Analysis of Gene Expression Data from Normal Human Tissues

Hila Benjamin
M.Sc. Thesis submitted to the Feinberg Graduate School
Weizmann Institute of Science

Research conducted under the supervision of
Prof. Eytan Domany and **Prof. Doron Lancet**

February 2004

אנליזה של ביטוי גנים ברקמות נורמליות באדם

הילה בנימין

תזה לשם קבלת התואר מוסמך למדעים
מוגש למועצה המדעית של מכון ויצמן למדע

בהדרכת
פרופסור איתן דומאני ופרופסור דורון לנצט

פברואר 2004

Acknowledgements

I would like to take this opportunity to express my gratitude to all the people who were involved in this work. First, I would like to thank my supervisors, Prof. Eytan Domany and Prof. Doron Lancet, for their guidance throughout this research. Eytan is a true teacher, his door is always open for every question, ready to share thoughts and ideas. Doron has shown me a different and fascinating way of thinking, which gave me a new approach to many problems. It has been a pleasure to work with both of them. I would also like to thank Dr. Itai Yanai for his invaluable collaboration. Itai is a talented scientist; working with him was a wonderful and highly educational experience, the results of which comprise a large part of this thesis. I am sincerely grateful to Dr. Orit Shmueli who helped me during my first days in the lab and from whom I learned a lot professionally. I would also like to express my gratitude to the members of the 'Domany group' for creating such a wonderful working atmosphere, for helping each other at all times and for stimulating discussions. I am especially grateful to Yuval Tabach for his faith and encouragement from the day we both joined Eytan's group. I am happy to thank Nimrod Dorfman and Tal Shay for reading this manuscript and sharing their enlightening observations with me. Their help was invaluable. Last, I would like to thank my family for their everlasting support.

Contents

ABSTRACT **2**

1 INTRODUCTION **3**

1.1 BIOLOGICAL BACKGROUND 3
1.2 DNA MICROARRAYS 15
1.3 CLUSTERING OF GENE EXPRESSION DATA 21
1.4 GENEANNOT – MICROARRAY GENE ANNOTATION 26

2 METHODS **28**

2.1 RNA SAMPLES AND ARRAY HYBRIDIZATIONS 28
2.2 EXPRESSION DATA PREPROCESSING 28
2.3 CENTERING AND NORMALIZATION 29
2.4 CLUSTERING THE ENTIRE DATASET 29
2.5 STATISTICAL ANALYSIS OF DIFFERENTIAL EXPRESSION 30
2.6 SIGNAL QUANTILIZATION 30
2.7 FILTERING USING THE 'GAP' CRITERION 31
2.8 UNSUPERVISED CLUSTERING USING SPC 31
2.9 BINARY CLASSIFICATION 31
2.10 TISSUE SPECIFICITY INDEX (T) 32
2.11 EXPRESSION PROFILES OF GENES 32
2.12 ANALYSIS OF MOST DIVERGENT GENES 33
2.13 COMPARISON OF RESULTS TO A PUBLISHED DATASET [17] 33

3 RESULTS **35**

3.1 CLUSTERING THE ENTIRE DATASET 35
3.2 CLUSTERING THE MINGAP SET 38
3.3 EXPRESSION PROFILES OF GENES 47
3.4 ANALYSIS OF MOST DIVERGENT GENES 51
3.5 COMPARISON OF RESULTS TO A PUBLISHED DATASET [17] 52

4 CONCLUSIONS **56**

4.1 FUTURE DIRECTIONS 58

REFERENCES **59**

Abstract

The ontogeny of complex multicellular organisms is enabled by the differential expression of genes across various cell types. Understanding this process requires a comprehensive, whole-genome view of gene expression patterns. In this study we aim to investigate gene expression patterns in normal human tissues. For this purpose we analyzed the results of whole-genome microarray experiments of 12 normal human tissues, using cluster analysis and binary classification. In the past, genes have often been characterized dichotomously as housekeeping or one-tissue specific. However, many more patterns of gene expression were found in normal human tissues. Specifically, we found a tendency to either expression or suppression in a relatively small number of tissues. Clusters of tissues with related function and embryonic origin were found: brain and spinal cord; skeletal muscle and heart; bone marrow, spleen and thymus; and liver and kidney. Previous studies treated the expression profiles of probe sets as the expression profiles of genes. There are, however, cases in which several probe sets represent the same gene. Furthermore, sometimes these probe sets have different expression profiles. In order to shift from the level of probe set expression to the level of gene expression, we developed a method that considers the probes' sequences and the probe set's expression profile. Probe sets' expression profiles were averaged if their probe sequences matched the same mRNA sequence, and if they were correlated in terms of their expression profiles in the current study. This method was used to create a set of expression profiles characteristic of known genes, resulting in a reduction from a set of 23,689 probe sets representing known genes, to a set of 17,118 gene expression profiles. Cluster analysis was applied to a subset of these gene expression profiles, revealing four major clusters of genes, each over-expressed in a group of tissues related to the tissue clusters identified above.

1 Introduction

The genomes of an increasing number of organisms, including the human genome, have been completely sequenced. Despite this fact, the functional role of most genes is still unknown, and cannot always be inferred from its nucleotide sequence [1]. Knowledge of differential gene expression patterns across a variety of cell types and conditions (for example, diseases, cell cycle) may contribute to the understanding of gene function. This study examines gene expression in various normal human tissues. It is based on the analysis of DNA microarray results from 12 normal human tissues, using methods of cluster analysis. This introduction includes four parts: biological background, DNA microarray technology, cluster analysis algorithms and the GeneAnnot tool.

1.1 Biological background

1.1.1 Differentiation

Multicellular organisms have evolved ways to form an organized array of a variety of cell types. The generation of cellular diversity is called *differentiation* [2]. Once a cell in a multicellular organism has committed to differentiate into a specific cell type, the decision is maintained through many subsequent generations [3]. This phenomenon of "cell memory" is required in order to create organized tissues and organs, and to maintain stable differentiated cell types.

1.1.2 Morphogenesis

Differentiated cells are not randomly distributed; rather, they are organized in tissues and organs. The creation of an ordered form is termed *morphogenesis* [2]. This term refers to the cell and tissue movements that give the organ or organism its shape in three dimensions [4]. From a morphological point of view, embryo cells can be divided into two groups: *epithelial cells* and *mesenchymal cells*. These terms relate to cell shape and behavior and not to embryonic origin (discussed in section 1.1.4), both epithelia and mesenchyme can arise from all three germ layers. Epithelial cells are tightly connected to one another in sheets or tubes, arranged on a basement membrane. Each cell is joined to its neighbors by specialized junctions, and shows distinct polarity. Mesenchymal

cells are unconnected and operate as individual units. They fill up much of the embryo and later form fibroblasts, adipose tissue, smooth muscle and skeletal tissues.

There are several cellular processes that bring about morphogenesis: (1) direction and number of cell divisions; (2) cell shape changes; (3) cell movement; (4) cell death; (5) cell growth; (6) changes in composition of cell membrane and extracellular matrix [2].

1.1.3 Cells regulate development by gene expression

Almost every cell in a multicellular organism contains the entire genomic information. The set of proteins present in each cell, however, is not identical in different cell types and along the developmental axis. The proteins that exist in a cell at a given time are determined by the genes expressed in the cell at that time: a gene is transcribed to messenger RNA (mRNA), and the mRNA is translated into a protein.

The simplest changes in gene expression are transient, and appear in all organisms, from prokaryotes to multicellular developed eukaryotes. The control over gene expression in eukaryotes is combinatorial: multiple gene-regulatory proteins act in combination to regulate the expression of a single gene, and each regulatory protein contributes to the regulation of many genes. Most gene-regulatory proteins are switched on in several different cell types, and at various time points during development [5].

There are many levels of control in the process of gene expression. Although control on the initiation of gene transcription is the predominant form of regulation for most genes, other kinds of control can act later in the pathway from RNA to protein to modulate the amount of gene product [3]. The main control mechanisms are as follows:

1. *Transcription attenuation* - expression of certain genes is inhibited by premature termination of transcription. The new RNA chain adopts a structure that causes it to interact with the RNA polymerase in such a way as to abort its transcription [3, 6].

2. *RNA splicing* - many genes are first transcribed as long mRNA precursors that are then shortened by a series of processing steps to produce the mature mRNA molecule. One of these steps is RNA splicing, in which intron sequences are removed from the mRNA precursor. This procedure is mostly used for alternative RNA splicing, creating different mRNA molecules, which result in different proteins, from the same gene sequence. In some cases, however, alternative RNA splicing is used to switch from the production of a nonfunctional protein to the production of a functional one [3, 7].

3. *RNA transport from the nucleus* - The process of mRNA transport from the nucleus, through the nuclear pores, into the cytoplasm is an active process. This process is under regulation by specific RNA-binding proteins [3, 8].

4. *Localization in the cytoplasm* – in some cases, mRNAs are directed to specific intracellular locations by signals in the mRNA sequence itself, before the sequence has been translated into an amino acid sequence [3].

5. *Negative translation control* - translation of some mRNA molecules is blocked by specific translation repressor proteins that bind near the 5' end of the mRNAs [3].

6. *Regulated mRNA stability* – some mRNAs are unstable because they contain specific sequences that stimulate their degradation. The stability of an mRNA can be changed in response to extracellular signals, through binding of these sequences by specific proteins that enhance the stability of the mRNA molecule [3].

7. *Poly-A length control* - once in the cytoplasm, the 200-nucleotide-long poly-A tails of most mRNAs gradually shorten over the course of days. Tails shorter than about 30 nucleotides, however, are not observed, and therefore mRNAs with short tails are not translated. The poly-A tail length of some mRNAs is specifically controlled - either by selective poly-A addition or by selective poly-A removal [3].

8. *Posttranslational modifications* - several changes can take place after translation is complete that determine whether or not the protein will be active. Some proteins are inactive without the cleaving away of certain inhibitory sections; other proteins must be localized to specific intracellular destinations in order to function; another group of proteins needs to assemble with other proteins to form a functional unit; and last, proteins that are not active unless they bind a specific ion, or are modified by a covalent addition of a phosphate or acetate group [2].

The collection of genes that are transcribed from genomic DNA in a certain cell at a given time (called “expression profile” or “transcriptome”) can be considered a measure of cellular phenotype [1]. Different mRNA expression profiles characterize specialized cell types; however, they do not reflect the full range of differences between protein production profiles, due to posttranscriptional control of protein expression discussed above. Nonetheless, mRNA expression is still considered an indicator of gene function [9-13].

The ontogeny of complex multicellular organisms is enabled by the differential expression of genes across various cell types and at different levels in the developmental process. Some genes are expressed in all cell types and are considered

housekeeping or maintenance genes [14, 15], whereas others are expressed in a restricted selection of tissues and are hence identified as specific [16, 17]. Only a small percentage of the genome is being expressed in a given cell at a given time, and a portion of these genes is specific to that cell type, and is not expressed in other cell types [2]. In previous research on the tissue specificity of genes, emphasis has mainly been on the extremes of one-tissue specific [16, 17] and housekeeping genes [15, 18, 19]. However, many genes may show midrange expression patterns, i.e. are expressed only in a subset of the tissues.

A multicellular organism arises by a relatively slow process of progressive developmental changes. The main stages of embryonic development from a fertilized egg to a complete multicellular organism are detailed below.

1.1.4 Stages of embryonic development

The life of a new individual is initiated by *Fertilization*, the fusion of genetic material from the sperm and the egg, to form a fertilized egg or a *zygote*. The fertilized egg develops to a complete organism through the process of *embryogenesis*.

Embryogenesis includes several stages [2, 4]:

a) *Cleavage* – extremely rapid mitotic divisions. The enormous volume of the zygote cytoplasm is divided into numerous smaller cells, called *blastomeres*. In contrast to regular cell divisions, cleavage does not include a growth phase between successive divisions. By the end of cleavage, the blastomeres form a sphere, called *Blastula*. Commitment of a cell to a differentiated cell type starts at this stage, when the cleavage planes separate qualitatively different regions of the polar zygote cytoplasm into different daughter cells.

b) *Gastrulation* – cells rearrangements. Blastomeres undergo dramatic movements wherein they change their position relative to one another. Three cell regions are formed, called *germ layers*.

The three-layered structure formed in this stage is called the *gastrula*. The three germ layers are:

1. Ectoderm – outer layer. Produces the cells of the epidermis (skin) and the nervous system.
2. Mesoderm – middle layer. Cells that later form muscles, connective tissues, excretory organs and gonads.

3. Endoderm – inner layer. Produces the lining of the epithelial tissues, digestive tube and associated organs (pancreas, liver, lung etc.).

c) *Organogenesis* – the cells interact with one another and rearrange themselves to produce tissues and organs. Many organs contain cells from more than one germ layer.

Separation of somatic cells from germ cells is often one of the first differentiations to occur during animal development. The germ cells are not considered as belonging to any of the three germ layers identified above [4]. *Gametogenesis*, the forming of mature gametes, usually does not occur until the organism is mature.

1.1.5 Primary tissues

The adult human body is composed of over 250 different cell types. A tissue is a functional aggregation of similar cells and their intercellular materials. An organ usually contains several tissue types, originating from different embryonic cell lineage and arranged to fulfill a common function. Classical histology distinguishes between four different primary classes of tissues: epithelia, connective tissues, muscles and neural tissues [20].

1.1.5.1 Epithelial tissues

Epithelium has diverse functions in different tissues: as epidermis, the epithelium covers the exterior of the body, protecting it from mechanical trauma, loss of moisture, and harmful substances in the environment; in the digestive track, the epithelium has an absorptive action; the digestion procedure involves epithelial enzymes; hormones regulating endocrine functions are secreted by epithelial cells; in the kidney, the epithelium has an excretory function; the senses of hearing, seeing and smelling are mediated by neuroepithelium.

Epithelium forms continuous layers that cover surfaced (skin) and line cavities of the body. Epithelial cells are derived from all three embryonic germ layers: the epithelial cells of the digestive and respiratory systems arise from the endoderm, the epithelium lining the oral and nasal cavities and the anus originate from the ectoderm, and the cells lining the pleural, pericardial and peritoneal cavities, the kidneys, gonads, liver and pancreas are of mesodermal origin.

1.1.5.2 Connective tissues

Connective tissues form the supporting framework for all tissues and organs of the body. They also provide means for anchoring and binding organs together, as well as forming the packing tissue between them. Connective tissues are all mesodermal derivatives and they include mainly blood, bone, lymph, fibroblasts, macrophage and fat cells.

There is considerable morphological variation between connective tissue types, ranging from the hard, calcified bone to the circulating blood. In spite of their diversity, all connective tissues have an intercellular matrix, composed of an amorphous ground substance and extracellular fibers.

Some of the major functions performed by connective tissues are: (1) Bony and cartilagenous framework for all other organs and tissues; (2) O_2 and nutrient transport; CO_2 and metabolite removal; (3) Lipid storage by adipose tissue; (4) Insulation against heat loss by fat cells.

1.1.5.3 Muscular tissues

Muscle is the primary tissue of action and motion, with the highest degree of contractility. Muscle tissue includes, in addition to muscle cells, connective tissue fibers, nerve cells and blood capillaries lined with epithelial cells.

There are three types of muscle fibers: *smooth* – involuntary, nonstriated, found in the walls of viscera; *cardiac* - involuntary, striated, found in the heart; *skeletal* - voluntary, striated, usually attached to bones or skin. All muscle cells are of mesodermal origin.

1.1.5.4 Neural tissues

The neural tissues are the ones to alert the organism to changes in the external as well as the internal environments. Nerve cells coordinate and integrate the functioning of all tissues and organs of the body.

The neural tissues are divided into three systems: the central nervous system (CNS), including the spinal cord and brain; the peripheral nervous system (PNS), formed by the nerves that arise from the brain and spinal cord to pass to other parts of the body; the autonomic nervous system (ANS), including many small ganglia and nerve fibers, innervating the organs of the body. The ANS carries nerve impulses to smooth muscles such as blood vessels.

There are two main cell types in the neural tissues: neurons and glial (supportive) cells. Most cells are of ectodermal origin, except for microglia (a sub-type of glial cell), which is mesodermal. A neuron is typically composed of three parts: (1) sell body, containing a large nucleus; (2) short dendrites, which respond to stimuli and convey impulses towards the cell body; (3) long axon that transmits signals away from the cell body, to terminate on a target cell: a muscle, a gland or another neuron. The glial cells bind together all elements of the nervous tissue within the CNS, functioning similarly to connective tissue cells.

The CNS – "gray matter" and "white matter": Some nerve fibers are covered with a sheath of myelin, around the core of the axon. Both brain and spinal cord are built of "gray" and "white" matter. Gray matter includes nerve cell bodies and unmyelinated fibers. The cell bodies are separated by dense fibrous network consisting of dendrites, axons and glia, and permeated by a capillary bed. It is these areas where most synapses occur. The gray matter has an extra-cellular compartment, which is a space comprising 20-30% of the tissue volume. White matter, on the other hand, consists of parallel bundles of myelinated axons. There are relatively few capillaries and very little extracellular space. The function of the white matter is largely conductive, and therefore it has no synapses, no dendrites and limited blood supply [20].

1.1.6 Tissues tested in current study

Some recent high-throughput DNA arrays studies of gene expression have been aimed at characterizing healthy tissue transcription patterns. One of these examined the transcription profiles in 40 human tissues (including 28 tissues of normal state and 12 cancer tissues and cell lines), utilizing 12,000 oligonucleotide probe sets. cDNA arrays have also been used to examine expression of over 23,000 genes across normal human tissues [21]. These, as well as other surveys on normal tissues [16, 22] were limited to only the more well-characterized genes. Studies on a more complete gene set focused on a comparison between diseased and non-diseased states [23-25].

In this study, we investigated and analyzed the gene expression profile of normal human tissues. Specifically, we used the Affymetrix' GeneChip technology to monitor the mRNA abundance of 62,839 gene and EST representations (Affymetrix arrays HG-U95(A-E)), in 12 normal human tissues (see section 2.1). Next, we will describe the main characteristics of each of the tissues tested in the current study. The location of each of the tissues in the human body is shown in Figure 1.

