Semi-automated annotation of cell type-specific protein expression patterns in human testis based on immunohistochemistry

Biraja Ghoshal
Brunel University London https://orcid.org/0000-0001-5456-2197
Allan Tucker
Brunel University
Feria Hikmet Norradin
Uppsala University
Cecilia Lindskog (cecilia.lindskog@igp.uu.se)
Uppsala University https://orcid.org/0000-0001-5611-1015

Article

Keywords: immunohistochemistry, human cell type-specific proteome, human testis

Posted Date: August 5th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-51610/v1

Abstract

Immunohistochemistry (IHC) provides the basis for cell type-specific localization of protein expression patterns in human tissues. Manual annotation of complex IHC images is however expensive and may lead to errors or inter-observer variability. Artificial Intelligence holds much promise for efficient and accurate pattern recognition, but confidence in prediction needs to be addressed. To present a reliable model for annotation of IHC images, we developed a semi-automated framework for multi-label classification of 7,848 human testis samples stained with IHC, and manually annotated in situ protein expression in eight different cell types. The dataset was used as a basis for training and testing a proposed Hybrid Bayesian Neural Network. By combining the deep learning model with a novel uncertainty metric, the average diagnostic performance improved from 86.9% to 96.3%. The streamlined workflow has important implications for accurate large-scale efforts mapping the human cell type-specific proteome in health and disease.

Background

Proteins are the essential building blocks of life, and for a full understanding of their functional role in health and disease, it is necessary to resolve their spatial distribution at an organ, tissue, cellular and subcellular level. One initiative mapping the entire human proteome is the Human Protein Atlas (HPA) project [1-5]. This large-scale initiative focuses on an integrated omics approach, by combining antibody-based proteomics using immunohistochemistry (IHC) with quantitative assessment of corresponding mRNA levels. IHC is the standard approach for spatial proteomics and an important method for identifying cell type-specific protein expression patterns to deepen the understanding of human biology.

The standard approach for evaluation of IHC staining patterns is manual annotation. While a manual observer has the advantage of identifying technical staining errors or artifacts, it is both time-consuming and costly. Additionally, manual annotation is error-prone and poorly reproducible, as it may lead to fatigue or mislabeling of images due to lack of experience in detecting the correct cell types or structures, or technological challenges related to staining intensity or identification of small objects. Manual annotation is commonly faced with two types of errors, i) false negatives where true positive staining is missed or neglected, and ii) false positives where lack of protein expression is falsely interpreted as positive. Histological samples consist of a mixture of different cell types that can be challenging to distinguish even by a trained eye, and setting a manual threshold of what is regarded as negative/positive is tedious and highly difficult.

To increase accuracy and speed up the process of manual interpretation, the application of Artificial Intelligence (AI) in the evaluation of medical images has received increased attention both in research and diagnostics. AI has been used for many different classification problems across various medical fields, especially in cancer and other diseases [6-10]. While an AI-driven approach holds much promise for efficient and accurate pattern recognition of histological images, few efforts were based on IHC images, and no previous study has used AI for distinguishing between cell type specific protein expression patterns in human IHC samples [11-12].

One of the challenges of the automated annotation of IHC images based on complex tissue samples is that tissues consist of multiple cell types of various shapes and sizes that can express a protein in different combinations. Additionally, a protein may not only be expressed in certain cell types, but could also be localized to different subcellular compartments, e.g. cytoplasm or nucleus, or be expressed at different levels. As a result, training an algorithm to distinguish cell type-specific localization of proteins based on IHC is a multi-label task. Since each class is not mutually exclusive, both the manual observer and the trained model must consider every possible label separately. Different approaches to address multi-label classification problems have been developed previously [13], but none of these have been applied to IHC images. Another challenge is correctly addressing the accuracy of automated predictions, which requires a large dataset of manually annotated images, but also a method to score the confidence in the prediction. Such methods are not currently considered by many state-of-the-art algorithms.

Here, we focused on one particular organ – the testis. Testis has the highest number of tissue elevated genes [14-15], and considered to be one of the most complex organs in the human body due to the spermatogenesis process that requires activation and suppression of thousands of genes and proteins, out of which a large proportion have an unknown function [16-18]. Spermatogenesis is built on a continuous interplay between multiple cell types and cell stages leading to sperm maturation, and for a full understanding of the molecular mechanisms behind infertility and reproductive disorders, it is necessary to study the exact localization of all proteins related to testis specific-functions.

The goal of the present investigation was to automate annotation [19-20] of protein expression patterns based on IHC samples and improve the manual classification accuracy by identifying images with high classification uncertainty. The analysis included a large dataset of 7,848 human testis histology images, corresponding to IHC stainings of 2,794 different proteins, generated as part of the HPA project. All images were manually evaluated in eight different cell types. To address classification confidence, we introduced a bias-corrected uncertainty metric, the Ghoshal-Tucker-Lindskog (GTL) score, enabling ranking of prediction confidence. Such a workflow in estimating uncertainty gives the possibility to discard highly uncertain predictions to guarantee a higher accuracy of the remaining predictions. It also has the advantage to identify unfamiliar patterns corresponding to outliers in the data distribution.

Here, we propose a semi-automated annotation framework together with confidence metrics for multi-label classification of cell type-specific protein expression patterns in testis, based on a Hybrid Bayesian Neural Network (HBNet). To the best of our knowledge, this is the first study combining deep learning of multi-label IHC images with uncertainty measures. The model has important implications for unbiased high-throughput annotation of IHC images and will aid in gaining important biological insights within the field of spatial proteomics, ultimately leading to further understanding of human cell biology in health and disease.

Results

Generation of a semi-automated image annotation framework

A total of 7,848 IHC stained high-resolution images of human testis, corresponding to 3,046 different antibody stainings and 2,794 unique proteins there divided into three different sets: a training set (5,411 images), a validation set (1,063 images) and a test set (1,374 images). All images were annotated manually in five germ cell types (spermatogonia, preleptotene spermatocytes, pachytene spermatocytes, round/early spermatids and elongated/late spermatids), and three somatic cell types (Sertoli cells, Leydig cells and peritubular cells), taking into consideration staining intensity (negative, weak, moderate, strong) and subcellular localization of the staining (cytoplasm, nucleus, membrane). The manually scored images formed the basis for a semi-automated image annotation framework, as presented in Figure 1.

Cell type-specific expression based on manual annotation

To determine the relationship between different cell types based on protein expression as determined by manual annotation, a correlation matrix was generated using Pearson's correlation and Ward's hierarchical clustering (Figure 2a). As expected based on functional characteristics [18], there were three main clusters: i) somatic cells (Sertoli cells, Leydig cells and peritubular cells), ii) premeiotic cells (spermatogonia and preleptotene spermatocytes) and iii) meiotic/post-meiotic cells (pachytene spermatocytes, round/early spermatids and elongated/late spermatids). Of the 7,848 images analyzed, only 815 (10%) showed immunoreactivity in one cell type only, while most of the images were positive in two to five cell types (Figure 2b). In 35 images, the human observer had marked all cell types as negative. When separated, the datasets showed slightly different proportions of the number of positive cell types (Figure 2c), where the test set comprised of more cell type-specific images and validation set contained a higher proportion of images with five to eight cell types that had been labeled (Figure 2c). There were large differences in the presence of different cell type labels (Figure 2d), with Leydig cells being labeled in as many as 5,218 (66%) of the images, while peritubular cells represented the most unusual staining pattern, positive in only 755 (10%) of the images. The staining was mostly localized to the cytoplasm, both cytoplasm and the plasma membrane, or the nucleus, but there were clear differences between cell types. Sertoli cells more often showed positivity in plasma membrane or a combination of nucleus + membrane, in most cases referred to as the nuclear membrane. A majority of the staining observed in Leydig cells was cytoplasmic (Figure 2d).

Training of neural network and overall model performance

The manually annotated images from the training set of 5,411 images and the validation set of 1,063 images were used for training a Hybrid Bayesian Neural Network (HBNet) model, exploiting DropWeights and combining the features from a standard deep neural network (DNN) with handcrafted features. The output of the neural network is an 8-dimensional probability vector, where each dimension indicates how likely each cell type in a given image expresses the protein. The neural network was then applied to the test set of 1,374 images, for which the accuracy was evaluated.

Evaluation metrics for multi-label classification performances are different from those used in binary or multi-class classification [21]. In multi-label classification, a miss-classification is no longer a definite right or wrong, since a correct prediction containing a subset of the actual labels is considered better than a prediction containing none of them. Here, four different metrics were used for evaluating the multi-label classification performance: i) Hamming loss, ii) F1-score, iii) Exact Match ratio, and iv) mean-Average Precision (mAP). Table 1 presents the statistics for each of these metrics both for standard DNN and the proposed HBNet. Hamming loss is the most common evaluation metric in multi-label classification, which takes into account both prediction errors (false positives) and missed predictions (false negatives), normalized over the total number of classes and total number of samples analyzed. The smaller the value of Hamming loss (closer to 0), the better the performance of the learning algorithm. F1 score is the harmonic mean of recall and precision, where Macro F1 score calculates the metric independently for each label and then takes an average, and Micro F1 score aggregates the contributions of all labels when calculating the average metric. The Exact Match ratio is the strictest metric, indicating the percentage of all analyzed samples that have all their labels classified correctly. Mean Average Precision (mAP) takes into account both the average precision (AP) separately for each label and the average over the class. It provides a measure of quality across recall levels, and has shown to be stable and able to distinguish between cell types. The higher the mAP (closer to 100), the better the quality. In the present investigation, there was considerable improvement using HBNet across all metrics used (Table 1). Based on HBNet, the Exact Match ratio showed that 67% of the 1,374 images were correctly classified in all eight cell types.

Cell type-specific model performance

Next, we evaluated the model's performance on a cell type-specific level. In Figure 3, a confusion matrix is shown, comparing the output of the neural network with the manual observer, summarising the false positives and negatives of the DNN and the HBNet for each cell type. For all cell types, HBNet had a higher accuracy than DNN, with >80% overall accuracy, and >90% for Sertoli cells and peritubular cells. The largest difference between DNN and HBNet was seen for pachytene spermatocytes and round/early spermatids, where the accuracy improved from 75.6 to 82.6%, and 69.3 to 80.5%, respectively. HBNet dramatically reduced the number of false negatives compared to DNN, but also showed a decrease in the number of false positives. The total number of false positives (n=444) across all cell types was lower compared to the number of false negatives (n=993), indicating that the model performed better at accurately detecting positive labels, but more often differed with the human observer in classifying cell types as negative. This is expected, due to the human observer deliberately neglecting very weak staining patterns that can be considered unspecific or being due to artifacts. The ratios between false positives and false negatives were however opposite for Sertoli cells and peritubular cells, for which false negatives were rare. Positivity in these cell types was not only less common in general (Figure 2d), but also to a larger extent cell type-specific and not as often showing simultaneous staining in other cell types (Figure 2a). This suggests that positivity in these cell types was mostly considered as specific by the human observer.

