F=64, D=1) + ReLU + BN
	Conv(K=1, F=8, D=1) + ReLU + BN
	Conv(K=1, F=1, D=1) + Hard Sigmoid
(a)	(b)
Tissue CNN	**Correction CNN**
Input(F=3)	Input(F=3)
Conv(K=5, F=32, D=1) + ReLU + MaxPool(P=3) + BN	Conv(K=5, F=32, D=1) + ReLU + MaxPool(P=3) + BN
Conv(K=5, F=48, D=2) + ReLU + MaxPool(P=5) + BN	Conv(K=5, F=48, D=2) + ReLU + MaxPool(P=5) + BN
Conv(K=3, F=64, D=4) + ReLU + MaxPool(P=9) + BN	Conv(K=3, F=64, D=4) + ReLU + MaxPool(P=9) + BN

	Conv(K=3, F=98, D=8) + ReLU + MaxPool(P=17) + BN		Conv(K=3, F=98, D=8) + ReLU + MaxPool(P=17) + BN
	Conv(K=3, F=256, D=16) + ReLU + BN		Conv(K=3, F=256, D=16) + ReLU + BN
	Conv(K=1, F=64, D=1) + ReLU + BN		Conv(K=1, F=64, D=1) + ReLU + BN
	Conv(K=1, F=8, D=1) + ReLU + BN		Conv(K=1, F=8, D=1) + ReLU + BN
	Conv(K=1, F=1, D=1) + ReLU		Conv(K=1, F=1, D=1)
(c)		(d)	

Table CNN. Description of the four CNN architectures used by the DeepCytometer pipeline. (See nomenclature in Table NOM).



(a)

Threshold (%)	FPR (%)	TPR (%)
35%	5.15	99.19
40%	3.87	98.85
45%	3.76	98.23
50%	1.80	97.71
55%	1.80	96.58
60%	1.76	95.10
65%	1.75	93.85

(b)

Fig. CLASS_ROC. Tissue CNN classification validation. (a) Receiver Operating Characteristic (ROC) curve. (b) Some ROC curve numerical values. Object classification error (white adipocyte vs. non white adipocyte) weighted by number of pixels. (Weighting used as the hand traced data set contains many more white adipocyte objects, but non white adipocyte objects can be very large.)

Genotype effect

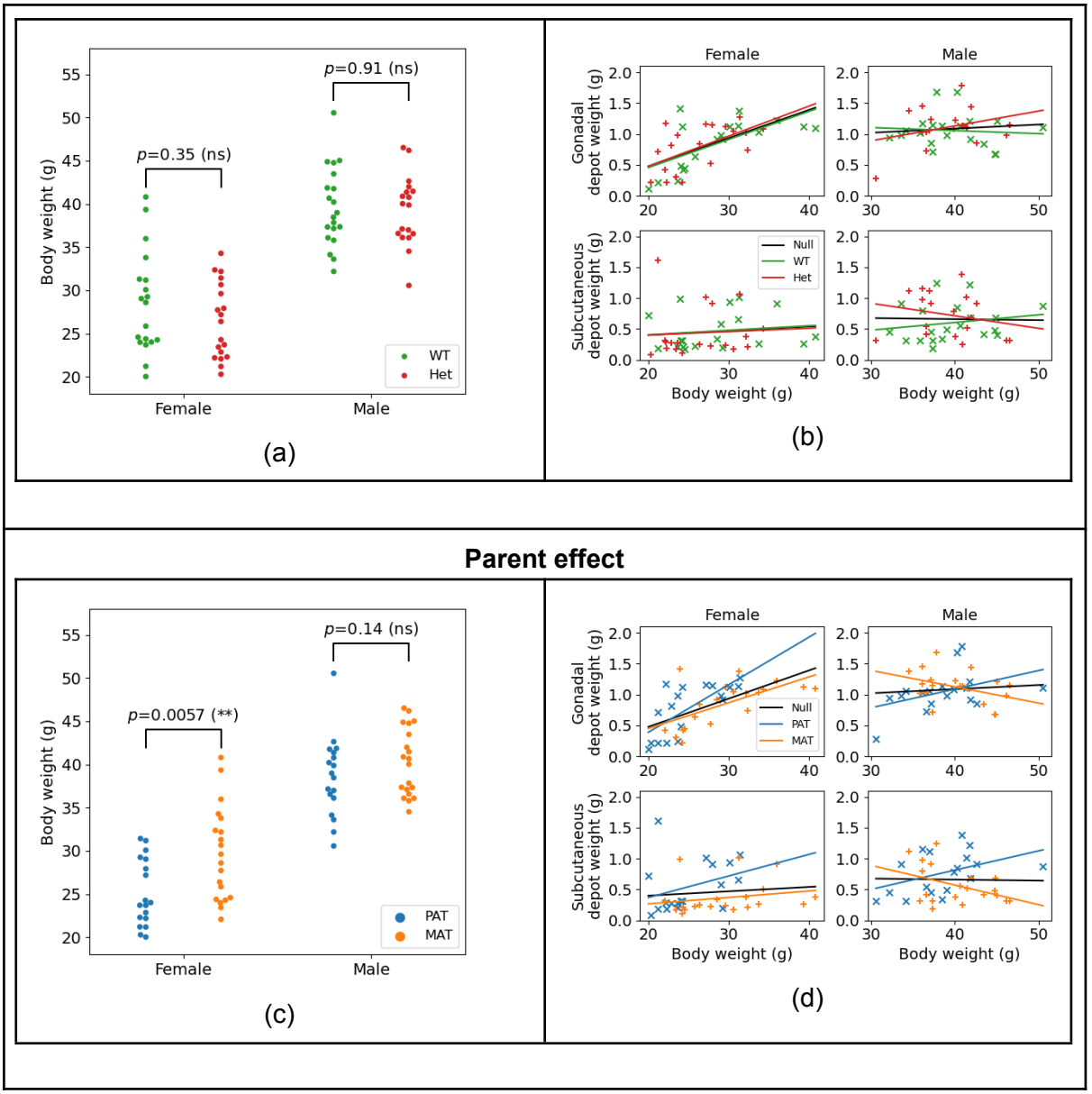


Fig. BW. Mouse body weight (BW) and depot weight (DW) analysis. (a), (c) BW swarm plots stratified by sex and genotype/parent. P-values computed from OLS models (BW ~ genotype) or (BW ~ parent), respectively. (b), (d) DW OLS models (DW ~ BW * genotype) and (DW ~ BW * parent), respectively, stratified by sex, genotype/parent and depot. Null models are (DW ~ BW).

