Journal of Proteomics & Bioinformatics - Open Access

Research Article
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000153
JPB/Vol.3 Issue 10

Tissue-Specific Protein Expression in Human Cells, Tissues and Organs

Marcus Gry[1,2,3], Per Oksvold[1], Fredrik Pontén[2] and Mathias Uhlén[1,4]*

[1]*AlbaNova University Center, Royal Institute of Technology (KTH), Stockholm, Sweden*
[2]*Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, Uppsala Sweden*
[3]*Astrazeneca, Safety Assessment, Molecular toxicology, Södertälje Sweden*
[4]*Science for Life Laboratory, Royal Institute of Technology (KTH), Stockholm, Sweden*

Abstract

An important part of understanding human biology is the study of tissue-specific expression both at the gene and protein level. In this study, the analysis of tissue specific protein expression was performed based on tissue micro array data available on the public Human Protein Atlas database (www.proteinatlas.org). An analysis of human proteins, corresponding to approximately one third of the protein-encoding genes, was carried out in 65 human tissues and cell types. The spatial distribution and relative abundance of 6,678 human proteins, were analyzed in different cell populations from various organs and tissues in the human body using unsupervised methods, such as hierarchical clustering and principal component analysis, as well as with supervised methods (Breiman, 2001). Well-known markers, such as neuromodulin for the central nervous system, keratin 20 for gastrointestinal tract and CD45 for hematopoietic cells, were identified as tissue-specific. Proteins expressed in a tissue-specific manner were identified for cells in all of the investigated tissues, including the central nervous system, hematopoietic system, squamous epithelium, mesenchymal cells and cells from the gastrointestinal tract. Several proteins not yet associated with tissue-specificity were identified, providing starting points for further studies to explore tissue-specific functions. This includes proteins with no known function, such as ZNF509 expressed in CNS and C1orf201 expressed in the gastro-intestinal tract. In general, the majority of the gene products are expressed in a ubiquitous manner and few proteins are detected exclusively in cells from a particular tissue class, as exemplified by less than 1% of the analyzed proteins found only in the brain.

Introduction

One of the largest challenges in the post-genome era (Check, 2007; Lander et al., 2001; Venter et al., 2001) is to map the protein–based molecular architecture of the human body (Uhlen, 2007). An important quest is to define the protein abundance in various cells, tissues and organs, including various disease states, and to characterize and to annotate the proteins expressed in a ubiquitous manner, as well as those found only in fractions of the cells or tissues. A hurdle for extrapolation of RNA-based data to protein levels is the plasticity of RNA levels, including regulatory aspects of splice variants and RNAi (Yelin et al., 2003; Katayama et al., 2005) and there is therefore a need to explore the protein expression based on direct measurements of protein levels in cells and tissues.

We have recently (Berglund et al., 2008) described the high-throughput generation of antibodies and subsequent creation of a Human Protein Atlas (www.proteinatlas.org) using immunohistochemistry-based tissue microarrays based on 65 major cell types in 45 different normal tissues. The version 6.0 of this publically available database portal contains the protein profile form 8,400 proteins corresponding to approximately 42% of the estimated number of human protein coding genes (Clamp et al., 2007). A global analysis of the protein expression data revealed that very few of the analyzed proteins were detected in only a single cell type and that a large fraction of the proteins were expressed in any given cell type in the human body (Ponten et al., 2009).Despite this ubiquitous expression, hierarchical cluster analysis revealed that the analyzed cells could be subdivided into categories according to current concepts of differentiation.

Here, we have extended the global study of protein expression to further explore the presence of proteins expressed in a tissue-type specific manner, with an emphasis on fundamental human tissues, such as the central nervous system (CNS), hematopoietic cells (blood), squamous epithelia, mesenchymal cells and cells lining the gastrointestinal tract (GI-tract). The spatial distribution and relative abundance of proteins in the different cell populations from various organs and tissues in the human body were analyzed using unsupervised methods, such as hierarchical clustering, and principal component analysis as well as supervised methods like random forests. Proteins expressed in a tissue-type specific manner were identified for all investigated tissues and the identification of these proteins should allow further studies to explore their specific role and function in cell biology.

Results

Annotation of the immunohistochemistry images

The investigation presented here is based on the results of 8,830 antibodies, corresponding to 6,678 proteins coding genes, which has been used to generate more than 1.7 million immunohistochemistry images. All images are publicly available through the Human Protein Atlas (www.proteinatlas.org). Each antibody has been used to stain 65 normal cell types from 48 different human tissues. The protocol for immunohistochemistry, including concentration of primary antibodies, is titrated to allow the dynamic range of the particular

***Corresponding author:** Mathias Uhlén, AlbaNova University Center, Royal Institute of Technology (KTH), Stockholm, Sweden; Science for Life Laboratory, Royal Institute of Technology (KTH), Stockholm, Sweden, E-mail: mathias.uhlen@scilife.se

Received September 24, 2010; **Accepted** October 19, 2010; **Published** October 21, 2010

Citation: Gry M, Oksvold P, Pontén F, Uhlén M (2010)Tissue-Specific Protein Expression in Human Cells, Tissues and Organs. J Proteomics Bioinform 3: 286-293. doi:10.4172/jpb.1000153

protein target to be captured, regardless of the absolute concentration in the multitude of tissues included in the tissue microarrays. The annotations of the tissues and organs were performed by certified pathologists according to a standardized web-based interface. The annotation scoring, based on the intensity of immunoreactivity and fraction of immunoreactive cells is weighted by using an algorithm and translated into a four-color code and these annotation scores were here converted to numerical values reflecting the relative protein abundance between the different staining levels. The scoring for a given protein is a combined evaluation based on the expression pattern observed in tissues from three different individuals.