Estimation of model certainty

To rank all images based on model confidence over eight cell types, each prediction included an uncertainty measurement, presented as a GTL Score. Supplementary Table 1 shows the predictions per cell type for each of the 1,374 images in the test set, along with GTL Score and manual annotation. The GTL Scores ranged from zero to one for each HBNet prediction over the eight cell types. All predictions were then plotted in confidence maps (Figure 4), where images for which the model agreed with the human observer, i.e. the cell type was truly positive or truly negative, were marked in green, whilst images with disagreement between the model and the human observer were marked in red. Images suggested to be misclassified tend to have lower GTL Scores, compared to correctly classified images. The shape of the GTL curves varies for each cell type, and the curves for Sertoli cells and peritubular cells stood out as having a higher proportion of images with low GTL Scores than the other cell types. This is because staining in these cell types was less common (Figure 2d), and cell types classified as lacking staining often have low GTL Scores. The spread of misclassifications determined the cutoff for reliable classification, which was marked as a blue line. Note that this cutoff was set at a GTL Score between 0.0 and 0.11 for all types except pachytene spermatocytes, round/early spermatids and elongated/late spermatids, for which it was set at 0.22, 0.78 and 0.22, respectively. The protein expression patterns of these three cell types showed a high correlation (Figure 2a), suggesting that many proteins were co-expressed in these cells. Since they were not mutually exclusive, this may explain why the model would have more difficulties to distinguish these cell types from each other.

When only considering thresholded samples above the GTL cutoff, including classifications of high reliability, the classification accuracy of the HBNet model was substantially improved (Table 2). The HBNet GTL-thresholded accuracy was >92% for all cell types

except for round/early spermatids, which had an accuracy of 83.5%. For most cell types, approximately 30 to 39% of the images were below the GTL cutoff, except for peritubular cells where only 1.3% of the images were discarded, and Sertoli cells, where none were. Predictions above cutoff can be considered reliably annotated by the model, which means that manual annotation is only needed for on average 28.1% of the predictions. Note that there is a direct tradeoff for choice of GTL threshold between accuracy and number of discarded images (Supplementary Figure 1).

Evaluation of correctly classified and misclassified images

The GTL confidence metric allowed us to identify both correctly classified images, as well as images where the model disagreed with the human observer for one or several cell types. In Figure 5, examples of correctly classified images are provided, i.e. these images were among the 67% that according to the Exact Match Ratio had all eight cell types annotated as either true positive or true negative. The images show that the model performed well both for proteins with distinct and selective staining and for more complex images where the protein was expressed in several cell types of varying intensity and staining patterns. The IHC stained images are presented along with heatmaps [22] highlighting which area of the images that the model focused on for making the labeling decision. For the correctly classified images, it is evident that the model focused on several different areas within the image, including areas where cells were intact and well-represented.

Misclassified predictions included both falsely positive and falsely negative images, and could be further divided into cases with high certainty (high GTL Score) and low certainty (low GTL Score). Several misclassified predictions represented clear errors made by the manual observer (Figure 6a). Such misclassifications often had high GTL Scores, and in these cases, the model can be used for identifying manual mistakes. Other misclassified predictions were due to unspecific staining deliberately neglected by the human observer (Figure 6b). Such stainings in need of further protocol optimization were often represented by false negative predictions with high GTL Scores, indicating that the model performed a correct prediction, but based on experience, the positivity was interpreted as unspecific by the human observer. Some misclassified images corresponded to proteins expressed in small structures including nuclear membranes, nucleoli or centrosomes (Figure 6c). Such staining patterns are rare, and may be particularly challenging for the model to interpret due to limitations in the current pixel resolution. These predictions were often false positives with low GTL Scores. Finally, some misclassified images contained artifacts, such as damaged tissue sections, or sections that contained areas where the testicular samples were not completely healthy (Figure 6d). Such misclassifications, both false positives and false negatives, often had low GTL Scores and it was evident from the model heatmaps that the labeling decisions were mostly made on areas of the images where not all cell types were clearly represented, or the image/visible cells had poor quality.

Model performance based on subcellular localization and staining intensity

The manual annotation of the cell type-specific protein expression did not only take into consideration which cell types that were positive, but also in which subcellular organelle the staining was observed. In Table 3, the GTL-thresholded model performance in the test dataset is presented on a subcellular level. Similarly, as in the whole dataset, (Figure 2d), it was clear that some organelles were more common in certain testicular cell types, which may affect the overall accuracy, but it should also be noted that the patterns of different subcellular localizations appear differently in the various cell types based on the cell shape. In total, the best accuracy was found for staining patterns where all subcellular localizations (cytoplasmic, membranous and nuclear) were present. This is not surprising, as clear outlining of each cell structure increases the likelihood of the model identifying the correct cell types. Sertoli cells had lower accuracy of certain subcellular localizations compared to other cell types. Staining of Sertoli cells is challenging to interpret as these cells are situated in the interspace between the germ cells, and staining may be difficult to distinguish from other cell types.

In addition to cell type-specific pattern and subcellular localization of the staining, the human observer also takes into consideration the intensity of the staining. This rather subjective measurement that determines the brown saturation level, is considered to represent the amount of protein expression ranging from low levels (weak staining/beige color), through moderate levels (medium brown) to high levels (dark brown/black). As seen in Table 4, it is evident that the GTL-thresholded accuracy did not depend on staining intensity, and there was no significant improvement in predictions performed on distinctly stained cells compared to those that showed more faint positivity.

Discussion

IHC constitutes the standard approach for spatial localization of proteins at a cell type-specific level. The technology originates from the early 1940s [23] and has emerged as a quick, simple and cost-effective method applicable to both diagnostic routine as well as basic and clinical research. The output of the IHC staining is typically a tissue section manually evaluated under a microscope, but with advances in digital pathology, large-scale digitization of stained sections is becoming more common. This allows for the development of algorithms to automatically predict the IHC staining. There is a widely acknowledged demand for implementation of such algorithms both in healthcare and research for more accurate and fast annotation. Until date, there are however no previous studies suggesting how such frameworks can be implemented for high-throughput annotation of complex tissue samples stained with IHC. Despite impressive reported accuracy, deep learning models tend to require huge training sample image sets. Furthermore, they tend to make overconfident predictions and lack the ability to report “I don’t know” for ambiguous or unknown cases. It is therefore not sufficient to depend on prediction scores alone from deep learning models, but critical to estimate bias-reduced uncertainty as an additional insight to the prediction.

Combining IHC with the TMA technology where a large number of tissue samples are assembled into one array allows for large-scale mapping efforts studying the entire human proteome. Built upon this strategy, the HPA project has characterized >15,000 different proteins across >40 different normal tissues and organs, and 20 types of cancer [1-2]. The publicly available database www.proteinatlas.org contains >10 million high-resolution images that have been manually annotated, thereby constituting a major resource for machine learning algorithms. In the present investigation, we focused on images of normal testis, due to the complex architecture of this organ built up by several different cell types that are challenging to interpret, and the unique nature of this tissue harbouring a large number of specific proteins of unknown function that are interesting to characterize further [16-17].

We here successfully associated deep learning-based predictions on cell type-specific protein expression patterns in histological sections stained with IHC. The predictions were combined with hybrid image features, DropWeights methods and approximate BNN to compute a bias-reduced uncertainty score as a vital additional measure, generating an uncertainty measurement defined as GTL Score. The proposed HBNet architecture showed outstanding performance in both simple images with clear cell type-specific staining, and more complex images where several cell types showed positivity of varying intensity and staining patterns. The novel GTL Score adds another level of insight, particularly important for challenging cases where uncertain predictions can be highlighted. This unique workflow of image annotation allows for dividing the dataset into images that are reliably classified by the model, and images that need to be annotated by the manual observer, thereby introducing a semi-automated high-accuracy framework that reduces the manual burden.

Manual annotation of testis samples is a tedious task, as the germ cell lineage constitutes a continuum, where the stem cells undergo several steps of mitosis and meiosis before being developed into mature sperm. This complex process involves thousands of genes and proteins activated and repressed at certain time points, which means that different proteins are expressed in certain combinations of these cell types. Some proteins may be expressed in just one subset, while others are more ubiquitously expressed. Many proteins are increased or decreased during differentiation, seen as a gradient in expression. It should also be noted that despite the germ cells can be divided into distinct stages and cell types, some proteins may be expressed just between these stages during a short period in time, which means that such proteins may be found only in the seminiferous ducts that are in the correct stage, i.e. variations may occur within the same image. Furthermore, depending on the stage and how the section was taken, not all ducts within an image may contain the end product of the spermatogenesis - the mature sperm. A manual observer needs to take into consideration all cells for each of the eight cell types that are present within an image, and make an average decision that corresponds to the overall expression pattern.

The manual annotation is not only based on visual examination of staining intensity, but to a large extent also relies on experience, where the manual observer takes into consideration staining protocol, overall image quality, artifacts and previous literature on the protein being analyzed. An optimized staining protocol shows only specific antibody binding, i.e. brown color, in structures expressing the protein. It is however extremely challenging to retrieve the right balance of specific staining vs. unspecific antibody binding [24], and despite the HPA spends considerable efforts on antibody validation [25], many IHC images contain weak staining that may be considered unspecific. Such off-target binding means that the antibody when present in high enough concentration binds to structures that do not express the protein. This staining is of lower intensity than the true protein expression, and can be neglected by the human observer, especially when accompanied with distinct staining in other structures that more likely represents the true protein expression. Such experience-driven decisions take into consideration the overall positivity and general staining pattern in the whole image. If certain cells show strong nuclear staining, it is likely that faint cytoplasmic staining seen in other cells is unspecific and regarded as background,

even if it is of an intensity that in other images would be considered above threshold. Furthermore, the manual observer can consult available literature on proteins that are challenging to interpret, which may guide in which cell types or subcellular localization the protein should be present. Finally, the manual observer is better at detecting artifacts. The edge of a tissue core may more often attract unspecific binding, and staining only observed near this border would therefore often be neglected. Despite all tissue samples having undergone quality control before processing, it is possible that a fraction of the samples or a certain part within an image contain cells that are not healthy. If representative cells are still available, the human observer would only consider these in the annotation process and neglect areas with artifacts. Images may also contain mechanical artifacts related to the tissue processing such as folds, scratches or damaged structures, or parts of the images may be unfocused during digitization.

Despite the challenges related to tissue processing, IHC staining and manual annotation, our proposed HBNet showed high accuracy for all eight cell types, especially after applying a GTL Score threshold. When examining images above and below this threshold, it was evident that many images for which the model faced challenges constituted images expected to be particularly difficult, often due to the reasons described above. Three cell types needed a higher GTL Score threshold for reliable prediction: pachytene spermatocytes, round/early spermatids and elongated/late spermatids. This is not surprising, as these cells correspond to the most common combination for proteins co-expressed in more than one testicular cell type, as described previously [18]. This means that proteins are more commonly expressed in combinations of two to three of these cell types than solely expressed in just one of them. As a result, a high proportion of the images used for model training corresponded to proteins co-expressed in these cells, thereby leading to challenges in distinguishing them.

