Functional profiling of gene expression in the human and chimpanzee brain

Diplom thesis

Bjoern Muetzel

1. Introduction

1.1. Background and aim

"Man adapts his environment to his genes more frequently and efficiently than his genes to his environment"
(T. Dobzhansky , 1900 – 75)

As pointed out in this quotation by one of the most influential population geneticists of this century, the cultural transfer of information from one generation to the next, has, in the course of hominisation, reached a velocity, that surpasses the biological evolution of man by many orders of magnitude. This acceleration in the cultural progression is the product of the unique ability and power of the human brain, that enabled man adapt environment to his needs.

Today many scientific disciplines, like psychology, neurobiology and anthropology address the question, how the human brain works and what makes it different than the one of other species. Like pieces of a puzzle, an enormous amount of knowledge has accumulated in the course of this research. Still, this scientific progress, which is itself a part of the human cultural evolution, fails to answer these questions, which refer to its own source and origin.

With the advances in molecular biology, researchers in this field are now able to add pieces to these questions. Hereby one promising approach is the large scale analysis of gene expression in the brain. To get a better understanding of its functionality and uniqueness, such an analysis was performed in two related experiments proceeding this work. In the first, genes differentially expressed between four different brain regions were identified. In the second, genes were determined, that differ in expression in the brain of humans and chimpanzees, their closest living relatives.

Before, from these lists of thousands of differentially expressed genes, a global picture of the underlying biological processes was attained, three tasks had to be accomplished. First, the expressed and differentially expressed genes from these experiments had to be classified into functional groups. Therefore a tool was developed, that automatically translates these two lists of genes into the functional categories provided in the Gene Ontology (GO) database, which classifies genes according to their biochemical activity, cellular component and biological function.

Second, a statistical test had to be applied to furthermore determine single categories, in which more or less differentially expressed genes than expected have accumulated.
Third, to correct for multiple testing of the thousands of functional categories tested, two different tests had to be developed, to assure, that the overall distribution of changed genes into the functional categories differs from a distribution given by chance.
Therefore two aims were pursued in this thesis. The first was, to shed light on the uniqueness and function of the human brain by determining groups of genes differentially expressed among brain regions and between brains of humans and chimpanzees. The second was, to develop a data mining tool, with which the underlying functional differences in the comparison of any large scale gene expression study can be rigorously examined.

Before the results of these efforts are described in Chapter 3, I will first give a short introduction, in which the background knowledge, necessary for the understanding of these results will be given. In this first part of this thesis I will describe some of the known differences between different brain regions and differences between human and chimpanzee brains. Then the principle of GeneChips$^{®}$, the platform used in this large scale analysis of gene expression and the factors influencing the measurements will be explained. Finally I will give an overview of the structure and concept of the GO database used to classifiy the genes into functional categories and the statistical tests, which can be used, to asses the significance of a category from the numbers of included detected and changed genes.
In the following materials and methods part, first the tissue samples and microarray data used for the analysis will be described. Then the databases, statistical tests and algorithms used in the course of the functional profiling of the expression data will be mentioned and explained.
After the results and discussion of the analysis, I will finish with an outlook on further developments and applications of the established data mining tool.

1.2. Differences between brain regions

A complete list of differences between brain regions would include all functional, anatomical and morphological, as well as molecular biological differences. Especially concerning the functional and molecular biological differences, such a list is not available at the current state of research.

I will therefore focus on a short anatomical and functional overview of the structure of the human brain and then present some of the known morphological and molecular biological differences of neurons in different brain regions.

The human brain is divided into six major regions, which can be further subdivided into several anatomically and functionally different areas. These regions are the medulla, pons, midbrain, diencephalons, cerebellum and the cerebral hemispheres, which constitute the telencephalon (Kandel 2000). These regions are depicted in Figure 1.1.

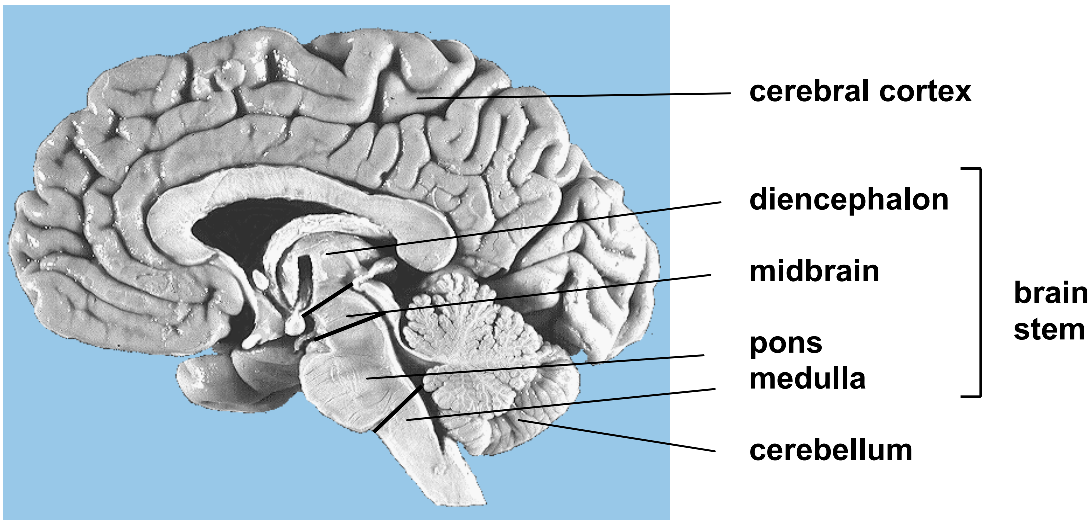


Figure 1.1. Regions of the human brain

The first four of these mentioned regions form the brain stem. The medulla is the direct rostral extension of the spinal cord, which it resembles both in organization and function. It is involved in the regulation of basic physiological processes like blood pressure and respiration. Situated rostral to the medulla is the pons. The ventral portion of the pons relais information from the cerebral cortex to the cerebellum. Above the pons lies the midbrain. The midbrain links components of the motor system, situated in the cerebellum, the basal ganglia and the cerebral hemispheres of the cortex. Additionally it contains components of the auditory and visual system.