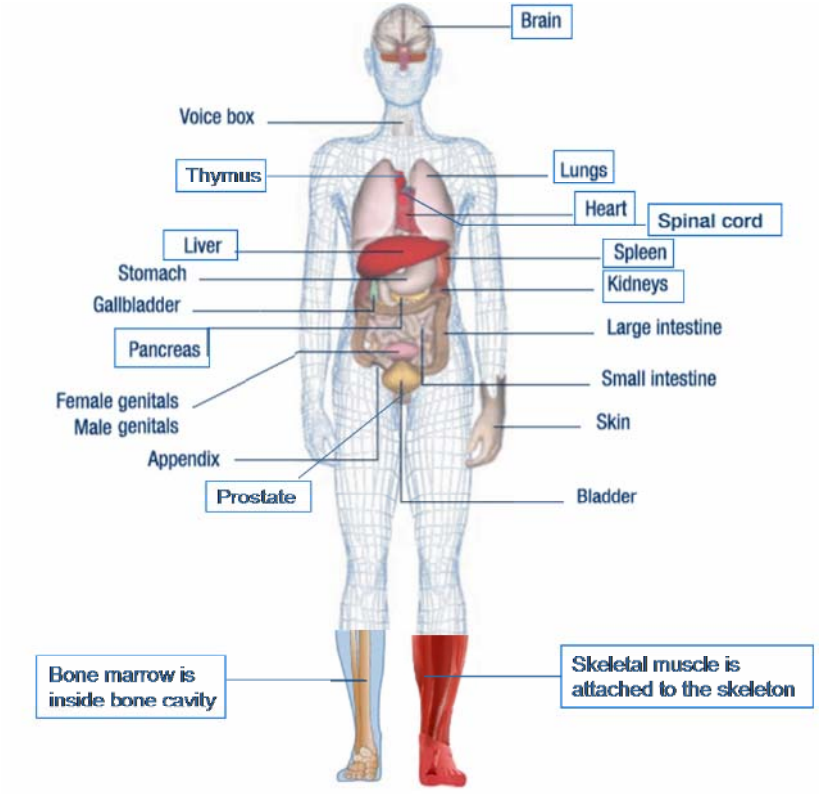


Figure 1: Location of the 12 tested tissues in the human body. Note that the spinal cord is behind the seen organs.

1.1.6.1 Brain

As mentioned above, the brain receives and integrates sensory information regarding the internal and external environments of the organism, and it responds by transmitting appropriate signals to effector organs.

The brain is largely divided into three parts: the cerebrum, the cerebellum and the brain stem. These three main parts can be further divided into smaller subparts. The cerebrum and the cerebellum are composed of two hemispheres. The outer layer of the brain, the cortex, is composed of gray matter (see section 1.1.5.4). The cortex is highly convoluted, which increases the surface area. The brain stem includes the main sensory and motor tracts of the brain, concentrated in a cylindrical mass of white matter. These tracts are tapered caudally to form the spinal cord [4, 20].

1.1.6.2 Spinal cord

The spinal cord consists of bundles of axons having specific functions, either motor or sensory, for example, pain, touch etc. The spinal cord consists of an outer layer of white matter, and an H-shaped inner layer of gray matter (see section 1.1.5.4). In the center of the inner layer, there is a small canal lined with glial cells [20].

1.1.6.3 Skeletal muscle

Most skeletal muscles are attached to the skeleton, and respond to conscious control. The skeletal muscle fiber is typically a giant, multinucleated cylindrical cell, enclosed in a cell membrane, called sarcolemma. The fiber is filled with myofibrils, which are fibers that contain contractile units. A contractile unit is made of a cylindrical column including 1000-2000 filaments of actin (thin filaments) or myosin (thick filaments). The myofibrils create the striation pattern of the skeletal muscle. Fiber lengths range between 1-40mm, and the diameter varies from 10-100μm. There are two types of skeletal muscles: slow-acting and fast-acting. The slow-acting muscles have more mitochondria and a richer vascularity than the fast-acting muscles. They are slow and do not fatigue easily, in contrast to the fast-acting muscles that fatigue rather quickly. In humans, all muscles contain both muscle-types, with one type being dominant [20].

1.1.6.4 Heart

The heart, acting as a pump, supplies the propelling force for the circulation of the blood. The heart is built of four segments: two atria and two ventricles. All segments of the heart are lined with three layers of cells: the endocardium, composed of endothelial cells and connective tissue; the myocardium, a circular layer of cardiac muscle; and the epicardium, composed of connective tissue fibers. Four fiborous valves guard the cavities of the heart. The valves are actually folds of endocardium enclosing a central core of collagenous and elastic fibers.

The cardiac muscle contracts to pump blood through the cardiovascular system. In comparison to skeletal muscle, the myofibrils of the cardiac muscle are more delicate, making the striations of the cardiac muscle less prominent. Cardiac fibers are relatively short (50-100μm), and branch to form a complicated network. The walls of the heart are rich in blood capillaries, located between the individual cardiac fibers. The high metabolic needs of the heart are also reflected in the abundance of mitochondria in cardiac cells [20].

1.1.6.5 Bone marrow

In adults, red and white blood cells (all of mesodermal origin) are formed in the bone marrow. The marrow in the cavities of most long bones becomes inactive by the age of 20, and is infiltrated with fat. About 75% of the stem cells in the marrow are white-cell producing. The bone marrow contains multipotent uncommitted stem cells

that differentiate into committed stem cells, and committed stem cells that differentiate into mature blood cells [26].

1.1.6.6 Spleen

The spleen, acting as a pump, is a blood filter that removes abnormal red blood cells from the bloodstream. It also plays a significant role in the immune system, housing many macrophages that cleanse the blood of cellular debris, parasites and pathogenic bacteria. The spleen is a hematopoietic organ, producing lymphocytes and monocytes in the adult. Last, the spleen is a reservoir of blood and platelets.

The spleen is the largest lymphoid organ in the human body. The capsule of the spleen consists of a dense collagenous shell, rich with elastic fibers and smooth muscle components. The splenic pulp has two areas: the white pulp and the red pulp. The white pulp consists mainly of small clusters of lymphocytes arranged around germinal centers. The red pulp consists of large, thin-walled sinuses filled with blood and cords of lymphoid tissue [20, 26].

1.1.6.7 Thymus

The principal function of the thymus gland is T-cell production, especially in late fetal life and early childhood. In adulthood, most of the thymus turns into adipose tissue (a specialized connective tissue that functions as a storage site for fat). The thymus does not contain germinal centers. T-cells are seeded by the lymph and blood streams into other lymphoid organs, where they proliferate.

The thymus is composed of two lobes, surrounded by a thin connective tissue capsule. Blood and lymph vessels and nerves penetrate into the gland. The parenchyma (the secretory part of the thymus) of each lobe is divided into the cortex, composed mainly of lymphocytes, and the inner medulla, with a different cell composition and large blood vessels. The inner medulla contains less lymphocytes and more epithelial-reticular cells: cells that have features of epithelial cells, and of connective tissue. Morphologically, the epithelial-reticular cells resemble connective tissue (reticular) cells of mesenchymal origin, however, they arise from endoderm. Both the medulla and the cortex contain a scattering of macrophages, mast cells and in adulthood, adipose cells [20, 26].

1.1.6.8 Kidney

The kidney is a bean-shaped organ, performing excretory, homeostatic and endocrine functions. There are two kidneys in the human body. In the kidneys, the plasma is filtered, such that its volume is reduced and its composition is altered. As endocrine organs, the kidneys secrete renin (causing an increase of blood pressure) and erythropoietin (accelerating erythropoiesis). Kidneys are of mesodermal origin.

The human kidney contains around one million functional units called nephrons. Each nephron is composed of an individual renal tubule and a glomerulus. The nephron includes at least 12 different cell types, and over 10,000 cells [2]. The glomerular capsule is a double layered membrane composed of two layers of squamous epithelium, with a subcapsular space between them, where the urine accumulates before it drains into the renal tubule. The renal tubule involves the processes of reabsorption and secretion. It is a continuous tube, consisting of different cell types in different areas. The epithelium lining the lumen of the tubule has a brush border of microvilli [20].

1.1.6.9 Liver

The liver is the largest gland in the human body, about 2% of the total body weight. The liver is both endocrine and exocrine. Its exocrine secretion is bile. Its endocrine secretions, released into the bloodstream, include glucose (largely derived from glycogen), lipoproteins and plasma proteins. The liver creates an out-pocketing to the primitive gut via the common bile duct that empties into the duodenum [20].

The liver is organized in lobules, within which blood flows past hepatocytes (liver cells). The hepatic lobules are the anatomical units of the liver. The hepatocytes are polygonal cells, capable of varied functions, including synthesis and secretion of bile; storage of glucose, glycogen, fats etc.; detoxification of metabolic wastes and more [20]. Each hepatocyte is apposed to several bile canaliculi (minute canals). The canaliculi drain into intralobular bile ducts to form the hepatic ducts. These ducts join together outside the liver to form the common bile duct [26].

The endocrine activity of the liver involves mainly protein production. Several proteins are produced by hepatocytes, among which are fibrinogen, prothrombin and albumin. These proteins are secreted into the bloodstream continuously, in contrast to other glands, in which the secreted substances are stored in secretory granules within the cells.

1.1.6.10 Pancreas

The pancreas is a large, soft gland located in the upper abdominal cavity. The pancreas is both an exocrine and an endocrine gland. Most of the gland consists of acini – the exocrine portion; small, scattered clusters of endocrine cells, called the islets of Langerhans, comprise about 2% of the pancreas' volume.

The exocrine pancreas composes a system of excretory ducts that join the common bile duct and reaches the deodenum. The ducts are lined with an epithelial tissue, and a thin layer of smooth muscle cells. The pancreas also contains glandular acini, each containing epithelial membrane cells surrounding a small lumen. The acinic cells secrete a variety of enzymes into the lumen of the acini by exocytosis.

The endocrine pancreas, the islets of Langerhans, secretes two hormones: insulin and glucagon, both involved in carbohydrate metabolism by regulating blood sugar level. A typical islet is composed of a few hundreds of cells. The islets are penetrated by rich capillaries, since the secretions enter the bloodstream [20].

1.1.6.11 Prostate

The prostate is a chestnut sized gland in the male reproductive system. It produces and stores prostatic fluid, which makes up most of the semen. The prostate is enclosed by a fibromascular capsule. The stroma of the prostate is mostly smooth muscle fibers, also rich with collagenous and elastic fibers. The parenchyma is divided into 30 or more tubuloalveolar glands, whose ducts empty to the urethra [20].

1.1.6.12 Lung

The lung is a gas-exchanging organ, part of the respiratory system. There are two lungs in the human body, each composed of bronchi, bronchioles, alveolar ducts, alveolar sacs and alveoli, accompanied by blood and nerve supplies. The primary bronchi enter the lung and branch into secondary and tertiary bronchioles. The bronchi and bronchioles are generally composed of the same cell types. They contain, among other cell types, respiratory epithelium, collagenous connective tissue bands, cartilage rings and smooth muscle. Each terminal bronchiole divides into two or more respiratory bronchioles. The walls of the respiratory bronchioles are interrupted with alveoli. This is the site of O_2 and CO_2 exchange.

The alveolus is the functional unit of gas exchange. There are about 300 million alveoli in humans, surrounded by pulmonary capillaries. The total area of the alveolar

walls in contact with capillaries in both lungs is $70m^2$. The alveoli are lined with two types of epithelial cells. Type I cells are flat, with large cytoplasmic extensions, and are primary lining cells. Type II cells (granular pneumocytes) are thicker and secrete *surfactant*, a stabilizing surface-active material. The lung also contains other special types of epithelial cells, pulmonary alveolar macrophages (PAMs), lymphocytes, plasma cells and mast cells [20, 26].

1.2 DNA microarrays

DNA microarrays are powerful tools to study gene expression, genotypes, gene mutations and location analysis [27, 28]. The primary use of DNA array technologies is gene expression monitoring [1, 29]. Arrays of nucleic acids have been used for many years, but it is only in the last few years that it has become possible to miniaturize nucleic acid arrays, and monitor the abundance of tens of thousands of mRNA molecules simultaneously. The ability to look at an enormous number of genes in parallel gives a broad viewpoint, allowing one to inspect non-trivial hypotheses. There are generally two DNA array technologies: high-density oligonucleotide arrays and complementary DNA (cDNA). We will focus on the first array type, which was used in this study.

1.2.1 High-density oligonucleotides arrays

1.2.1.1 The array

An oligonucleotide array is composed of hundreds of thousands of probe cells, placed on a glass surface. Each probe cell contains copies of an oligonucleotide probe, representing a specific gene. The probes are designed to be specific to the gene they represent.

Probe cell - each cell consists of $\sim 10^7$ copies of an oligonucleotide probe, typically 25 nucleotides in length, that are synthesized base by base in a defined area on the glass surface [30]. Probe arrays are manufactured in a series of cycles, by photolithography (Figure 2). Initially, a glass substrate is coated with linkers containing photolabile protecting groups. Then, a mask is applied that exposes selected portions of the probe array to ultraviolet light. Illumination removes the photolabile protecting groups enabling selective nucleoside phosphoramidite addition only at the previously exposed sites. Next, a different mask is applied and the cycle of illumination and chemical

coupling is performed again. By repeating this cycle, a specific set of oligonucleotide probes is synthesized with each probe type in a known location. Through successive steps, any sequence can be built up in any position on the chip. The number of cycles is determined by the length of the oligonucleotide probe.

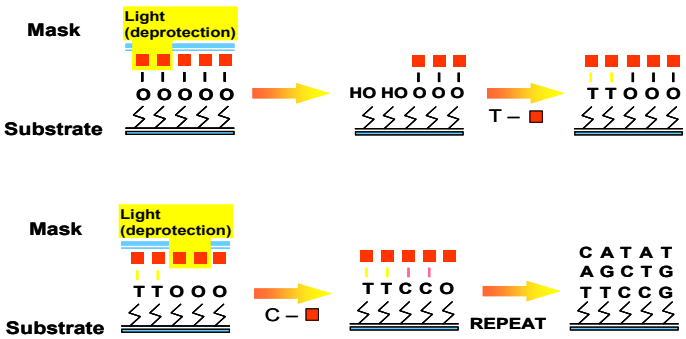


Figure 2: Synthesis of oligonucleotide arrays using photolithography. The glass surface is covered by a protective layer that is susceptible to light. Light is directed through a mask to activate selected sites, and a specific nucleotide is added. The process is repeated for the four bases, and for the successive positions on the array, until the sequence of the oligonucleotide is completed.

Figure 3: Probe set design. The probe sets are taken from the 3' end of the mRNA sequence. A probe set is typically 16-20 probe pairs, where each pair is composed of a PM probe, complementary to the mRNA sequence, and a MM probe, that has a mismatch in the central position of the sequence. Each probe is typically 25 nucleotides long.

There are two kinds of probe cells [31]:

Perfect Match (PM) – a probe that was designed to be complementary to the reference sequence. The PM probes are taken from the 3' end of the gene.

Mismatch (MM) - a probe that was designed to be complementary to the reference sequence, except for one mismatch at the central position. The MM is used for assessing non-specific hybridization to the sequence. On the array, the PM cell is located directly above the MM cell.

A PM and its corresponding MM probe are called a Probe Pair (PP). A Probe Set includes a series of probe pairs (usually 16-20 probe pairs) and represents an expressed transcript (Figure 3).

1.2.1.2 Target preparation

The target cRNA is prepared as follows (Figure 4):

- RNA is isolated from a sample of a certain cell type or tissue (either total RNA or mRNA).
- mRNA is reverse transcribed into complementary DNA (cDNA).
- The complementary strand is synthesized to create a double stranded cDNA (DScDNA).
- An *in vitro* transcription reaction (IVT) using biotinylated nucleotides is done to both amplify and label the transcripts, resulting in biotin–labeled cRNA.
- The cRNA is fragmented in order to get a more efficient hybridization (the goal being fragments of 50-200 base pairs in length [29]).

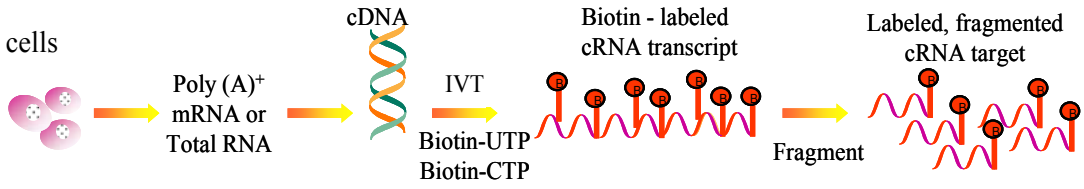


Figure 4: Target preparation. RNA is extracted from the cells and reverse transcribed to cDNA, and then to DScDNA. Biotinylated nucleotides are used to amplify and label the transcripts, resulting in biotin-labeled cRNA. The labeled cRNA is fragmented to create 50-200 base pairs long molecules.

1.2.1.3 Hybridization and scanning

The labeled cRNA target is hybridized to the array (Figure 5); the array is stained by a fluorescent dye and scanned to get a quantitative fluorescence image.

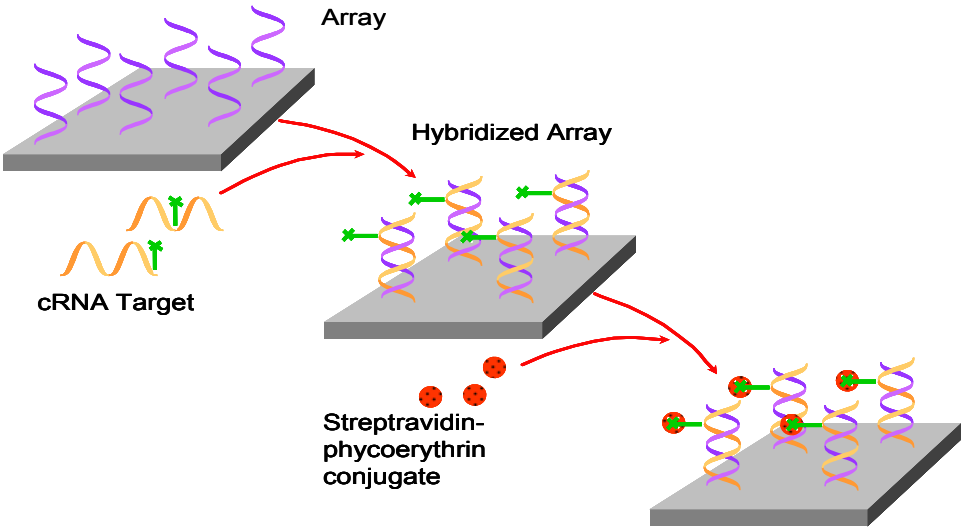


Figure 5: Hybridization procedure. The labeled and fragmented cRNA target is hybridized to the array. The hybridized array is stained with a fluorescent dye.