Previous multi-level classification studies, including a recent Kaggle challenge [26] have used immunofluorescence (IF) images of human cell lines, where antibody staining determined different subcellular localizations of the protein, related to the Subcellular Atlas of the HPA [5, 27]. There have however been fewer automated classification studies using human tissue images stained with IHC [11, 28-30]. Previous studies conducted on histological images mainly focused on tissue type classification for disease detection [12]. Here, we present the first approach in multi-label classification based on antibody-based proteomics, to recognize cell type-specific protein expression patterns in eight different testicular cell types.

In addition to highlighting which images that need to be examined by the manual observer and allow for large-scale IHC efforts, our proposed workflow incorporates a confidence metric that has important implications for identifying images with manual annotation errors, and thereby improving the overall accuracy. This is applicable to both research and clinical routine and may replace the otherwise common manual annotation workflow by which one observer first annotates each image, followed by quality control by a second observer. It may also be used for teaching purposes in the training of manual observers that have less experience, which saves both time and money as less quality control is needed from experienced personnel.

In the present investigation, healthy samples from one particular tissue – testis – were used. Based on the encouraging performance of our proposed model for what constitutes a particularly challenging tissue, we believe that the approach is applicable also on other tissues. Similar workflows can be used in projects focusing on distinguishing between healthy and diseased tissues, widely applicable to e.g. cancer research but also routine diagnostics. The daily pathology workflow largely depends on manual microscopic evaluation of tissue sections, which may not only lead to a delayed disease diagnosis with potential worsened patient prognosis but also to a false diagnosis [31]. Further advances in automated annotation of histological sections are therefore clearly warranted.

In the explosive era of “big data”, the emerging field of single cell RNA-seq (scRNA-seq) has received increased attention during the last few years. This novel technology allows for quantitative measurements of single cell transcriptomes across different human tissues and cell types [32]. Further advances in this field will lead to the possibility to identify sets of genes and proteins elevated in certain cell types, e.g. a gene may not only be defined as elevated in testis in comparison to other organs, but robust data suggests that the gene is elevated in e.g. spermatogonia. In order to translate such findings on the transcriptomic level to functionally relevant information, it is necessary to complement the data with studies on the protein level, as the proteins constitute the functional representation of the genome. With IHC constituting the main approach for cell type-specific localization of proteins in human tissues, this will likely result in an increased interest for *in situ*-based techniques such as IHC and other antibody-based technologies, both for studying one protein at a time, and multiplex efforts where several different proteins are labeled simultaneously in one tissue section. For implementation of such large-scale efforts, machine learning approaches that can save both time and money and lead to more accurate image annotations which is highly desirable.

Here, we present a comprehensive strategy for semi-automated annotation of IHC sections combined with an uncertainty metric. The suggested streamlined workflow constitutes an important approach for accurate large-scale efforts mapping the human proteome at a cell type-specific level, and holds promise for both research and diagnostics aiming at analyzing the spatio-temporal expression of human proteins in health and disease.

Methods

Tissues and protein profiling

Human tissue samples for IHC analysis were collected and handled in accordance with Swedish laws and regulations. Tissues were obtained from the Clinical Pathology department, Uppsala University Hospital, Sweden and collected within the Uppsala Biobank organization. All samples were anonymized for personal identity by following the approval and advisory report from the Uppsala Ethical Review Board (Ref # 2002-577, 2005-388, 2007-159). Informed consent was obtained from all subjects in the study. Generation of tissue microarrays (TMAs), IHC staining and digitization of stained TMA slides was performed essentially as previously described [33]. In brief, formalin-fixed, paraffin-embedded (FFPE) tissue blocks were assembled into TMAs based on 1 mm cores from 44 different normal tissue types corresponding to three individuals per tissue, including normal testis samples from adult individuals. TMA blocks were cut in 4 µm sections, dried overnight at room temperature (RT), and baked at 50°C for at least 12h. Automated IHC was performed by using Lab Vision Autostainer 480S Module (ThermoFisher Scientific, Freemont, CA), as described in detail previously. The stained slides were digitized with ScanScope AT2 (Leica Aperio, Vista, CA) using a 20x objective. All digital images corresponding to antibody data that passed HPA quality criteria were made publicly available on www.proteinatlas.org

Dataset and manual annotation

High-resolution digital images of IHC stained testis TMA cores corresponding to 512 testis elevated proteins [18], publicly available on the HPA version 18 (v18.proteinatlas.org), were downloaded along with images from 2,282 proteins published in the current version 19 (v19.proteinatlas.org) that previously had been manually annotated as showing IHC staining of moderate intensity in at least a subset of cells in testis. All proteins were analyzed with at least one antibody that was approved according to Human Protein Atlas criteria for antibody validation. For most of the proteins, three different images were available, and the total dataset comprised 7,848 images corresponding to 2,794 unique human proteins. Each antibody staining was manually re-annotated in eight different testicular cell types, including five germ cell types (spermatogonia, preleptotene spermatocytes, pachytene spermatocytes, round/early spermatids and elongated/late spermatids), and three somatic cell types (Sertoli cells, Leydig cells and peritubular cells). The annotation considered staining intensity (negative, weak, moderate, strong) and subcellular localization (cytoplasmic, nuclear, membranous, or a combination of those). The entire dataset was divided into three sets: A training set of 5,411 images (manually annotated by one observer), a validation set of 1,063 images (manually annotated by one observer and quality controlled by two independent observers, as previously described [18]), and a test set of 1,374 images (manually annotated by one observer and quality controlled by one independent observer).

The Hybrid Bayesian Neural Network (HBNet)

The models were trained and evaluated using Keras with a Tensorflow backend. For hybrid feature extraction, we used a combination of hand-crafted feature extraction and a convolutional neural network (CNN) approach [34-35]. The original JPEG images of 3000x3000 pixels were resized to 1024x1024 pixels using a bicubic interpolation over a 4x4 pixel neighborhood.

We used a generic building block containing the VGG16 [36-37] network to extract deep image features and generate heatmaps. The output of the final pooling layer was the CNN feature. We introduced fully connected layers on top of the VGG16 convolutional base to keep the main parts of VGG16 architecture and connect hand-crafted features to the end of CNN feature as the input of the fully connected layers.

Hand-crafted features were extracted separately complementary to the CNN feature. The hand-crafted approaches used were Histogram of Oriented Gradients (HOG) [36], Haralick [38] and HU Moments [39]. HOG was applied to all images equally, with eight orientation bins, 8x8 pixels forming a single cell, and those cells organized in 8x8 formation to form a block. This feature vector

containing the image descriptions is the input into the feature selection and classification algorithm. A hybrid feature vector increases the dimensionality of image features. We therefore extracted a 3,732-component feature vector by using the HU, Haralick, HOG method and a 256-component feature vector using the CNN method. Thus, we used the subspace method to reduce the dimensionality of the hybrid feature vector using PCA to classify and estimate uncertainty in classification.

DropWeights followed by a sigmoid activated layer was then applied to the network in the fully connected layer as an approximation to the Gaussian Process (GP), to cast it as approximate Bayesian inference for meaningful estimation of model uncertainty.

In HBNet, overfitting was reduced by using DropWeights with a rate of 0.3, which means that during both training and inference, approximately one-third of all weights were turned off and set to 0. To train the model in our study, we used an Adam optimizer with the default learning rate of 0.001. The training process was conducted in 250 epochs, with mini-batch size 32. We monitored the validation accuracy after every epoch and saved the model with the best accuracy on the validation dataset. During test time, DropWeights were active and Monte Carlo (MC) sampling was performed by feeding the input image with 1000 MC samples through the HBNet. This in turn, allowed us to apply variational DropWeights during testing [40]. For every tested image, the model provided not only its predicted class but also a measure of uncertainty estimated using variational DropWeights (see GTL confidence below).

The cell type labels in multi-label datasets may be correlated and a prediction for a cell type is not mutually exclusive. Therefore, we utilized label correlation information during classification. For the cost function for multi-label classification, we selected the sigmoid function with the addition of binary cross-entropy. A grid search scheme was adopted based on Matthews Correlation Coefficients (MCC) to determine the optimal thresholds for each dimension on the model outcome, which improves the accuracy of the model.

Approximate Bayesian Neural Networks (BNN)

Bayesian Neural Networks (BNN) provide a natural framework for modeling uncertainty. BNN methods are however intractable in computing the posterior of a network's parameters. The most common approach to estimate uncertainty in deep learning places distributions over each of the network's weight parameters. There are many methods proposed for quantifying uncertainty or confidence estimates approximated by MC dropout, including Laplace approximation, Markov chain MC (MCMC) methods, stochastic gradient MCMC variants such as Langevin Dynamics, Hamiltonian methods including Multiplicative Normalising Flows, Stochastic Batch Normalization, Maximum Softmax Probability, Heteroscedastic Classifier, and Learned Confidence Estimates including Deep Ensembles [41].

Given a dataset $X = \{x_1, x_2 \ldots . x_N\}$ and the corresponding labels $Y = \{y_1, y_2 \ldots y_N\}$ where $X \in R^d$ be a d-dimensioned input vector and $Y \in \{1 \ldots \ldots C\}$ with $y_i \in \{1 \ldots \ldots C\}$, C class label, a set of independent and identically distributed (i.i.d.) training samples size $N \{x_i, y_i\}$ for $i = 1 \; to \; N$, the task is to find a function $f: X \to Y$ using weights of neural net parameters w as close as possible to the original function that has generated the outputs Y. The principled predictive distribution of an unknown label $\hat{y}$ of a test input data $\hat{x}$ by marginalizing the parameters:

$$P(\hat{y}|\hat{x}, X, Y) = \int_w P(\hat{y}|\hat{x}, w) P(w|X, Y) df$$

The expectation of $\hat{y}$ is called the predictive mean of the model, and its variance is called the predictive uncertainty.

Unfortunately, finding the posterior distribution $P(w|X, Y)$ is often computationally intractable. Recently, Gal [41] proved that a gradient-based optimization procedure on the dropout neural network is equivalent to a specific variational approximation on an HBNet. Following Gal [41], Ghoshal et al [42] also showed similar results for neural networks with MC DropWeights (MCDW). The model uncertainty was approximated by averaging stochastic feed forward MC sampling during inference. During test time, the unseen samples were passed through the network before the Softmax predictions were analyzed. Practically, the expectation of is called the predictive mean of the model. The predictive mean μ_{pred} over the MC iterations is then used as the final prediction on the test sample: where $\mu_{pred} = \frac{1}{T} \sum_{i=1}^{T} P(\hat{y}|\hat{x}, w)$. For each test sample $\hat{x}$, the class with the largest predictive mean μ_{pred} is selected as the predictive probabilities.