The cerebellum is situated above the pons. It is involved in coordinating movements and in learning motor skills and therefore connected to somatosensory input from the cerebral cortex, the spinal cord and the vestibular organs. The density of neurons in the cerebellum is greater than in any other subdivision of the brain.

The diencephalon contains two major subdivisions, the thalamus and the hypothalamus. The thalamus has a gate keeping function for sensory information. It determines, which sensory signals reach awareness in the neocortex. The hypothalamus is essential for the homeostasis of the organism and controls basic body functions by regulating hormonal secretion in the pituary gland.

The telencephalon is the largest region of the human brain. It consists out of the cerebral cortex, white matter and inner structures. These consist out of amygdala, hippocampus and basal ganglia including caudate nucleus and putamen.

The cortical regions are the sites of higher cognitive functions, which are located functionally in specific areas. The two hemispheres are interconnected by the corpus callosum. The cerebral cortex consists out of six distinct layers that vary in their composition of different types of neurons. The amygdala is involved in the expression of emotion, the hippocampus in the processing of memory and the basal ganglia in regulating fine movement.

The brain is composed of neurons and protecting and nourishing sourrounding glia cells.

Neurons in general can be broadly defined as projection neurons or interneurons. Projection neurons are pyramidically shaped and use glutamate as their primary transmitter whereas local interneurons use inhibitory GABA. Several types of GABA neurons based on their pattern of connections and the cotransmitters can be identified.

It is known, that functionally different cortical regions differ in the thickness of the composing layers, which differ in the distribution of inter and projection neurons. Many neurons, especially of the brain stem and the basal nucleus, which is located beneath the basal ganglia, modulate attention and arousal. The axons of these neurons project through the whole brain and specific transmitters are released at the synapses. One prominent example are modulatory systems of dopaminergic neurons in the midbrain, which are concerned with reward systems. It is furthermore known that different functional cortical regions exhibit distinct receptor profiles for transmitter binding (Zilles, Palomero-Gallagher et al. 2002) .

Studies analyzing gene expression differences between different brain regions have already been conducted in mice (Sandberg, Yasuda et al. 2000; Carter, Del Rio et al. 2001; de Chaldee, Gaillard et al. 2003).

1.3. Differences between human and chimpanzee brains

As the chimpanzee is our closest living relative, the differences found between humans and chimpanzees are important to determine human uniqueness. Differences in cognitive abilities in apes and humans (Tomasello and Call 1997), anatomic brain comparisons (Rilling and Insel 1999), as well as molecular biological differences in human and chimpanzee brains (Buxhoeveden, Switala et al. 2001) have been used to so far to examine the distinctiveness of the human brain. As functional differences in gene expression patterns between human and chimpanzee brains might be related to these known differences, some of these will be explained in the following. As in general also genotypic differences might be related to these expression differences, these will be mentioned here as well. Figure 1.2. shows a human and a chimpanzee brain.

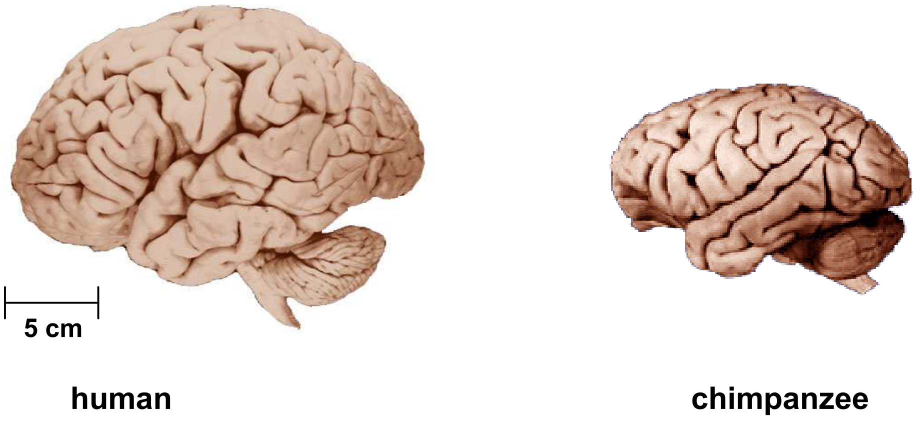


Figure 1.2. Human and chimpanzee brain

Using fossils and genetic data it was estimated that the split between humans and chimpanzees occurred ca. 4.8 - 6.4 million years ago (Chen and Li 2001). The number of observed nucleotide differences at single sites was estimated as 0.6 % for coding and 1.6 % for non coding sequences (Wildman, Uddin et al. 2003).

One of the few biochemical differences found so far between humans and chimpanzees is concerned with a deletion in the CMP-sialic acid hydroxylase gene. This deletion occurred on the human lineage and therefore humans can not turn the sialic acid N-acetylneuraminic acid (Neu5Ac) into N-glycolylneuraminic acid (Neu5Gc)(Chou, Takematsu et al. 1998; Irie, Koyama et al. 1998) .

Concerning anatomical differences between the human and chimpanzee brain, the most apparent difference is the size of the human brain. It is supposed to be 3.4 times larger than expected for an anthropoid primate of the same size, especially larger than expected is the neocortex (Rilling and Insel 1999). The neocortex of the human brain is also finer structured and contains more windings (gyri) and foldings (sulci) (Semendeferi, Lu et al. 2002). Morphologically, the minicolumns, basic anatomical units of many areas of the brain are supposed to differ in their structure between humans and chimpanzees (Buxhoeveden, Switala et al. 2001). One of the few known developmental differences concerns the migration of neurons from a structure in the telencephalon to the dorsal thalamus. This migration occurs in humans, but probably not in monkeys and mice (Letinic and Rakic 2001).

1.4. The principle of Gene Chips

The expression data analysed in this thesis was produced in microarray experiments. In such an experiment thousands of different gene specific DNA probes are positioned at a high density to distinct spots on a glass surface. Labeled RNA is allowed to hybridise with the probes and the intensity signal, which corresponds to the RNA quantity attached to a certain spot, is measured. This way the gene expression for thousands of genes can be measured and compared simultaneously (Zhao, Hashida et al. 1995).