1.2.2 Probe set summary by Micro Array Suite 5.0

In order to derive biologically meaningful results from the hybridization intensities measured by the probe array, the intensity values of each probe set must be summarized into one number, representing the amount of bound mRNA transcript that was measured in the experiment. Several low-level analysis methods for Affymetrix' GeneChips results have been proposed [27, 32-34]. Affymetrix' Micro Array Suite, version 5 (MAS 5.0) [30] software is the most commonly used.

We will next describe the main steps taken by the MAS 5.0 software.

1.2.2.1 Probe cell intensity

The intensity of a probe cell represents the hybridization level of the target. The image of each probe cell is composed of ~7*7 pixels. To calculate probe cell intensity, the bordering 24 pixels of the cell are excluded, and the intensity value associated with the 75 percentile of the remaining pixels is used as the probe cell intensity.

1.2.2.2 Background calculation and subtraction

The background is the signal intensity caused by autofluorescence of the array surface and non-specific binding of target or stain molecules. The background establishes a "floor" to be subtracted from each probe cell intensity value. The array is divided into K equally spaced zones (default K=16). The cells are ranked according to their intensity and the lowest 2% are chosen as the background *b* for that zone (bZ_k). The average background is assigned to the center of each zone. Distances are computed from each cell (*x,y*) on the chip to the various zone centers. A weighted sum is then calculated as follows:

$$w_k(x,y) = \frac{1}{d_k^2(x,y) + smooth}$$

where d_k is the distance between the probe cell (*x,y*) and the center of zone *k*, and *smooth* is a small factor added to d_k^2 to ensure that the value will never be zero. The default value of *smooth* is 100.

The background *b* value to be used for cell (*x,y*) is therefore given by:

$$b(x,y) = \frac{1}{\sum_{k=1}^{K} w_k(x,y)} \sum_{k=1}^{K} w_k(x,y) \times bZ_k$$

The background *b* is subtracted from each probe cell intensity value [30].

1.2.2.3 Noise correction

Noise results from small local variations in the signal observed by the scanner as it samples the probe array's surface. For noise correction, a local noise value *n* based on the standard deviation of the lowest 2% of the background in that zone (nZ_k) is calculated and weighted for background values. The noise *n* value to be used for cell (*x,y*) is therefore:

$$n(x,y) = \frac{1}{\sum_{k=1}^{K} w_k(x,y)} \sum_{k=1}^{K} w_k(x,y) \times nZ_k$$

In the next step, a threshold is set at some fraction *NoiseFrac* of the local noise value (default *NoiseFrac* = 0.5), so that no value is adjusted below that threshold. That is, for a cell intensity *I'*(*x,y*) at chip coordinates (*x,y*), we compute an adjusted intensity *A(x,y)*:

$A(x, y) = \max(I'(x, y) - b(x, y), NoiseFrac*n(x, y))$

where $b(x, y)$ is the background level, $n(x, y)$ is the local noise level and *NoiseFrac* is the selected fraction of the local noise value [30].

1.2.2.4 The Expression Value (Signal)

The MAS 5.0 algorithm distinguishes array-wide constant background from stray signal. The stray signal is the non-specific hybridization to a probe on the array. It is unique to each PM probe, and is estimated by the MM probe [33]. The MM probe provides a value that comprises most of the background cross-hybridization and stray signal affecting the PM probe. The ideal situation would be that when a transcript is absent, the intensity of the PM probe would equal to that of the MM probe, and when a transcript is present, the intensity of the PM probe would be higher than that of the MM probe, and proportional to the concentration of the mRNA transcript. However, this is not always the case. There are MM probes with intensity higher than that of their corresponding PM probes. In such cases it does not make sense to use the MM to estimate the amount of stray signal in the PM intensity. Instead, an idealized value (Ideal Mismatch - IM) can be estimated, based on the whole probe set or on the behavior of probes in general.

To calculate a specific background ratio representative for the probe set, the one-step Tukey biweight algorithm (T_{bi}) is used [30]. The Tukey biweight is a robust average that is unaffected by outliers.

The biweight specific background (*SB*) for probe set *i* is:

$$SB_i = T_{bi}(\log_2(PM_{i,j}) - \log_2(MM_{i,j}) : j = 1,..., n_i)$$

where n_i is the number of probe pairs in probe set *i*.

If SB_i is large, then the probe set values are generally reliable, and SB_i can be used to construct the ideal mismatch (*IM*) for a probe pair if needed. If SB_i is small (less than an arbitrary defined *contrast* τ), a value based on the PM intensity is used as the ideal mismatch. The three cases of determining ideal mismatch *IM* for probe pair *j* in probe set *i* are described by the following formula:

$$IM_{i,j} = \begin{cases} MM_{i,j}, & MM_{i,j} < PM_{i,j} \\ \dfrac{PM_{i,j}}{2^{(SB_i)}}, & MM_{i,j} \geq PM_{i,j} \text{ and } SB_i > constant\ \tau \\ \dfrac{PM_{i,j}}{2^{\left(\frac{constant\ \tau}{1+\left(\frac{constant\ \tau - SB_i}{scale\ \tau}\right)}\right)}}, & MM_{i,j} \geq PM_{i,j} \text{ and } SB_i \leq constant\ \tau \end{cases}$$

Scale τ is the cutoff that describes the variability of the probe pairs in the probe set. default *contrast* τ = 0.03, default *scale* τ = 10

Given the ideal mismatch value, the probe value (*PV*) is calculated as follows:

$V_{i,j} = \max((PM_{i,j} - IM_{i,j}), \delta)$ default $\delta = 2^{(-20)}$

$PV_{i,j} = \log_2(V_{i,j})$

The absolute expression value for probe set *i* is then computed as the one-step biweight estimate of the adjusted probe values:

$SignalLogValue = T_{bi}(PV_{i,1}, PV_{i,2}, ..., PV_{i,n_i})$

1.3 Clustering of Gene Expression Data

1.3.1 Cluster analysis

Cluster analysis deals with the problem of identifying the underlying structure of a set of data points by partitioning it into groups. The clustering problem may be stated as follows: given *N* points in a *d*-dimentional space, determine the partition of the points into *M* groups, called "clusters" such that points that belong to the same cluster are more similar to each other than to points that belong to different clusters [35]. Note that by its nature, the clustering problem is not well defined. This gives rise to various clustering methods.

Clustering is a special kind of classification. Classification methods can be divided into two general classes, designated supervised and unsupervised. In supervised classification, category labels denoting *a priory* partition of the data points are used, and the problem is to establish a discriminant that separates the data points according to their categories. This discriminant is then used to partition new unfamiliar data. In unsupervised classification, also termed unsupervised clustering, no category labels are used [36, 37].

Unsupervised clustering methods can be further classified into hierarchical and non-hierarchical (partitional) methods according to the type of structure imposed on the data. A hierarchical clustering method organizes the data into a nested sequence of groups according to their similarity. The resulting tree-shaped graph (dendrogram) enables one to see how data points are being merged into clusters at successive levels of proximity. A partitional classification, on the other hand, is a single partition of the points in an attempt to recover natural groups present in the data. Partitional clustering methods assume that the data can be divided into a given number of clusters and that the clusters are well separated.

The main advantage of hierarchical clustering methods is the ability to view the data at different resolutions when there is no *a priory* knowledge of the number of expected clusters, or when one is interested in several resolution levels. The main advantage of using partitional clustering methods is that large datasets can be clustered much faster than by using hierarchical clustering [37].

1.3.2 Clustering of DNA microarrays results

The enormous amount of data obtained using the DNA microarray technology poses a challenge of interpreting the results. There are two approaches to this problem. When there are specific questions that a researcher is interested in, a statistical approach of hypothesis testing is appropriate. However, when trying to generate new hypotheses, one is interested in the underlying structure of the data. In this case, cluster analysis is more suitable.

There are two ways to cluster gene expression data: one is to cluster the genes according to their expression over the different samples, and the other is to cluster the samples according to their expression profiles over the set of genes. The main problem in the analysis of microarray data is that each of the biological processes of interest may involve a relatively small subset of genes, while the remaining genes behave in a way that is uncorrelated with the signal of this small subset. The contribution of the relevant genes is often dominated by the random signal of the larger irrelevant set, resulting in low *signal-to-noise ratio*. The same may apply to samples, where the cellular process studied may take place in only a subset of the samples [38].

Another difficulty is that the number of clusters is usually not known in advance. Clusters may also be of irregular shapes, and there may be interesting clusters at different resolutions.

We will next describe clustering algorithms used in the current study: K-means, Hierarchical Clustering and Super Paramagnetic Clustering (SPC).

1.3.3 K-means Clustering

K-means is a partitional clustering algorithm; it finds a partition of a given set of data points into K clusters. The value of K is determined *a priory*, and a clustering criterion must be adopted [37]. The selected criterion depends on one's notion of what constitutes a cluster. There is no unique best criterion since the clustering problem itself is not well defined. Clusters can be of arbitrary shapes and sizes in a multidimensional data space. The most common partitional clustering strategy is to minimize the sum of squared distances between all data vectors and the center of their cluster (centroid).

The K-means algorithm finds k centroids $\mu_1, \mu_2, \ldots, \mu_k$ representing the K clusters [39]. Let us denote the given set of data points as $x_1,\ldots,x_n$ and the desired number of clusters K. The set of populated clusters will be $\{C_1,\ldots,C_k\}$. The algorithm is then:

```
1  initialize μ1, μ2, …, μk to the values of random data points
2      repeat
3          for each data point x
4              find i such that distance(x, μi) is minimal
5              move x to Ci
6          for each centroid μi
7              let μi be the center of Ci
8      until no change in μi
9  return μ1, μ2, …, μk, {C1, C2, …, Ck}
```

In practice, the number of iterations is generally much less than the number of points. The iterative procedure minimizes the chosen criterion, such as the squared-error criterion function:

$$E^2 = \sum_{k=1}^{K} \sum_{i=1}^{n_k} || x_i - \mu_k ||^2$$

where in the inner sum the index i runs over those x that were assigned to C_k.

1.3.4 Hierarchical clustering

As mentioned above, hierarchical clustering divides a set of data points into a sequence of nested clusters. Suppose we have *n* data points. Let us consider a sequence of partitions of these points into clusters. At the highest resolution, we will partition the data points to *n* clusters, each containing a single point. We can get a partition into *n-1* clusters by merging the two nearest clusters. Another merge will give a partition into *n-2* clusters and so on. The n^{th} partition will be a single cluster that contains all the data points. The most natural way to represent a hierarchical clustering is a tree that shows the cluster merging steps (a *dendrogram*).

Hierarchical clustering algorithms can be divided to two classes: agglomerative (bottom-up) and divisive (top-down). Agglomerative algorithms start with *n* clusters and keep merging them until a single cluster is obtained. Divisive algorithms start with a single cluster that contains all data points, and keep splitting clusters until each cluster consists of a single point. In this study, agglomerative hierarchical clustering was used.

Let us denote the given set of *n* data points $x_1, \ldots, x_n$. The algorithm is as follows:

1 set $\boldsymbol{C_i}$ to be $\{\boldsymbol{x_i}\}, (\boldsymbol{i = 1, \ldots, n})$
2 repeat $\boldsymbol{n-1}$ times
3 find $\boldsymbol{i, j}$ such that distance($\boldsymbol{C_i}$, $\boldsymbol{C_j}$) is minimal
4 merge $\boldsymbol{C_i}$ and $\boldsymbol{C_j}$ to form the next partition

The distance between the clusters may be defined in different ways. In the present work, *average linkage* was employed as a distance function. Average linkage is defined as the average Euclidean distance between all pairs of points in cluster c_i and cluster c_j by the following formula:

$$dist(C_i, C_j) = \frac{1}{n_{C_i} n_{C_j}} \sum_{\substack{x \in C_i \\ y \in Cj}} \| x - y \|$$

where n_{C_i} and n_{C_j} are the number of data points in clusters C_i and C_j, respectively.

1.3.5 Super-Paramagnetic Clustering (SPC)

SPC is an unsupervised hierarchical clustering method based on physical properties of inhomogeneous ferromagnets [35]. We will briefly present the main characteristics of SPC. Full details of the algorithm and underlying physical model are beyond the scope of the present work, and are described elsewhere [35, 40].

The SPC algorithm assigns a small magnetic element (a *potts spin*) to each data point. It introduces an interaction between neighboring points, whose strength is a decreasing function of the distance between them. The spin-spin correlation function is used to partition the spins and the corresponding data points into clusters. The magnetic system exhibits three temperature (T) dependent phases. At T = 0, the system is completely ordered; all spins are aligned and all data points form a single cluster. At a very high temperature (T_{max}), the system does not exhibit any order, and each data point forms its own cluster. In an intermediate temperature regime, clusters of strongly interacting spins are ordered, while spins of different clusters are uncorrelated. In this phase, meaningful clusters may be found, reflecting the inherent structure of the data [40].

The range of temperatures ΔT over which a cluster remains unchanged serves as a measure for the relative stability of clusters. The threshold value for ΔT above which a cluster is considered stable should be a significant fraction of T_{max}.

Using SPC in microarray analyses has several advantages: the number of clusters is not required in advance, as is needed for partitional clustering algorithms; SPC relies on proximity between points and does not assume a particular shape of a cluster (in contrast to K-means, for example, that by using cluster-centroids assumes a spherical shape of the clusters); SPC is stable against noise; it generates a dendrogram, which allows finding meaningful partitions of the data at different resolutions; SPC can rate the stability of a cluster.

1.4 GeneAnnot – Microarray Gene Annotation

Each probe set on the Affymetrix GeneChip® array was designed to represent a certain gene sequence. In some cases, several probe sets on the array correspond to the same gene. On the other hand, some probe sets are not specific and can be aligned to more than one gene. The result is many-to-many relationships between probe sets and genes [41].

Affymetrix array set U95A-E was designed to include gene representations of the entire human genome. Some of the genes are well characterized, while others are novel genes for which little information is available. Many of the probe sets representing such novel genes are derived from Expressed Sequence Tags (ESTs), which are not always reliable indicators of mRNA identity. Probe set annotation provided by Affymetrix includes information about the sequence from which the probes were taken, but the specificity of the probes to the gene they represent is not provided [42].

The GeneAnnot system [41, 42] was developed to explore and document the many-to-many relationships between probe sets and genes' sequences (Figure 6). GeneAnnot deals with the challenge of improving the annotation of DNA microarrays and providing qualitative assessment to the various probe sets [41, 42].

The GeneAnnot procedure uses direct sequence comparison of all the Perfect Match (PM) probes on the array to RefSeq, Ensembl and GenBank mRNA sequences (allowing one mismatch), using the Blat algorithm [43] (Figure 6). Each probe set / gene pair receives a score indicating the sensitivity (*Sn*) and specificity (*Sp*) of the relation [41] (Figure 7).

The sensitivity score gives a measure to how well a probe set aligns to a gene. The sensitivity score *Sn* for probe set *i* and gene *j* is calculated as follows:

$$Sn_{i,j} = \frac{Nm_{i,j}}{Nprobes_i}$$

where $Nm_{i,j}$ is the number of probes in probe set *i* that matched gene *j*, and $Nprobes_i$ is the total number of probes in probe set *i*.

The specificity score gives a measure to the exclusivity of the alignment of a probe set to a gene. The specificity score *Sp* for probe set *i* and gene *j* is calculated as follows:

$$Sp_{i,j} = \frac{1}{Nm_i} \sum_{k=1}^{Nm_{i,j}} \frac{1}{Nm\_g_k}$$

where Nm_i is the number of probes in probe set *i* that had any matching gene, and

Nm_g_k is the number of genes matched to probe k of probe set i. The sum runs over the probes $k = 1,...,Nm_{i,j}$ that matched gene j.

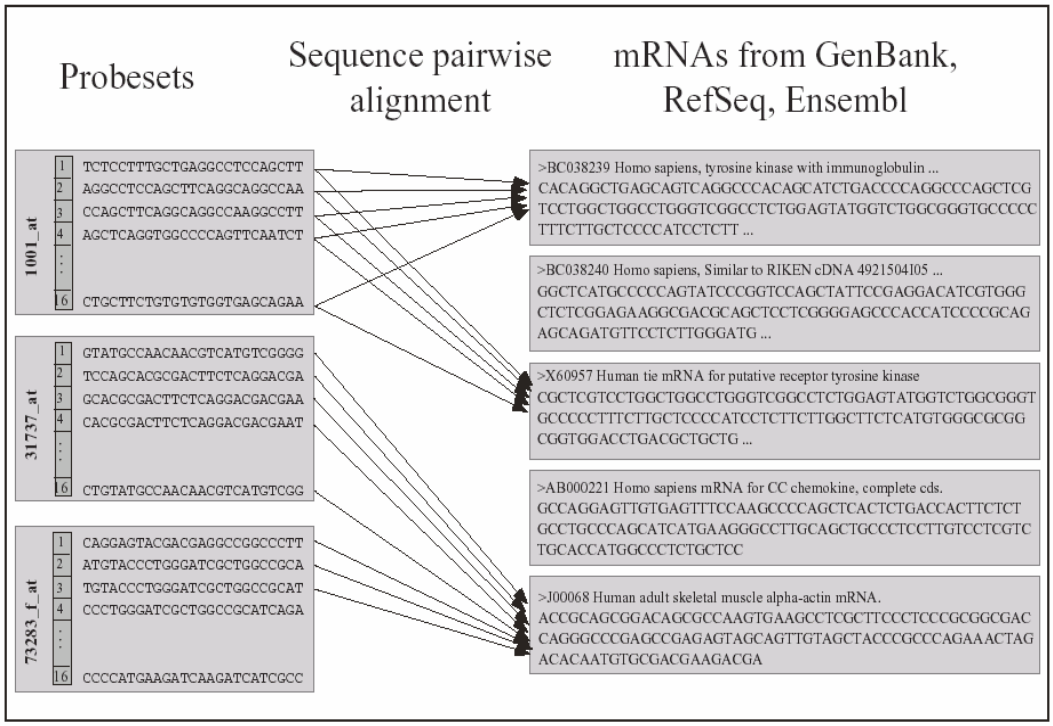
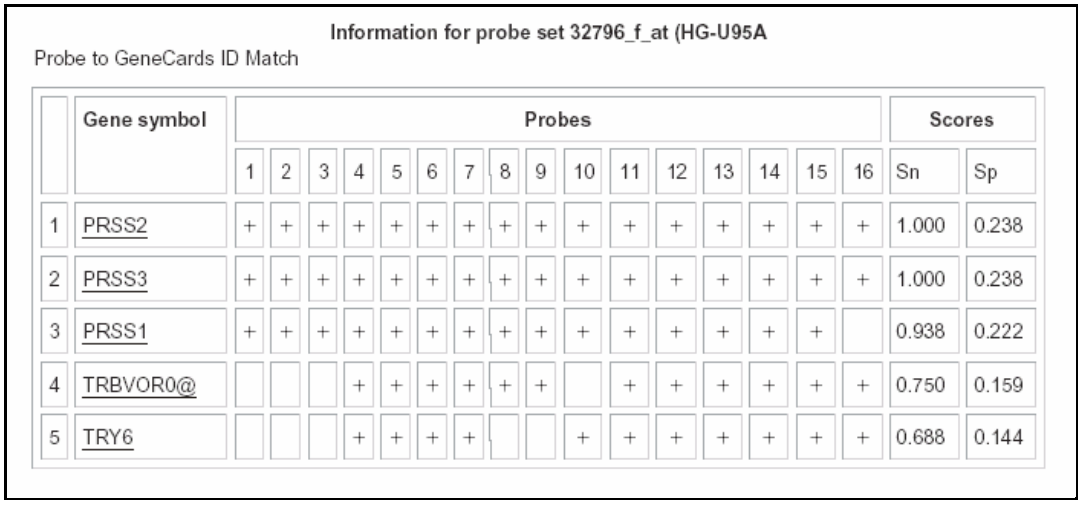


Figure 6: Direct sequence comparison of Affymetrix PM probes and public transcript sequences reveals many-to-many relationships between probe sets and genes. Sequence alignment was performed using the Blat algorithm [43], allowing up to one mismatch [41].