Overall protein expression

The tissue profiles were analyzed using heat map visualization combined with dendrograms based on hierarchical clustering procedures using a correlation metric of all protein expression values (n=8,830). The analysis shows that the normal cells cluster into an expected pattern with most of the cells divided into six major groups (i) cells of the central nervous system (CNS), (ii) hematopoietic cells, (iii) mesenchymal cells, (iv) cells with squamous differentiation, (v) endocrine cells and (vi) glandular and transitional epithelial cells (Figure 1A) similar to the results shown recently (Ponten, Gry et al. 2009). The pattern of tissue groups are supported by a principal component analysis (PCA) (Joliffe, 2002) of the protein expression data set (Figure 1B). The patterns in the PCA analysis are accentuated when a filtration step based on variance is applied, implying that there is a subset of proteins that are important for tissue function and cellular determination (data not shown). To further validate the independence of the clusters found in the hierarchical clustering and the PCA, an unsupervised random forest evaluation was applied

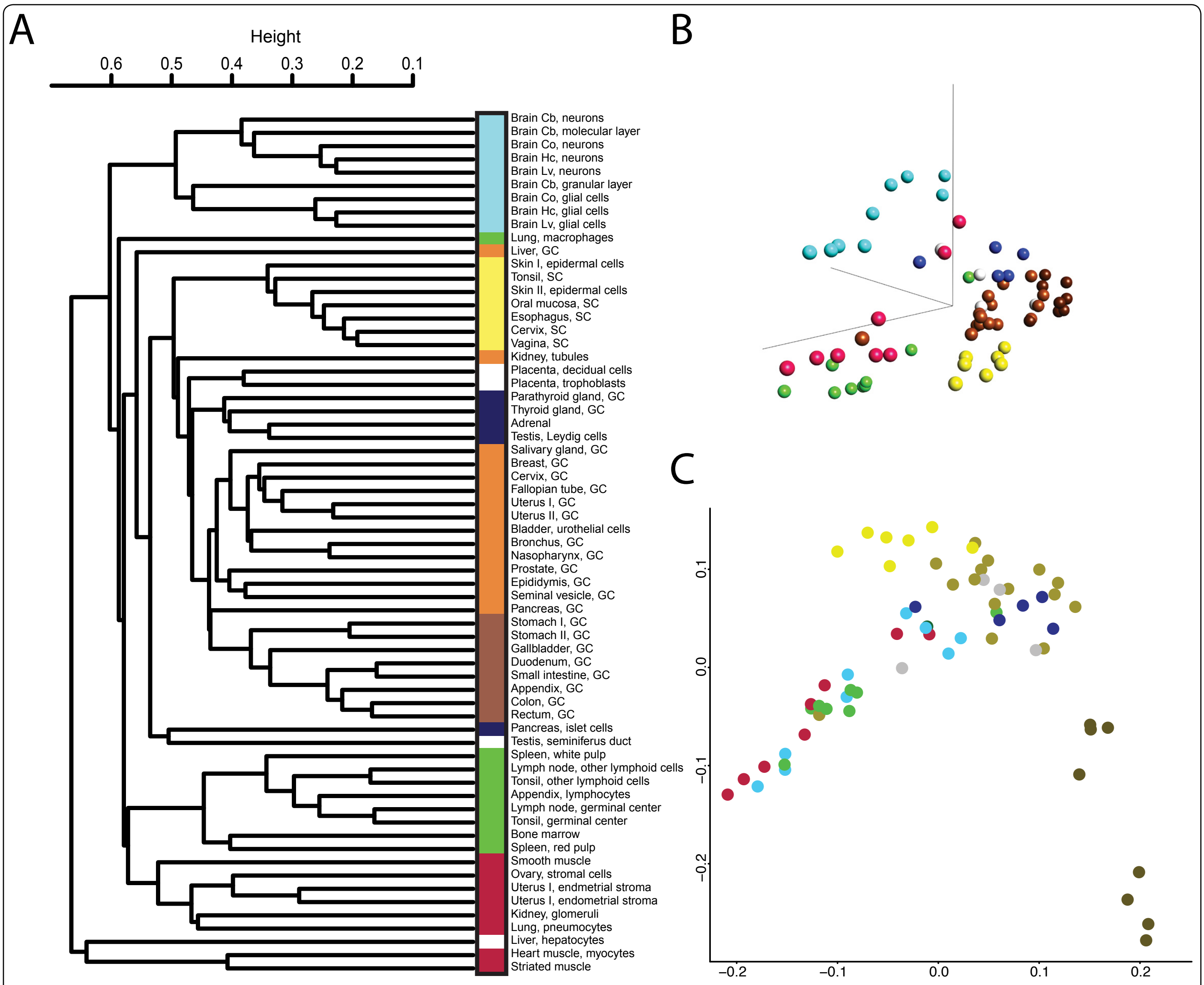

Figure 1: Clusters based on 6,600 proteins and 65 normal tissues. The dendrograms shows how normal human cells fall into distinct categories using hierarchical clustering based on a correlation metric (A). The dendrograms shows that the cells in different tissues are arranged in groups that contain cells with similar functions. The different categories have been colour-coded according to CNS (light blue), hematopoietic system (green), mesenchymal cells (pink), squamous epithelia (yellow), endocrine cells (blue) and glandular and transitional epithelia (light and dark brown). The resulting dendrograms from the hierarchical clustering was confirmed by additional unsupervised methods. In (B) a principal component analysis (PCA) was performed and in (C) a random forest distance calculation was applied followed by a multi dimensional scaling, represented in two dimensions.