Many platforms have been developed on which such a study can be conducted. The platform used in these experiments are GeneChips®, developed by Affymetrix® (Lockhart, Dong et al. 1996).

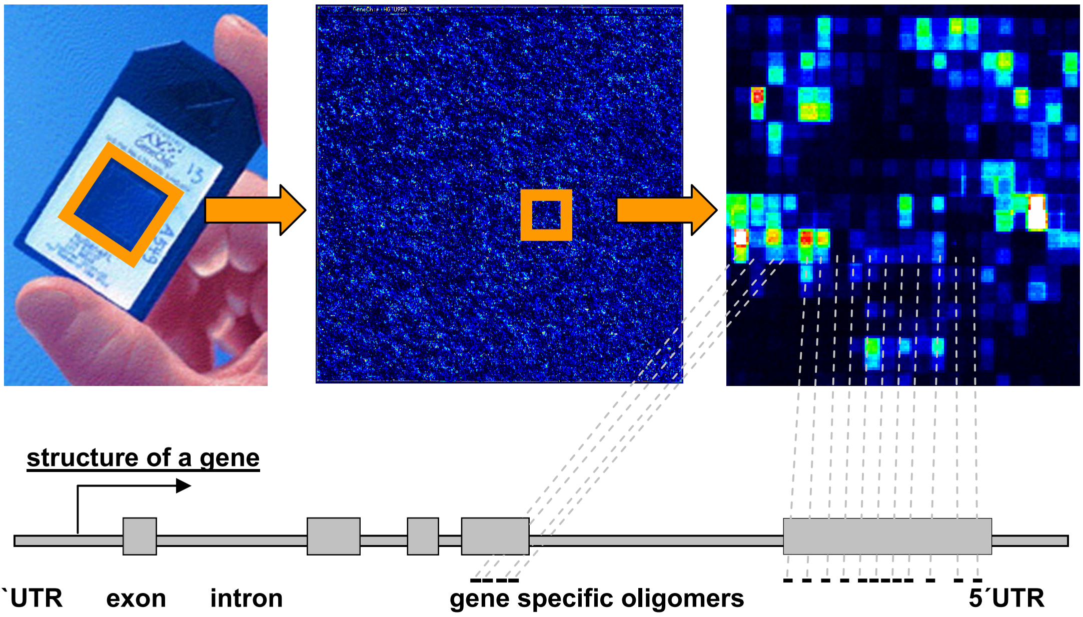


Figure 1.3. Design of a GeneChip®. The original size of a gene chip is 1.6 cm². In an amplification of the chip the small spots with fluorescent labeled RNA become visible. At an even higher resolution, these fluorescent 24x24 µm squares become clearly visible. The different intensities of the spots correspond to the amount of RNA attached. 16 different 25-mers, each spotted at high density to a single square are derived from the last exons of a gene.

As shown in Figure 1.3., a GeneChip® consists of a glass carrier on which DNA oligonucleotides are spotted on small squares. On each square of 24 x 24 µm in size, identical oligomers are attached at a high density. Sets of 16 oligomers (probesets) are complementary to a single mRNA from a human gene. The number of probesets is redundant, so that the 12.599 probesets on a chip correspond to about 8.800 genes.

In a microarray experiment using this platform, an RNA mixture, usually extracted from tissue samples or cell extracts, is labeled with a fluorescent marker and injected into the GeneChip® (Wodicka, Dong et al. 1997). The RNA is allowed to hybridise to the DNA on the chip and the RNA which did not attach to the oligomers is subsequently washed off. The intensity of the fluorescent signal is measured for each square. As the signal intensity is correlated with the RNA amount attached to the gene specific oligomers on the spots, RNA quantities can be measured for each of the squares and consequently for thousands of genes at a time.

To analyse the measurements, software provided by Affymetrix® was used. Implemented in this programs are statistical tests, that indicate first, whether the mRNA quantity for a certain gene was detectable beyond a certain threshold and second, if the expression of a gene is different in a comparison of two hybridisation experiments.

1.5. Factors influencing the microarray experiments

For the results of the data analysis, in which the detected and differentially expressed (changed) genes from the microarray experiments are grouped into functional categories it is crucial, how much noise the data contains. If many genes, due to measurement errors or other sources of variation are erroneously identified as differentially expressed, also a functional group will contain many false-positive differentially expressed genes just by chance. The probability of observing significantly more or less differentially expressed genes in such a category than expected, will subsequently be lower and the result of the study less meaningful.

In the following, possible sources of variation will be explained. The effect of some of these factors influencing the measurements has been estimated (see (Nadon and Shoemaker 2002)).

The signal intensity measured can be influenced by many factors, ranging from the quality and design of the oligomers, RNA isolation to scanning. The experimental variance can be estimated with the help of a replicated experiment for an individual (duplicate). At certain selection criteria, the number of differentially expressed genes identified between two duplicates can be compared to the number of differentially expressed genes found between two individuals. The ratio between these two numbers gives an estimate of the rate of false-positive differentially expressed genes caused by measurement errors. This rate is also an estimate of the experimental variance.

To determine genes differentially expressed between two species, the pairwise comparisons between each individual of the two species is used. In this comparison between two species, the variance due to intra-species differences can be estimated in a similar fashion as the experimental variance. This biological variance was assesed using a permutation test. An example, that illustrates, how this test works, is given in Figure 1.4.

Another source of variation, if two separate species are compared are sequence differences between these species. As the probesets on the microarray are usually designed from the genes of just one species used in the comparison, differences in the coding sequence of these two species, cause a different hybridization pattern between the RNA pools of the species.