Ghoshal-Tucker-Lindskog (GTL) Confidence Score

Based on the input sample, a network can be certain with high or low confidence of its decision, indicated by the predictive posterior distribution. Traditionally, it has been difficult to implement model validation under epistemic uncertainty. Thus, we predicted that epistemic uncertainty could inform model uncertainty. One of the measures of model uncertainty is predictive entropy H of the predictive distribution:

$$H(\hat{y}|\hat{x},X,Y) = \sum_{c=1}^{C} P(\hat{y}=c|\hat{x},X,Y)\ log\ P(\hat{y}=c|\hat{x},X,Y)$$

where C ranges over all class labels. In general, the range of the obtained uncertainty values is dependent on e.g. the dataset, network architectures and the number of MC samples. Therefore, we normalized the estimated uncertainty to report our results and facilitate comparison across various sets and configurations. Estimation of entropy from the finite set of data suffers from a severe downward bias when the data is under-sampled. Even small biases can result in significant inaccuracies when estimating entropy. We leveraged the plug-in estimate of entropy and the Jackknife resampling method to calculate bias-reduced entropy [40, 43-45]. The entropy was based on maximizing mutual information between the model posterior density function and the prediction density function, approximated as the difference between the entropy of the predictive distribution and the mean entropy of predictions across samples. Test points that maximize mutual information are points over which the model is uncertain on average, but there are model parameters that produce erroneous predictions with high confidence. This is equivalent to points with high variance in the input to the sigmoid layer (the logits). Thus, each stochastic forward pass through the model would have the highest probability assigned to a different class.

Each prediction from our trained model returned a set of labels. We calculated the GTL Score for each label. We employed the maximum class predictive probability distance (CPPD), which is the difference between the probability values of the highest and the second highest predictive probability value as a measure of a representativeness heuristic. The vector of class probabilities $\hat{y}_t = f^{\hat{w}_t}(\hat{x})$ obtained after the t the stochastic forward pass is denoted $(\hat{y}_t|\hat{x},\hat{w}_t)$, where $\hat{w}_t$ denotes the sampled parameters resulting from DropWeights. Thus, the class probabilities of estimates are given by $\frac{1}{T}\sum_{i=1}^{T} P(\hat{y}|\hat{x},\hat{w}_t)$. We obtain the Class Predictive Probability Distance (CPPD):

$$CPPD\ (x_i) = argmin\ (\tfrac{1}{T}\sum_{i=1}^{T} P(\hat{y}_{Best}|\hat{x},\hat{w}_t) - \tfrac{1}{T}\sum_{i=1}^{T} P(\hat{y}_{NextBest}|\hat{x},\hat{w}_t)).$$

The MCDW estimate of the vector of class probabilities aimed to decompose the source of uncertainty. The main idea was to select samples that were not only highly uncertain but also highly representative. Based on this strategy, we defined the GTL Score as an approximation of semi-automated sample selection as below:

$GTL = \frac{CPPD(x_i)}{\hat{H}_J}$, where $\hat{H}_J$ is bias-corrected entropy using the Jackknife method.

We ranked all unlabelled samples in ascending order of GTL Score. The formulation for the sample selection measure can be given as $x_{GTL} = argsort\ \{GTL_x\}\ [:sample\ size]$. The higher the GTL Score, the higher the information content of the corresponding sample images, which should represent certainty in predictions. The GTL Score was used along with the predictive probabilities, to identify and discard images for which specific cell types did not express a particular protein, as well as images that expressed the protein with high confidence.

Multi-Label Cross Validation

A Multi-label Stratified Shuffle Split cross-validation merge of Multi-label Stratified KFold and Shuffle Split [46] were used for returning stratified, randomized folds for multi-label data. The folds were made by preserving the percentage of samples for each label repeated ten times in the process of 10-fold cross-validation, with different randomization in each repetition.

Declarations

Data Availability

JPEG files of all 7,848 images used in the present investigation are available on v19.proteinatlas.org. The manually annotated protein expression in eight different cell types will be available in the upcoming version 20 of the HPA, released in October 2020. Manual errors identified as part of this study have been corrected, which means that some of the presented protein expression data on the HPA will differ from the input data used for model training.

Upon acceptance of the paper, all codes will be made available in GitHub. The code will also be available to editors and referees upon request.

Acknowledgements

The project was funded by the Knut and Alice Wallenberg Foundation. Pathologists and staff at the Department of Clinical Pathology, Uppsala University Hospital, are acknowledged for providing the tissues used for immunohistochemistry. The authors would also like to thank all staff of the Human Protein Atlas for their work.

Author Contributions Statement

C.L. conceived the study. C.L. and A.T. supervised the experiments. B.G. conducted the machine learning and data analysis. F.H. and C.L. generated the immunohistochemistry data. All authors analyzed the results and reviewed the manuscript.

Competing Interests

The authors declare no competing interests.

References

1. Uhlen, M.; Fagerberg, L.; Hallstrom, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, A.; Kampf, C.; Sjostedt, E.; Asplund, A.; Olsson, I.; Edlund, K.; Lundberg, E.; Navani, S.; Szigyarto, C. A.; Odeberg, J.; Djureinovic, D.; Takanen, J. O.; Hober, S.; Alm, T.; Edqvist, P. H.; Berling, H.; Tegel, H.; Mulder, J.; Rockberg, J.; Nilsson, P.; Schwenk, J. M.; Hamsten, M.; von Feilitzen, K.; Forsberg, M.; Persson, L.; Johansson, F.; Zwahlen, M.; von Heijne, G.; Nielsen, J.; Ponten, F., Proteomics. Tissue-based map of the human proteome. *Science* **2015,** *347* (6220), 1260419.
2. Uhlen, M.; Zhang, C.; Lee, S.; Sjostedt, E.; Fagerberg, L.; Bidkhori, G.; Benfeitas, R.; Arif, M.; Liu, Z.; Edfors, F.; Sanli, K.; von Feilitzen, K.; Oksvold, P.; Lundberg, E.; Hober, S.; Nilsson, P.; Mattsson, J.; Schwenk, J. M.; Brunnstrom, H.; Glimelius, B.; Sjoblom, T.; Edqvist, P. H.; Djureinovic, D.; Micke, P.; Lindskog, C.; Mardinoglu, A.; Ponten, F., A pathology atlas of the human cancer transcriptome. *Science* **2017,** *357* (6352).
3. Uhlen, M.; Karlsson, M. J.; Zhong, W.; Tebani, A.; Pou, C.; Mikes, J.; Lakshmikanth, T.; Forsstrom, B.; Edfors, F.; Odeberg, J.; Mardinoglu, A.; Zhang, C.; von Feilitzen, K.; Mulder, J.; Sjostedt, E.; Hober, A.; Oksvold, P.; Zwahlen, M.; Ponten, F.; Lindskog, C.; Sivertsson, A.; Fagerberg, L.; Brodin, P., A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. *Science* **2019,** *366* (6472).
4. Sjostedt, E.; Zhong, W.; Fagerberg, L.; Karlsson, M.; Mitsios, N.; Adori, C.; Oksvold, P.; Edfors, F.; Limiszewska, A.; Hikmet, F.; Huang, J.; Du, Y.; Lin, L.; Dong, Z.; Yang, L.; Liu, X.; Jiang, H.; Xu, X.; Wang, J.; Yang, H.; Bolund, L.; Mardinoglu, A.; Zhang, C.; von Feilitzen, K.; Lindskog, C.; Ponten, F.; Luo, Y.; Hokfelt, T.; Uhlen, M.; Mulder, J., An atlas of the protein-coding genes in the human, pig, and mouse brain. *Science* **2020,** *367* (6482).
5. Thul, P. J.; Akesson, L.; Wiking, M.; Mahdessian, D.; Geladaki, A.; Ait Blal, H.; Alm, T.; Asplund, A.; Bjork, L.; Breckels, L. M.; Backstrom, A.; Danielsson, F.; Fagerberg, L.; Fall, J.; Gatto, L.; Gnann, C.; Hober, S.; Hjelmare, M.; Johansson, F.; Lee, S.; Lindskog, C.; Mulder, J.; Mulvey, C. M.; Nilsson, P.; Oksvold, P.; Rockberg, J.; Schutten, R.; Schwenk, J. M.; Sivertsson, A.; Sjostedt, E.; Skogs, M.; Stadler, C.; Sullivan, D. P.; Tegel, H.; Winsnes, C.; Zhang, C.; Zwahlen, M.; Mardinoglu, A.; Ponten, F.; von Feilitzen, K.; Lilley, K. S.; Uhlen, M.; Lundberg, E., A subcellular map of the human proteome. *Science* **2017,** *356* (6340).