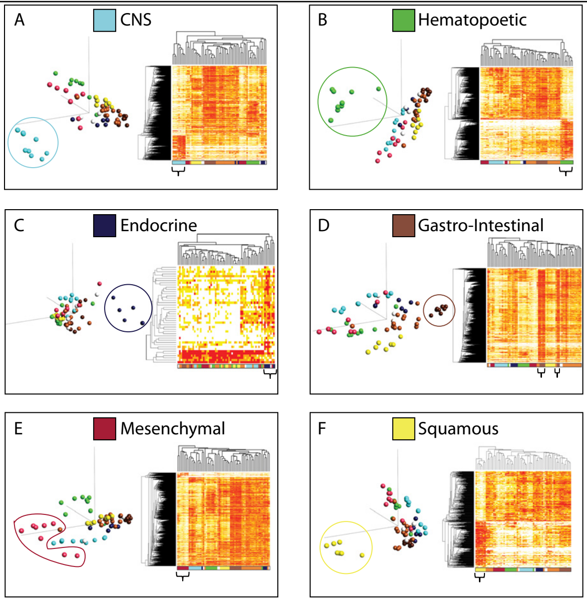

Figure 2: Significantly expressed proteins in specific tissue groups. A student's t-test was conducted to find significantly expressed proteins between the specific groups and all other tissues. The subset of proteins was used to calculate a PCA and a correlation metric based dendrograms in two dimensions. For the CNS group, 1836 proteins were found to be differentially expressed and the subset of differentially expressed proteins is used in a PCA and hierarchical clustering analysis **(A)**. The group of CNS cells are clearly separated and within the CNS tissue group, there is a smaller separation between the neurological and non- neurological cells, 1371 proteins were differentially expressed in the hematopoietic lineage tissues, and this group of tissues is also separated from all other tissues in the PCA plot **(B)**. Lung macrophages and bone marrow are two tissues that are a bit separated from the other tissues in the hematopoietic group. For the endocrine tissues **(C)** the group of cells show less homogeny and a larger spread between the cells within the endocrine group is obvious. The number of differentially expressed genes for the endocrine tissue group is 46. The tissues in the gastro-intestinal groups had 2800 differentially expressed proteins and cell types in this group show a low intrinsic spread and the cluster is clearly defined in the PCA plot **(D)**. The tissues within the mesenchymal tissue group had 2536 differentially expressed proteins and within this group of cells, myocytes from striated muscle and heart form a distinct subgroup in the PCA plot **(E)**. The squamous tissue group is distinct in the PCA plot were the number of differentially expressed genes was 581 **(F)**. The most deviant cell type within the squamous group is surface epithelia of the tonsil. In the accompanying heat maps there is an even distribution between high and low protein expression in the significantly differentially expressed proteins except for the gastro-intestinal tract where all differentially expressed proteins are more highly expressed and the mesenchymal/endocrine class where almost every proteins is more lowly expressed than all other tissues.

(Breiman, 2001; Shi et al., 2005). Also in this case, the various cells belonging to similar tissues clustered together (Figure 1C).

Classification of tissues

The use of supervised classification procedures makes it possible to determine the proteins that contribute to successful classification and the proteins can be ranked based on the contribution to various tissues. There exists a plethora of different methods to perform such a classification, such as random forest, recursive support vector machines, neural networks etc, where every method is suitable for different applications. Here, we used a random forest approach since this method has a documented success with ordinal/categorical data (Shi et al., 2005) and the proteins that contribute to a more accurate tissue classification procedure could be investigated by a ranking procedure of the input variables. A significant overlap between proteins found as highly important in the random forest classification and proteins found as significantly different in the student's t-test computation (Additional file 1) were found for the CNS, endothelial, glandular and haematological tissue groups. The least overlap between the two analysis approaches were for the gastrointestinal, squamous and mesenchymal tissue groups, where the classification had misclassified the endocrine tissues thyroid and parathyroid as glandular and the pancreas islet cells as CNS, the gastro-intestinal tract tissue gallbladder and the squamous tissue tonsil as glandular. In many of the misclassified tissues the tissue is found to be close to the wrongly classified class in the space spanning the principal components (Figure 1).

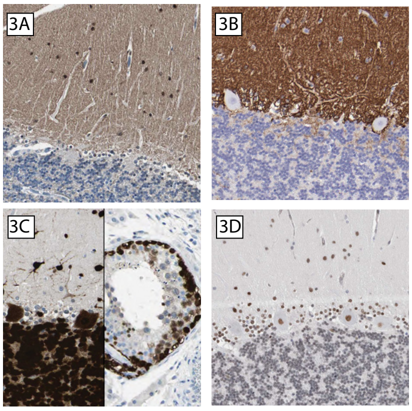

Figure 3: Proteins specifically expressed in the central nervous system. Four examples of four different proteins expressed in cerebellum, and in addition, an example of ELAVL4 expression in testis. The examples cover known proteins involved in neuro-transmission and neurological development such as GAP43 **(A)** and SLC1A3 **(B)** showing strong expression in the neuropili corresponding to the molecular layer of cerebellum. Less well known proteins with enhanced but not exclusive expression in the CNS are ELAVL4 showing strong expression in glial cells mainly in the granular layer of cerebellum and in early spermatogonia in testis **(C)** and a protein with essentially unknown function, ZNF509 showing a nuclear expression pattern mainly in Purkinje cells **(D)**.