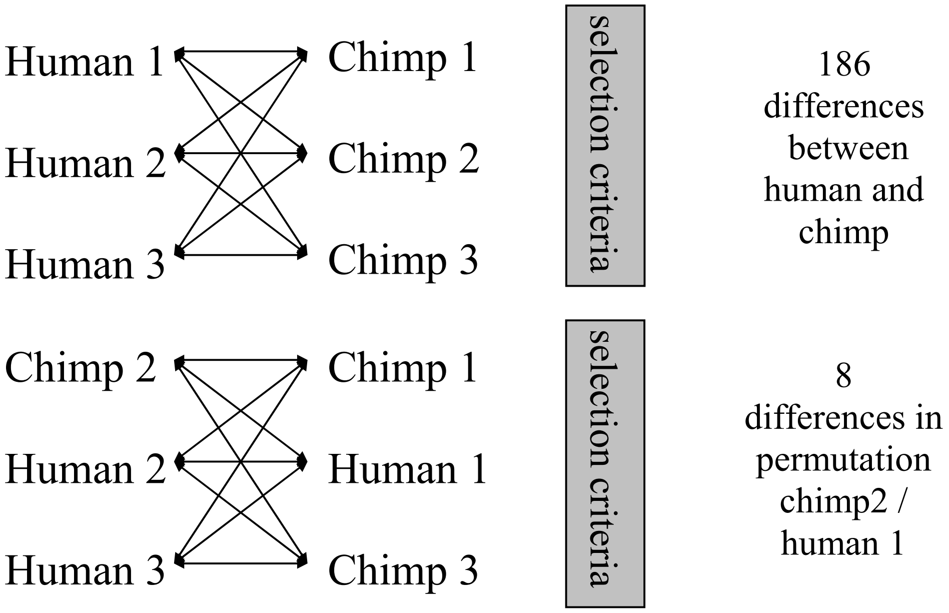


Figure 1.4. Estimation of the biological variance for a comparison between humans and chimpanzees. Through the nine pairwise comparisons between the three humans and three chimps, genes differentially expressed between humans and chimpanzees are determined using certain selection criteria. In the permutation test, the measurements for chimpanzee 2 and human 1 are exchanged and the genes differentially expressed between the species are determined in this comparison. Whereas 186 differences between humans and chimpanzees are found in the first setup, 8 are identified in the permutation. The average number of genes differentially expressed in all such possible permutations divided by the number of changed genes in the human-chimpanzee comparison gives an estimate of the biological variance.

The variation caused by sequence differences can be estimated by removing measurements from all oligomers of a probeset, which are known to contain sequence differences between the two species.

Other factors that influence the measurements are age of the individuals, condition of death and nutrition. Though in this study, we tried to control for these factors as well as possible, due to the limited amount of sample material, this was only possible to a limited extend.

1.6. The Gene Ontology Database

To group genes according to their function we used the Gene Ontology (GO) database (Ashburner, Ball et al. 2000). This curated database includes and structures the available knowledge about the genes from different genetic model organisms, like yeast, human and mouse. The advantage of the Gene Ontology over other databases is, that it organizes and describes the biological knowledge at all the different stages of completeness at which it is available.

This is achieved by structuring the data in a hierarchical classification tree, meaning, that the further downstream a gene can be classified the more detailed information is available for the corresponding gene. To simplify the search and identification of homologues proteins among different species, the Gene Ontology was designed to produce a dynamic, controlled vocabulary that can be applied to all eukaryotes.

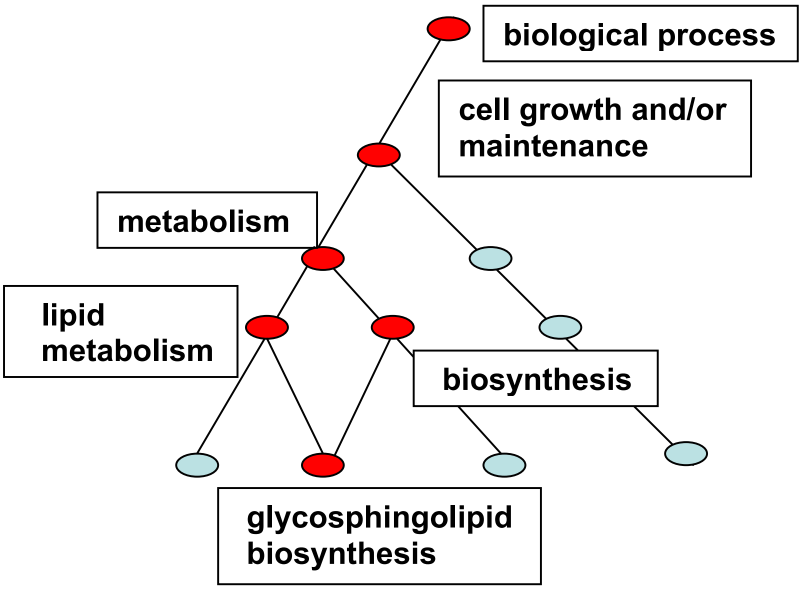

biological process
cell growth and/or maintenance
metabolism
lipid metabolism
biosynthesis
glycosphingolipid biosynthesis

Figure 1.5. Subtree of the *biological process* taxonomy. In this example the hierarchical structure of the classification tree is shown in a clipping of the *biological process* taxonomy. If a gene is known to be involved in *metabolism*, it will be classified accordingly. If more detailed information is available, it will be classified further downstream. *Glycosphingolipid biosynthesis* is both a subgroup of *biosynthesis* and *lipid metabolism*. In this part of the tree, the structure is reticle-like.

Furthermore an evidence code for each gene is provided, as the evidence, that a gene has a specific function can be experimental, inferred from sequence similarities or predicted.

In the used version of the GO database about 7.250 human genes are annotated, probably comprising about one forth of all genes of the human genome. The genes are classified in the three different taxonomies *molecular function* (*function*), *cellular component* (*component*) and *biological process* (*process*). In the version used in this work, in these taxonomies 5.306, 1.152 and 6.895 categories exist.

In the *molecular function* taxonomy, genes are grouped according to their biochemical activity (including specific binding to ligands or structures). Examples of unspecific terms on a high level are *enzyme*, *transporter* or *ligand*. Examples of low level classifications are *adenylate cyclase* or *serine-type peptidase.*

Cellular component refers to the place in the cell, where a gene product is located. This taxonomy includes such terms as *membrane* or *intracellular*. More specific localizations are *mitochondrial inner membrane* or *proteasome*.