6. Nagpal, K.; Foote, D.; Liu, Y.; Chen, P. C.; Wulczyn, E.; Tan, F.; Olson, N.; Smith, J. L.; Mohtashamian, A.; Wren, J. H.; Corrado, G. S.; MacDonald, R.; Peng, L. H.; Amin, M. B.; Evans, A. J.; Sangoi, A. R.; Mermel, C. H.; Hipp, J. D.; Stumpe, M. C., Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit Med* **2019,** *2*, 48.
7. Ehteshami Bejnordi, B.; Veta, M.; Johannes van Diest, P.; van Ginneken, B.; Karssemeijer, N.; Litjens, G.; van der Laak, J.; the, C. C.; Hermsen, M.; Manson, Q. F.; Balkenhol, M.; Geessink, O.; Stathonikos, N.; van Dijk, M. C.; Bult, P.; Beca, F.; Beck, A. H.; Wang, D.; Khosla, A.; Gargeya, R.; Irshad, H.; Zhong, A.; Dou, Q.; Li, Q.; Chen, H.; Lin, H. J.; Heng, P. A.; Hass, C.; Bruni, E.; Wong, Q.; Halici, U.; Oner, M. U.; Cetin-Atalay, R.; Berseth, M.; Khvatkov, V.; Vylegzhanin, A.; Kraus, O.; Shaban, M.; Rajpoot, N.; Awan, R.; Sirinukunwattana, K.; Qaiser, T.; Tsang, Y. W.; Tellez, D.; Annuscheit, J.; Hufnagl, P.; Valkonen, M.; Kartasalo, K.; Latonen, L.; Ruusuvuori, P.; Liimatainen, K.; Albarqouni, S.; Mungal, B.; George, A.; Demirci, S.; Navab, N.; Watanabe, S.; Seno, S.; Takenaka, Y.; Matsuda, H.; Ahmady Phoulady, H.; Kovalev, V.; Kalinovsky, A.; Liauchuk, V.; Bueno, G.; Fernandez-Carrobles, M. M.; Serrano, I.; Deniz, O.; Racoceanu, D.; Venancio, R., Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA* **2017,** *318* (22), 2199-2210.
8. Esteva, A.; Kuprel, B.; Novoa, R. A.; Ko, J.; Swetter, S. M.; Blau, H. M.; Thrun, S., Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017,** *542* (7639), 115-118.
9. Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M. C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; Kim, R.; Raman, R.; Nelson, P. C.; Mega, J. L.; Webster, D. R., Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **2016,** *316* (22), 2402-2410.
10. Jackson, C. R.; Sriharan, A.; Vaickus, L. J., A machine learning algorithm for simulating immunohistochemistry: development of SOX10 virtual IHC and evaluation on primarily melanocytic neoplasms. *Mod Pathol* **2020**.
11. Long, W.; Yang, Y.; Shen, H. B., ImPLoc: a multi-instance deep learning model for the prediction of protein subcellular localization based on immunohistochemistry images. *Bioinformatics* **2020,** *36* (7), 2244-2250.
12. Raczkowski, L.; Mozejko, M.; Zambonelli, J.; Szczurek, E., ARA: accurate, reliable and active histopathological image classification framework with Bayesian deep learning. *Sci Rep* **2019,** *9* (1), 14347.
13. Gonzalez-Lopez, J.; Ventura, S.; Cano, A., Distributed Selection of Continuous Features in Multilabel Classification Using Mutual Information. *IEEE Trans Neural Netw Learn Syst* **2020,** *31* (7), 2280-2293.
14. Djureinovic, D.; Fagerberg, L.; Hallstrom, B.; Danielsson, A.; Lindskog, C.; Uhlen, M.; Ponten, F., The human testis-specific proteome defined by transcriptomics and antibody-based profiling. *Mol Hum Reprod* **2014,** *20* (6), 476-88.
15. Fagerberg, L.; Hallstrom, B. M.; Oksvold, P.; Kampf, C.; Djureinovic, D.; Odeberg, J.; Habuka, M.; Tahmasebpoor, S.; Danielsson, A.; Edlund, K.; Asplund, A.; Sjostedt, E.; Lundberg, E.; Szigyarto, C. A.; Skogs, M.; Takanen, J. O.; Berling, H.; Tegel, H.; Mulder, J.; Nilsson, P.; Schwenk, J. M.; Lindskog, C.; Danielsson, F.; Mardinoglu, A.; Sivertsson, A.; von Feilitzen, K.; Forsberg, M.; Zwahlen, M.; Olsson, I.; Navani, S.; Huss, M.; Nielsen, J.; Ponten, F.; Uhlen, M., Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics* **2014,** *13* (2), 397-406.
16. Jumeau, F.; Com, E.; Lane, L.; Duek, P.; Lagarrigue, M.; Lavigne, R.; Guillot, L.; Rondel, K.; Gateau, A.; Melaine, N.; Guevel, B.; Sergeant, N.; Mitchell, V.; Pineau, C., Human Spermatozoa as a Model for Detecting Missing Proteins in the Context of the Chromosome-Centric Human Proteome Project. *J Proteome Res* **2015,** *14* (9), 3606-20.
17. Vandenbrouck, Y.; Lane, L.; Carapito, C.; Duek, P.; Rondel, K.; Bruley, C.; Macron, C.; Gonzalez de Peredo, A.; Coute, Y.; Chaoui, K.; Com, E.; Gateau, A.; Hesse, A. M.; Marcellin, M.; Mear, L.; Mouton-Barbosa, E.; Robin, T.; Burlet-Schiltz, O.; Cianferani, S.; Ferro, M.; Freour, T.; Lindskog, C.; Garin, J.; Pineau, C., Looking for Missing Proteins in the Proteome of Human Spermatozoa: An Update. *J Proteome Res* **2016,** *15* (11), 3998-4019.
18. Pineau, C.; Hikmet, F.; Zhang, C.; Oksvold, P.; Chen, S.; Fagerberg, L.; Uhlen, M.; Lindskog, C., Cell Type-Specific Expression of Testis Elevated Genes Based on Transcriptomics and Antibody-Based Proteomics. *J Proteome Res* **2019,** *18* (12), 4215-4230.
19. Cheng, Q.; Zhang, Q.; Fu, P.; Tu, C.; Li, S., A survey and analysis on automatic image annotation. *Pattern Recognition* **2018,** *79*, 242-259.
20. Bhagat, P.; Choudhary, P., Image annotation: Then and now. *Image and Vision Computing* **2018,** *80*, 1-23.
21. Wu, X.-Z.; Zhou, Z.-H. In *A unified view of multi-label performance measures*, International Conference on Machine Learning, 2017; pp 3780-3788.
22. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. In *Learning deep features for discriminative localization*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016; pp 2921-2929.

23. Coons, A. H., Creech, H.J., Jones, R.N., Immunological properties of an antibody containing a fluorescent group. *Proc. Soc. Exp. Biol. Med.* **1941,** *47*, 200-202.
24. O'Hurley, G.; Sjostedt, E.; Rahman, A.; Li, B.; Kampf, C.; Ponten, F.; Gallagher, W. M.; Lindskog, C., Garbage in, garbage out: a critical evaluation of strategies used for validation of immunohistochemical biomarkers. *Mol Oncol* **2014,** *8* (4), 783-98.
25. Edfors, F.; Hober, A.; Linderback, K.; Maddalo, G.; Azimi, A.; Sivertsson, A.; Tegel, H.; Hober, S.; Szigyarto, C. A.; Fagerberg, L.; von Feilitzen, K.; Oksvold, P.; Lindskog, C.; Forsstrom, B.; Uhlen, M., Enhanced validation of antibodies for research applications. *Nat Commun* **2018,** *9* (1), 4130.
26. Ouyang, W.; Winsnes, C. F.; Hjelmare, M.; Cesnik, A. J.; Akesson, L.; Xu, H.; Sullivan, D. P.; Dai, S.; Lan, J.; Jinmo, P.; Galib, S. M.; Henkel, C.; Hwang, K.; Poplavskiy, D.; Tunguz, B.; Wolfinger, R. D.; Gu, Y.; Li, C.; Xie, J.; Buslov, D.; Fironov, S.; Kiselev, A.; Panchenko, D.; Cao, X.; Wei, R.; Wu, Y.; Zhu, X.; Tseng, K. L.; Gao, Z.; Ju, C.; Yi, X.; Zheng, H.; Kappel, C.; Lundberg, E., Analysis of the Human Protein Atlas Image Classification competition. *Nat Methods* **2019,** *16* (12), 1254-1261.
27. Sullivan, D. P.; Winsnes, C. F.; Akesson, L.; Hjelmare, M.; Wiking, M.; Schutten, R.; Campbell, L.; Leifsson, H.; Rhodes, S.; Nordgren, A.; Smith, K.; Revaz, B.; Finnbogason, B.; Szantner, A.; Lundberg, E., Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nat Biotechnol* **2018,** *36* (9), 820-828.
28. Kumar, A.; Rao, A.; Bhavani, S.; Newberg, J. Y.; Murphy, R. F., Automated analysis of immunohistochemistry images identifies candidate location biomarkers for cancers. *Proc Natl Acad Sci U S A* **2014,** *111* (51), 18249-54.
29. Xu, Y. Y.; Yang, F.; Zhang, Y.; Shen, H. B., An image-based multi-label human protein subcellular localization predictor (iLocator) reveals protein mislocalizations in cancer tissues. *Bioinformatics* **2013,** *29* (16), 2032-40.
30. Newberg, J.; Murphy, R. F., A framework for the automated analysis of subcellular patterns in human protein atlas images. *J Proteome Res* **2008,** *7* (6), 2300-8.
31. Goodman, M.; Ward, K. C.; Osunkoya, A. O.; Datta, M. W.; Luthringer, D.; Young, A. N.; Marks, K.; Cohen, V.; Kennedy, J. C.; Haber, M. J.; Amin, M. B., Frequency and determinants of disagreement and error in gleason scores: a population-based study of prostate cancer. *Prostate* **2012,** *72* (13), 1389-98.
32. Regev, A.; Teichmann, S. A.; Lander, E. S.; Amit, I.; Benoist, C.; Birney, E.; Bodenmiller, B.; Campbell, P.; Carninci, P.; Clatworthy, M.; Clevers, H.; Deplancke, B.; Dunham, I.; Eberwine, J.; Eils, R.; Enard, W.; Farmer, A.; Fugger, L.; Gottgens, B.; Hacohen, N.; Haniffa, M.; Hemberg, M.; Kim, S.; Klenerman, P.; Kriegstein, A.; Lein, E.; Linnarsson, S.; Lundberg, E.; Lundeberg, J.; Majumder, P.; Marioni, J. C.; Merad, M.; Mhlanga, M.; Nawijn, M.; Netea, M.; Nolan, G.; Pe'er, D.; Phillipakis, A.; Ponting, C. P.; Quake, S.; Reik, W.; Rozenblatt-Rosen, O.; Sanes, J.; Satija, R.; Schumacher, T. N.; Shalek, A.; Shapiro, E.; Sharma, P.; Shin, J. W.; Stegle, O.; Stratton, M.; Stubbington, M. J. T.; Theis, F. J.; Uhlen, M.; van Oudenaarden, A.; Wagner, A.; Watt, F.; Weissman, J.; Wold, B.; Xavier, R.; Yosef, N.; Human Cell Atlas Meeting, P., The Human Cell Atlas. *Elife* **2017,** *6*.
33. Kampf, C.; Olsson, I.; Ryberg, U.; Sjostedt, E.; Ponten, F., Production of tissue microarrays, immunohistochemistry staining and digitalization within the human protein atlas. *J Vis Exp* **2012,** (63).
34. LeCun, Y.; Bengio, Y.; Hinton, G., Deep learning. *Nature* **2015,** *521* (7553), 436-44.
35. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R., Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* **2014,** *15* (1), 1929-1958.
36. Dalal, N.; Triggs, B. In *Histograms of oriented gradients for human detection*, 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), IEEE: 2005; pp 886-893.
37. Simonyan, K.; Zisserman, A., Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.
38. Haralick, R. M., Statistical and structural approaches to texture. *Proceedings of the IEEE* **1979,** *67* (5), 786-804.
39. Hu, M.-K., Visual pattern recognition by moment invariants. *IRE transactions on information theory* **1962,** *8* (2), 179-187.
40. Ghoshal, B.; Lindskog, C.; Tucker, A. In *Estimating Uncertainty in Deep Learning for Reporting Confidence: An Application on Cell Type Prediction in Testes Based on Proteomics*, International Symposium on Intelligent Data Analysis, Springer: 2020; pp 223-234.
41. Gal, Y., Uncertainty in deep learning. *University of Cambridge* **2016,** *1* (3).
42. Ghoshal, B.; Tucker, A.; Sanghera, B.; Wong, W. L. In *Estimating uncertainty in deep learning for reporting confidence to clinicians when segmenting nuclei image data*, 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), IEEE: 2019; pp 318-324.
43. Shannon, C. E., A mathematical theory of communication. *The Bell system technical journal* **1948,** *27* (3), 379-423.