Genotype effect					
Depot	Genotype	Intercept (95% CI)		p-value	
Female					
Gonadal	WT	-0.45	(-1.24, 0.33)	0.24	ns
	Het	-0.49	(-1.44, 0.45)	0.29	ns

Subcut.	WT	0.26	(-0.51, 1.03)	0.49	ns
	Het	0.29	(-1.00, 1.59)	0.64	ns
Male					
Gonadal	WT	1.25	(0.09, 2.42)	0.036	*
	Het	0.17	(-1.39, 1.73)	0.82	ns
Subcut.	WT	0.11	(-1.16, 1.38	0.86	ns
	Het	1.53	(-0.27, 3.32)	0.091	ns

(a)

Depot	Genotype	β(BW/$\overline{BW}$) (95% CI)		p-value		Corrected p-value	
Female							
Gonadal	WT	1.52	(0.62, 2.42)	0.0024	**	0.015	*
	Het	1.63	(0.45, 2.81)	0.0098	**	0.031	*
Subcut.	WT	0.25	(-0.64, 1.14)	0.57	ns	0.60	ns
	Het	0.18	(-1.43, 1.80)	0.81	ns	0.64	ns
Male							
Gonadal	WT	-0.16	(-1.14, 0.81)	0.73	ns	0.64	ns
	Het	0.80	(-0.52, 2.13)	0.22	ns	0.46	ns
Subcut.	WT	0.41	(-0.65, 1.48)	0.42	ns	0.53	ns
	Het	-0.68	(-2.20, 0.84)	0.36	ns	0.53	ns

(b)

Table DW_BW_RLM_GENOTYPE. Coefficients and p-values from OLS models (DW ~ BW/$\overline{BW}$) fitted to data stratified by sex, depot and genotype in Fig. BW(b). (a) Intercept. (b) Slope = β(BW/$\overline{BW}$).

Parent effect					
Depot	Parent	Intercept (95% CI)		p-value	
Female					
Gonadal	PAT	-1.15	(-2.16, -0.15)	0.027	*
	MAT	-0.39	(-1.17, 0.39)	0.31	ns

Subcut.	PAT	-0.33	(-1.71, 1.06)	0.62	ns
	MAT	0.060	(-0.69, 0.81)	0.87	ns
Male					
Gonadal	PAT	-0.11	(-1.42, 1.19)	0.85	ns
	MAT	2.18	(0.95, 3.42)	0.0016	**
Subcut.	PAT	-0.43	(-1.77, 0.91)	0.51	ns
	MAT	1.84	(0.43, 3.26)	0.014	*

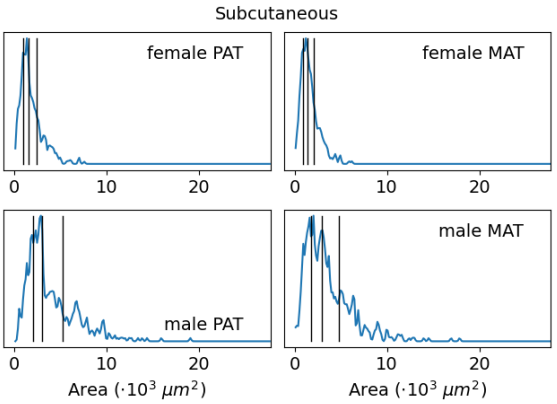
(a)

Depot	Parent	β(BW/$\overline{BW}$) (95% CI)		p-value		Corrected p-value	
Female							
Gonadal	PAT	2.58	(1.26, 3.90)	0.00077	***	0.0048	**
	MAT	1.40	(0.53, 2.27)	0.0033	**	0.010	*
Subcut.	PAT	1.17	(-0.66, 2.99)	0.19	ns	0.17	ns
	MAT	0.35	(-0.48, 1.18)	0.39	ns	0.31	ns
Male							
Gonadal	PAT	1.01	(-0.12, 2.13)	0.076	ns	0.090	ns
	MAT	-0.88	(-1.90, 0.14)	0.086	ns	0.090	ns
Subcut.	PAT	1.04	(-0.11, 2.20)	0.074	ns	0.090	ns
	MAT	-1.06	(-2.23, 0.11)	0.073	ns	0.090	ns

(b)

Table DW_BW_RLM_PARENT. Coefficients and p-values from OLS models (DW ~ BW/$\overline{BW}$) fitted to data stratified by sex, depot and parent in Fig. BW(d). (a) Intercept. (b) Slope = β(BW/$\overline{BW}$).

Hand traced segmentations of training windows for population studies

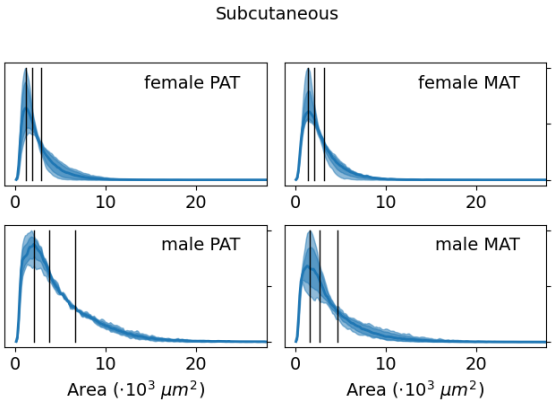


(a)

Subcutaneous			
Area, 95%-CI (10^3 μm^2)		PAT	MAT
f	Q1	0.97 (0.90, 1.05)	0.92 (0.83, 1.01)
	Q2	1.48 (1.38, 1.58)	1.41 (1.30, 1.52)
	Q3	2.37 (2.18, 2.57)	2.09 (1.92, 2.27)
m	Q1	2.06 (1.90, 2.22)	1.73 (1.58, 1.89)
	Q2	2.99 (2.83, 3.14)	2.94 (2.78, 3.10)
	Q3	5.21 (4.69, 5.73)	4.77 (4.39, 5.15)

(b)

DeepCytometer segmentations of whole slides that training windows were extracted from



(c)

Subcutaneous			
Area, 95%-CI (10^3 μm^2)		PAT	MAT
f	Q1	1.18 (1.18, 1.19)	1.42 (1.42, 1.43)
	Q2	1.87 (1.86, 1.88)	2.11 (2.10, 2.12)
	Q3	2.84 (2.83, 2.85)	3.17 (3.16, 3.19)
m	Q1	2.04 (2.03, 2.05)	1.64 (1.63, 1.64)
	Q2	3.70 (3.68, 3.72)	2.74 (2.73, 2.76)

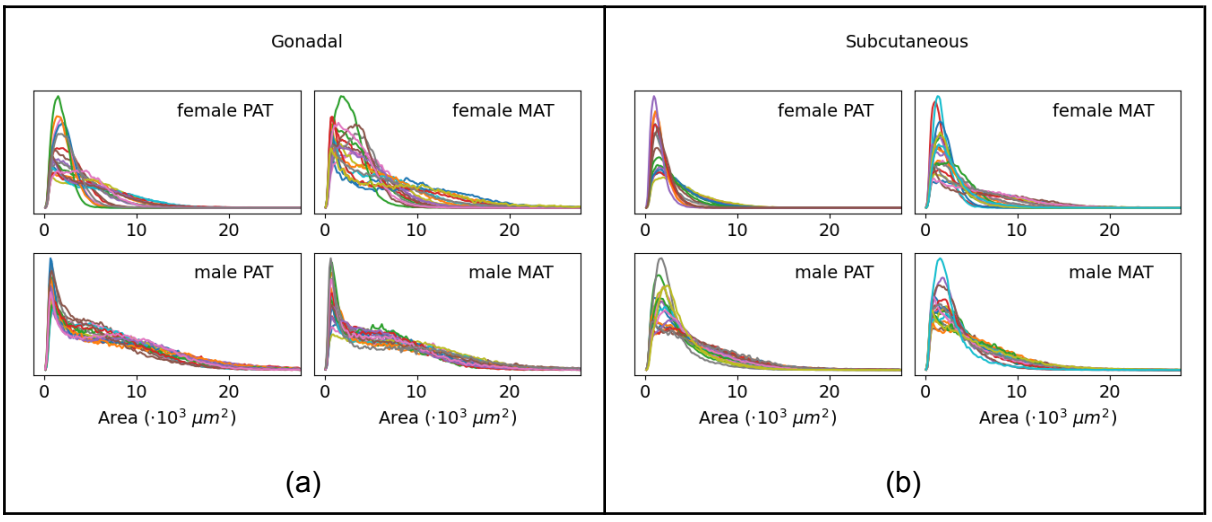
	Q3	6.60 (6.57, 6.63)	4.69 (4.67, 4.71)