In the *biological process* taxonomy genes are classified according to their biological function. Examples of broad functional terms are *physiological process* or *signal transduction*, more specific categories are *hearing*, *glycosphingolipid biosynthesis* or *sexual reproduction*.

An example of a sub tree of the *biological process* taxonomy is given in Figure 1.5. Notably, the taxonomy is not entirely tree-like, but may contain reticle-like structures.

1.7. Statistical tests

The result of the classification in the GO tree are the numbers of detected and differentially expressed or changed genes in each group of a taxonomy. The frequency of detected and changed genes in a group is the number of these genes in a group (including genes classified further downstream) divided by the total number of detected or changed genes in the corresponding taxonomy. If the differentially expressed genes were randomly distributed among the groups of a taxonomy, one would expect the observed frequency of changed genes to equal the frequency of detected genes in each category.

Table 1.1. *P*-value for a functional category in the test using the hypergeometric distribution

group name	N	M	F-det	K	x	F-chan	*x-exp*	*P*-l	*P*-r
glycosphingolipid biosynthesis	2789	7	0.0025	360	6	0.0167	1	0.99	0.00003

N = number of genes annotated in the taxonomy *biological process*
M = number of detected genes in the functional group
F-det = frequency of detected genes in the functional group
K = number of changed genes annotated in the taxonomy *biological process*
x = number of changed genes in the functional group
F-chan = frequency of changed genes in the functional group
x-exp = number of changed genes expected from the frequency of detected genes
P-l = *P*-value for conservation in the test using the hypergeometric distribution
P-r = *P*-value for change in the test using the hypergeometric distribution
In Table 1.1. the HG test is used to determine a *P*-value for change or conservation of a functional group. In the example 7 genes, 0.25 %, out of 2789 in the process taxonomy are classified in this group. Among the 360 differentially expressed genes, 6 genes, 1.7 % are classified in this group. If the changed genes were randomly distributed among the functional categories, one would expect about the same frequency of changed genes as detected genes to occur in this category. Therefore, 1 gene would be expected in this category. Using the hypergeometric distribution, one can determine the *P*-value for observing 6 changed genes instead of 1in this category, which is 0.00003.

With a test using the hypergeometric distribution (HG test) or a Chi square test for the equality of proportions, it can be assessed, if a significantly higher or lower number of differentially expressed genes than expected from the frequency of detected genes in a group occur (Doniger, Salomonis et al. 2003; Draghici, Khatri et al. 2003). A group can be defined as a changed group, or a group changed in its expression profile, if significantly more differentially genes occurred in this group using the HG test. A group can be defined as conserved or conserved in its expression profile, if significantly less than expected differentially expressed genes occurred in this group.

2. Material and Methods

2.1. Tissue samples and microarray data collection

Brain tissue was sampled from three male humans who were 45, 45, and 70 years old, had no history of brain related diseases and had suffered sudden deaths without associated brain damage. Tissue was dissected from Broca's area, dorsolateral prefrontal cortex, premotor cortex, primary visual, anterior cingulate cortex, the caudate nucleus and Vermis cerebelli (see Figure 3.1.).

Brains were similarly removed at autopsies from three male chimpanzees who were 12, 12 and approximately 40 years old and had also all died from natural causes. From these brains, the same brain regions as in the humans were removed in a similar fashion.

Expression data was collected using Affymetrix® HG U95Av2 arrays as well as Affymetrix® HG U95B, C, D, and E arrays and analyzed with Affymetrix Microarray Suite v 5.0 using default parameters. For analysis all arrays were scaled to the same average intensity using all probes on the array. Altogether 62.000 probesets spotted on the five arrays, have been used in the studies (Khaitovich 2003).

Genes differently expressed or changed between brain regions were determined using the comparisons within each individual separately according to certain selection criteria. These criteria were established using sets of duplicate experiments for three brain regions. Each set consisted out of an independently prepared and hybridized RNA probe of the brain tissue from three individuals. The genes, that were stated as changed in an array, but not in its duplicate, were used to estimate the false-positive rate of differentially expressed genes due to experimental variance. The applied selection criteria resulted in a false-positive rate of differentially expressed genes of less than 1 %. Under this condition 19466 probesets were stated as detected and 3817 were identified as differentially expressed.

Gene expression levels were compared in each brain region separately in all nine possible pairwise comparisons among the three individuals of each species. Again, the rate of false-positives due to experimental variance measured with the duplicates was less than 1 %. As the criteria for detection was the same in this comparison, 19466 detected probesets were identified. Among these, 2577 were stated as differentially expressed.

2.2. Sources of variation in gene expression and masking for sequence differences

In order to evaluate to what extent three human and three chimpanzee individuals are enough to evaluate inter-species gene expression differences, the expression measurements for each gene were randomized with respect to the individual (irrespective of species affiliation) in which it occurred for each of the six brain regions. For 54 such data sets on average 14 genes differed significantly between the two groups of three individuals in all nine possible comparisons. Since on average 302 differences are found in the non-randomized data, about 5% of the observed differences between the species are expected to be caused by the variation among individuals within the two species. (see Chapter 1.5.)

In order to test to what extent nucleotide sequence differences between humans and chimpanzees may influence the results, oligonucleotides among the 16 probes targeting a gene were excluded from the analysis, if they contained sequence differences. Using published chimpanzee sequence data, it was estimated, that approximately 22% of the genes classified as differently expressed between humans and chimpanzees are caused by nucleotide sequence differences between the species.

Using a masking algorithm, that identifies outliers among the measured RNA quantities for the 16 probes targeting a gene, we could remove 45 % of all DNA probes with sequence differences on the arrays. After removal of these probesets, 18522 probesets were determined as detected and 2014 as changed.

2.3. Databases involved in the annotation process

- Affymetrix NetAffx database (www.affymetrix.com - Oct. 2002 release) (Liu, Loraine et al. 2003)
- Unigene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene - Jan. 2003 release) (Wheeler, Church et al. 2003)
- LocusLink, the LL_tmpl file (ftp.ncbi.nih.gov/refseq/LocusLink/ - release from Feb. 6th 2003) (Pruitt and Maglott 2001)
- GeneOntology (www.godatabase.org/dev/database/archive/ - Jan. 2003 release) (Ashburner, Ball et al. 2000).