44. Quenouille, M. H., Notes on bias in estimation. *Biometrika* **1956,** *43* (3/4), 353-360.
45. Yeung, R. W., A new outlook on Shannon's information measures. *IEEE transactions on information theory* **1991,** *37* (3), 466-474.
46. Sechidis, K.; Tsoumakas, G.; Vlahavas, I. In *On the stratification of multi-label data*, Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer: 2011; pp 145-158.

Tables

Table 1. Overall model performance. Evaluation of classification performance for a standard deep neural network (DNN) and the proposed hybrid Bayesian neural network (HBNet), based on five different metrics. The results for each metric are shown as a percent.

Metrics	DNN (%)	HBNet (%)
Hamming Loss	17.0	13.0
Macro F1 Score	81.0	84.0
Micro F1 Score	80.0	84.0
Exact Match ratio	48.0	67.0
mean-Average Precision (mAP)	71.0	76.0

Table 2. Model performance on a cell type-specific level. The % accuracy for predicting the labels for each cell type is shown for standard deep neural network (DNN), hybrid Bayesian neural network (HBNet), and GTL-thresholded HBNet (HBNet - GTL) along with the percentage of discarded images based on low GTL confidence.

	Model Performance Accuracy (%)			Discard Tradeoff
Cell Types (GTL Threshold)	DNN	HBNet	HBNet- GTL	HBNet - GTL Percentage Discarded
Spermatogonia (0.11)	84.9	85.7	99.4	37.2%
Preleptotene spermatocytes (0.11)	80.7	84.9	99.2	37.2%
Pachytene spermatocytes (0.22)	75.6	82.6	99.2	31.7%
Round/early spermatids (0.78)	69.3	80.5	83.5	39.1%
Elongated/late spermatids (0.22)	80.9	85.2	98.7	30.1%
Sertoli cells (1.00E-10)	91.6	92.2	92.2	0.0%
Leydig cells (0.11)	84.5	85.7	99.2	38.3%
Peritubular cells (1.00E-10)	98.1	98.6	98.7	1.3%

Table 3. Model performance based on subcellular localization.

Cell type	#GTL-thresholded labels / #actual labels with subcellular localization)	HBNet - GTL % accuracy (#labels)						
		Cyt	Cyt, Mem	Mem	Nucl	Nucl, Cyt	Nucl, Cyt, Mem	Nucl, Mem
Spermatogonia	518/521	99.5 (204/205)	97.6 (41/42)	100.0 (5/5)	99.5 (201/202)	100.0 (40/40)	100.0 (25/25)	100.0(2/2)
Preleptotene spermatocytes	357/360	100.0 (121/121)	100.0 (30/30)	100.0(2/2)	98.17 (161/164)	100.0(15/15)	100.0(28/28)	0
Pachytene spermatocytes	388/391	99.3 (145/146)	100.0 (66/66)	100.0 (4/4)	98.5 (135/137)	100.0 (9/9)	100.0 (29/29)	0
Round/early spermatids	361/362	99.2 (131/132)	100.0 (57/57)	0	100.0 (147/147)	100.0 (10/10)	100.0 (16/16)	0
Elongated/late spermatids	405/409	98.2 (215/219)	100.0 (83/83)	0	100.0 (67/67)	100.0 (22/22)	100.0 (18/18)	0
Sertoli cells	225/231	97.8 (87/89)	100.0 (31/31)	100.0 (6/6)	95.2 (79/83)	100.0 (1/1)	100.0 (11/11)	100.0 (10/10)
Leydig cells	466/470	98.9 (277/280)	100.0 (71/71)	100.0 (5/5)	100.0 (81/81)	100.0 (25/25)	100.0 (7/7)	0
Peritubular cells	105/120	89.3 (50/56)	100.0 (9/9)	83.7 (46/55)	0	0	0	0
Average all cell types	*2825/2864*	*97.78*	*99.7*	*97.28*	*98.77*	*100*	*100*	*100*

Table 4. Model performance based on staining intensity.

Cell type	HBNet - GTL % accuracy (#GTL-thresholded labels / #actual labels)		
	Only weak labels (intensity =1)	Only moderate labels (intensity = 2)	Only strong labels (intensity = 3)
Spermatogonia	100.0 (27/27)	99.3 (142/143)	99.4 (349/351)
Preleptotene spermatocytes	100.0 (49/49)	100.0 (150/150)	98.14 (158/161)
Pachytene spermatocytes	100.0 (70/70)	99.3 (141/142)	98.9 (177/179)
Round/early spermatids	100.0 (53/53	100.0 (102/102)	99.6 (206/207)
Elongated/late Spermatids	100.0 (41/41)	97.3 (145/149)	100.0 (219/219)
Sertoli cells	98.6 (72/73)	86.1 (31/36)	100.0 (122/122)
Leydig cells	98.9 (172/174)	99.5 (202/203)	100.0 (92/92)
Peritubular cells	100.0 (17/17)	84.5 (49/58)	86.7 (39/45)
Average all cell types	*99.7*	*95.8*	*97.9*

Figures

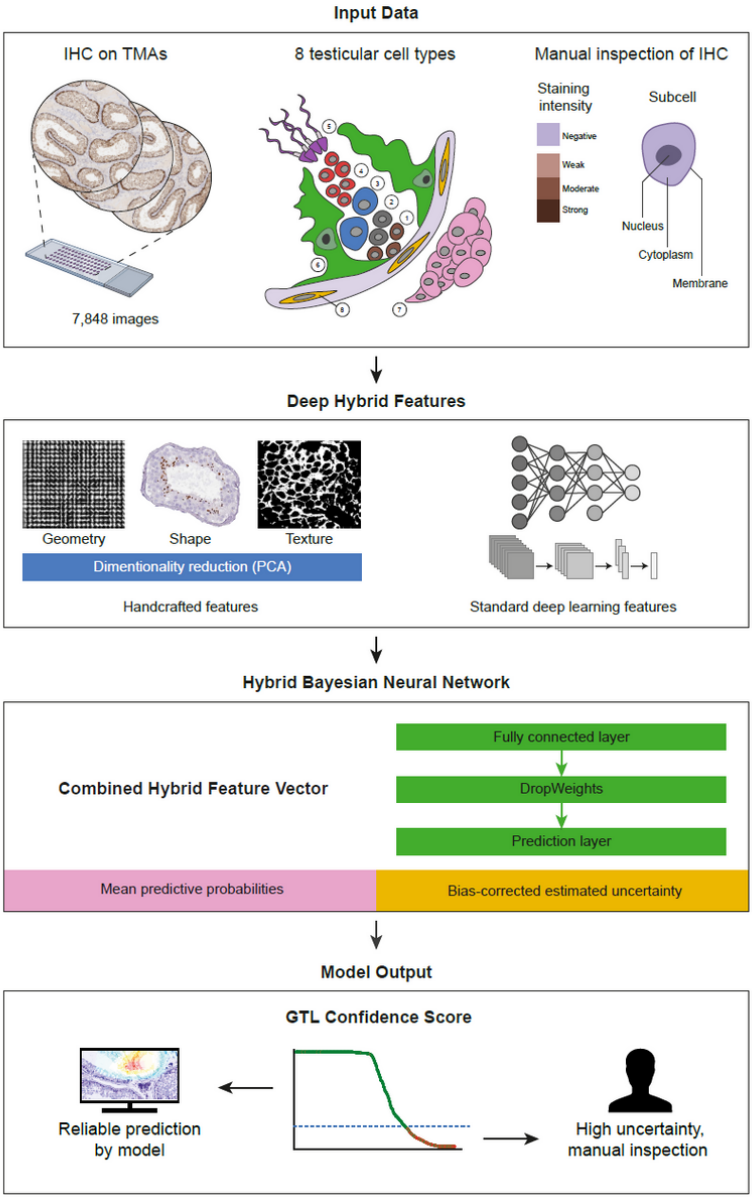

Figure 1

Overview of the image annotation framework. A Hybrid Bayesian Neural Network (HBNet) model was trained taking into consideration both handcrafted features and deep learning features. The input IHC high-resolution images consisted of 1-3 human testis TMA punch-outs for each antibody comprising a total of 7,848 images. For each antibody, eight different cell types were manually inspected with regards to staining intensity (negative, weak, moderate, strong) and subcellular location (cytoplasm, nucleus, membrane); 1: Spermatogonia; 2: Preleptotene spermatocytes; 3: Pachytene spermatocytes; 4: Round/early spermatids; 5: Elongated/late spermatids; 6: Sertoli cells; 7: Leydig cells; 8: Peritubular cells. The manual data was used as a basis for machine learning, combining handcrafted features with standard deep learning features. The mean predictive probability and bias-corrected estimated uncertainty were used for generation of Ghoshal-Tucker-Lindskog (GTL) confidence score, which allowed for dividing the images into those that were reliably predicted by the model, and those of high uncertainty that need manual inspection.