Information for probe set 32796_f_at (HG-U95A

Probe to GeneCards ID Match

	Gene symbol	Probes															Scores	
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Sn	Sp
1	PRSS2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	1.000	0.238
2	PRSS3	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	1.000	0.238
3	PRSS1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		0.938	0.222
4	TRBVOR0@				+	+	+	+	+	+		+	+	+	+	+	+	0.750	0.159
5	TRY6				+	+	+	+			+	+	+	+	+	+	+	0.688	0.144

Figure 7: Genes associated with the probe set 32796_f_at and their sensitivity (*Sn*) and specificity (*Sp*) scores. A plus sign (+) indicates that the probe matched at least one of the mRNA sequences associated with the gene [41].

2 Methods

Sections 2.7, 2.9-2.11 describe methods developed in our lab: the 'gap' criterion (section 2.7) and binary classification (section 2.9) were developed by Dr. Itai Yanai; the tissue specificity index (τ) (section 2.10) was developed by Arren Bar-Even; a method for determining expression profiles of genes was developed in the course of this study (section 2.11).

2.1 RNA samples and array hybridizations

The expression intensity of mRNA from 12 normal human tissues was assayed by 62,839 probe sets across five microarrays (Affymetrix GeneChips U95A-E), in replicates. RNA samples from human tissues were purchased from Clontech (Palo Alto, CA). This collection of major human tissues includes: bone marrow, brain, heart, kidney, liver, lung, pancreas, prostate, skeletal muscle, spinal cord, spleen and thymus. Each RNA sample was typically composed of a pool of 10-25 individuals. Replicate experiments were done independently, mostly from RNA of identical lot numbers. Exceptions are kidney, pancreas, and prostate. Aliquots of each sample (12 μg cRNA in 200 μl hybridization mix) were hybridized to a GeneChip Human Genome U95A-E array set (Affymetrix, Santa Clara, CA, USA). Preparation and hybridization of cRNA were done according to the manufacture's instructions [30].

The following abbreviations were used for tissue names: bone marrow- BMR; brain- BRN; heart- HRT; kidney- KDN; liver- LVR; lung- LNG; pancreas- PNC; prostate- PST; skeletal muscle- MSL; spinal cord- SPC; spleen- SPL; and thymus- TMS.

2.2 Expression data preprocessing

The expression value for each gene was determined using the MicrroArray Suite version 5.0 (MAS 5.0) software [27, 30, 33] with default parameters. Affymetrix MAS 5.0 intensity values (I) were normalized as follows: first, the intensity values were $\log_{10}$ transformed (MAS 5.0 intensity values ranged roughly between zero and 20,000. Intensity values of zero were substituted by 0.1). Then, the mean for the particular array was subtracted from all measurements of that array. Finally, the total (log transformed) experimental mean was added to each measurement [44]. The procedure can be formulated as follows:

$$S_{i_0,j_0} = \log_{10}(I_{i_0,j_0}) - \frac{1}{M}\sum_{i=1}^{M}\log_{10}(I_{i,j_0}) + \frac{1}{N \times M}\sum_{i=1}^{M}\sum_{j=1}^{N}\log_{10}(I_{i,j})$$

where I_{i_0,j_o} is the intensity value in probe set i_0 in array j_0 , M is the number of probe sets and N is the number of arrays.

2.3 Centering and normalization

For each probe set (or gene expression profile), the mean normalized intensity of the probe set was subtracted from each measurement such that the probe set mean was zero. The centered intensity for probe set i_0 in sample j_0 (CS_{i_0,j_0}) is given by:

$$CS_{i_0,j_0} = S_{i_0,j_0} - \frac{1}{N_s}\sum_{j=1}^{N_s} S_{i_0,j}$$

where S_{i_0,j_0} is the preprocessed signal of probe set i_0 in sample j_0 and N_s is the number of samples. Then, every centered measurement CS_{i_0,j_0} was divided by the square root of the sum of squares of the probe set, such that its norm became one:

$$NS_{i_0,j_0} = \frac{CS_{i_0,j_0}}{\sqrt{\sum_{j=1}^{N_s} CS_{i_0,j}^2}}$$

2.4 Clustering the entire dataset

Prior to the clustering procedure, the probe sets were preprocessed as described in section 2.2. Probe sets whose normalized intensity was below $\log_{10}30$ in all samples were considered "not expressed" and were filtered out. There were 16,341 such probe sets (26%). Then, each of the remaining 46,498 probe sets was centered and normalized as described in section 2.3.

As mentioned earlier, partitional clustering methods allow one to cluster a large amount of data in a relatively short time. In the current study, the K-means algorithm was used. In order to find the most appropriate value of K, the Sum of Squares of Errors (SSE) was calculated for different K values, ranging from two to 150. For each value, 20 repeats were done. Naturally, as K increases, SSE decreases. However, if there is a "natural" number of clusters in the data, we would expect to see a "break" in the graph, i.e. a change in the rate of decrease.

Figure 8 shows SSE for different values of *K* in the current study. As can be seen from the graph, there was no clear break in SSE, but a gradual decrease. Several values of *K* were tested, and *K*=60 was chosen. K-means clustering was preformed with K=60. Each of the 60 clusters was represented by its centroid. Two-way hierarchical clustering was then applied to the centroids, using the average linkage as a clustering algorithm.

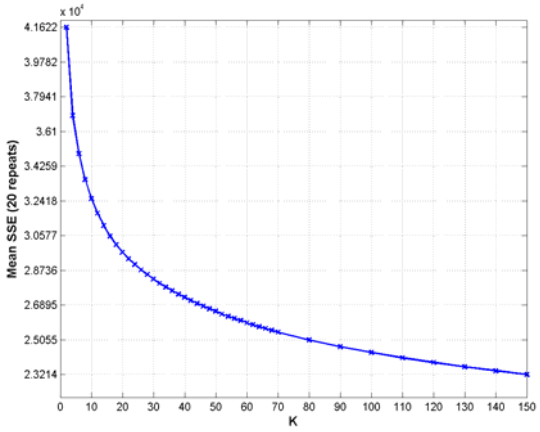


Figure 8: Sum of Squares of Error (SSE) as a function of the parameter *K*. In order to choose a suitable value for *K*, the number of clusters, we tested values of *K* ranging from 2-150, and calculated SSE. For each value of *K*, 20 repeats were done. As can be seen from the graph, no clear break in the decrease of SSE is evident.

2.5 Statistical analysis of differential expression

Single-classification ANOVA with equal sample sizes was employed on the preprocessed 24-element expression vector composed of 12 tissues in replicates. First, a threshold was used: all normalized intensities below $\log_{10}30$ were set to $\log_{10}30$. Then, for each probe set, the sum of the squares of the differences between the replicates was compared with the sum of the squares of the differences between the averages of the tissue expressions. To account for the multiple comparison problem inherent in calculating the P-values for all 62,839 probe sets, we calculated the false discovery rate of the P-values [45]. We chose a P-value cutoff of 0.0036 which estimates a 1% error rate. This resulted in 22,936 probe sets that were defined as "differentially expressed".

2.6 Signal Quantilization

The MAS 5.0 preprocessed intensities (see section 2.2) were converted into a quantile scale. The expression intensities for each tissue, averaged over the two replicates, were divided into 11 quantiles as follows. A "zero quantile" included the low intensity values, lower than or equal to $\log_{10}30$. The remaining intensities were divided into 10 "equal-area" bins, such that each bin contained the same number of intensity values.

2.7 Filtering using the 'gap' criterion

We first defined the 'gap' index for the expression profile of each probe set as the maximum difference between two neighboring values in the sorted quantile vector. When the same 'gap' was found more than once in a profile, the first gap, between the smaller neighboring values with that gap, was taken. The 'gap' criterion was used to further filter the 22,936 differentially expressed probe sets, identified using ANOVA. Those differentially expressed probe sets with a 'gap' of at least 3 were included in our analysis and are henceforth referred to as 'Mingap set'. The Mingap set was composed of 8,224 probe sets that passed the filtering procedure.

2.8 Unsupervised clustering using SPC

The Super-Paramagnetic Clustering (SPC) algorithm [35] was applied to the Mingap set (see section 1.3.5). Before clustering, each profile was centered and normalized as described in section 2.3. The SPC parameters used are detailed in Table 1.

Table 1: **Parameters used in SPC clustering of the Mingap set.**

	G1(S1)	S1(G1)
K (nearest neighbors)	27*	4*
Minimal Temperature	0	0
Maximal Temperature	0.25	0.3
Delta T	0.004	0.003
Cycles	3000	3000
Growth	TRUE	TRUE
Stable delta T	6	4
Ignore dropout size	3	1

* The parameter K was determined using the homogeneity order parameter [46, 47]

2.9 Binary classification

The 'gap' criterion was used to convert expression profiles into binary form. For each probe set in the Mingap set, tissues in which expression was detected above a minimal gap (gap = 3) were interpreted as over-expression (1), and the rest were referred to as under-expression (0). In this manner each expression profile was classified to a particular binary pattern. Next, the remaining 14,712 differentially expressed profiles (out of the 22,936 profiles identified in section 2.5) were classified to the best matching binary pattern detected by 'gap' as follows. The Euclidean distance was calculated between each of the 14,712 profiles and the mean expression profile of each of the binary classes. The class to which this distance was smallest was selected as the matching binary pattern for the profile. The binary index, I_B, corresponding to each binary pattern, was defined as the number of 1's in the pattern [48].

2.10 Tissue specificity index (τ)

The index τ was defined as: $$\tau = \frac{\sum_{i=1}^{N}(1 - x_i)}{N - 1}$$

where N is the number of tissues and x_i is the expression profile component normalized by the maximal component value [48].

2.11 Expression profiles of genes

GeneAnnot [41, 42] was used to annotate Affymetrix U95A-E probe set sequences (see section 1.4). First, probe sets with low sensitivity and specificity scores, according to GeneAnnot, were filtered out. Probe sets with $Sn \geq 9/16$ and $Sp = 1$ were kept. The 23,689 probe sets that passed our filter (38% of the 62,839 probe sets on the arrays) were preprocessed as described in section 2.2, and tissue replicates were averaged. In the next step, the genes represented by the filtered probe sets were inspected. If a gene was associated with one probe set, the expression profile of the probe set was taken as the expression profile of the gene. If a gene was associated with more than one probe set, the mRNA sequences that matched the probe sets, and the similarities between the expression profiles of the probe sets were examined as follows: each probe set may match one or more mRNA sequences. If two probe sets matched the same mRNA sequences from RefSeq / Ensembl / GenBank, or if the lists of matched sequences from these sources were overlapping, the probe sets were considered to be associated with the same mRNA transcript. The similarities between expression profiles of probe sets were evaluated using Pearson correlation coefficient (r). All pairwise correlations between probe sets representing the same gene were calculated. Probe sets' expression profiles were averaged if the correlation between them was at least 0.5, and if they were associated with the same mRNA transcript. If a gene was associated with, say, three probe sets matching the same mRNA sequence, and the correlations between them were: $r_{ps1,ps2} \geq 0.5$, $r_{ps2,\ ps3} \geq 0.5$ but $r_{ps1,ps3} < 0.5$, all three probe sets were averaged. Probe sets that were below the threshold of $\log_{10}30$ in all tissues were not used if there were other probe sets representing the gene that were above the indicated threshold in at least one tissue. The resulting matrix included 17,118 gene expression profiles.

2.12 Analysis of most divergent genes

A filtering procedure was applied to the 17,118 preprocessed gene expression profiles (see sections 2.2, 2.11). Next, a threshold was employed: all normalized intensities below $\log_{10}30$ were set to $\log_{10}30$. Two filtering criteria were used to select the most divergent profiles: 1. Standard deviation of at least 0.3; 2. Range of at least 1. Profiles that met both criteria were taken for further analysis. There were 1,950 gene expression profiles that passed the filtering procedure.

The Super-Paramagnetic Clustering (SPC) algorithm [35] was applied to the filtered set of gene expression profiles. Prior to clustering, each profile was centered and normalized as explained in section 2.3. The SPC parameters used are detailed in Table 2.

Table 2: Parameters used in SPC clustering of the most divergent genes.

	G1(S1)	S1(G1)
K (nearest neighbors)	15*	4
Minimal Temperature	0	0
Maximal Temperature	0.3	0.3
Delta T	0.004	0.004
Cycles	**	3000
Growth	TRUE	TRUE
Stable delta T	6	4
Ignore dropout size	3	1

* The parameter K was determined using the homogeneity order parameter [46, 47].
** The mean field approximation [49] was used for the operation G1(S1).

2.13 Comparison of results to a published dataset [17]

In order to test the generality of our findings, our results were compared to a published dataset, Su et al. (2002) [17], including 40 human tissues (normal tissues, cancer tissues and cell lines) and 45 mouse tissues. There were 28 human tissues of normal state: Adrenal Gland, Amygdala, Blood, Brain, Caudate nucleus, Cerebellum, Corpus callosum, Cortex, Fetal Brain, Fetal Liver, Heart, Kidney, Liver, Lung, Ovary, Pancreas, Pituitary gland, Placenta, Prostate, Salivary gland, Spinal cord, Spleen, Testis, Thalamus, Thymus, Thyroid, Trachea, Uterus. There were 86 human samples, tested using Affymetrix GeneChip U95A.

Affymetrix' raw data were analyzed and scaled by MAS 5.0 with default parameters, except for the scaling factor which was set to 200. MAS 5.0 intensity values ranged roughly between zero and 60,000. Intensity values of zero were substituted by 0.1. The scaled intensity values were $\log_{10}$ transformed.

In order to compare the results of the two datasets, the GeneNote results were preprocessed (as described in section 2.2) for chip A separately.

There were 10 tissues common to the two experiments: Brain, Heart, Kidney, Liver, Lung, Pancreas, Prostate, Spinal cord, Spleen and Thymus. There were 12,533 probe sets common to the two datasets (chip U95A).

3 Results

3.1 Clustering the entire dataset

As mentioned earlier, clustering deals with the problem of identifying the structure of data by partitioning it into groups. In order to identify groups of genes that would allow us to distinguish between different human tissues, and in order to find groups of tissues that are related in terms of their expression profiles, two-way clustering was performed. We used a two-step procedure to cluster the 46,498 probe sets that were expressed in at least one tissue (see section 2.4): in the first step, we used the K-means algorithm with K=60 to cluster the probe sets into 60 groups. Each cluster was represented by its centroid. In the second step, two-way hierarchical clustering was used on the 60-centroid representations of the clusters found by K-means. The reordered expression matrix according to the clustering result is presented in Figure 9A.

The dendrogram of samples according to the hierarchical clustering is presented in Figure 9B. Tissue replicates were the most similar, and were clustered together for all tissues. Four groups of tissues were found (marked in colored arrows): 1. brain and spinal cord (nervous system, red), 2. skeletal muscle and heart (muscle groups, orange), 3. spleen and thymus (blood-related, gray) and 4. liver and kidney (magenta). These tissue-groups are biologically meaningful, as will be discussed in section 4.

Cluster size ranged between 502 and 1,149 probe sets per cluster. Four types of clusters were found (Figure 10), identified by simply observing the expression patterns of each cluster. In a later stage, we will use a more quantitative way to assess the clusters' expression patterns. The first cluster type included probe sets that were higher in one tissue relative to the other tissues (Figure 10A). These probe sets represent one-tissue specific genes. There were 10 clusters of one-tissue specific probe sets, one for each tissue tested except for spleen and spinal cord. In the second cluster type, probe sets were expressed in several specific tissues but not in others. There were 32 such clusters, representing group-specific genes (Figure 10B, C). A major sub-type of the group-specific clusters was two-tissue specific, composed of probe sets highly expressed in two tissues relative to the others (Figure 10C). There were five two-tissue specific clusters (out of the 32 group-specific clusters). Biologically relevant pairs of tissues created such clusters: brain and spinal cord, skeletal muscle and heart and

thymus and spleen. These pairs of tissues were also proximate in the sample dendrogram (Figure 9B).