Relevant information from these databases was downloaded and stored locally in a relational database:

- MySQL, Database Management System (www.mysql.com)

2.4. Chi square test for the significance of the overall distribution into categories

To assess the significance of the set of differentially expressed genes, we calculated the Chi square distance between the detected and changed numbers of genes for each subgroup of a taxonomy, given by the following formulas :

First, q_t, the probability of a detected gene to occur in group t is defined by

$$q_t = \frac{M_t}{\sum_{s \in T} M_s} \qquad (1)$$

where T is the set of groups, that are subgroups of a certain taxonomy and M_t is the number of detected genes in such a group t. In $\sum_{s \in T} M_s$ the detected genes of all groups of a taxonomy are summed up. In this sum, genes that belong to a group are counted again in the corresponding parent group. We applied a Chi square test to the table of detected and changed genes among the functional categories of a taxonomy. The definition of the probability of a detected gene to occur in category t is therefore not the frequency of detected genes in a group, but defined as in (1).

Accordingly, the expected number of differentially expressed genes of a certain taxonomy to occur in group t is then :

$$nq_t = q_t \cdot \sum_{s \in T} x_s \qquad (2)$$

where x_t is the observed number of differentially expressed genes in group t of the taxonomy T. The sum of the observed number of changed genes in all groups of a taxonomy ($\sum_{s \in T} x_s$) times the probability of a gene to occur in group t is the expected number of differentially expressed genes to occur in group t.

The Chi square distance d_t between the changed and detected number of genes in a single group t of a taxonomy is then:

$$d_t = \frac{(x_t - nq_t)^2}{nq_t} \qquad (3)$$

Note, that, as the square between the expected and observed number of differentially expressed genes is taken in this formula, this test does not distinct between groups with a higher and groups with a lower number of changed genes than expected.

Finally the Chi square distance, that gives a probability weighted distance between the distribution of changed and detected genes across the functional categories dist_T is given by the sum of Chi square distances of each category in this taxonomy T:

$$dist_T = \sum_{teT} d_t \qquad (4)$$

This Chi square distance calculated for the changed set of genes was subsequently calculated for 10.000 random sets sampled from the list of detected genes. A random set had been generated by sampling a random set of genes of the same size of the changed set of GO annotated genes from the detected set of GO annotated genes. To test, whether the difference between the distribution of changed and detected genes into functional groups in a taxonomy is significantly higher than the difference between the distribution of randomly sampled and detected genes, a *P*-value was assigned as the proportion of random sets with a higher or equal distance than the one observed in the data set.

2.5. Refinement of the groups significant in the hypergeometric test

If a taxonomy was significant in the Chi square test, consequently the numbers of significantly changed and conserved groups in such a taxonomy were determined at different significance levels with a test using the hypergeometric distribution (see Chapter 1.7.). These numbers of conserved and changed groups were equally determined for the 10.000 random sets. At each significance level the average of significantly changed and conserved groups in the 10.000 random sets represents the number of false-positive significant groups that can be found by chance due to the multiple testing of the thousands of GO categories. For the further analysis a significance level was chosen where this number of false-positive groups was significantly low in the observed data. (for details see Chapter 3.2.3.2.)

Still, a group on a higher level in a taxonomy might be significant in the hypergeometric test, due to the fact, that it contains a high proportion of genes from significant subgroups. If the genes in these subgroups were not contained in this group, it might not be significant any more. This means, that the remaining genes in this group do not provide further information about the change or conservation of this group. As the subgroups provide more exact information about the categorized genes, the parent group can now be removed from the list of significant or informative groups. The algorithm we developed to refine the sub tree S of groups significant in the hypergeometric test is based on this idea. It is recursively defined on a procedure that works in five steps in this sub tree S:

In step one the significant groups, that do not contain any sub nodes are identified. In step two all significant groups that are direct parents of these in the sub tree of significant groups and contain exclusively groups on the last hierarchical level of the tree are identified. In step three, for each of these parent groups the distinct set of detected and changed genes in all the subgroups is evaluated and subtracted from the observed number of detected and changed genes in the according parent group. In step four, all these parent nodes are tested for significance on a 5% level using the hypergeometric distribution with the numbers of remaining detected and changed genes from the group as parameters. If a parent node is not significant any more in this test, the group is taken out from the list of significant groups S in step five. In this step it is also removed from the list S′, the updated list of significant groups used for the refinement in the following round of the recursion. If it is still significant in step five, its sub groups are removed from the list S´, but stay in the list of significant groups S. The recursion starts again in step one with the updated list S´ of significant groups defined by its proceeding step. The algorithm is used recursively until all significant groups have been tested, yielding a subset of refined significant groups S.

3. Results

3.1. Data sets

In the microarray experiments proceeding this analysis brain tissue was sampled from three adult male humans and three adult male chimpanzees (Khaitovich 2003). For each individual, samples were collected from four cortical regions, cerebellum and the caudate nucleus. (see Figure 3.1.)