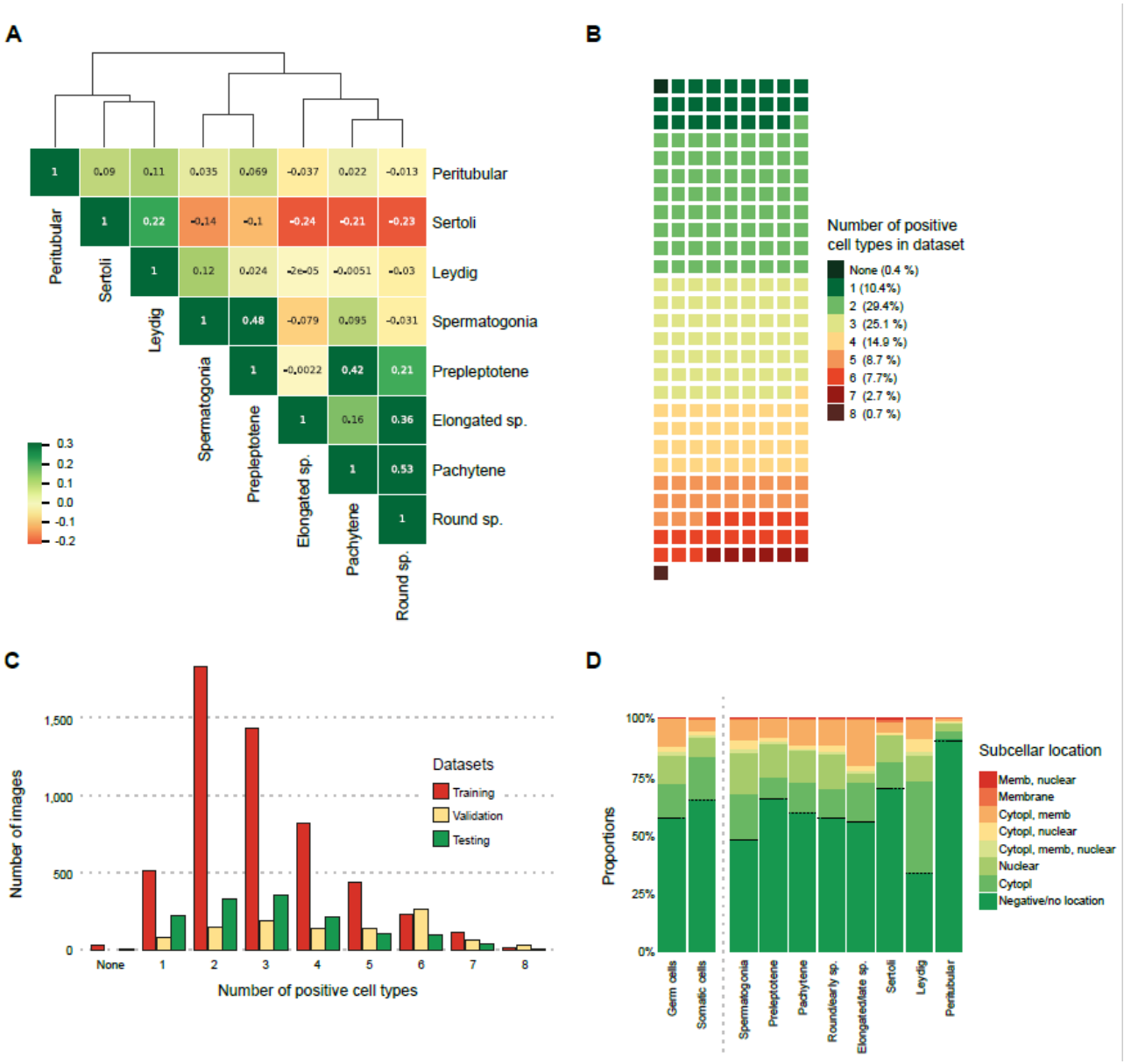

Figure 2

Input image data distribution based on manual annotation. (A) Heatmap and cluster analysis of testicular cell types. (B) All 7,848 images were grouped based on the number of positive cell types (or lack of positive cell types), and visualized as a waffle distribution plot, which shows that most images contain 2-5 positive cell types. In (C), the number of positive cell types is visualized separately by each dataset. The training set consisted of 5,411 images, validation set 1,063 images, and testing set 1,374 images. (D) The distribution of subcellular location (and lack of subcellular location due to no antibody staining) for each cell type in all 7,848 images showed that Leydig cells more often showed cytoplasmic staining, while Sertoli cells and peritubular cells had the highest proportion of images that were negative/lacked protein expression in these cell types.

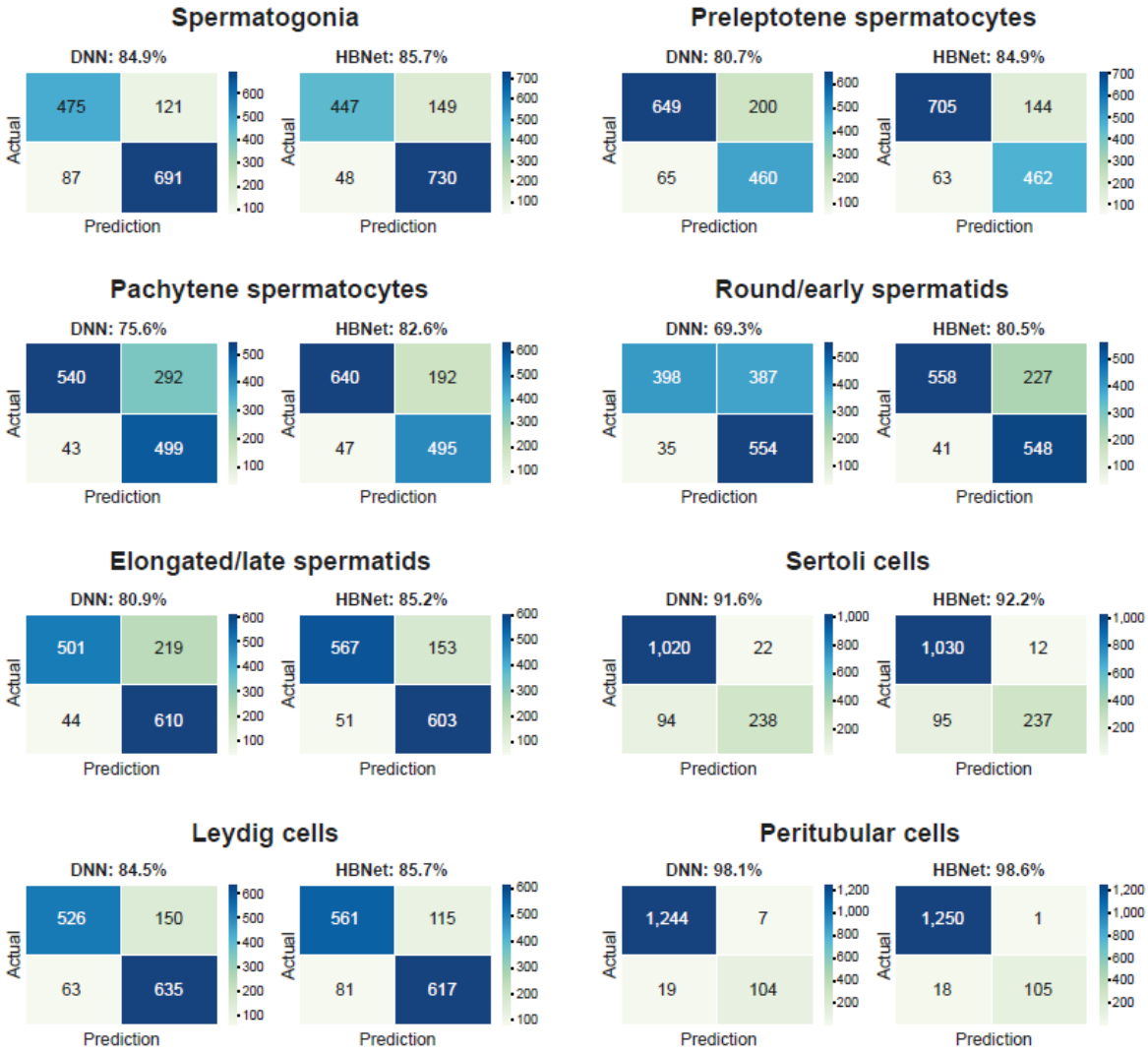

Figure 3

Confusion matrix for each of the eight testicular cell types based on standard deep neural network (DNN) and hybrid Bayesian neural network (HBNet). Each quadrant shows the number of images that were true negative (upper left), false negative (upper right), false positive (bottom left) and true positive (bottom right), color-coded based on the number of images.

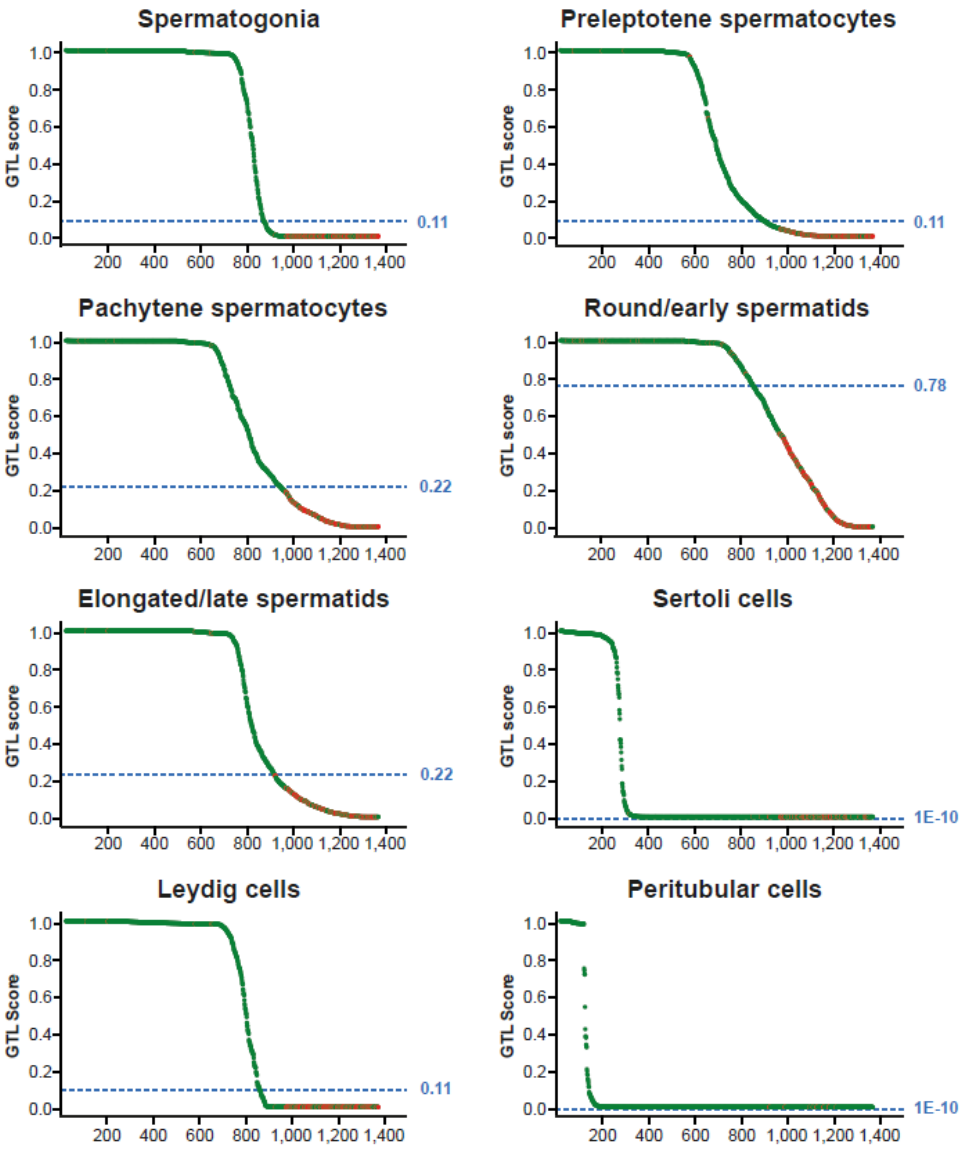

Figure 4

Confidence maps of all automated predictions for each of the eight cell types. Each dot corresponds to one prediction, with green = correct and red = incorrect. The predictions were sorted based on their GTL Score, showing the confidence in prediction. The blue lines depict the determined cut-off for each cell type where classification is considered too unreliable.