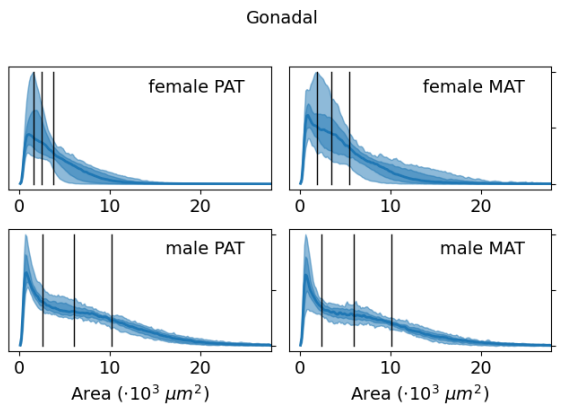
(d)

Subcutaneous			
Area difference (%)		PAT	MAT
f	Q1	+21.18	+54.67
	Q2	+26.18	+49.60
	Q3	+19.67	+51.51
m	Q1	-0.69	-5.49
	Q2	+23.99	-6.59
	Q3	+23.68	-1.58

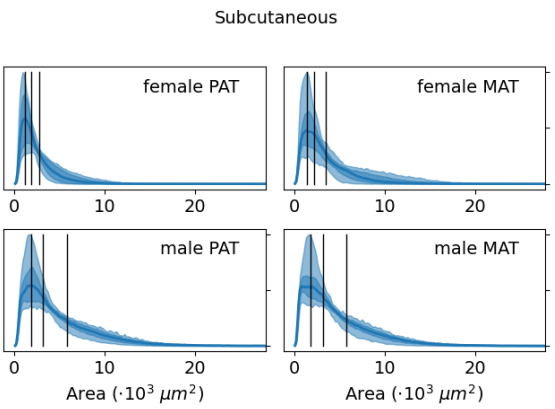
(e)

Fig. MANUAL_POPULATION_HISTOS. Cell populations of the hand traced dataset stratified by sex, parent and depot. (a), (c) Kernel Density estimation of cell population (blue) and Harrell-Davis (HD) quartiles (Q1, Q2, Q3) (vertical black lines). (b), (d) Numerical values for HD quartiles and 95%-CI for the quartile estimate. (c) Also showing 95%-range (light shaded area), interquartile range (dark shaded area). (e) Area quartile difference (%) from the hand traced to DeepCytometer whole slide segmentations (subcutaneous slides).





(c)



(d)

Gonadal			
Area, 95%-CI (10^3 µm^2)		PAT	MAT
f	Q1	1.53 (1.52, 1.53)	1.95 (1.94, 1.96)
	Q2	2.46 (2.45, 2.46)	3.53 (3.52, 3.54)
	Q3	3.72 (3.72, 3.73)	5.45 (5.44, 5.46)
m	Q1	2.56 (2.54, 2.57)	2.45 (2.43, 2.46)
	Q2	5.98 (5.96, 5.99)	6.00 (5.98, 6.02)
	Q3	10.15 (10.13, 10.17)	10.08 (10.05, 10.10)

(e)

Subcutaneous			
Area, 95%-CI (10^3 µm^2)		PAT	MAT
f	Q1	1.15 (1.15, 1.16)	1.45 (1.45, 1.46)
	Q2	1.85 (1.84, 1.85)	2.23 (2.23, 2.24)
	Q3	2.73 (2.72, 2.73)	3.46 (3.45, 3.47)
m	Q1	1.88 (1.87, 1.88)	1.83 (1.83, 1.84)
	Q2	3.17 (3.17, 3.18)	3.24 (3.23, 3.25)
	Q3	5.85 (5.84, 5.87)	5.74 (5.73, 5.75)

(f)

Subcutaneous			
Area difference (%)		PAT	MAT
f	Q1	-2.27	+1.92
	Q2	-1.23	+5.71

	Q3	-4.10	+9.00
m	Q1	-8.08	+12.05
	Q2	-14.28	+18.18
	Q3	-11.35	+22.36

(g)

Fig. SEG_POPULATION_HISTOS. DeepCytometer Corrected cell populations. (a)-(b): One histogram per mouse. Colours used to better differentiate between histograms. (c)-(d): 95%-range (light shaded area), interquartile range (dark shaded area) and median (solid curve) computed for each histogram bin. No adjustment for body or depot weight. Vertical black lines represent combined Q1, Q2, Q3; obtained by computing cell area Q1, Q2, Q3 from each mouse and combining them using the inverse-variance method. (e)-(f) Numerical values and 95%-CIs for the combined Q1, Q2, Q3 of cell areas from each mouse in that stratum, for females (f) and males (m). (g) Area quartile difference (%) from DeepCytometer segmentations in the 20 whole slides used for hand traced windows and DeepCytometer 75 whole slide segmentations (subcutaneous slides).