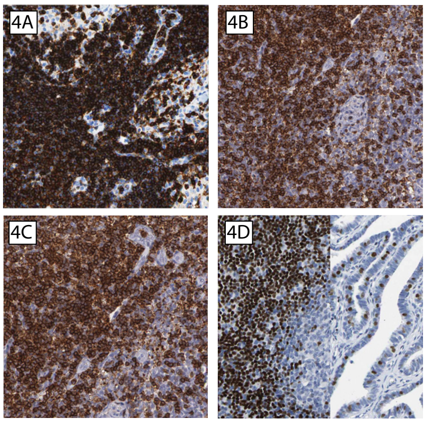

Figure 4: Proteins specifically expressed in the hematopoietic lineage. Examples of proteins expressed in the hematopoietic lineage tissues. Among the significantly up-regulated proteins in the hematopoietic lineage tissues are well known CD markers such as PTPRC (CD-45). CD-45 is expressed in virtually all lymphoid cells, exemplified by two different antibodies from the top 100 list showing expression in lymph node tissues (**A** and **C**). The tyrosine protein kinase BTK, a protein known to play a crucial role in B-cell ontogeny, shows distinct expression in lymphoid cells of the spleen **(B)**. The less well-characterized TCF-7, suggested as a T-cell transcription factor shows expression in lymphoid cells of T-cell areas as well as in subsets of certain epithelia including the fallopian tube **(D)**.

Tissue-specific proteins

The presence of proteins expressed predominantly in a certain subset of cells with proposed similarities is interesting to explore, such as cells which are part of organ systems, e.g. CNS and GI-tract or cells with shared morphological features e.g. squamous epithelia and lymphocytes in various locations. Using PCA on the protein expression data set, proteins that contribute to the patterns for individual tissues and groups of tissues can be found. By filtering proteins that differ significantly in a two-group comparison (Student's t-test) (Student, 1908), it could be investigated whether the remaining proteins would emphasize the independence of a specific group of tissues in the "tissue space" made up by the principal components (Figure 2 A-F, left panels). In addition, hierarchical cluster analysis (Ward, 1963; Eisen et al., 1998) could be used to visualize the expression profiles across all tissues for the subgroups of proteins that were selected using Student's t-tests (Figure 2 A-F, right panels). The number of significantly expressed proteins varied across the different tissues, ranging from 46 (endocrine tissues) to 2800 (Gastro-Intestinal tract). The most significant tissue-specific proteins for CNS, hematopoietic cells, mesenchymal cells, squamous epithelia and GI-tract are listed in the additional material (Additional files 2 to 7). Due to an overall higher number of proteins detected in glandular cells (Ponten et al., 2009), most proteins identified are up-regulated (Figure 2D), while the opposite is found for mesenchymal cells (Figure 2E). In the following, we will focus on the proteins up regulated in the various protein classes shown in (Figure 2).

Proteins specific for cells in the central nervous system

An interesting group of proteins are those predominately expressed in the brain. A list of the 100 most significantly up regulated in CNS can be found in Additional file 2 with four examples displayed in (Figure 3). The group of CNS tissues included neuronal and glial cells from four different regions in the brain (cerebral cortex, hippocampus, adjacent lateral ventricle and cerebellum). Proteins up-regulated in the brain included several proteins known to be highly specific for CNS tissues and used in clinical diagnostics, e.g. GFAP, ENO2, SNCB and S-100 proteins as well as other proteins involved in neuro-transmission and neurological development, e.g. Gap43, OMG, INA, SLC1A3 and MAP2. Out of the four examples shown in (Figure 3), the two most significantly up-regulated proteins, Gap43 (Figure 3A) and SLC1A3 (Figure 3B) show no expression in the other cells, while the other two ELAVL4 (Figure 3C) and ZNF509 (Figure 3D) also show expression in other cells, including testis and pancreas. A group of proteins with an overall high expression level across many different tissues and cell types, but diminished expression in the brain, e.g. SHC1 and B4GALT1 could also be found using the student's t-test. A comparison between glial cells and neuronal cells show that proteins such as CNP and UCHL1 have a glial and neurological specific expression, respectively. A subset of proteins related to neuroendocrine secretion, such as SNAP-25, SV2A and SCG3, were also found to be predominately expressed in brain tissue. In addition, proteins with essentially unknown function were found e.g. ZNF509 (Figure 3D) and Zink-finger SWIM containing protein 5.

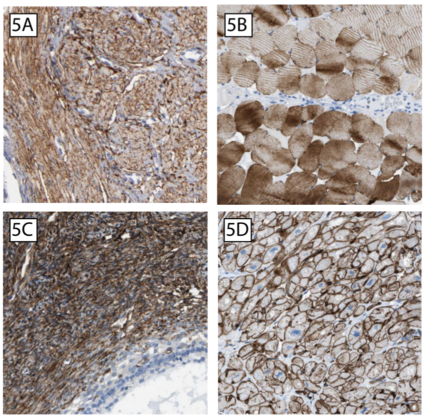

Figure 5: Proteins specifically expressed in the mesenchymal cells. Examples of proteins expressed in the group of mesenchymal tissues. This group of tissues has in general a more heterogeneous expression across the tissues constituting the cluster found in the analysis. Emilin, an extracellular matrix glycoprotein that associates with elastic fibers, shows cytoplasmic expression in smooth muscle cells **(A)**. Desmin, a well-known marker for striated muscle fibers used in clinical diagnostics for distinguishing rhabdomyosarcomas, shows a distinct positivity in striated muscle fibres **(B)**. Less well-known proteins included in the mesenchymal group included TMEM43 expressed in ovarian stroma cells **(C)** and SLC22A showing a distinct membranous expression pattern in cardiomyocytes from a normal heart **(D)**.