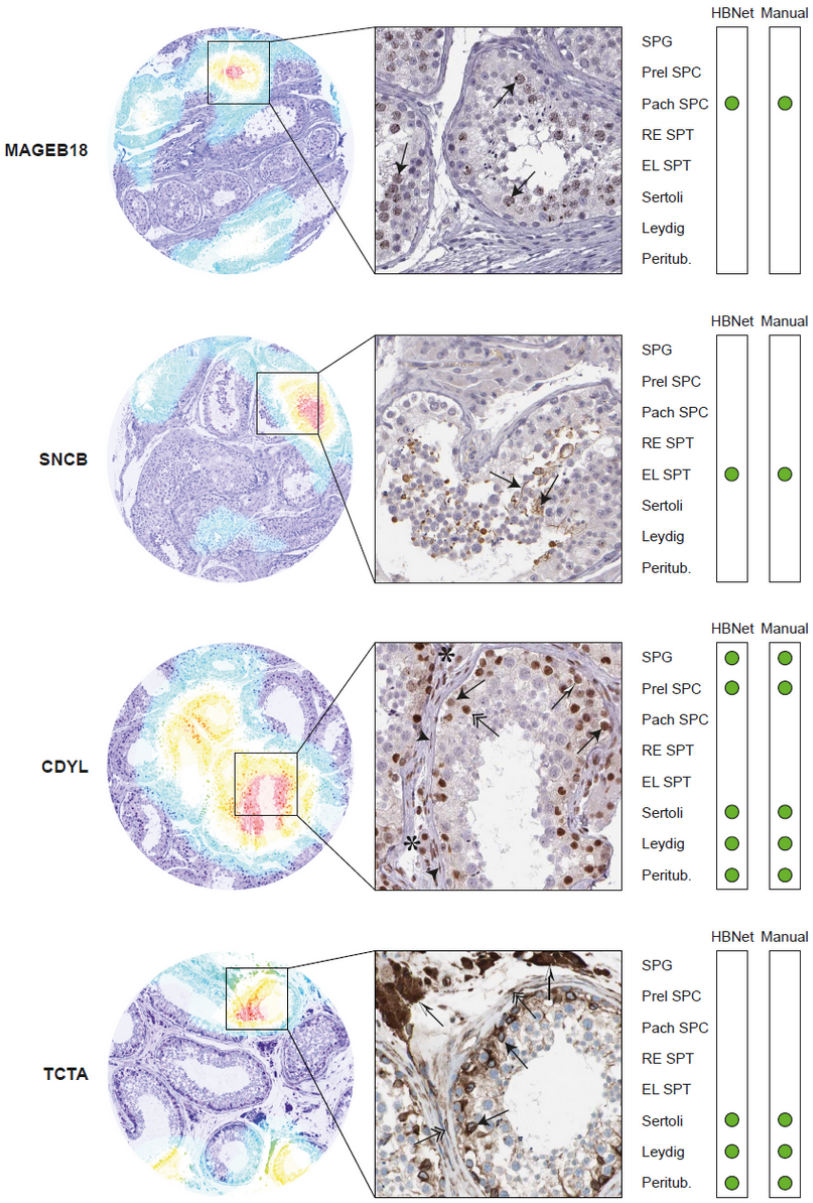

Figure 5

Examples of correctly classified images. Heatmaps (left), IHC staining patterns (middle), with an overview of HBNet prediction and manual annotation of the eight different cell types (right). The colors of the heatmaps indicate where the HBNet model focuses on making a labeling decision from purple (no activation) through blue, green, yellow, to red (high activation). IHC images show positive staining in brown (protein expressed) and counterstaining in blue (protein not expressed). Cell type names: Spermatogonia (SPG), preleptotene spermatocytes (Prel SPC), pachytene spermatocytes (Pach SPC), round/early spermatids (RE SPT), elongated/late spermatids (EL SPT), Sertoli cells (Sertoli), Leydig cells (Leydig) and peritubular cells (Peritub.). Green dots: Correct classification. Melanoma-associated antigen B18 (MAGEB18) and Synuclein beta (SNCB) showed selective expression in one cell type only, while the Chromodomain Y like protein (CDYL) and T cell leukemia translocation altered protein (TCTA) were expressed in several testicular cell types. MAGEB18 showed a speckled nuclear staining pattern in pachytene spermatocytes (arrows), with clearly visible nucleoli. SNCB was positive in elongated/late spermatids and sperm flagella (arrows), seen in the lumen of seminiferous ducts. CDYL displayed nuclear staining in spermatogonia (black arrows), preleptotene spermatocytes (white/black arrow), Sertoli cells (double-headed arrow), Leydig cells (asterisks) and peritubular cells (arrowheads). TCTA showed mainly cytoplasmic staining in Sertoli cells (arrows), Leydig cells (white/black arrows) and peritubular cells (double-headed arrows), in Sertoli cells accompanied with distinct positivity of nuclear membranes.

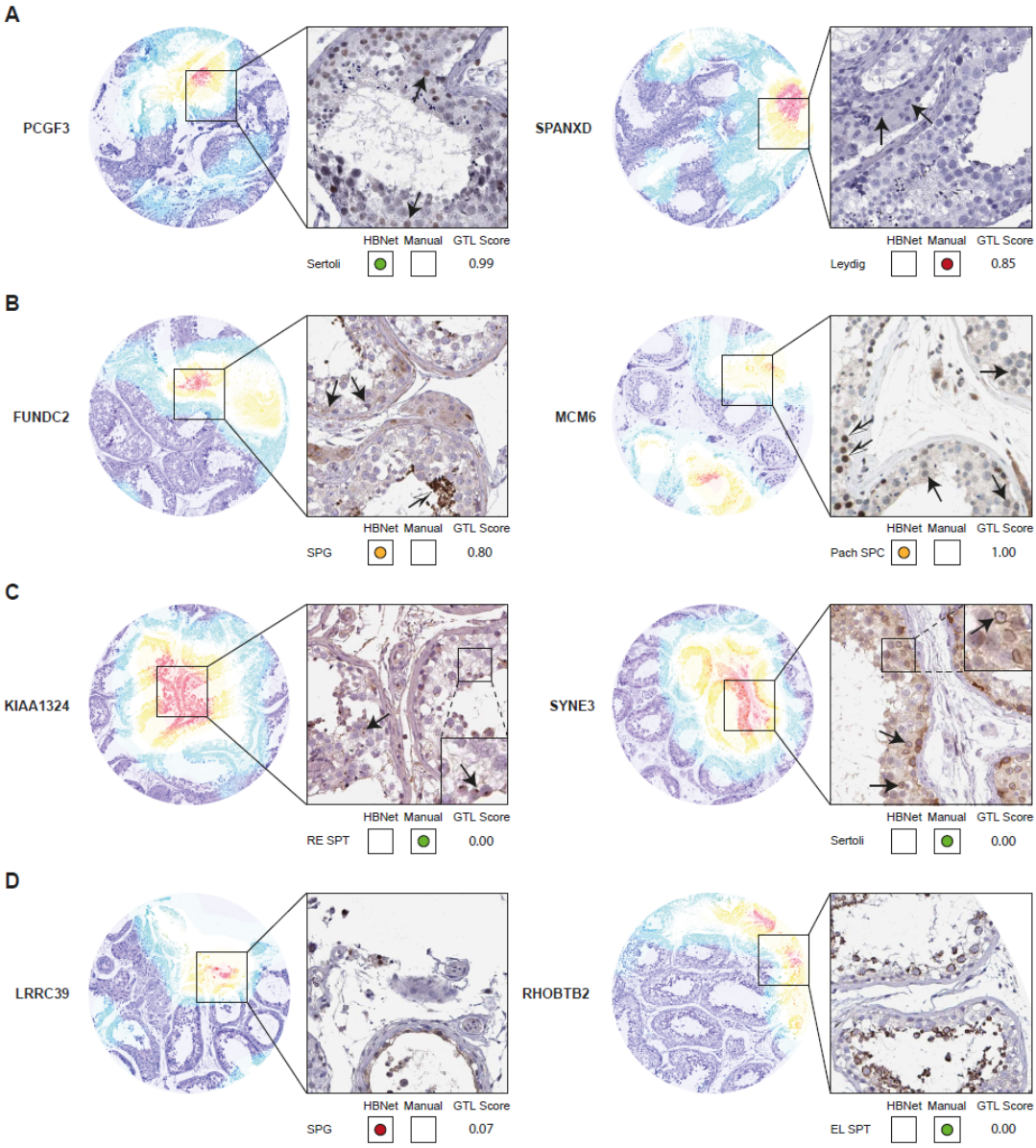

Figure 6

Examples of misclassified images. Heatmaps (left) and IHC staining patterns (right), exemplified by one cell type each where HBNet prediction and manual annotation disagreed. The colors of the heatmaps indicate where the HBNet model focuses on making a labeling decision from purple (no activation) through blue, green, yellow, to red (high activation). IHC images show positive staining in brown (protein expressed) and counterstaining in blue (protein not expressed). Cell type names: Spermatogonia (SPG), pachytene spermatocytes (Pach SPC), round/early spermatids (RE SPT), elongated/late spermatids (EL SPT), Sertoli cells (Sertoli) and Leydig cells (Leydig). Green dots: Correct classification. Orange dots: Correct classification, but can be considered incorrect based on human knowledge. Red dots: Incorrect classification. (A) Polycomb group ring finger 3 (PCGF3) and SPANX family member D SPANXD represent manual errors. For PCGF3, the manual observer missed Sertoli cells that showed clear nuclear staining (arrows), while for SPANDX, Leydig cells had been annotated as positive, despite being completely negative (arrows). (B) FUN14 domain containing 2 (FUNDC2) and Minichromosome maintenance complex component 6 MCM6 showed staining neglected by the human observer. FUNDC2 displayed weak cytoplasmic positivity in spermatogonia (arrows), but due to strong staining in elongated/late spermatids (white/black arrow), the spermatogonia staining was considered unspecific. Similarly, MCM6 showed weak nuclear staining in pachytene spermatocytes, and considered unspecific compared to the strongly positive preleptotene spermatocytes (white/black arrows). (C) The uncharacterized protein KIAA1324 and Spectrin repeat containing nuclear envelope family member 3 (SYNE3) were stained in small structures missed by the HBNet prediction. KIAA1324 showed positivity in small perinuclear structures of round/early spermatids most likely representing centrosomes (arrows). SYNE3 was stained in nuclear membranes of Sertoli cells (arrows). (D)

Leucine-rich repeat containing 39 (LRRC39) and Rho related BTB domain containing 2 (RHOBTB2) correspond to images of poor quality. The area for which the HBNet model focused on for prediction of LRRC39 staining only contained unhealthy seminiferous ducts without the correct cell types. Similarly, RHOBTB2 had damaged seminiferous ducts where the cells had been separated from each other and several cell types were missing.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigure1.pdf
- SupplementaryTable1.xlsx