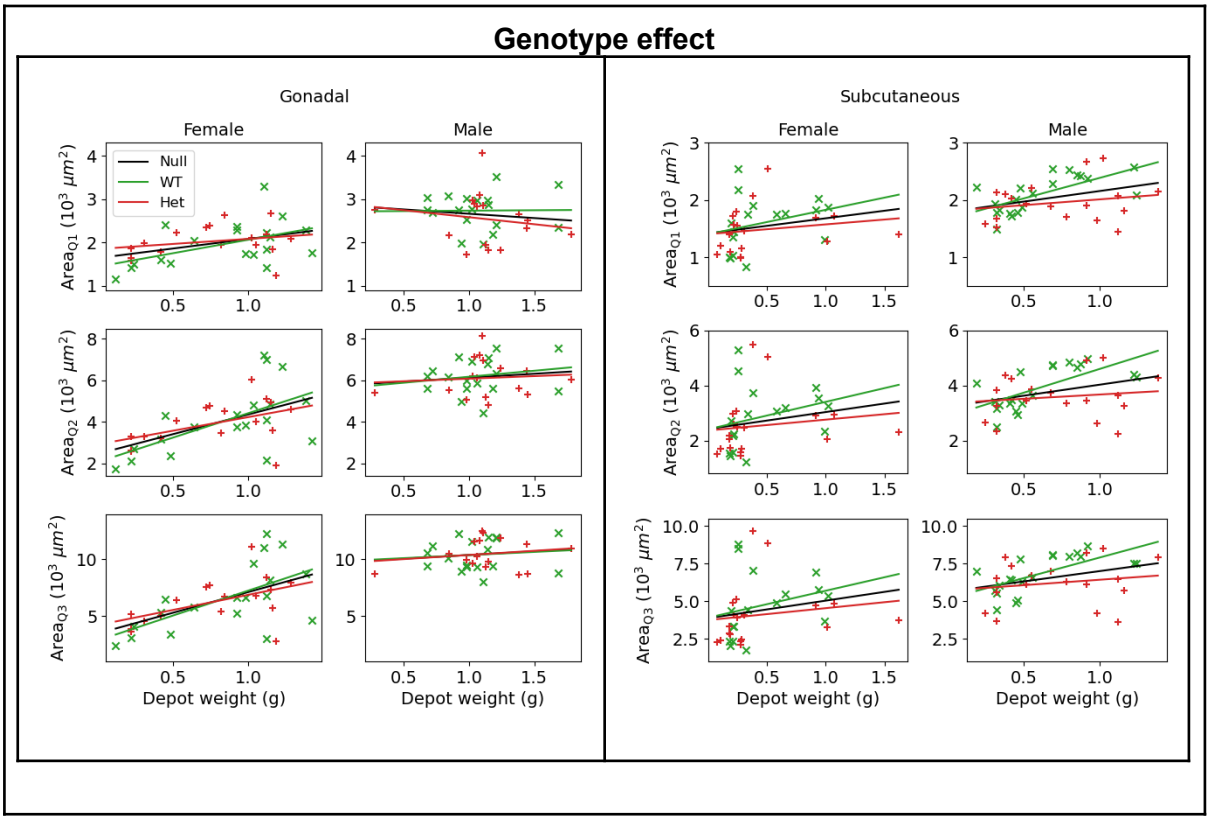


Fig. AREAQ_DW_GENOTYPE_LINREG. Cell quantile area vs. depot weight (DW) and genotype effect model. Scatter plots and fitted robust linear models ($area_q$ ~ DW) for cell population mode and quartiles (Q1, Q2, Q3), stratified by genotype, depot and sex. Each point corresponds to a mouse.

Genotype effect

	q	LR	p-value	
Female				
Gonadal	Q1	1.72	0.19	ns
	Q2	1.11	0.29	ns
	Q3	0.92	0.34	ns
Subcut.	Q1	1.11	0.29	ns
	Q2	1.13	0.29	ns
	Q3	1.08	0.30	ns
Male				
Gonadal	Q1	1.44	0.23	ns
	Q2	0.28	059	ns
	Q3	0.02	0.89	ns
Subcut.	Q1	6.80	0.0091	**
	Q2	7.62	0.0058	**
	Q3	6.06	0.014	*

(a)

Parent effect

	q	LR	p-value	
Female				
Gonadal	Q1	6.64	0.01	**
	Q2	9.30	0.0023	**
	Q3	9.33	0.0022	**
Subcut.	Q1	11.20	0.00082	***
	Q2	10.06	0.0015	**
	Q3	9.48	0.0021	**
Male				
Gonadal	Q1	1.09	0.30	ns
	Q2	1.78	0.18	ns
	Q3	4.15	0.042	*
Subcut.	Q1	5.26	0.022	*
	Q2	5.63	0.018	*
	Q3	6.59	0.01	*

(b)

Table AREAQ_DW_LRT. Likelihood Ratio Tests (LRT) for null models ($area_q$ ~ DW) and alternative models stratified by sex and depot: (a) ($area_q$ ~ genotype ∗ DW). (b) ($area_q$ ~ parent ∗ DW).

Genotype effect

Depot	Parent	Intercept (95% CI)		p-value	
Female					
Gonadal WT	Q1	1436.9	(907.7, 1966.1)	2.46E-05	****
	Q2	2055.6	(526.3, 3584.9)	0.011	*
	Q3	2832.2	(59.8, 5604.5)	0.046	*

Gonadal Het	Q1	1845.5	(1403.7, 2287.4)	2.25E-07	****
	Q2	2907.4	(1743.4, 4071.3)	8.50E-05	****
	Q3	4194.8	(1985.5, 6404.2)	0.0011	**
Subcut. WT	Q1	1399.1	(964.3, 1833.9)	5.43E-06	****
	Q2	2406.4	(1350.8, 3462)	0.00021	***
	Q3	3893.7	(1874.6, 5912.9)	0.00093	***
Subcut. Het	Q1	1402.7	(1115.8, 1689.5)	9.80E-09	****
	Q2	2364.7	(1563.8, 3165.6)	9.16E-06	****
	Q3	3737.6	(2232.3, 5243)	6.67E-05	****
Male					
Gonadal WT	Q1	2713.1	(1816.6, 3609.6)	8.54E-06	****
	Q2	5591.6	(3839, 7344.2)	4.55E-06	****
	Q3	9836.4	(6987.4, 12685.4)	1.72E-06	****
Gonadal Het	Q1	2906.8	(1650.8, 4162.9)	0.00021	***
	Q2	5829.0	(3880.6, 7777.3)	1.61E-05	****
	Q3	9656.1	(6945.7, 12366.4)	2.33E-06	****
Subcut. WT	Q1	1666.0	(1401.3, 1930.8)	1.05E-10	****
	Q2	2871.1	(2247.4, 3494.9)	1.49E-08	****
	Q3	5155.0	(4109.7, 6200.4)	5.16E-09	****
Subcut.	Q1	1807.2	(1389.5, 2224.9)	9.01E-08	****

Het	Q2	3360.3	(2403.3, 4317.3)	1.39E-06	****
	Q3	5698.0	(3895.5, 7500.5)	5.09E-06	****

(a)

Depot	Parent	β(DW) (95% CI)		p-value		Corrected p-value	
Female							
Gonadal WT	Q1	624.9	(57.6, 1192.2)	0.033	*	0.11	ns
	Q2	2360.7	(721.4, 4000)	0.0074	**	0.030	*
	Q3	4420.8	(1449, 7392.6)	0.0060	**	0.030	*
Gonadal Het	Q1	235.4	(-285.6, 756.3)	0.35	ns	0.54	ns
	Q2	1319.4	(-52.9, 2691.7)	0.058	ns	0.15	ns
	Q3	2683.4	(78.4, 5288.4)	0.044	*	0.13	ns
Subcut. WT	Q1	428.1	(-353.2, 1209.4)	0.26	ns	0.54	ns
	Q2	1000.2	(-896.6, 2897)	0.28	ns	0.54	ns
	Q3	1794.5	(-1833.7, 5422.7)	0.31	ns	0.54	ns
Subcut. Het	Q1	170.2	(-312.5, 653)	0.47	ns	0.54	ns
	Q2	395.7	(-952, 1743.5)	0.54	ns	0.54	ns
	Q3	794.0	(-1739.3, 3327.2)	0.52	ns	0.54	ns
Male							
Gonadal WT	Q1	19.1	(-795.9, 834)	0.96	ns	0.80	ns
	Q2	579.9	(-1013.3, 2173.1)	0.45	ns	0.54	ns
	Q3	549.9	(-2039.9, 3139.7)	0.66	ns	0.60	ns