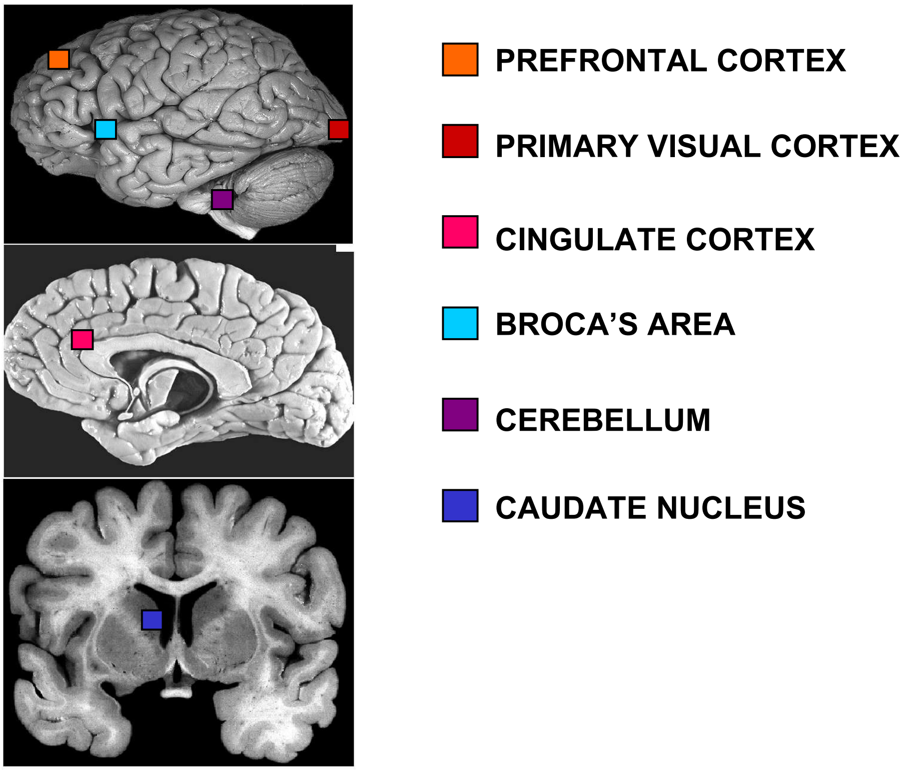


Figure 3.1. Different brain regions sampled in the study

RNA extracted from these samples was hybridized to the Affymetrix® array HG U95Av2 and to the arrays HG U95B, C, D, and E, which contain approximately 62.000 probesets corresponding to 43.300 Unigene cluster or human genes (Wheeler, Church et al. 2003). In order to identify genes, which differ between brain regions, all pairwise comparisons between the four different brain regions cingulated cortex, Broca´s area, caudate nucleus and the cerebellum within each human individual were made. Genes were defined as changed in expression between the brain regions, if they were, according to certain selection criteria, differentially expressed between these regions in all three individuals.

In the second study, RNA quantities were compared in each brain region separately in all pairwise comparisons among the three individuals of each species. (see Figure 1.4.)

Since in the first study, expression differences among the brain regions of one individual are compared, only experimental variance affects the result of the genes determined as differentially expressed. With the help of replicate experiments, this variance could be determined as consistently low, resulting in a false-positive rate of differentially expressed genes of less than 1 % in both studies. Yet, about 5 % of the observed differences between the species were due to biological variance and about 22 % are due to sequence differences between humans and chimpanzees (see Chapter 2.2.).

To reduce the erroneous measurements introduced by sequence differences, a masking algorithm was applied to the probesets on the Affymetrix chips. Hereby, measurements from DNA probes were removed from the probeset designed for each human gene, if they hybridise significantly different than the other probes. This way about 45 % of the probes containing sequence differences were removed.

3.2. Algorithms used in the functional profiling

3.2.1. Annotation process

After the lists of detected and differentially expressed or changed probesets had been determined, the corresponding genes were classified and annotated in the Gene Ontology database to derive the underlying biological functions involved in the condition under study. As shown in Figure 3.2., four different databases are involved in this annotation procedure, which consists out of three steps.

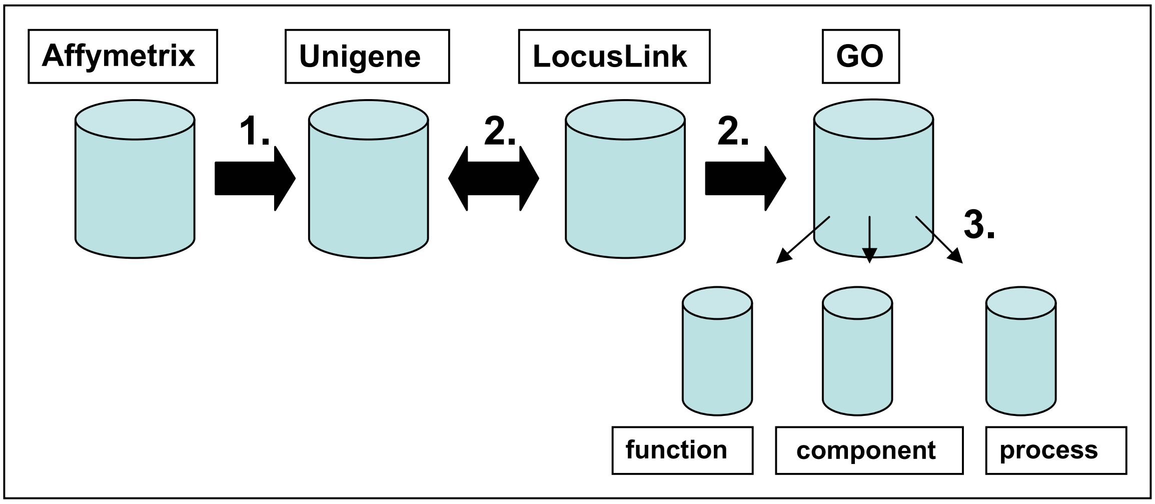


Figure 3.2. Different steps of the annotation process. Shown are the different databases - Affymetrix, Unigene, Locuslink and GO database - which are used in the three steps of the annotation process (1.–3.). The *function*, *component* and *process* taxonomies belong to the Gene Ontology database.

In the first step, the probeset names are linked to the corresponding Unigene cluster via their common Genebank Accession number using the information provided by NetAffx. Unigene cluster are meant to represent single genes and more than one probeset can match to a single Unigene cluster. A Unigene cluster was defined as differentially expressed, if at least one associated probeset was differentially expressed.

The obtained Unigene cluster are linked to their GO annotation – if available – via LocusLink in the second step, which results in a subset of GO annotated cluster. Step three navigates through the three taxonomies for *molecular function (function)*, *biological process (process)* and *cellular component (component)*, respectively, and assigns the cluster to their specific GO function. Note that a gene might belong to several functional categories, even among a single taxonomy. The categories, whose cardinalities are used in the following statistical tests are groups. A group is the list of genes belonging to a functional category and the associated sub nodes.