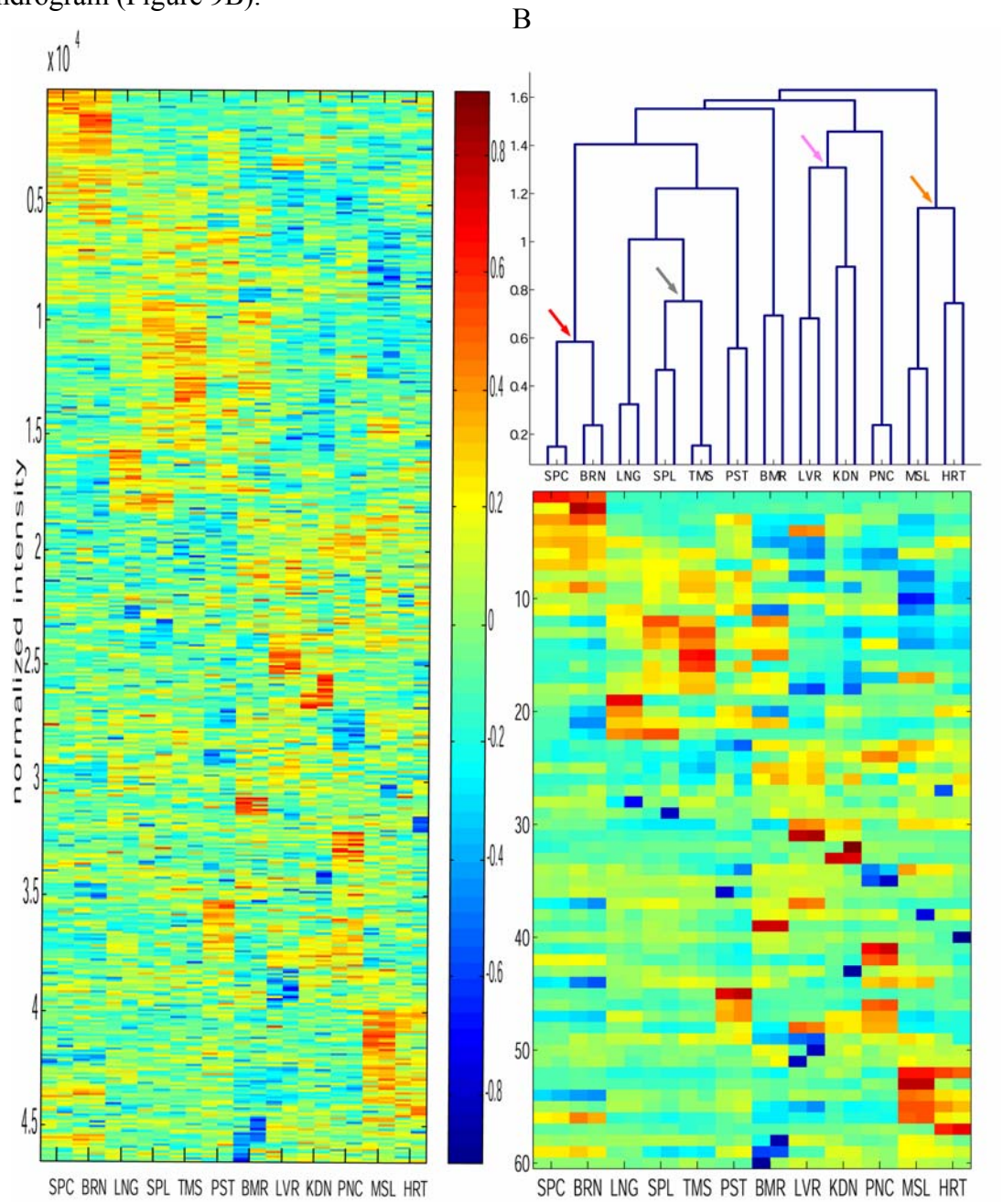


Figure 9: Clustering the entire dataset. K-means with *K*=60 was done in order to divide the probe sets into 60 groups according to their expression profiles. Each of the 60 clusters was represented by a centroid (the mean vector of all cluster members). Two-way clustering was performed on the centroids using average linkage. **A.** Expression matrix, reordered according to the clustering result. Each row represents a probe set, and each column represents a sample. The colors represent the centered and normalized expression levels according to the colorbar on the right. The probe sets were ordered as follows: probe sets that were in the same cluster are presented in adjacent rows. The 60 clusters and the samples are ordered according to the hierarchical clustering. **B.** Bottom, expression profiles of centroids. Each row represents a centroid, and each column represents a sample. Color code is as in (A). Top, dendrogram of samples based on the expression profiles of centroids. Four biologically relevant groups are marked by colored arrows: brain and spinal cord (red), spleen and thymus (gray) skeletal muscle and heart (orange) and liver and kidney (magenta).

The third cluster type identified included clusters in which probe sets were suppressed in one tissue relative to the rest of the tissues (Figure 10D). The tissues for which such clusters were found are liver, prostate, bone marrow and pancreas. The last cluster type included probe sets that were differentially expressed within replicates (Figure 10E). Since we could not determine the real expression level of the genes represented by these probe sets in one or more tissues, we considered these clusters biologically insignificant. A histogram of the number of probe sets in each cluster type is shown in Figure 11.

The clustering results strongly indicate specific expression and suppression in one or more tissues. Such patterns may imply a complex regulation system, related to ontogeny. In the next step, we will look more carefully into the specificity patterns that exist in normal human tissues.

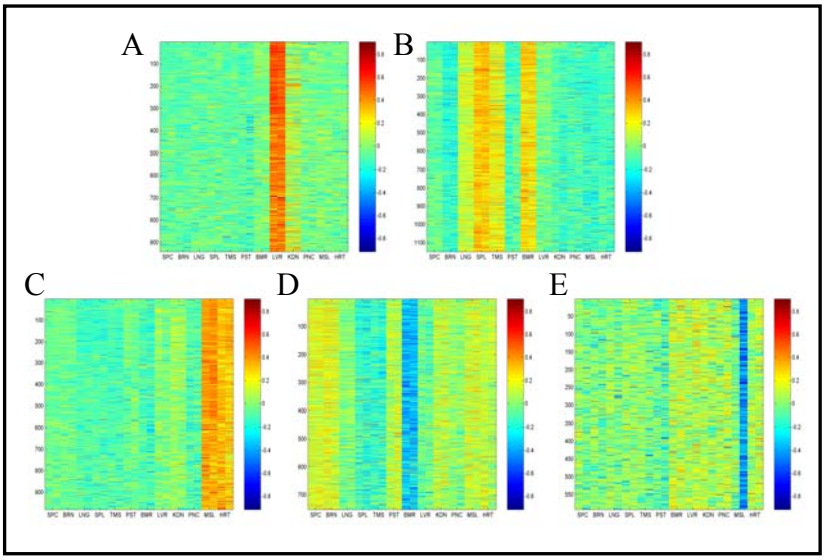


Figure 10: Four cluster types were found: **A**. One-tissue specific. The shown cluster is liver specific (941 probe sets); **B**, **C**. Group-specific. The cluster presented in B is high in the blood-related tissues (1149 probe sets). The cluster shown in C is high in skeletal muscle and heart (981 probe sets). This cluster is two-tissue specific, a sub-type of group-specific clusters; **D**. Suppressed in one tissue. The current cluster is suppressed in bone marrow (752 probe sets); **E**. Biologically insignificant. The given example is low in one replicate of skeletal muscle (592 probe sets).

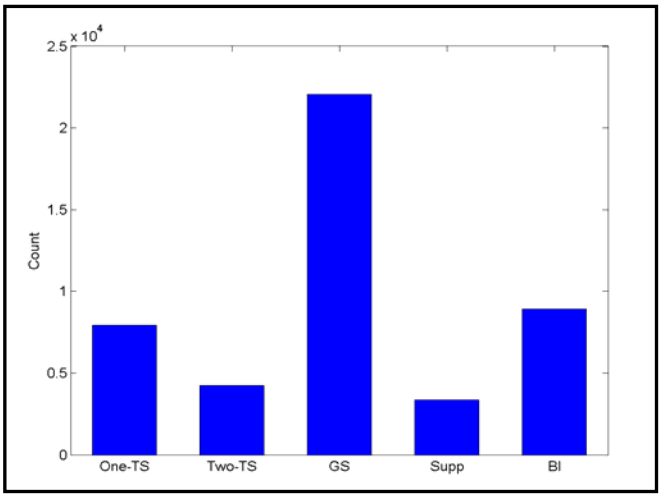


Figure 11: Distribution of the number of probe sets in each cluster-type (see **Figure 10**): one-tissue specific (One-TS), two-tissue specific (Two-TS), group specific (GS), suppressed in one tissue (Supp) and biologically insignificant (BI).

3.2 Clustering the Mingap set

A major problem identified by clustering of the entire dataset was the biologically insignificant clusters that were differentially expressed within replicate experiments. Another problem may arise from centering and normalization of probe sets that are expressed at low levels in all samples (even if above the threshold). Normalization of such low expression profiles magnifies the observed pattern in cases where there is no real change in expression. These probe sets introduce noise into the analysis, and make it hard to focus on the more significant features of the data.

Overcoming the problems mentioned above is possible using a filtering procedure; keeping only probe sets whose expression profile meets some relevant criteria. In the present study, such filtering procedure was used, selecting probe sets that were differentially expressed among different tissues, but not within replicates, and whose differences between tissues were large enough (see sections 2.5-2.7).

Usually, standard deviation is used to filter gene expression data, since one is interested in genes that show a variation in expression level along the experiment. In the current study, however, this approach would have missed many of the one-tissue specific genes that are by definition constant in most tissues and are differentially expressed in only one tissue. These probe sets have a small standard deviation and would not have passed a standard deviation based filter.

In the current study, an ANOVA procedure was applied to the preprocessed probe sets (see section 2.5) to select expression profiles whose variation between tissues is greater than their variation within tissue-replicates. Then, a signal quantilization procedure was employed, dividing the expression spectrum to 11 bins (see section 2.6). In the next step, the 'gap' criterion was used (see section 2.7) to select for probe sets whose differences between tissues are sufficient.

We used two different methods to analyze the Mingap set: an unsupervised method and a supervised one. For unsupervised analysis we used SPC to cluster both the probe sets and the tissues. The supervised method, developed and implemented by Itai Yanai [48] classified each probe set into a binary pattern.

3.2.1 Unsupervised clustering of the Mingap set

Two-way clustering was applied to the expression profiles of the Mingap set, using SPC (see sections 1.3.5, 2.8). Focusing on a set of differentially expressed profiles, we wish to look into the expression profiles of genes that contribute most to the expression diversity of normal human tissues. The reordered expression matrix and the tissue dendrogram are shown in Figure 12.

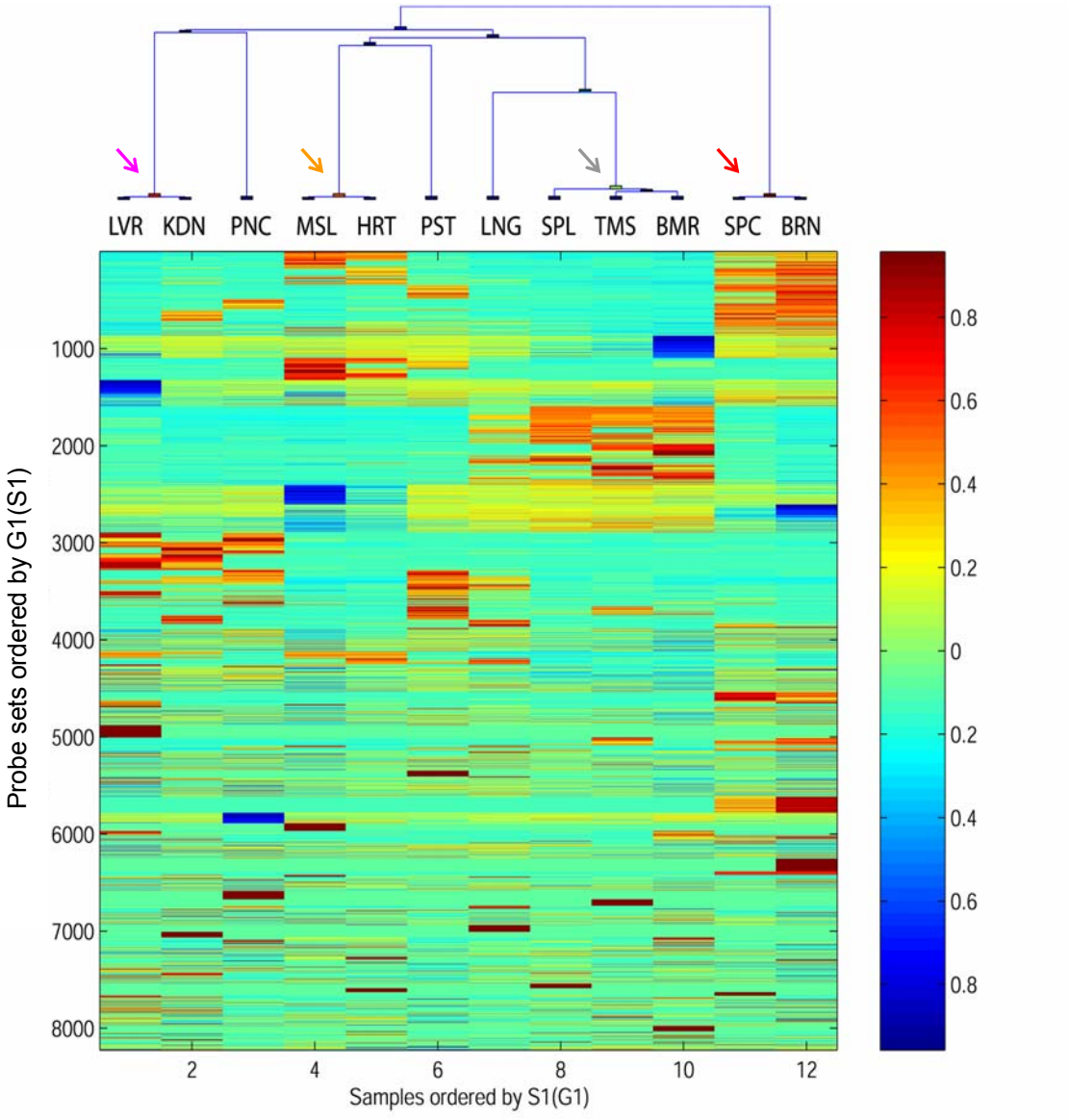


Figure 12: Reordered expression matrix and tissue dendrogram according to SPC of the Mingap set. Bottom, the reordered expression matrix according to the clustering result, where each row represents a probe set and each column represents a tissue. The clustering operation imposes a linear ordering of the data points, according to which the rows and columns were ordered. Top, a tissue dendrogram based on the Mingap set reveals four groups of tissues, marked by colored arrows: brain and spinal cord (red), spleen, thymus and bone marrow (gray), skeletal muscle and heart (orange), and liver and kidney (magenta). All these clusters were identified as stable by SPC.

Four tissue groups were identified: brain and spinal cord; spleen, thymus and bone marrow; skeletal muscle and heart; and liver and kidney (Figure 12). These groups are in correspondence with the groups identified using the clustering of centroids (see section 3.1), except that bone marrow is now closer to spleen. In addition, the groups of tissues are now much tighter, since we used a more restricted set of relevant expression profiles. The nervous system tissues (brain and spinal cord) form a separate branch from all other tissues. These are the tissues whose expression profiles are relatively unique. Inspecting the correlations between tissues (Figure 13), based on the Mingap set, we see that the two most similar tissues are, indeed, brain and spinal cord, and the most different ones are brain and liver.

All biologically relevant gene expression cluster types identified in the analysis of the full set were found in the current analysis. There was a one-tissue specific cluster for each tissue tested (including spinal cord and spleen, for which no such cluster was found using the K-means clustering). There were 45 group-specific clusters, among which 32 were two-tissue specific. Within these, complex behaviors were found (see Figure 14), where the expression levels of the two highly expressed tissues were not the same. For example, there were three gene clusters expressed in both brain and spinal cord (G8, G10 and G65). In G8, the expression levels in brain and spinal cord were similar. In G10 – genes were more highly expressed in brain than in spinal cord and in G65 – genes were more highly expressed in spinal cord than in brain (Figure 14). This behavior reveals more complex relationships between tissues in terms of expression. The remaining 13 group-specific clusters contained mainly three-tissue and four-tissue specific groups, but there were also clusters of five-tissue, and seven-tissue specific expressions. These groups typically contained high expressions in all members of one of the tissue groups identified above, in addition to one or more other tissues (for example: kidney, liver and pancreas). Patterns of suppression were found for brain, pancreas and liver.

Figure 14 reveals refined patterns of two-tissue expression, enlarging the repertoire of observed expression patterns in normal tissue gene expression. In the analysis of the entire dataset, we found four cluster types. Clustering the Mingap set using SPC showed that the same pattern of expression, for example, over-expression in two tissues, may also be composed of several sub-patterns.

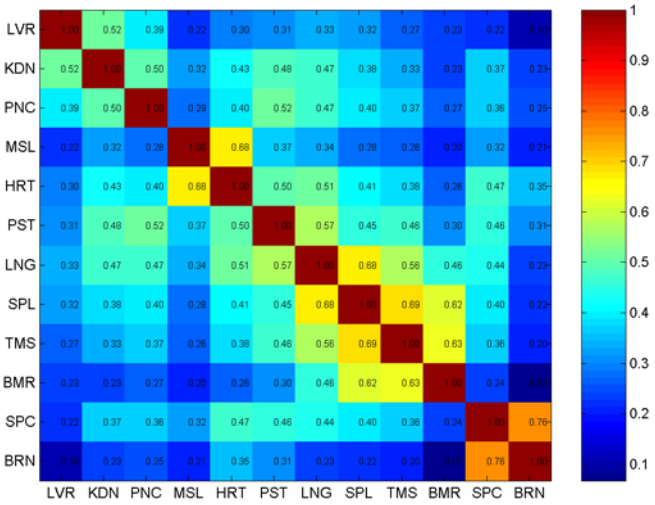


Figure 13: Correlation matrix between tissues. Pearson correlations were calculated between all tissue-pairs, based on quantiled expression of the Mingap set (see sections 2.5-2.7). The colors indicate Pearson correlation between each pair of tissues, according to the colorbar on the right. Highest correlation was found between brain and spinal cord; lowest correlation was found between brain and liver.