Gonadal Het	Q1	-325.9	(-1396.5, 744.8)	0.52	ns	0.54	ns
	Q2	248.8	(-1412, 1909.5)	0.75	ns	0.65	ns
	Q3	738.1	(-1572.2, 3048.3)	0.50	ns	0.54	ns
Subcut. WT	Q1	720.9	(324.9, 1116.9)	0.0012	**	0.011	*
	Q2	1735.2	(802.2, 2668.1)	0.0010	**	0.011	*
	Q3	2744.8	(1181.3, 4308.4)	0.0017	**	0.011	*
Subcut. Het	Q1	203.7	(-313.3, 720.7)	0.42	ns	0.54	ns
	Q2	313.9	(-870.6, 1498.3)	0.58	ns	0.55	ns
	Q3	722.9	(-1508.1, 2954)	0.50	ns	0.54	ns

(b)

Table AREAQ_DW_GENOTYPE_LINREG. Coefficients of OLS model ($area_q$ ~ DW) stratified by genotype and depot for females.

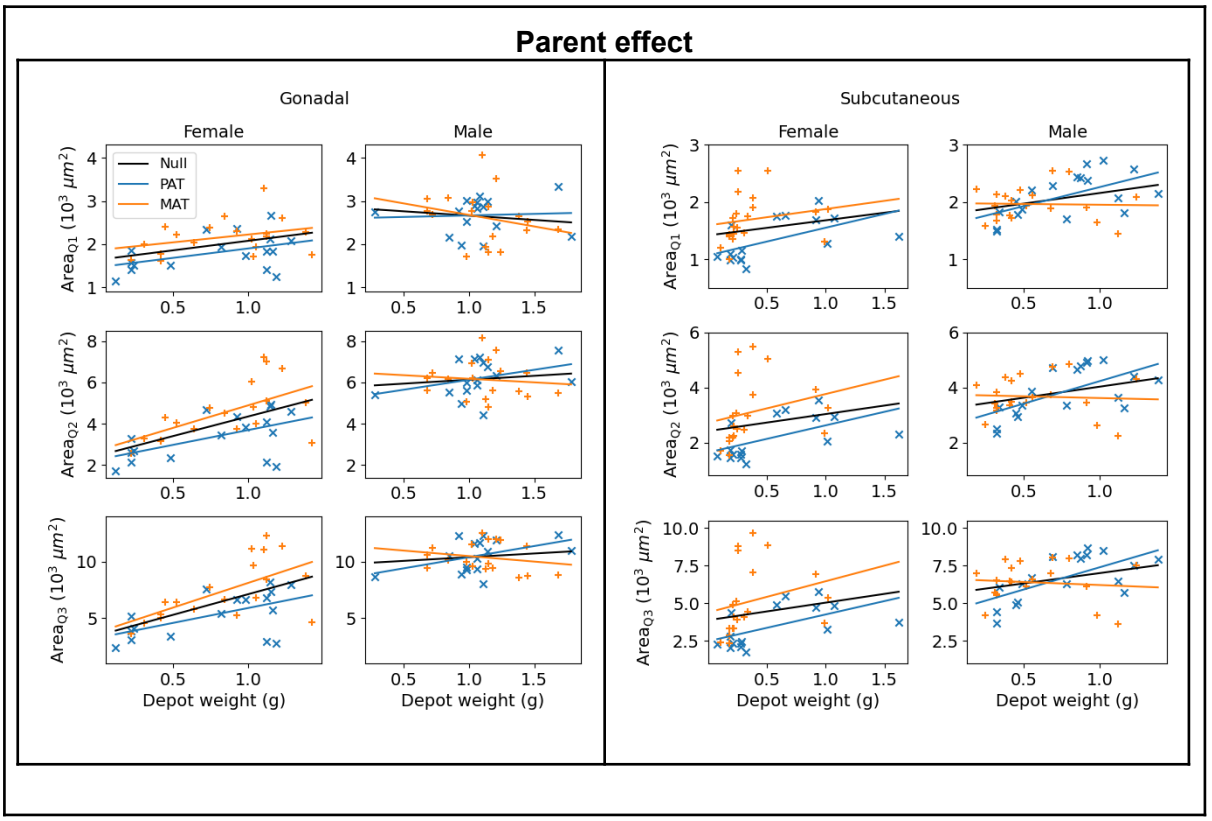


Fig. AREAQ_DW_PARENT_LINREG. Cell quantile area vs. depot weight (DW) and parent effect model. Scatter plots and fitted robust linear models ($area_q$ ~ DW) for cell population mode and quartiles (Q1, Q2, Q3), stratified by depot and sex. Each point corresponds to a mouse.

Parent effect					
Depot	Parent	Intercept (95% CI)		p-value	
Female					
Gonadal PAT	Q1	1463.9	(1037.8, 1890)	2.51E-06	****
	Q2	2258.1	(1212.4, 3303.7)	0.00035	***
	Q3	3248.2	(1385.5, 5110.9)	0.0021	**
Gonadal MAT	Q1	1860.4	(1364.7, 2356.1)	4.20E-07	****
	Q2	2702.4	(1291.6, 4113.3)	0.00085	***
	Q3	3734.5	(1101.8, 6367.2)	0.0082	**
Subcut. PAT	Q1	1064.9	(790.2, 1339.6)	8.73E-07	****
	Q2	1639.0	(1071, 2207.1)	2.36E-05	****
	Q3	2464.4	(1460.6, 3468.2)	0.00012	***
Subcut. MAT	Q1	1589.5	(1257.8, 1921.1)	8.04E-09	****
	Q2	2718.2	(1770.6, 3665.8)	1.07E-05	****
	Q3	4372.6	(2533.5, 6211.7)	9.38E-05	****
Male					
Gonadal PAT	Q1	2593.9	(1780.8, 3407)	8.04E-06	****
	Q2	5158.4	(3541.1, 6775.8)	8.06E-06	****

	Q3	8414.9	(5933.2, 10896.5)	4.08E-06	****
Gonadal MAT	Q1	3210.8	(1891.6, 4530.1)	9.50E-05	****
	Q2	6518.4	(4530.9, 8505.9)	3.26E-06	****
	Q3	11465.5	(8709.7, 14221.3)	1.53E-07	****
Subcut. PAT	Q1	1592.5	(1179.7, 2005.3)	4.17E-07	****
	Q2	2594.9	(1670.2, 3519.5)	2.04E-05	****
	Q3	4419.7	(2872.9, 5966.6)	1.66E-05	****
Subcut. MAT	Q1	1979.5	(1673.5, 2285.4)	6.62E-11	****
	Q2	3745.8	(3007.4, 4484.2)	3.32E-09	****
	Q3	6608.1	(5264.8, 7951.4)	5.37E-09	****