Proteins specific for cells of hematopoietic lineages

Many proteins have been identified as specific in hematopoietic cells, in particular surface proteins defined as CD-markers (Zola et al., 2007). Lymphoid cells from lymph nodes, tonsil tissue and appendix as well as hematopoietic cells in the bone marrow and monocyte-derived macrophages constitute the hematopoietic tissue group. We found a large number of hematopoietic cell type-specific proteins and a list of the most significantly up regulated can be found in Additional file 3. As expected, several of the proteins were known CD-markers, including the leukocyte common antigen CD-45 (H1 and H10, Figure 4A and 4C), B-cell markers CD-79A and CD-72, T-cell markers CD-5, CD-8A and TCL1A and monocyte/macrophage marker CD-68. In addition, there were several other intracellular markers of activation and differentiation within cells of the immune system, e.g. the B-cell specific transcription factor PAX-5, neutrophil cytosolic factor NCF-1 and BTK/LCK, a tyrosine protein kinase (Figure 4B) and TCF7, named as a T-cell specific transcription factor but that also shows expression in distinct subsets of cells in the fallopian tube (Figure 4D). Furthermore, proteins less well characterized as markers of hematopoietic subpopulations were found in the analysis, including NUP205, DOK2 and DOCK5. The usefulness of the three latter proteins as markers for subpopulations of cells within hematopoietic lineages needs to be investigated further. Moreover, proteins with differential expression within lymphoid tissues could be identified, e.g. MAP4K1 with enhanced expression in cells comprising the reaction centers (MAP4K1) and PSTPIP1, TNFRSF13C showing enhanced expression in lymphoid cells surrounding reaction centers.

Proteins specific for mesenchymal cells

The mesenchymal cells constitute a relatively heterogeneous group of cells, mainly consisting of muscle cells and stroma cells. No proteins were up-regulated in all the mesenchymal cells, although for the most significantly up-regulated proteins, such as Emelin (Figure 5A) and Desmin (Figure 5B), very little expression was found in non-mesenchymal cells. For most proteins in this group, expression was also detected in other tissues, as exemplified by the presence in skin and testis for the membrane-protein TMem43 (Figure 5C) and by the presence in testis for the putative protein SLC22A15 (Figure 5D). Although only a subset of mesenchymal cell types were included in the analysis, several of the up-regulated proteins in this group were proteins involved in specific soft tissues, including muscle (actins), fat (ADIPOQ), bone (DCN) and connective tissue (PDGFRB). In addition,

Research Article
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000153
JPB/Vol.3 Issue 10

extra-cellular matrix proteins, e.g. laminins, collagens and Galectin-1 (LGALS1) were well represented. Interestingly, this cluster also contained proteins (NES and CD34) known to be implicated in stem cell development.

Proteins specific for squamous epithelia cells

Stratified squamous epithelium consists of multi-layered cells covering the surface of our body, i.e. skin consisting of cornified squamous epithelium, and mucosal surfaces in the upper GI-tract, i.e. oral mucosa and oesophagus, and female genital tract, i.e. vagina and exocervix, consisting of glycogen rich non-keratinizing squamous epithelium. Some examples of tissues are shown in (Figure 6). As expected many proteins from the keratin family are specifically expressed in this group of tissues, such as keratin-1, keratin-14 and keratin-36, as well as well characterized proteins involved in desmosomal function (desmocollin and desmogleins) and the cornification (INV). A protein showing a very specific squamous expression is SPINK5 (Figure 6A), which is involved in hair formation and was found to be expressed at the outermost layers in the squamous differentiated tissues. Since expression data from the skin also included melanocytes, several melanocyte-specific proteins were also found in this group of tissues, e.g. members of the S100 protein family like S100A7 and S100A2, despite that these proteins are not expressed in cells with squamous differentiation. However, other S100 protein family members like S100A7 and S100A2 in this group showed a restricted pattern of cytoplasmic expression in defined stages of squamous differentiation corresponding to layers of squamous epithelium, e.g. suprabasal and basal cells respectively. Proteins of other families like Galectin-7 and interleukin-1 receptor antagonist protein (IL1RN) (Figure 6B) was found to have a predominantly nuclear expression pattern. Also present in this group of proteins is the well-studied transcription factor p63, a transcription factor implicated in development and maintenance of stratified epithelial tissues. Another protein, highly specific for squamous epithelia and showing expression in suprabasal cells was CRNN (Figure 6C), also known as squamous epithelial heat shock protein 53, suggested to play a role in the mucosal/epithelial immune response and epidermal differentiation. The hyaluronic acid receptor CD44 was found to be mainly expressed in squamous epithelia (Figure 6D), consistent with earlier reports on the isoform 10 (epithelial isoform), known to be expressed by cells of epithelium.