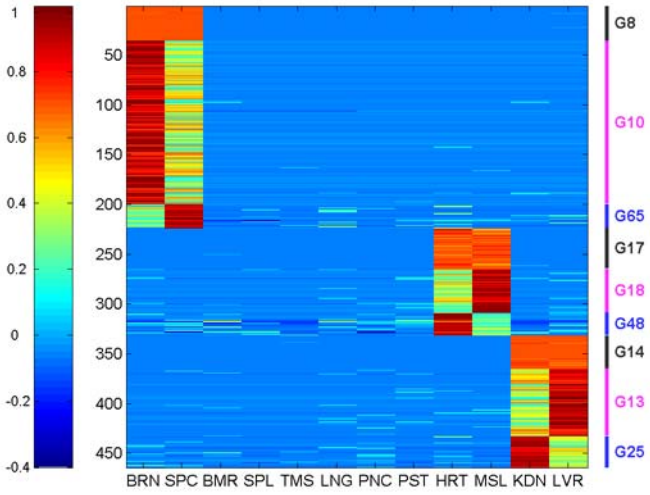


Figure 14: Augmented view of nine individual SPC clusters. Centered and normalized quantiled expression profiles of the clusters' members are shown, according to the colorbar on the left. The clusters G8, G10 and G65 manifest expression in both brain and spinal cord; G8 is equally expressed in both tissues, G10 is higher in brain than in spinal cord, and G65 is higher in spinal cord than in brain. Similar relations are seen in the cluster triplets G17, G18 and G48, expressed in heart and skeletal muscle, and G14, G13 and G25, expressed in liver and kidney.

3.2.2 Binary classification of probe sets [48]

A procedure that converts a gene expression profile into a binary pattern was applied to the Mingap set (see section 2.7). The conversion to binary patterns is a reduction of the expression profiles to only two possible levels of expression. We saw in the previous section that there may be more than two levels of expression in the same expression profile; a binarization procedure may, however, be useful for characterizing the normal human expression repertoire systematically. The 'gap' parameter is particularly suitable for the binarization process, providing a dynamic criterion for over- and under- expression (see section 2.9). The quantiled expression profiles were mapped from a space of 11^{12} patterns (12 tissues in 11 quantiles) to a reduced set of $2^{12} = 4{,}096$ possible binary patterns. Of the possible 4,096 binary patterns, 861 were actually observed in this set, including the all-0 and all-1 patterns (the all-0 and all-1 patterns were added to this analysis, they were not part of the Mingap set).

The binarization procedure was used to classify the probe sets into the space of 4,096 binary patterns. The results of the classification are shown in Figure 15 and Figure 16. Figure 15 shows the quantile expression profiles, classified into 12 groups according to the number of tissues that were over-expressed in each profile. The different panels 15.*i* in Figure 15 (i = 1 to 12) have profiles with over-expression in i tissues and under-expression in 12 minus i tissues. Panel 15.12 contains the strictly-defined 4,216 housekeeping profiles. A probe set was defined as housekeeping if it was not differentially expressed (did not pass the ANOVA, see section 2.5) and if it had a very small standard deviation (less than one quantile unit). In panel 15.1 (one-tissue specificity), brain, bone marrow, pancreas, skeletal muscle and liver are more highly represented, while spinal cord, kidney, heart, and spleen have relatively few profiles. In panel 15.2, prevalent two-tissue specific patterns are brain and spinal cord, heart and skeletal muscle, bone marrow and thymus, and kidney and liver. Bone marrow, spleen, and thymus define a major three-tissue pattern in panel 15.3. Panels 15.9 to 15.11 depict profiles with expression in all but 3, 2 or 1 tissue(s), respectively. Notably, the same five tissues with the most single tissue specific profiles (brain, bone marrow, pancreas, muscle, and liver) also have the greatest number of single tissue suppressed profiles. The distribution of the number of probe sets with different numbers of over-expressed tissues is presented in Figure 16.

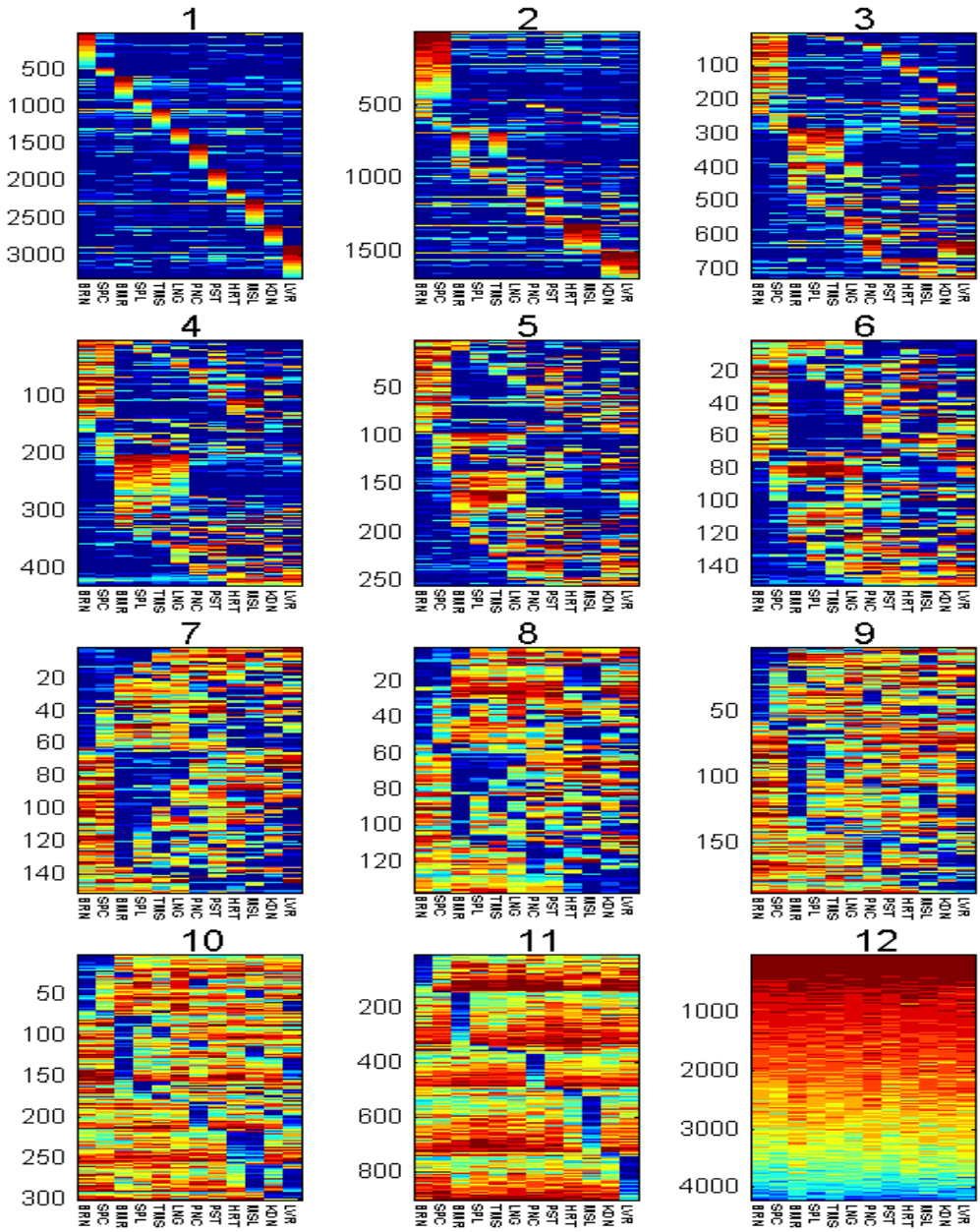


Figure 15: Profile classification in 12 sets according to binary expression patterns. In panels 1 to 11, the profiles in the Mingap set were classified into the 2^{12}-2 possible binary tissue expression patterns. The profiles were categorized into 11 sets according to the number of tissues (*i*) with a binary score of 1, and sorted by their binary patterns. The profiles of each pattern were further sorted according to the mean expression of the non-zero elements. Housekeeping profiles are shown in panel 12, sorted according to their mean expression level. The color code denoting gene expression in quantile units ranges from 0 (dark blue) to 10 (dark red).

We next inspected the 99 most populated binary patterns. These included the housekeeping (all 1's) and null (all 0's) patterns, and the 97 most populated binary patterns among the differentially expressed profiles (see section 2.9), having at least 25 probe set profiles (Figure 17A). We grouped these 99 patterns according to their binary index I_B (see section 2.9). The number of populated binary patterns in each binary index

showed a clear bimodal distribution (Figure 17B), with peaks at binary index values of 2 and 10. Whereas all 12 one-tissue specific patterns were included, only about one third of the two-tissue expressed patterns ($I_B = 2$) and about a quarter of the two-tissue repressed patterns ($I_B = 10$) were included in this set, suggesting biases towards specific oligo-tissue combinations.

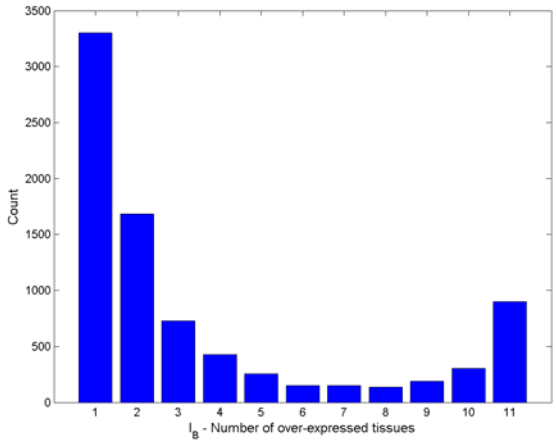


Figure 16: Distribution of the number of probe sets from the Mingap set with different I_B values (see section 2.9).

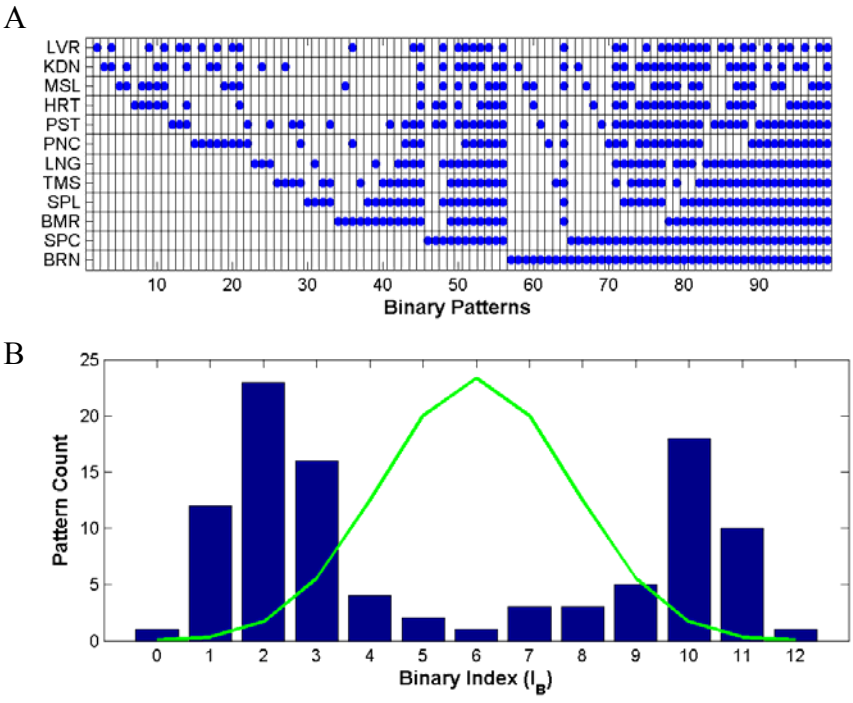


Figure 17: Expression pattern repertoire. **A.** A summary representation of the most populated binary patterns (columns), where blue circles indicate over-expression. The patterns are sorted according to binary value. **B.** The frequency distribution of I_B values of the binary patterns shown in (A). The green curve indicates the expected distribution following a random binomial model.

3.2.3 Comparison between binary classification and SPC clustering

We next checked the correspondence between the clusters found by SPC and the 95 most populated differentially expressed binary patterns among the Mingap set (excluding all-0 and all-1 patterns). For every SPC cluster and binary pattern, an association score was calculated for the degree of correspondence, as follows: the number of shared probe sets between the two groups (the intersection) was divided by the number of probe sets in the smaller of the two groups. The score ranges between 0 and 1. The identified 70 SPC clusters showed a strong correlation with the 95 most populated binary classes. For every SPC cluster, the maximal score was found. The maximal scores ranged between 0.125 – 1, where 54 out of the 70 clusters had a score of at least 0.7. The association between the two methods is presented in Figure 18. Some binary patterns corresponded to multiple SPC clusters, thus the SPC clusters further refine the relevant binary patterns (see Figure 14, Figure 18).

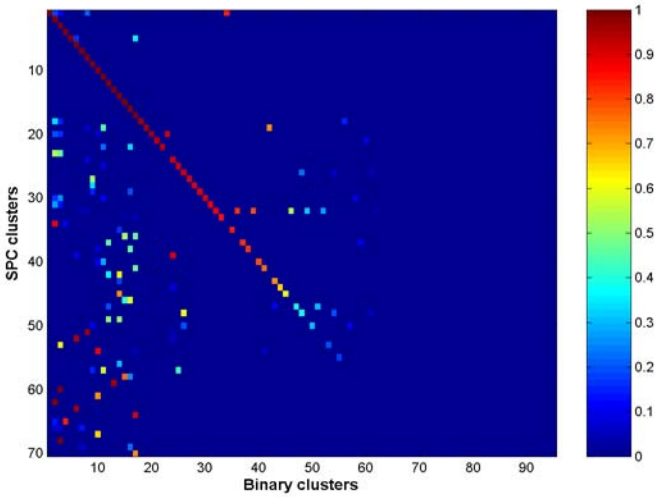


Figure 18: Correspondence between SPC clusters and binary patterns. Each element in the matrix is a comparison between the profiles of an SPC cluster and those of a binary pattern. Binary patterns were obtained for the 8,224 profiles of the Mingap set, and only the patterns with at least 10 profiles are shown. The color-coded association score (right bar), ranging between 0 and 1, was calculated as the number of shared probe sets between the two clusters, divided by the size of the smaller of the two clusters.

3.2.4 Tissue Specificity Index (τ) of binary patterns

We used a Tissue Specificity Index, τ, which is a quantitative, graded scalar measure of the specificity of an expression profile (see section 2.10). τ values range between 0, for housekeeping genes, and 1 for strictly one-tissue specific genes. Figure 19 shows the distribution of τ values for the 22,936 differentially expressed (see section

2.5) and the 4,216 housekeeping profiles (see section 3.2.2). It can be observed that τ values near 0 and 1 tend to be more probable than the intermediate values, generating a U-shaped distribution. However, 57% of all profiles have intermediate specificities: $0.15 \leq \tau \leq 0.85$, constituting a very large and significant group of genes.

To further verify our results, we checked the τ scores of each of the 12 sets in Figure 15. As can be seen in Figure 20, there is a gradual decrease in τ scores as a function of I_B, the number of over-expressed tissues in the binary pattern. This result shows the correspondence between the different methods. The U-shaped distribution of τ values indicates the observed tendency to either low or high tissue specificity, also observed in the other methods: there is a small number of SPC clusters and binary patterns with intermediate numbers (around I_B=5-9) of over-expressed tissues. The human tissue expression repertoire tends to specific oligo-tissue combinations, including both specific expressions and specific suppressions.

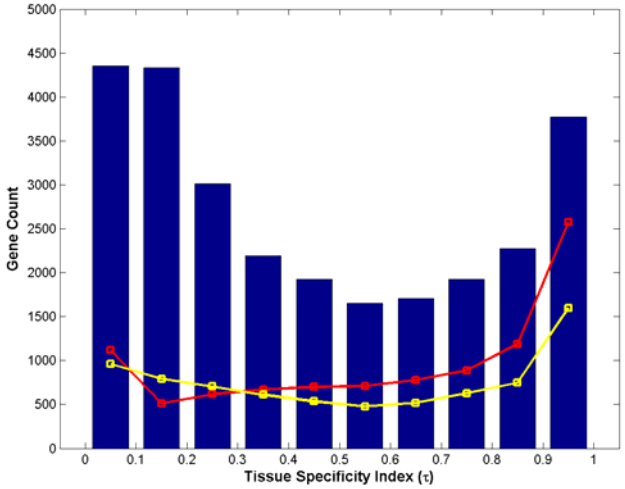


Figure 19: Distribution of Tissue Specificity Index (τ) scores. Distribution of τ values for 27,152 profiles which include the 22,936 differentially expressed and 4,216 housekeeping profiles (bars). The τ distributions are also shown for the 12,626 profiles across 27 human tissues (red curve), and 12,654 across 45 mouse tissues (yellow curve) from a published study [17] [48].

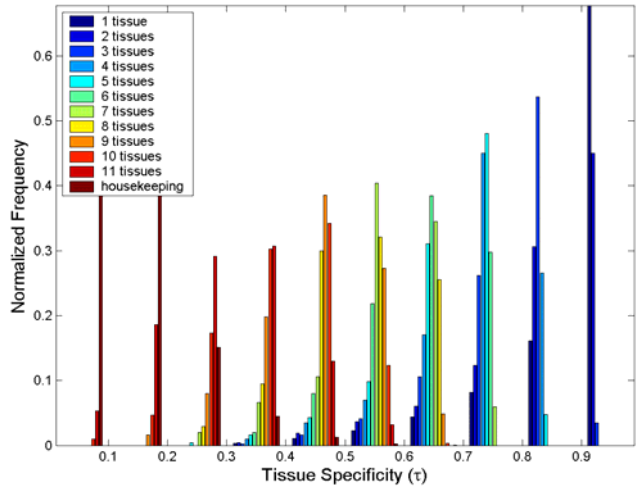


Figure 20: Normalized frequency distributions of the Tissue Specificity Index (τ) for each of the 12 sets of Figure 15, each represented by a different color. An overall correspondence is observed between the classification of the 12 sets and the τ values, whereby low τ is seen for high I_B values (see 2.9) and vice versa.

3.3 Expression profiles of genes

The raw results of a microarray experiment are the hybridization intensities measured by the probe array. These results are summarized such that the hybridization intensity of every probe set, indicating the expression of some mRNA transcript, is given by one number (see section 1.2.2). However, we would like to shift from the level of probe set expression to the level of gene expression. For this purpose, we focused on probe sets that were associated with known genes, with high specificity and sensitivity (see section 1.4).