(a)

Depot	Parent	β(DW) (95% CI)		p-value		Corrected p-value	
Female							
Gonadal PAT	Q1	437.8	(-50.4, 926)	0.075	ns	0.16	ns
	Q2	1439.1	(241.2, 2637)	0.022	*	0.055	ns
	Q3	2653.0	(519, 4787)	0.018	*	0.055	ns
Gonadal MAT	Q1	363.4	(-180.1, 906.8)	0.18	ns	0.32	ns
	Q2	2190.4	(643.8, 3736.9)	0.0083	**	0.052	ns
	Q3	4390.7	(1504.7, 7276.6)	0.0051	**	0.052	ns
Subcut. PAT	Q1	484.5	(91, 878.1)	0.019	*	0.055	ns

	Q2	986.2	(172.5, 1799.9)	0.021	*	0.055	ns
	Q3	1781.6	(343.7, 3219.5)	0.019	*	0.055	ns
Subcut. MAT	Q1	287.1	(-438.7, 1012.8)	0.42	ns	0.55	ns
	Q2	1046.1	(-1027.4, 3119.7)	0.30	ns	0.48	ns
	Q3	2080.3	(-1944, 6104.6)	0.29	ns	0.48	ns
Male							
Gonadal PAT	Q1	71.5	(-652.8, 795.8)	0.84	ns	0.92	ns
	Q2	965.6	(-475, 2406.3)	0.17	ns	0.32	ns
	Q3	1976.6	(-233.9, 4187.1)	0.076	ns	0.16	ns
Gonadal MAT	Q1	-534.2	(-1685.4, 617)	0.34	ns	0.50	ns
	Q2	-346.8	(-2081.1, 1387.4)	0.68	ns	0.83	ns
	Q3	-976.9	(-3381.6, 1427.8)	0.40	ns	0.55	ns
Subcut. PAT	Q1	668.5	(174, 1162.9)	0.011	*	0.055	ns
	Q2	1640.5	(533.1, 2747.9)	0.0063	**	0.052	ns
	Q3	2965.3	(1112.7, 4817.9)	0.0037	**	0.052	ns
Subcut. MAT	Q1	-27.7	(-507.2, 451.8)	0.90	ns	0.95	ns
	Q2	-127.0	(-1284.4, 1030.4)	0.82	ns	0.92	ns
	Q3	-406.5	(-2512, 1699)	0.69	ns	0.83	ns

(b)

Table AREAQ_DW_PARENT_LINREG. Coefficients of OLS model ($area_q$ ~ DW) stratified by parent and depot.

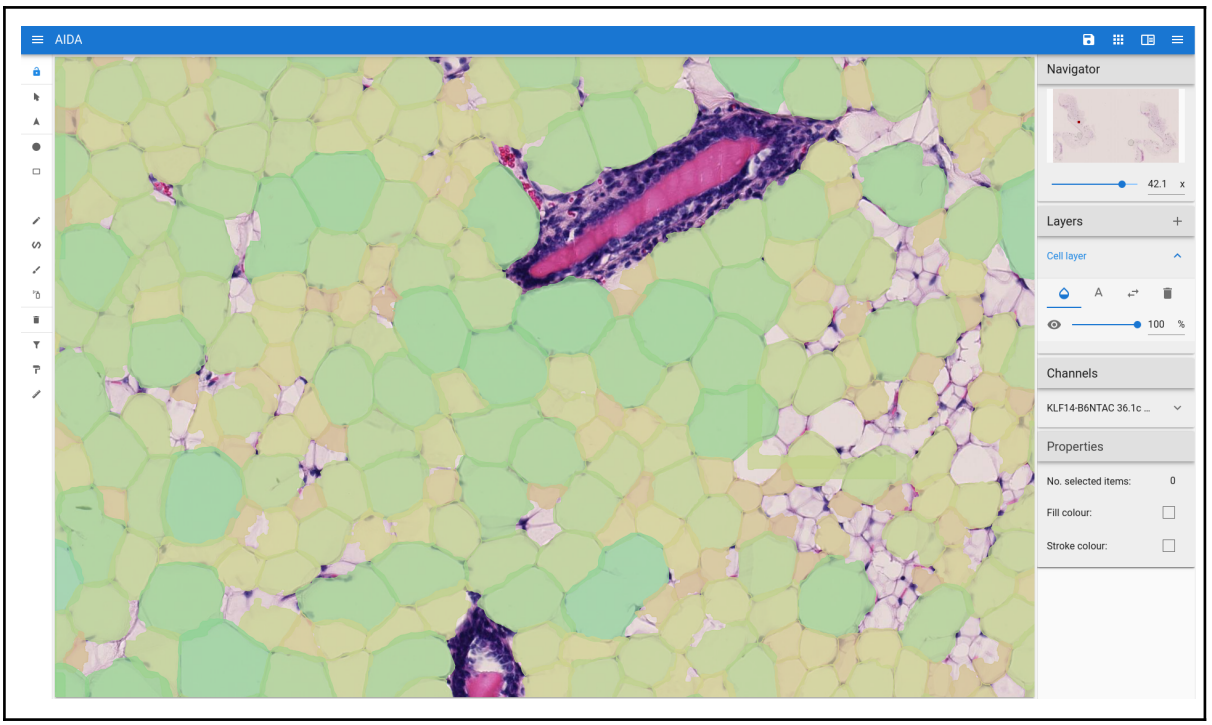


Fig. AIDA. AIDA web interface with DeepZoom navigation of NDPI histology file with DeepCytometer contours overlaid.

Acknowledgements

We would like to thank George Nicholson (Oxford Statistical Department) and Laura Bramley (IBME, University of Oxford) for their advice on Statistical methods; Stefano Malacrino (Dept. Engineering Science, University of Oxford), for his help with AIDA; and Liz Bentley, for her advice in the project.

Mouse Klf14$^{tm1(KOMP)Vlcg}$ adipose samples, body and depot weight data were obtained from a previous study that was supported by UK Medical Research Council grant MR/J010642/1. RDC is supported by UK Medical Research Council funding MC_U142661184.

Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Conflicts of interest

A.A. and J.R. are co-founders and equity holders of Ground Truth Labs; an AI and digital pathology company.