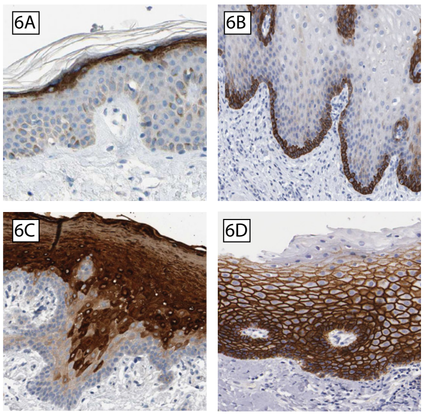

Figure 6: Proteins specifically expressed in squamous epithelia. Examples of proteins expressed in the group of squamous epithelia tissues. Several proteins found as specifically expressed in the group of squamous tissues are involved in squamous differentiation, cornification and desmosomal function. Examples shown include expression of SPINK5, known to be involved in hair formation, in the uppermost layer of normal skin **(A)**. The interleukin 1 receptor antagonist IL1RN showed a distinct nuclear expression pattern in squamous epithelia that was restricted to basal cells, exemplified in normal esophagus **(B)**. Cornulin (CRNN), suggested to play a role in squamous differentiation was expressed in suprabasal, differentiating squamous epithelia in the oral mucosa **(C)**. The membrane bound hyaloronic acid receptor was found to be selectively expressed in all layers of squamous epithelia of the outer portions of the cervix **(D)**.

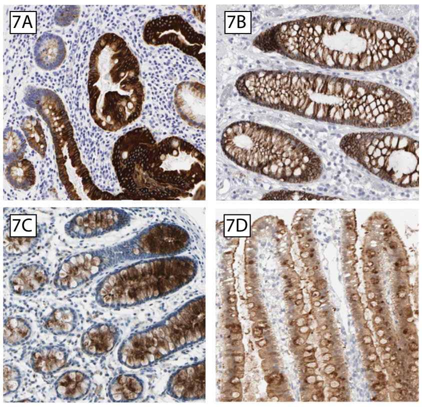

Figure 7: Proteins specific for the gastrointestinal tract (GI-tract). Examples of proteins expressed in the group of gastrointestinal tissues. The examples include known GI-tract specific proteins such as the clinically used marker for GI-tract tumours Keratin-20 **(A)** and CDH17, a member of the cadherin super family of calcium-dependent, membrane-associated glycoprotein's, showing membranous expression in normal colonic crypts **(B)**. Several proteins expressed in mucus producing glandular cells were in the top 100 list, including MUC5B showing expression in goblet cells of the rectal mucosa **(C)**. The unknown protein encoded by C1ORF201 showed a diffuse cytoplasmic expression pattern in the small intestine **(D)**.

Proteins specific for the gastrointenstinal tract (GI-tract)

The group of tissues representative of the GI-tract is comprised of glandular cells from the upper and lower gastric mucosa, duodenum, small intestine, appendix, colon and rectum. Some examples of tissues from this group are shown in Figure 7. Several known gastrointestinal tract specific proteins were found, such as keratin 20 (Figure 7A) and Cadherin-17 (Figure 7B). Other proteins selectively expressed included several ion channel proteins, e.g. SLC12A3, SLC9A1, KCNMB3 and FXYD4, as well as proteins mainly expressed in mucus producing glandular cells present in the GI-tract, e.g. Villin-1, MUC5B (Figure 7C) and serine proteases, such as SPINK4. More surprising is the presence of transcription factors required for the expression of several liver-specific genes, e.g. hepatocyte nuclear factor 4-alpha and nuclear hepatocyte factor 1-alpha. In addition to known proteins, the list also included totally unknown proteins such as C1ORF201, showing enhanced expression in cells from the GI-tract (Figure 7D).

Gene ontology analysis

Further, by using Gene Ontology (GO) analysis (Ashburner et al., 2000) for the proteins that were found as significantly expressed for a specific group of tissues, the significant ontology's that are important fore tissue formation and function could be investigated. In general, the Gene Ontology results were in correspondence with the current knowledge regarding the primary functions of the cells included in respective groups (Additional file 7). The tissues in the group of hematopoietic lineage were as expected related to the immune system, including lymphocyte and leukocyte regulation and the top ontology's for the squamous differentiated group of tissues were involved in ecto- and endoderm development and keratinocyte differentiation. The most significant ontology's for the CNS tissues were membrane docking and microtubule related processes and, although not significant using multiple adjusted p-values, vesicle related processes The endocrine tissues had several ontology's related to catabolic processes, although again not significant using multiple adjusted p-values, whereas the gastro intestinal tract tissues had phosphate, glycoprotein and carbohydrate metabolic processes as highly ranked results. Glandular tissues had several perceptic events as highly ranked results. The members of the mesenchymal tissue group resulted in ontology's related to several metabolic processes.

Discussion

Here, we show for the first time an analysis of tissue-specific proteins based on an anatomically comprehensive analysis of protein profiles in normal tissues and organs using annotation by certified pathologists of more than 1.7 million immunohistochemistry images. In addition to insights of more general nature, the unbiased analysis of protein expression in a large variety of normal cells allows for the identification of proteins with specific patterns of expression in certain groups of similar cell types. Proteins expressed in a narrow, well-defined set of cell types are most likely important for the function of the given cells. The identification of proteins with unknown functions specifically expressed in various cell types is important and further characterization with functional studies will broaden our current knowledge of cell biological pathways involved in cell type specific differentiation and homeostasis of corresponding tissues and organ types. The recognition of specific expression profiles is also of critical significance to understand pathological conditions and in an extended view; the detection of novel proteins will provide a fundament for the development of new diagnostic tools and targets for therapy for a wide spectrum of human diseases.