As mentioned above, often there is more than one probe set representing a gene. The expression profiles of the different probe sets corresponding to the same gene are not always similar. There are two possible explanations of this phenomenon: first, the probe sets may have been derived from different mRNA sequences, and the difference in mRNA expression may indicate alternative transcription. The second option is an experimental problem. For example, bad probe set design may result in non-specific hybridization or too weak an interaction of the mRNA with the probe set. In this case, the two profiles represent the same mRNA transcript, but the correct expression profile of that transcript cannot easily be determined from the experiment.

When looking at genes that are represented by more than one probe set, there are two levels of analysis: the sequence level and the expression level. By sequence level we refer to the sequence relations between probe sets and genes, according to GeneAnnot, whereas the expression level indicates the expression profiles found by the microarrays. We tried to create expression profiles for genes (see section 2.11), i.e. combine expression profiles of probe sets representing the same gene. Table 3 shows all possible combination categories of sequence and expression correspondence, and the resulting number of genes in each category.

If the probe sets match the same mRNA sequence and their expression profiles across the tested tissues are correlated, we would like to combine their expression profiles and remain with only one profile representing the gene. If, on the other hand, the probe sets match different mRNA sequences of the gene, and their expression profiles are not correlated, we would like to have several profiles representing this gene, one for each transcript (see Table 3, section 2.11).

When probe sets match different mRNA sequences, their expression profiles may be either similar or different; a different expression profile indicates that the two transcripts

are differentially expressed in the tested tissues, as explained above. A similar expression profile indicates that the two transcripts are not differentially expressed over the set of tissues tested in the current study. The two transcripts either have the same expression pattern in all tissues, or they are differentially expressed in other tissues, not examined in our study.

Table 3: Genes represented by several probe sets

Sequence - mRNA Expression correlation	Same	Different	Total
High	**Expected – combine** **2,982 genes**	**Possible – do not combine** **459 genes**	**3,237 genes**
Low	**Not expected - do not combine** **1,322 genes**	**Expected – do not combine** **788 genes**	**1,484 genes**
Total	**3,937 genes**	**1,121 genes**	**4,192 genes**

When there are several probe sets representing a unique gene, there are four possible relations between their mRNA and expression correspondence. We expect a high correlation between probe set expression when probe sets align the same mRNA. When probe sets match different mRNAs, we cannot predict the correlation between their expression profiles. There were 5,504 genes represented by two probe sets or more. Of these, 4,192 genes had more than one probe set above threshold in at least one tissue (if a gene was represented by several probe sets but only one of them was above threshold, only that probe set was used. See section 2.11). The number of genes in each category is marked. Note that the numbers do not sum up to the Total since a gene can appear in more than one category, if more than two probe sets are associated with it.

Affymetrix GeneChips U95A-E contain 62,839 probe sets. Of these, 23,689 (38%) passed our sensitivity and specificity thresholds of probe-set-to-gene sequence alignment (see section 2.11). These probe sets matched 15,112 gene entries in GeneCards [50].

The 15,112 genes represented by the filtered set of probe sets were examined. The distribution of the number of probe sets representing each gene is shown in Figure 21. Most genes (9,608 genes, 64%) were represented by one probe set. The remaining 5,504 genes (36%) were represented by 2-16 probe sets.

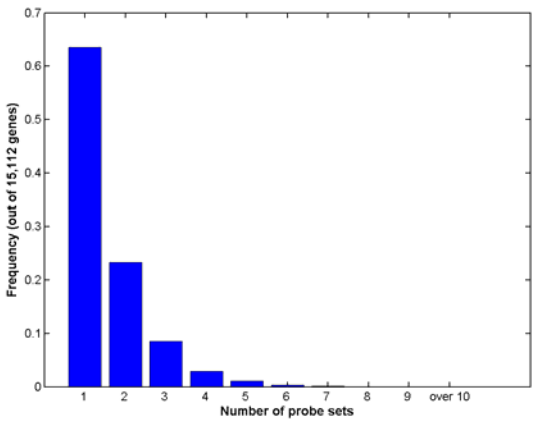


Figure 21: Distribution of number of probe sets per gene. Probe sets were filtered according to sensitivity and specificity scores ($Sn \geq 9/16$ and $Sp = 1$). There were 23,689 probe sets that passed the filtering procedure. These probe sets were associated with 15,112 genes. Out of these, 9,608 genes were represented by one probe set and 5,504 genes were represented by 2-16 probe sets.

In the next step, we applied a procedure designed to lower the dimensionality of our data by uniting probe sets that have correlated expression over the set of tested tissues and represent the same mRNA transcript (see section 2.11). The distribution of the number of profiles representing each gene is shown in Figure 22. After applying the averaging procedure, we were left with 13,364 genes (88%) represented by one expression profile. The remaining 1,748 genes were represented by 2-6 profiles.

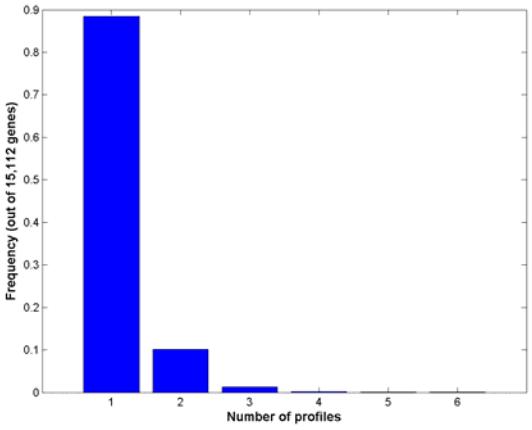


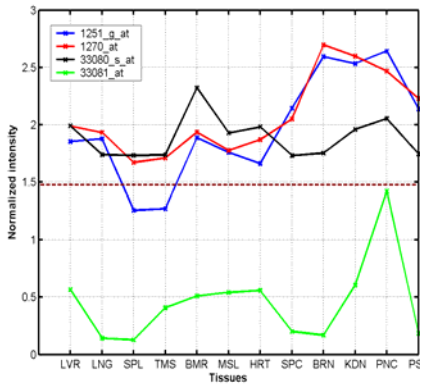
Figure 22: Distribution of number of profiles per gene. Probe sets were filtered according to sensitivity and specificity scores ($Sn \geq 9/16$ and $Sp = 1$). There were 23,689 probe sets that passed the filtering procedure, associated with 15,112 genes. Probe sets corresponding to the same gene were inspected in terms of their expression profiles and sequence. Probe sets that were aligned to the same mRNA sequences and were highly correlated in terms of expression across the tested tissues were averaged. The resulting 17,118 gene expression profiles include 13,364 genes represented by one profile and 1,748 genes represented by 2-6 profiles.

We present two examples for the results of our method. *RAP1GA1* is a GTPase activator for the nuclear protein RAP1A. RAP1A is a member of RAS oncogene family; it counteracts the mitogenetic function of the RAS gene. There were four probe sets (that passed our sensitivity and specificity thresholds) associated with RAP1GA1 in the current experiment (Figure 23A, C). Probe sets 1251_g_at and 1270_at were aligned to the same mRNA sequences from RefSeq and Ensembl (Figure 23C). The expression profiles of the two probe sets were highly correlated ($r = 0.94$). Therefore, their expression was averaged (Figure 23B). Probe sets 33080_s_at and 33081_at were aligned to a different mRNA sequence (GenBank), but 33081_at was below threshold in all tissues and therefore was not used. 33080_s_at had a low correlation with 1251_g_at and 1270_at ($r = 0.17$, $r = 0.05$ respectively), and it matched a different mRNA sequence, hence it added a profile to the gene RAP1GA1 (Figure 23B). It can be observed that the two resulting profiles for the gene RAP1GA1 differ in their expression in the brain, kidney, pancreas and prostate tissues.

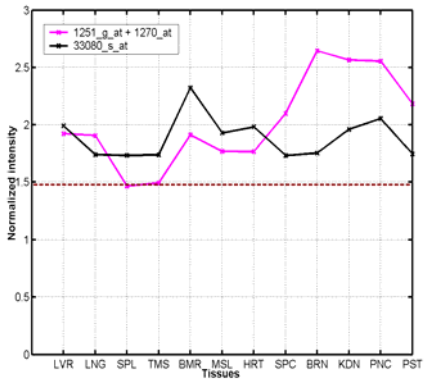
MTA3 is a metastasis-associated protein. There were four probe sets in the current experiment that were aligned to this gene with high specificity and sensitivity (Figure 23D, F). Probe sets 47932_at, 59743_at and 79737_at were aligned to the same mRNA (ENST00000282366, Figure 23F), and the correlations between them were above

threshold ($r_{47932\_at,\ 59743\_at} = 0.9$, $r_{59743\_at,\ 79737\_at} = 0.55$). Another probe set, 53186_s_at, matched a different mRNA, and its correlation with the other probe sets was below threshold. Therefore, this probe set remained as an additional profile for MTA3. The two resulting profiles differ mainly in their expression in brain and liver (Figure 23E).

A



B

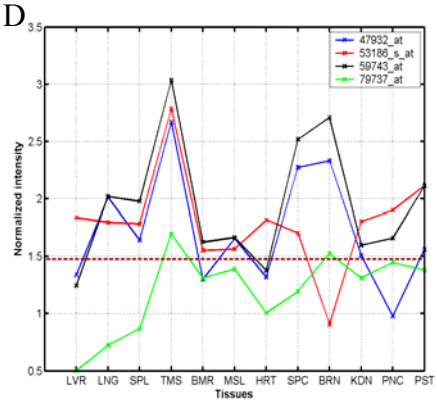


C

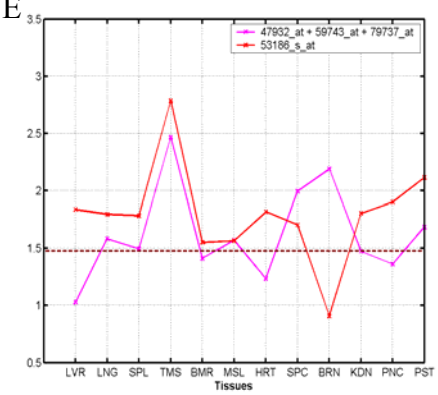
Matching mRNA:

Probe sets	mRNA
1251_g_at 1270_at	NM_002885 ENST00000317967 ENST00000264199
33080_s_at 33081_at	AB007943

D



E



F Matching mRNA:

Probe sets	mRNA
47932_at 59743_at 79737_at	ENST00000282366
53186_s_at	ENST00000315879 ENST00000282366 ENST00000263877 ENST00000315864

Figure 23: Expression of the genes RAP1GA1 (A-C) and MTA3 (D-F). A, B, D, E show normalized intensity of averaged tissue replicates. The dashed line indicates the intensity threshold ($\log_{10}30$). **A.** Expression profiles of probe sets associated with RAP1GA1. The gene RAP1GA1 was represented by four probe sets. It can be observed that the red and blue probe sets are correlated, whereas the black and green are not. The green probe set is below threshold in all tissues and therefore was excluded. **B.** Expression profiles of the gene RAP1GA1. After averaging the expression of the blue and red probe sets, and eliminating the green one, two distinct expression profiles remained for the gene RAP1GA1. **C.** Accession numbers of mRNA sequences that each of the probe sets from (A) aligned to. The first two probe sets, the blue and red, were aligned to the same mRNA sequences and were highly correlated in terms of expression, therefore, their expression profiles were averaged. **D.** Expression profiles of probe sets associated with MTA3. MTA3 was represented by four probe sets. For MTA3, the blue, black and green probe sets had similar expression profiles, whereas the red probe set had a different profile. **E.** Expression profiles of the gene MTA3. As for RAP1GA1, two distinct expression profiles were found for MTA3. **F.** Accession numbers of mRNA sequences that each of the probe sets aligned to. For MTA3, the blue, black and green probe sets from (D) were highly correlated, and aligned to the same mRNA sequence, therefore, their expression profiles were averaged.

3.4 Analysis of most divergent genes

In the former section, we created a set of expression profiles that were characteristic of known genes. Next, we analyzed the expression of a subset of these expression profiles that was most divergent (section 2.12). Our aim was to investigate the known genes that contribute the most to the differences between the tested tissues.

A filtering procedure was applied to the 17,118 profiles, using the standard deviation and range of the gene expression profiles (see section 2.12). The resulting set included 1,950 profiles. These profiles represented 1,913 genes, where 1,876 genes were represented by one profile and 37 genes were represented by two profiles.

As discussed in section 3.2, a standard deviation based filter such as the one used here, is not designed to grasp all one-tissue specific genes. However, in the current analysis we study the genes that show specificity to a group of tissues, revealing the relationships among the tested tissues.

Two-way clustering was applied to the filtered expression profiles, using SPC (see section 2.12). The reordered expression matrix and the tissue dendrogram are shown in Figure 24A. The reordered genes' distance matrix is presented in Figure 24B.

Four major clusters of genes were found, each over-expressed in one of the four groups of tissues identified in former analyses: skeletal muscle and heart (236 genes), spleen, thymus, bone marrow and lung (205 genes), liver and kidney (161 genes) and brain and spinal cord (149 genes). In addition, there were eight one-tissue specific clusters for the following tissues (number of genes in parentheses): liver (51), lung (45), prostate (31), kidney (30), skeletal muscle (25), pancreas (25), bone marrow (21) and heart (20). The tissue specific clusters here contain genes whose standard deviation and range were relatively high, implying that these genes are not only tissue specific, but that the difference in expression between the over-expressed tissue and the under-expressed ones is large.

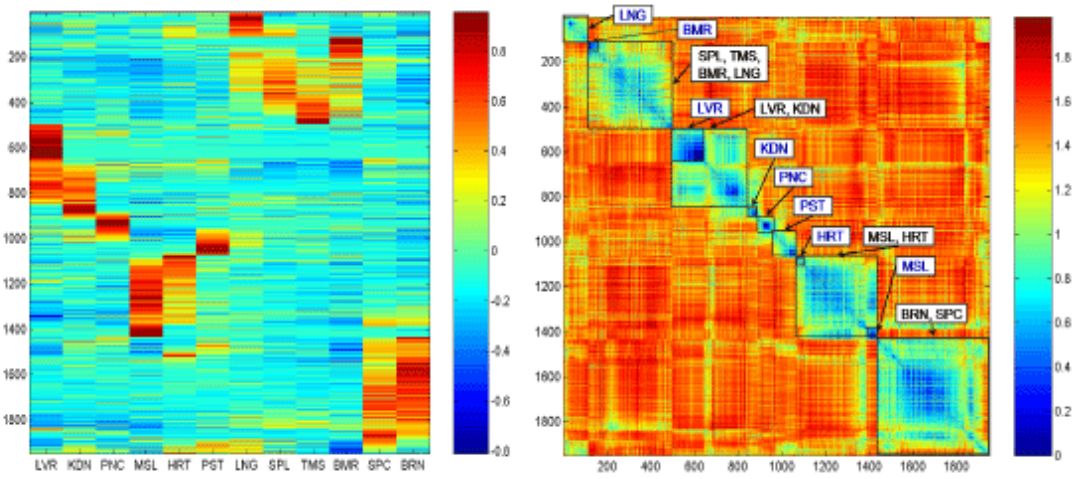


Figure 24: **A.** Reordered expression matrix. Each row represents a gene expression profile and each column represents a tissue. The normalized intensity levels are color coded, according to the colorbar on the right. The gene expression profiles were ordered by SPIN [51]. **B.** Reordered gene distance matrix. The order of the genes is as in (A). Distance between genes is color coded, according to the colorbar on the right. The marked boxes designate the gene groups identified. The four largest clusters of genes found in the current analysis were over-expressed in the following tissue groups: spleen, bone marrow, thymus and lung; liver and kidney; brain and spinal cord; skeletal muscle and heart.

3.5 Comparison of results to a published dataset [17]

In order to validate the findings of the current research, we compared our results to a published study of gene expression in human and mouse tissues, by Su et al. (2002) [17]. The human dataset included gene expression data for 86 samples of 40 human tissues, tested using Affymetrix GeneChip U95A (see section 2.13). In order to distinguish between the two sets, we will refer to our dataset as "GeneNote" (Gene Normal Tissue Expression) or "GN", and to other dataset as "Su et al." or "Su".

We tested the correlations between tissue expression profiles from the two datasets, as well as between expression profiles of probe sets tested in the two experiments (over the set of tissues in common to the two datasets).

Pearson correlation was calculated between all pairs of samples within each dataset, and between the two datasets (Figure 25). The two datasets had 10 tissues in common. In GeneNote, there were two replicates for each tissue. In Su et al., eight of the 10 tissues in common to GeneNote were tested in two replicates, and the other two, kidney and prostate, had three replicates each. The highest correlations (Figure 25) were found between replicate experiments in the GeneNote dataset (dark blue), i.e. two samples of the same tissue. Correlations between replicate experiments from different datasets

(pale blue) ranged roughly between 0.7 and 0.9. Note that high correlations between non-replicate experiments come from closely related tissues, such as brain and spinal cord. Generally, the results of the two experiments are highly correlated.

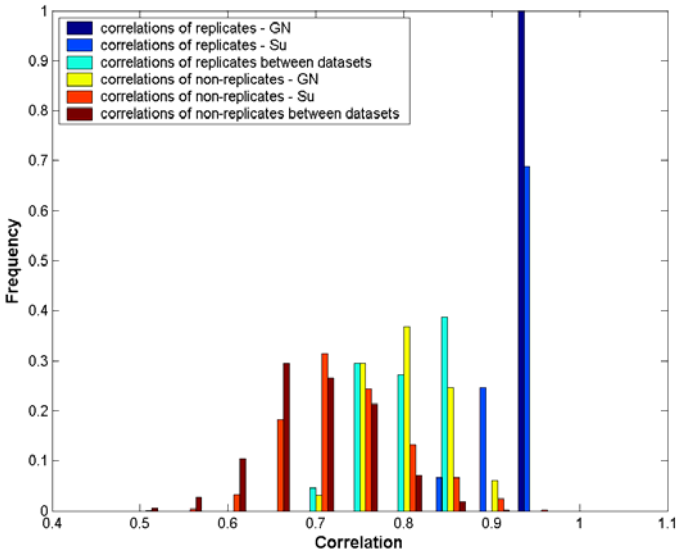


Figure 25: Distribution of correlations between samples. Pearson correlation was calculated between the gene expression profiles of human tissue samples. Dark blue, 12 pairwise correlations between replicate experiments in the GeneNote (GN) dataset; Light blue, 61 pairwise correlations between replicate experiments in Su et al.'s dataset; Pale blue, 44 pairwise correlations between replicate experiments, among the two datasets (for example, kidney from GeneNote and kidney from Su et al.); Yellow, 264 pairwise correlations between non-replicate samples in the GeneNote dataset; Orange, 3,594 pairwise correlations between non-replicate samples in the Su et al. dataset; Brown, 1,444 pairwise correlations between non-replicate samples among the datasets (for example, kidney from GeneNote and liver from Su et al.).