References

1. Bjørndal, B., Burri, L., Staalesen, V., Skorve, J. & Berge, R. K. Different Adipose Depots: Their Role in the Development of Metabolic Syndrome and Mitochondrial Response to Hypolipidemic Agents. *Journal of Obesity* https://www.hindawi.com/journals/jobe/2011/490650/ (2011) doi:10.1155/2011/490650.
2. Parlee, S. D., Lentz, S. I., Mori, H. & MacDougald, O. A. Quantifying Size and Number of Adipocytes in Adipose Tissue. in *Methods in Enzymology* (ed. Macdougald, O. A.) vol. 537 93–122 (Academic Press, 2014).
3. Eto, H. *et al.* Characterization of Structure and Cellular Components of Aspirated and Excised Adipose Tissue. *Plast. Reconstr. Surg.* **124**, 1087–1097 (2009).
4. Vernon, R. G. & Flint, D. J. ADIPOSE TISSUE | Structure and Function of White Adipose Tissue. in *Encyclopedia of Food Sciences and Nutrition (Second Edition)* (ed. Caballero, B.) 23–29 (Academic Press, 2003). doi:10.1016/B0-12-227055-X/00007-9.
5. Lenz, M., Arts, I. C. W., Peeters, R. L. M., de Kok, T. M. & Ertaylan, G. Adipose tissue in health and disease through the lens of its building blocks. *Sci. Rep.* **10**, 10433 (2020).
6. Nishimura, S. *et al.* Adipogenesis in Obesity Requires Close Interplay Between Differentiating Adipocytes, Stromal Cells, and Blood Vessels. *Diabetes* **56**, 1517–1526 (2007).
7. Rajbhandari, P. *et al.* Single Cell Analysis Reveals Immune Cell-Adipocyte Crosstalk Regulating the Transcription of Thermogenic Adipocytes. *bioRxiv* 669853 (2019)

doi:10.1101/669853.

8. Skelly, D. *et al.* Diet-driven changes in immune regulation of adipose tissue revealed by single cell transcriptomics. in *33rd International Mammalian Genome Conference* 66 (2019).
9. Ye, J. Adipose Tissue Vascularization: Its Role in Chronic Inflammation. *Curr. Diab. Rep.* **11**, 203–210 (2011).
10. Boumelhem, B. B., Assinder, S. J., Bell-Anderson, K. S. & Fraser, S. T. Flow cytometric single cell analysis reveals heterogeneity between adipose depots. *Adipocyte* **6**, 112–123 (2017).
11. 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).
12. Fang, L., Guo, F., Zhou, L., Stahl, R. & Grams, J. The cell size and distribution of adipocytes from subcutaneous and visceral fat is associated with type 2 diabetes mellitus in humans. *Adipocyte* **4**, 273–279 (2015).
13. Small, K. S. *et al.* Regulatory variants at KLF14 influence type 2 diabetes risk via a female-specific effect on adipocyte size and body composition. *Nat. Genet.* **50**, 572–580 (2018).
14. Hausman, D. B., DiGirolamo, M., Bartness, T. J., Hausman, G. J. & Martin, R. J. The biology of white adipocyte proliferation. *Obes. Rev.* **2**, 239–254 (2001).
15. Vazquez, G., Duval, S., Jacobs, D. R. & Silventoinen, K. Comparison of body mass index, waist circumference, and waist/hip ratio in predicting incident diabetes: a meta-analysis. *Epidemiol. Rev.* **29**, 115–128 (2007).
16. Myint, P. K., Kwok, C. S., Luben, R. N., Wareham, N. J. & Khaw, K.-T. Body fat percentage, body mass index and waist-to-hip ratio as predictors of mortality and cardiovascular disease. *Heart* **100**, 1613–1619 (2014).

17. Pulit, S. L. *et al.* Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum. Mol. Genet.* **28**, 166–174 (2019).
18. Verboven, K. *et al.* Abdominal subcutaneous and visceral adipocyte size, lipolysis and inflammation relate to insulin resistance in male obese humans. *Sci. Rep.* **8**, 4677 (2018).
19. Blüher, M. Obesity: global epidemiology and pathogenesis. *Nat. Rev. Endocrinol.* **15**, 288–298 (2019).
20. Kim, J.-Y. *et al.* Obesity-associated improvements in metabolic profile through expansion of adipose tissue. *J. Clin. Invest.* **117**, 2621–2637 (2007).
21. Kusminski, C. M. *et al.* MitoNEET-driven alterations in adipocyte mitochondrial activity reveal a crucial adaptive process that preserves insulin sensitivity in obesity. *Nat. Med.* **18**, 1539–1549 (2012).
22. Yang, Q. & Civelek, M. Transcription Factor KLF14 and Metabolic Syndrome. *Front. Cardiovasc. Med.* **7**, (2020).
23. Dale Caroline E. *et al.* Causal Associations of Adiposity and Body Fat Distribution With Coronary Heart Disease, Stroke Subtypes, and Type 2 Diabetes Mellitus. *Circulation* **135**, 2373–2388 (2017).
24. Winkler, T. W. *et al.* A joint view on genetic variants for adiposity differentiates subtypes with distinct metabolic implications. *Nat. Commun.* **9**, 1946 (2018).
25. Stamatoyannopoulos, J. A. *et al.* An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol.* **13**, 418 (2012).
26. Reed, D. R., Bachmanov, A. A. & Tordoff, M. G. Forty mouse strain survey of body composition. *Physiol. Behav.* **91**, 593–600 (2007).
27. Podrini, C. *et al.* High-fat feeding rapidly induces obesity and lipid derangements in C57BL/6N mice. *Mamm. Genome* **24**, 240–251 (2013).

28. van Beek, L. *et al.* The limited storage capacity of gonadal adipose tissue directs the development of metabolic disorders in male C57Bl/6J mice. *Diabetologia* **58**, 1601–1609 (2015).
29. Lutz, T. A. & Woods, S. C. Overview of Animal Models of Obesity. *Curr. Protoc. Pharmacol. Editor. Board SJ Enna Ed.--Chief Al* **CHAPTER**, Unit5.61 (2012).
30. Phan, J., Hickey, M. A., Zhang, P., Chesselet, M.-F. & Reue, K. Adipose tissue dysfunction tracks disease progression in two Huntington's disease mouse models. *Hum. Mol. Genet.* **18**, 1006–1016 (2009).
31. Williams-Dautovich, J. *et al.* The CRH-Transgenic Cushingoid Mouse Is a Model of Glucocorticoid-Induced Osteoporosis. *JBMR Plus* **1**, 46–57 (2017).
32. Parker-Katiraee, L. *et al.* Identification of the Imprinted KLF14 Transcription Factor Undergoing Human-Specific Accelerated Evolution. *PLOS Genet.* **3**, e65 (2007).
33. Di Girolamo, M., Mendlinger, S. & Fertig, J. A simple method to determine fat cell size and number in four mammalian species. *Am. J. Physiol.-Leg. Content* **221**, 850–858 (1971).
34. Tchoukalova, Y. D., Harteneck, D. A., Karwoski, R. A., Tarara, J. & Jensen, M. D. A quick, reliable, and automated method for fat cell sizing. *J. Lipid Res.* **44**, 1795–1801 (2003).
35. Bradshaw, A. D., Graves, D. C., Motamed, K. & Sage, E. H. SPARC-null mice exhibit increased adiposity without significant differences in overall body weight. *Proc Natl Acad Sci USA* **100**, 6045–6050 (2003).
36. Hirsch, J. & Gallian, E. Methods for the determination of adipose cell size in man and animals. *J. Lipid Res.* **9**, 110–119 (1968).
37. Majka, S. M. *et al.* Analysis and Isolation of Adipocytes by Flow Cytometry. *Methods Enzymol.* **537**, 281–296 (2014).
38. Chen, H. C. & Farese, R. V. Determination of adipocyte size by computer image

analysis. *J. Lipid Res.* **43**, 986–989 (2002).