In a recent paper (Ponten et al., 2009), we show evidence that a high fraction (>65%) of the human proteins are present in any single cell type in the human body. Consequently, few proteins (<2%) were found to be expressed in a single or only a few types of cell. These findings suggested that few proteins are expressed in a cell type-specific manner and we speculated that the phenotype of a cell is determined by localization, modifications and fluctuations in concentrations of a large portion of the proteome, and not by a mere "on/off" expression. These conclusions are supported by the results presented here regarding a more extended analysis including all the cells from a particular tissue type, such as CNS or mesenchymal cells. Few proteins are detected exclusively in cells from a particular tissue class, as exemplified by less than 1% of the analyzed proteins found only in the brain.

The analysis of global expression patterns using antibody-based proteomics imposes major challenges as compared to transcriptional profiling, including the presence of multiple isoforms, and the fact that the dynamic range of protein concentration can vary considerably in cells (Rimm, 2006) and more than 10^{10} in serum or plasma (Anderson and Anderson, 2002). In addition, the chemical space of the amino acids is much larger than for the corresponding nucleotides making sample handling and preparation more challenging for proteins than for RNA (Templin et al., 2003). The large number of isoforms, in the form of splice variants, proteolytic maturation and post-translation modifications makes the "proteome-space" large and complex. In this study, most antibodies were designed to recognize all isoforms of the protein target, and thus the expression patterns represents, in each case, the total number of isoforms of a particular gene product. Another major issue is due to the semi-quantitative nature of the enzyme-based amplification methods used for immunohistochemistry and the variability introduced by the individual experimental staining protocol, including choice of antibody dilution and antigen retrieval methods (Taylor and Levenson, 2006; Walker, 2006). We therefore establish an immunological protocol aimed to ensure that each protein target is analyzed at a concentration range relevant for that specific protein. This was accomplished by analysis of each individual protein target in parallel with a series of tissue microarrays (Kononen et al., 1998) using the same antibody dilution and retrieval methods. This procedure allows the determination of relative protein levels in various tissues for a particular protein target across a multitude of bio samples within the dynamic range of the particular protein target. However, it cannot be ruled out that some of the staining is due to background binding due to cross-reactivity of the antibodies used to stain the tissue microarrays. It is thus desirable to extend this study in the future with validations based on paired antibodies (ref) or knock-down of the genes based on siRNA methods (ref).

In conclusion, this systematic study covering approximately one third of the protein-encoded genes, suggest that relatively few proteins were expressed in a tissue-specific manner. Well-known markers, such as neuromodulin for CNS, keratin 20 for GI-tract and CD45 for hematopoietic cells, were identified as tissue-specific, but we also found many proteins not yet associated with tissue-specificity, providing starting points for further studies to explore tissue-specific functions. All antibodies used in this study are available to the public from a multitude of commercial vendors and these research reagents thus constitute a valuable resource to define the proteomic landscape in tissues, support the discovery of new diagnostic and therapeutic tools and enhance opportunities for basic biological and medical research.

Materials and Methods

Data collection and extraction

To determine the level of protein expression of each protein in this study, antibodies were used to immunohistochemically stain human tissues assembled in tissue microarrays (TMA) blocks. Tissue cores with 1mm diameter, sampled from 144 individuals, corresponding to 48 different normal human tissues types, were included in the study. Immunohistochemically stained sections from the TMA blocks were scanned in high-resolution scanners and separated to individual spot-images representing each core. For the TMAs, certified pathologists evaluated all images in a web-based annotation system to collect parameters regarding distribution, extent and level of protein expression (Oksvold and Björling, unpublished). Parameters from the annotation included staining intensity, fraction of stained cells in a defined cell population and sub-cellular localization of staining. The annotation was performed for selected cell types for each tissue as most tissue types include several defined cell phenotypes, e.g. neurons and glial cells in brain tissue and glomeruli and tubules in kidney. In total, ~8,800 antibodies with 298 annotations were assembled from the TMA measurements. The annotation parameters for intensity and quantity (fraction of positively stained cells) were combined into a four-grade scale represented by the colours white (negative), yellow (weak), orange (moderate), and red (Venter et al., 2001) level of protein expression. All data is presented in this format on the protein atlas (www.proteinatlas.org). For the data analysis regarding protein expression the colour codes representing the staining levels were converted to numerical values using a red to 4, orange to 3, yellow to 2 and white to 1, transformation. In cases where the protein expression value could not be derived, due to low image quality, a Not Available (NA) value was introduced. The data was ordered into a matrix with m (number of antibodies)* n (number of tissues (n=65) or cell lines (n=45)) dimensions. The number of tissues is a combined tissue and cell type parameter, where the number of tissues and cell types give rise to tissues * cell types number of parameters.

Imputations

Prior the principal component analysis and the random forest, an imputation of missing values were made. The imputations were calculated using median values across tissues.

Unsupervised hierarchical clustering

A correlation matrix based on Spearman's Rho (Spearman, 1904) was calculated for the protein expression data set. The correlation matrix was converted to a distance metric using a 1 -correlation value transformation. The data was clustered using unsupervised top down hierarchical clustering (Ward, 1963), where at each stage the distances between clusters are recomputed by the Lance-Williams dissimilarity update formula (Lance and Williams, 1966) according to average linkage. The antibodies with no defined correlation due to constant expression across all tissues or cell lines were removed in the clustering procedure.