Pearson correlation was also calculated between probe set expression profiles, for all 12,533 probe sets common to the two datasets, over the set of 10 common tissues (see section 2.13, Figure 26). Blue bars represent correlations between the same probe set in the two experiments, and green bars represent random pairs of probe sets: one from the GeneNote set and the other from the Su et al. set. The correlations between the random pairs were distributed normally around zero, indicating a random correlation pattern. The real distribution of probe set expression correlations (blue) was skewed towards positive correlations, but there were many probe sets for which the correlation was zero or lower.

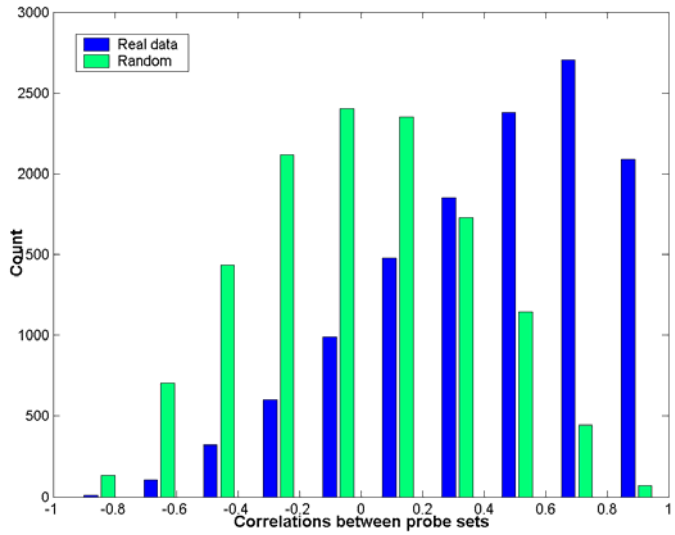


Figure 26: Distribution of 12,533 correlations between probe set expression profiles in the GeneNote (GN) data and the Su et al. data, over the set of 10 common tissues (see section 2.13). Blue, correlations of 12,533 probe sets common to the two datasets. Green, distribution of 12,533 random pairs of probe sets, one from the GeneNote dataset and one from the Su et al. dataset.

Next, the τ parameter (see section 2.10) was calculated according to each dataset separately, and the congruence between the results was tested (Figure 27). For the GeneNote dataset, the τ parameter was based on 12 tissues, whereas for the Su dataset, the parameter was based on 40 tissues. The overall correlation between the τ parameter results was 0.55. Figure 27 shows that there were many probe sets for which the τ score was similar in the two datasets. These are distributed around the diagonal, marked by a red line. There were, however, probe sets that were relatively tissue-specific in the Su et al. dataset and housekeeping in the GeneNote dataset (Figure 2, A). The opposite case, where a probe set received a high τ score according to the GeneNote data and a low one according to Su et al.'s data, is shown in Figure 27, B. Both groups were mostly composed of lowly expressed, constant probe sets. Group A included 986 probe sets. Out of these, 794 probe sets (81%) had a 'gap' (see section 2.7) smaller than 3. The remaining 192 probe sets had relatively high τ scores: 172 probe sets (90%) had a τ of 0.9 or higher. Out of the 76 probe sets of group B, 72 (95%) had a 'gap' smaller than 3.

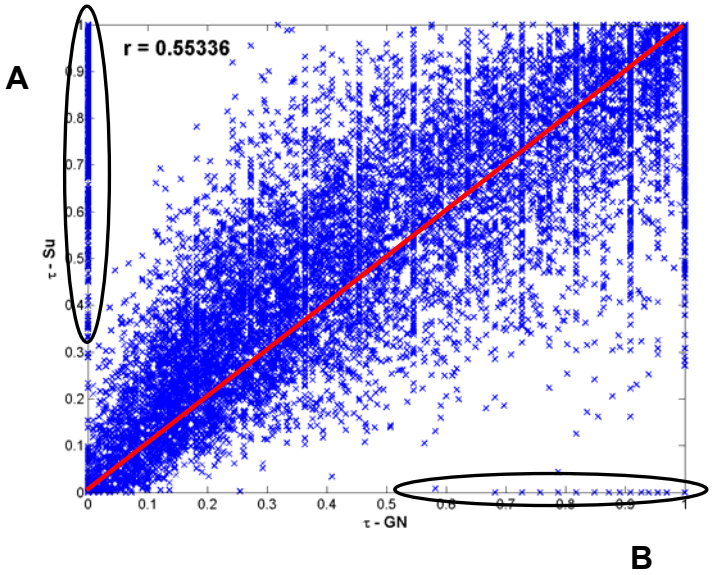


Figure 27: Scatter plot of τ scores for 12,533 probe sets in two datasets. The x axis is the τ score according to GeneNote (12 tissues), and the y axis is the τ score according to the Su et al. data (40 tissues). The overall correlation between the scores is 0.55. (A) Probe sets with high τ according to Su et al.'s data and low τ according to GeneNote. (B) Probe sets with high τ according to GeneNote and low τ according to Su et al.'s data.

4 Conclusions

In the present research we analyzed gene expression profiles of 12 normal human tissues. Our aim was to characterize the normal human gene expression pattern repertoire. In contrast to previous studies that focused on housekeeping genes [14, 15] and one-tissue specific genes [16, 17], we tried to find other groups of genes, expressed in a subset of the tested tissues. In addition, we studied the relations between the tested tissues in order to find groups of tissues that have similar gene expression profiles.

Binary classification analysis of gene expression profiles over the set of 12 normal human tissues revealed the underlying patterns of tissue specificity. Of the differentially expressed genes, one-tissue specific expression was the most common pattern. The number of genes decreased as the number of expressed tissues increased; however, at nine-tissue specific expression this trend reversed, yielding an increasing number of genes for nine-tissue, ten-tissue and eleven-tissue specific expression. Since *n*-tissue expression can be viewed as (*12*-*n*)-tissue suppression, we can formulate the following general rule: in normal human gene expression, there is a preference for either expression or suppression in a relatively small number of tissues.

For random gene expression patterns, the most common tissue specificity would be six-tissue expression/suppression. In contrast, our analysis indicates that the specification patterns present in normal human gene expression favor specificity of either expression or suppression in a small number of tissues.

Studying the binary classes of one-tissue specific genes, we found that the same tissues with the highest number of one-tissue specific genes (brain, bone marrow, pancreas, skeletal muscle, and liver) also have the greatest number of one-tissue suppressed genes.

The existence and wide extent of tissue specific suppressions, as well as the correlation between the numbers of over-expressed and under-expressed genes in the same tissue, implies that specific gene suppression, as well as specific gene expression is a significant mechanism in tissue specification.

Two-way clustering was applied to the differentially expressed probe sets, using SPC. Four groups of tissues were found: brain and spinal cord; skeletal muscle and heart; spleen, bone marrow and thymus; and liver and kidney. Considering the embryonic origin of these tissues may help to understand the relations between them:

- Brain and spinal cord are neural tissues, composing the CNS. Their embryonic origin is ectodermal (except for microglia, see section 1.1.5.4).
- Skeletal muscle and heart are both muscular tissues. All muscle cells are mesodermal in origin.
- Spleen, bone marrow and thymus are all related to blood cell formation and filtering. Blood is a connective tissue, of mesodermal origin.
- Liver and kidney come from different embryonic origin. The kidney is mesodermal, whereas the liver is endodermal. However, epithelium of both organs is of mesodermal origin.

The difference in embryonic origin of liver and kidney raises the question of why these two tissues were clustered together in our analyses. We would expect the kidney to be more similar to other organs of mesodermal origin, such as the skeletal muscle and heart. Several reasons may account for the similarity in gene expression of liver and kidney. First, liver and kidney are both glands, performing both endocrine and exocrine functions. The similarity between gene expression profiles of the two tissues may be functional rather than developmental. Second, as mentioned above, both liver and kidney have mesodermal epithelium. It is possible that the similarity stems from epithelial genes. Third, the joining of liver and kidney may have been based on their differences from the other tissues. If other tissues, more similar to the liver and kidney, were tested, they might have separated them, forming lower branches in the tissue tree. It should also be noted that whole organ expression patterns were tested in the present study; each organ contains several cell types. Nerve cells and blood supply exist, to some extent, in all the tested organs.

Many times several probe sets represent the same gene. The expression profiles of the different probe sets are not always similar. We developed a method to shift from the level of probe set expression to the level of gene expression, considering the sequences of the probes and the probe set's expression profile. This method was used to create a set of expression profiles characteristic of known genes.

We selected a subset of the known genes that had the most divergent expression profiles over the tested tissues. Clustering this subset using SPC identified four main clusters of genes. Three of these gene-clusters were equivalent to the tissue groups identified above (brain and spinal cord; skeletal muscle and heart; liver and kidney). The fourth cluster was over-expressed in the blood-related tissues (thymus, spleen and

bone marrow) and in the lung. The lung is of endodermal embryonic origin, whereas the other tissues are of mesodermal origin. However, the lung has a very rich vascularity, which may have caused the similarity in expression to the other blood-related tissues.

The similarity of lung to the blood-related tissues, along with the connection between liver and kidney suggests that gene expression patterns in normal human tissues are influenced by functional similarities as well as by embryonic origin of tissues.

Two main conclusions may be drawn from the present work. First, gene suppression is a major mechanism in normal human gene expression, playing an important role in tissue specification. Second, tissue expression patterns are influenced by functional relationships between tissues.

4.1 Future directions

- The dataset presented in the current study comprises of a set of gene expression profiles for a comprehensive list of genes, in a variety of normal human tissues. This dataset may serve as a baseline for past and future expression studies related to diseases.

- The present work presents a broad view of the human expression pattern repertoire. In the future, focusing on specific clusters of genes identified in the current study may be of interest. Specifically, characterizing genes within specific clusters in terms of their known tissue specificity, molecular functions, chromosomal locations and so forth. The characterization of tissue-suppressed genes would be of special interest, because it may contribute to our understanding of the role of gene suppression in normal tissues.

- The expression profiles of genes were attained using probe sets with very high sensitivity and specificity scores. Lowering these thresholds, will enable the use of many more probe sets, and the characterization of many more genes.

- In the current research we studied the expression profiles of known genes. The identified gene expression patterns may be used to classify expression profiles of uncharacterized genes and add information about the function of such genes.

References

1. Lockhart, D.J. and E.A. Winzeler, *Genomics, gene expression and DNA arrays.* Nature, 2000. **405**(6788): p. 827-36.
2. Gilbert, S.F., *Developmental Biology*. Fifth ed. 1997, Sanderland (MA): Sinauer Associates, Inc.
3. Alberts, B., et al., *Molecular Biology of the Cell*. Fourth ed. 2002, New York: Garland Science.
4. Slack, J.M.W., *Essential Developmental Biology*. 2001, Oxford: Blackwell Science Ltd.
5. Wolberger, C., *Multiprotein-DNA complexes in transcriptional regulation.* Annu Rev Biophys Biomol Struct, 1999. **28**: p. 29-56.
6. Jones, K.A., *HIV trans-activation and transcription control mechanisms.* New Biol, 1989. **1**(2): p. 127-35.
7. Black, D.L., *Activation of c-src neuron-specific splicing by an unusual RNA element in vivo and in vitro.* Cell, 1992. **69**(5): p. 795-807.
8. Dreyfuss, G., V.N. Kim, and N. Kataoka, *Messenger-RNA-binding proteins and the messages they carry.* Nat Rev Mol Cell Biol, 2002. **3**(3): p. 195-205.
9. Hughes, T.R., et al., *Functional discovery via a compendium of expression profiles.* Cell, 2000. **102**(1): p. 109-26.
10. Le Roch, K.G., et al., *Discovery of gene function by expression profiling of the malaria parasite life cycle.* Science, 2003. **301**(5639): p. 1503-8.
11. Arbeitman, M.N., et al., *Gene expression during the life cycle of Drosophila melanogaster.* Science, 2002. **297**(5590): p. 2270-5.
12. Golub, T.R., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**(5439): p. 531-7.
13. DeRisi, J.L., V.R. Iyer, and P.O. Brown, *Exploring the metabolic and genetic control of gene expression on a genomic scale.* Science, 1997. **278**(5338): p. 680-6.
14. Butte, A.J., V.J. Dzau, and S.B. Glueck, *Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues".* Physiol Genomics, 2001. **7**(2): p. 95-6.
15. Warrington, J.A., et al., *Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes.* Physiol Genomics, 2000. **2**(3): p. 143-7.
16. Hsiao, L.L., et al., *A compendium of gene expression in normal human tissues.* Physiol Genomics, 2001. **7**(2): p. 97-104.
17. Su, A.I., et al., *Large-scale analysis of the human and mouse transcriptomes.* Proc Natl Acad Sci U S A, 2002. **99**(7): p. 4465-70.
18. Lercher, M.J., A.O. Urrutia, and L.D. Hurst, *Clustering of housekeeping genes provides a unified model of gene order in the human genome.* Nat Genet, 2002. **31**(2): p. 180-3.
19. Eisenberg, E. and E.Y. Levanon, *Human housekeeping genes are compact.* Trends Genet, 2003. **19**(7): p. 362-5.
20. Telford, I.R. and C.F. Bridgman, *Introduction to Functional Histology*. 1990, New York: Harper & Row, Publishers, Inc.
21. Saito-Hisaminato, A., et al., *Genome-wide profiling of gene expression in 29 normal human tissues with a cDNA microarray.* DNA Res, 2002. **9**(2): p. 35-45.

22. Haverty, P.M., et al., *HugeIndex: a database with visualization tools for high-density oligonucleotide array data from normal human tissues.* Nucleic Acids Res, 2002. **30**(1): p. 214-7.
23. Mariani, T.J., et al., *A variable fold-change threshold determines significance for expression microarrays.* Faseb J, 2002.
24. Bakay, M., et al., *A web-accessible complete transcriptome of normal human and DMD muscle.* Neuromuscul Disord, 2002. **12 Suppl 1**: p. S125-41.
25. Iacobuzio-Donahue, C.A., et al., *Discovery of novel tumor markers of pancreatic cancer using global gene expression technology.* Am J Pathol, 2002. **160**(4): p. 1239-49.
26. Ganong, W.F., *Review of Medical Physiology*. Sixteenth ed. 1993, East Norwalk: Appleton & Lange.
27. Liu, W.M., et al., *Analysis of high density expression microarrays with signed-rank call algorithms.* Bioinformatics, 2002. **18**(12): p. 1593-9.
28. Lee, T.I., et al., *Transcriptional regulatory networks in Saccharomyces cerevisiae.* Science, 2002. **298**(5594): p. 799-804.
29. Lockhart, D.J., et al., *Expression monitoring by hybridization to high-density oligonucleotide arrays.* Nat Biotechnol, 1996. **14**(13): p. 1675-80.
30. Affymetrix, *Microarray Suite User Guide, Version 5*. Affymetrix, http://.com/support/technical/manuals.affx, 2001.
31. Affymetrix, *Statistical Algorithms Description Document.* Affymetrix, http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf, 2002.
32. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-64.
33. Hubbell, E., W.M. Liu, and R. Mei, *Robust estimators for expression analysis.* Bioinformatics, 2002. **18**(12): p. 1585-92.
34. Li, C. and W.H. Wong, *Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.* Proc Natl Acad Sci U S A, 2001. **98**(1): p. 31-6.
35. Blatt, M., S. Wiseman, and E. Domany, *Superparamagnetic clustering of data.* Phys Rev Lett, 1996. **76**(18): p. 3251-3254.
36. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.
37. Jain, A.K. and R.C. Dubes, *Algorithms for Clustering Data*. 1988, New Jersey: Prentice-Hall, Inc.
38. Getz, G., E. Levine, and E. Domany, *Coupled two-way clustering analysis of gene microarray data.* Proc Natl Acad Sci U S A, 2000. **97**(22): p. 12079-84.
39. Duda, R.O., P.E. Hart, and D.G. Stork, *Pattern Classification*. second ed. 2001, New York: John Wiley & sons, Inc.
40. Blatt, M., S. Wiseman, and E. Domany, *Data Clustering Using a Model Granular Magnet.* Neural Comput, 1997. **9**: p. 1805-1842.
41. Chalifa-Caspi, V., et al., *GeneAnnot: Interfacing GeneCards with high-throughput gene expression compendia.* Briefings in Bioinformatics, 2003. **4**(4): p. 1-12.
42. Chalifa-Caspi, V., et al. *GeneAnnot: Annotation of high-density oligunocleotide arrays and their linking with GeneCards*. in *ISMB*. 2003.
43. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002. **12**(4): p. 656-64.

44. Shmueli, O., et al., *GeneNote: whole genome expression profiles in normal human tissues.* Comptes rendus - Biologies, 2003. **326**(10-11): p. 1067-1072.
45. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society B, 1995. **57**: p. 289-300.
46. Agrawal, H. and E. Domany, *Potts ferromagnets on coexpressed gene networks: identifying maximally stable partitions.* Phys Rev Lett, 2003. **90**(15): p. 158102.
47. Agrawal, H., *Extreme self-organization in networks constructed from gene expression data.* Phys Rev Lett, 2002. **89**(26): p. 268702.
48. Yanai, I., et al., *Midrange genome-wide transcription profiles reveal expression/suppression relationships in human tissue specification. Submitted.* 2004.
49. Barad, O., *Advance clustering algorithms for gene expression analysis using statistical physics method*, in *Physics of Complex Systems*. 2003, Weizmann Institute of Science: Rehovot.
50. Safran, M., et al., *GeneCards 2002: towards a complete, object-oriented, human gene compendium.* Bioinformatics, 2002. **18**(11): p. 1542-3.
51. Tsafrir, I., et al. *Sorting Points Into Neighborhoods (SPIN)*. *In preparation*.