39. Zampirolli, F. de A., Stransky, B., Lorena, A. C. & Paulon, F. L. de M. Segmentation and Classification of Histological Images - Application of Graph Analysis and Machine Learning Methods. in *2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images* 331–338 (2010). doi:10.1109/SIBGRAPI.2010.51.
40. Galarraga, M. *et al.* Adiposoft: automated software for the analysis of white adipose tissue cellularity in histological sections. *J. Lipid Res.* **53**, 2791–2796 (2012).
41. Osman, O. S. *et al.* A novel automated image analysis method for accurate adipocyte quantification. *Adipocyte* **2**, 160–164 (2013).
42. Zhi, X. *et al.* AdipoCount: A New Software for Automatic Adipocyte Counting. *Front. Physiol.* **9**, (2018).
43. Arganda-Carreras, I. *et al.* Trainable Weka Segmentation: a machine learning tool for microscopy pixel classification. *Bioinformatics* **33**, 2424–2426 (2017).
44. Moen, E. *et al.* Deep learning for cellular image analysis. *Nat. Methods* **16**, 1233–1246 (2019).
45. Falk, T. *et al.* U-Net: deep learning for cell counting, detection, and morphometry. *Nat. Methods* **16**, 67–70 (2019).
46. Sirinukunwattana, K. *et al.* Artificial intelligence–based morphological fingerprinting of megakaryocytes: a new tool for assessing disease in MPN patients. *Blood Adv.* **4**, 3284–3294 (2020).
47. Van Valen, D. A. *et al.* Deep Learning Automates the Quantitative Analysis of Individual Cells in Live-Cell Imaging Experiments. *PLOS Comput. Biol.* **12**, e1005177 (2016).
48. Bannon, D. *et al.* Dynamic allocation of computational resources for deep learning-enabled cellular image analysis with Kubernetes. *bioRxiv* 505032 (2019) doi:10.1101/505032.
49. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic

segmentation. in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 3431–3440 (2015). doi:10.1109/CVPR.2015.7298965.

50. Wang, W. *et al.* Learn to segment single cells with deep distance estimator and deep cell detector. *Comput. Biol. Med.* **108**, 133–141 (2019).
51. Casero, R. *et al.* Cytometer: Computerised segmentation of white adipocytes in full size H&E histology images using convolutional neural networks. in *Procs. of 33rd International Mammalian Genome Conference* P12 (2019).
52. Zhang, J., Hu, Z., Han, G. & He, X. Segmentation of overlapping cells in cervical smears based on spatial relationship and Overlapping Translucency Light Transmission Model. *Pattern Recognit.* **60**, 286–295 (2016).
53. Böhm, A., Ücker, A., Jäger, T., Ronneberger, O. & Falk, T. ISOODL: Instance segmentation of overlapping biological objects using deep learning. in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* 1225–1229 (2018). doi:10.1109/ISBI.2018.8363792.
54. Kleczek, P., Jaworek-Korjakowska, J. & Gorgon, M. A novel method for tissue segmentation in high-resolution H&E-stained histopathological whole-slide images. *Comput. Med. Imaging Graph.* **79**, 101686 (2020).
55. Muñoz-Aguirre, M., Ntasis, V. F. & Guigó, R. *PyHIST: A Histological Image Segmentation Tool*. (Cold Spring Harbor Laboratory, 2020).
56. Coudray, N. *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
57. Benjamini, Y., Krieger, A. M. & Yekutieli, D. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93**, 491–507 (2006).
58. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 1137–1149 (2017).

59. He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask R-CNN. *ArXiv170306870 Cs* (2018).

60. Chen, X., Girshick, R., He, K. & Dollár, P. TensorMask: A Foundation for Dense Object Segmentation. in *International Conf. on Computer Vision (ICCV) 2019* (2019).

61. Satyanarayanan, M., Goode, A., Gilbert, B., Harkes, J. & Jukic, D. OpenSlide: A vendor-neutral software foundation for digital pathology. *J. Pathol. Inform.* **4**, 27 (2013).

62. Sherrah, J. Fully Convolutional Networks for Dense Semantic Labelling of High-Resolution Aerial Imagery. *ArXiv160602585 Cs* (2016).

63. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *ArXiv160600915 Cs* (2016).

64. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *ArXiv14127062 Cs* (2014).

65. Yu, F. & Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *ArXiv151107122 Cs* (2015).

66. Cientanni, G. & Casero, R. Developing a Deep Convolutional Neural Network to differentiate HFD and LFD adipocytes from histology Images. in (2018).

67. Bai, M. & Urtasun, R. Deep Watershed Transform for Instance Segmentation. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2858–2866 (2017). doi:10.1109/CVPR.2017.305.

68. Huang, C., Wu, Q. & Meng, F. QualityNet: Segmentation quality evaluation with deep convolutional networks. in *2016 Visual Communications and Image Processing (VCIP)* 1–4 (2016). doi:10.1109/VCIP.2016.7805585.

69. Luo, W., Li, Y., Urtasun, R. & Zemel, R. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems 29* (eds. Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I. & Garnett, R.)

4898–4906 (Curran Associates, Inc., 2016).

70. Lorensen, W. E. & Cline, H. E. Marching cubes: A high resolution 3D surface construction algorithm. in *Proceedings of the 14th annual conference on Computer graphics and interactive techniques - SIGGRAPH '87* 163–169 (ACM Press, 1987). doi:10.1145/37401.37422.
71. Caicedo, J. C. *et al.* Evaluation of Deep Learning Strategies for Nucleus Segmentation in Fluorescence Images. *Cytometry A* **95**, 952–965 (2019).
72. Cochran, W. G. Problems Arising in the Analysis of a Series of Similar Experiments. *Suppl. J. R. Stat. Soc.* **4**, 102–118 (1937).
73. Cochran, W. G. The Combination of Estimates from Different Experiments. *Biometrics* **10**, 101–129 (1954).
74. Oellrich, A. *et al.* Reporting phenotypes in mouse models when considering body size as a potential confounder. *J. Biomed. Semant.* **7**, (2016).
75. Karp, N. A., Melvin, D., Project, S. M. G. & Mott, R. F. Robust and Sensitive Analysis of Mouse Knockout Phenotypes. *PLOS ONE* **7**, e52410 (2012).
76. Rousselet, G. A., Pernet, C. R. & Wilcox, R. R. Beyond differences in means: robust graphical methods to compare two groups in neuroscience. *Eur. J. Neurosci.* **46**, 1738–1748 (2017).
77. Zeiler, M. D. ADADELTA: An Adaptive Learning Rate Method. *ArXiv12125701 Cs* (2012).
78. He, K., Zhang, X., Ren, S. & Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)* 1026–1034 (IEEE Computer Society, 2015). doi:10.1109/ICCV.2015.123.
79. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 318–327 (2020).
80. Smith, L. N. Cyclical Learning Rates for Training Neural Networks. in *2017 IEEE Winter*

Conference on Applications of Computer Vision (WACV) 464–472 (2017).

doi:10.1109/WACV.2017.58.