Unsupervised principal component analysis (PCA)

Principal component analysis is done using a transformation from the "feature" space (antibody space) to the space that is made by the principal components (Joliffe, 2002). In the analysis of the protein expression dataset using PCA, the number of components is consistently 3, which covers a large extent of the total variation. The variance filter applied in the analysis is an $sd_{specific} / sd_{total}$ cut-off where the specific variation is the variance for one individual antibody and the total variance in the variance for the complete data set.

Unsupervised random forest

The unsupervised random forest is constructed using 10000 trees (Breiman, 2001). The random forest was subjected to a multidimensional scaling, which is represented in two dimensions.

Supervised random forest

For the classification event using the random forest algorithm, the classes were determined using the output from the hierarchical clustering and the PCA. The number of trees was set to 10,000 which were determined based on different classification events using different number of trees (data not shown). The importance of the different variables was used to estimate the concordance with the differential expression analysis.

Supervised differential expression

To calculate the proteins that are most significant for a group of tissues, a Student's t-test (Student, 1908) with a common variance were applied between one chosen groups of tissues and the remaining tissues, as defines in the hierarchical clustering and the PCA. The p-values derived from the t-test are corrected using a Benjamini-Hochberg adjustment method (Benjamini and Hochberg, 1995).

Gene ontology analysis

The proteins found using the Student's t-test was analyzed using hyper-geometric tests based on the gene ontology's for each protein (Ashburner et al., 2000). The derived p-values are corrected using a false discovery rate adjustment (Benjamini and Hochberg, 1995), (Additional file 7).

Concordance analysis

The concordance between the importance ranking of the random forest classification event and the significant variables from the student's t-test was measured using a concordance at the top plot (Irizarry et al., 2005). The figures are the fraction of common protein ID:s among the number of protein ID:s that are investigated through each consecutive round.

Acknowledgements

This work was supported by grants from the Knut and Alice Wallenberg Foundation, the Swedish Research Council (VR) and the EU 7th framework program Prospects.

References

1. Anderson NL, Anderson NG (2002) The human plasma proteome: history, character, and diagnostic prospects. Mol Cell Proteomics 1: 845-867.
2. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.
3. Benjamini Y, Hochberg Y (1995) The control of the false discovery rate: A practical and powerful approach to multiple testing in multiple testing under dependency. J R Stat Soc Ser A 57: 289-300.
4. Berglund L, Björling E, Oksvold P, Fagerberg L, Asplund A, et al. (2008) A genecentric Human Protein Atlas for expression profiles based on antibodies. Mol Cell Proteomics 7: 2019-2027.
5. Breiman L (2001) Random Forests. Mach Learn 45: 5-32.
6. Check E (2007) Genome project turns up evolutionary surprises. Nature 447: 760-761.
7. Clamp M, Fry B, Kamal M, Xie X, Cuff J, et al. (2007) Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci U S A 104: 19428-19433.
8. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863-14868.
9. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, et al. (2005) Multiple-laboratory comparison of microarray platforms. Nat Methods 2: 345-350.
10. Joliffe I (2002) Principal Component Analysis. New York, Springer.
11. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, et al. (2005) Antisense transcription in the mammalian transcriptome. Science 309: 1564-1566.
12. Kononen J, Bubendorf L, Kallioniemi A, Bärlund M, Schraml P, et al. (1998) Tissue microarrays for high-throughput molecular profiling of tumor specimens. Nat Med 4: 844-847.
13. Lance GN, Williams WT (1966) A general theory of classifactory sorting strategies. Computer Journal 9: 373-380.
14. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.
15. Pontén F, Gry M, Fagerberg L, Lundberg E, Asplund A, et al. (2009) A global view of protein expression in human cells, tissues, and organs. Mol Syst Biol 5: 337.
16. Rimm DL (2006) What brown cannot do for you. Nat Biotechnol 24: 914-916.
17. Shi T, Seligson D, Belldegrun AS, Palotie A, Horvath S (2005) Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. Mod Pathol 18: 547-557.
18. Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15: 72-101.

19. Student (1908) The probable error of a mean. Biometrika 6: 1-25.

20. Taussig MJ, Stoevesandt O, Borrebaeck CA, Bradbury AR, Cahill D, et al. (2007) ProteomeBinders: planning a European resource of affinity reagents for analysis of the human proteome. Nat Methods 4: 13-17.

21. Taylor CR, Levenson RM (2006) Quantification of immunohistochemistry--issues concerning methods, utility and semiquantitative assessment II. Histopathology 49: 411-424.

22. Templin MF, Stoll D, Schwenk JM, Pötz O, Kramer S, et al. (2003) Protein microarrays: promising tools for proteomic research. Proteomics 3: 2155-2166.

23. Uhlen M (2007) Mapping the human proteome using antibodies. Mol Cell Proteomics 6: 1455-1456.

24. Venter JC, et al. (2001) The sequence of the human genome. Science 291: 1304-1351.

25. Walker RA (2006) Quantification of immunohistochemistry--issues concerning methods, utility and semiquantitative assessment I. Histopathology 49: 406-410.

26. Ward J (1963) Hierarchical Grouping to Optimize an Objective Function. J Am Stat Assoc 58: 236-244.

27. Yelin R, Dahary D, Sorek R, Levanon EY, Goldstein O, et al. (2003) Widespread occurrence of antisense transcription in the human genome. Nat Biotechnol 21: 379-386.

28. Zola H Bernadotte S et al. (2007) Leukocyte and Stromal Cell Molecules: THE CD MARKERS. Hoboken, Wiley.