3.2.2. Resampling

To assess, if the overall distribution of differentially expressed genes into functional categories differs significantly from the distribution of the detected genes, a resampling method was used. To simulate random distributions of the detected genes into functional categories, we constructed 10.000 random sets of genes, by sampling each time the number of changed genes in the GO tree from the set of detected genes in the ontology and classified them accordingly in the three taxonomies. An example of the resampling is given in Figure 3.3.

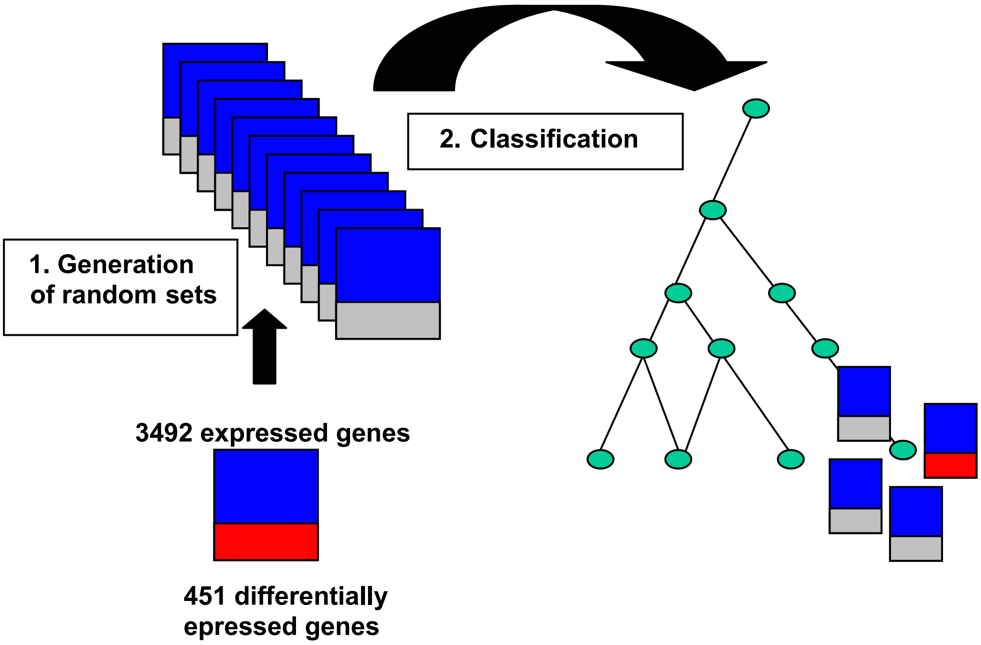

2. Classification
1. Generation of random sets
3492 expressed genes
451 differentially epressed genes

Figure 3.3. Resampling method. Shown is an example, how random distributions of detected genes across functional categories or random sets are generated from a data set. In this case, the data set consists out of 3492 expressed and 451 differentially expressed genes that have at least one annotation in the GO tree. In step one 10.000 random sets are generated by randomly picking 10.000 times 451 genes from the set of 3492 annotated genes. In step two these sets of randomly sampled genes are classified into the functional groups of the 3 taxonomies of the Gene Ontology. This way random distributions of detected genes into functional categories are produced.

3.2.3. Test for the overall significance of the distribution of changed genes into functional categories

With the test using the hypergeometric distribution, significantly changed or conserved categories are determined (Draghici, Khatri et al. 2003). Yet, it is unclear, if the groups identified are not a product of the multiple testing of the thousands of categories tested in the GO database.

The following statistical tests therefore asses, if the differentially expressed genes are, with respect to the distribution of the detected genes, randomly distributed across the functional categories of each of the three taxonomies or not. This significance was examined, by comparing their distribution to the distributions of randomly sampled expressed genes generated as just described. Two tests, first a Chi square test and then a test for the number of false-positive groups in the test using the hypergeometric distribution were used to asses the overall significance of the data.

3.2.3.1. Chi square test for a taxonomy

First, we applied a Chi square test to the table of detected and changed genes among the functional categories of a taxonomy. In this test, we calculated for each group of a taxonomy the Chi square distance between the observed number of changed genes and the number of differentially expressed genes expected from the number of detected genes in each subgroup of a taxonomy. The more these two numbers differ, the higher will be the Chi square distance in this category. The Chi square distance for a taxonomy is defined as the sum of Chi square distances of all its associated subgroups. This measurement gives a probability weighted distance between the distribution of changed and detected genes among functional categories of a taxonomy. (see Chapter 2.4.)

This Chi square distance calculated for the changed set of genes was subsequently calculated for the 10.000 random sets sampled from the list of detected genes. To test, whether the difference between the distribution of changed and detected genes into functional groups in a taxonomy is significantly higher than the difference between the distribution of randomly sampled and detected genes, a *P*-value was assigned as the proportion of random sets with a higher or equal distance than the one observed in the data set.

This test can also be modified, to test if the distribution of changed genes across functional categories clustering under a certain category is significant. This might be important, if, given a certain hypothesis, one is interested only in a particular category and its associated subgroups.

3.2.3.2. Test for the number of false-positive groups in a taxonomy

With the Chi square test a taxonomy is identified, in which the distribution of changed genes into functional categories differs significantly from the distribution of detected genes. Consequently we estimated for such a taxonomy the number of false-positive groups significant in the test using the hypergeometric distribution. We therefore calculated the mean number of significant groups in the observed data and in 10.000 random sets at the significance levels of 10 %, 5%, 1% and 0.1%. The mean number of significant groups in the 10.000 random sets gives an estimate of the number of false-positive groups. At each cutoff the significance of the data set was estimated from the proportion of random sets that contained an equal or larger number of significant groups than the data set. If less than 5 % of the random sets contained an equal or larger number of significant groups than in the experimental data at a significance level of 5 %, these groups were used for the following refinement algorithm. If this was not the case, the groups from the first significance level lower than 5%, where this condition holds, were chosen.

3.2.4. Groups significant in the hypergeometric test and refinement

From the list of groups significant in the HG test, only the significant groups under a significance level have been chosen, under which a low rate of false-positive groups is expected (see Chapter 2.5.).

Yet, a group on a higher level of taxonomy may be significant in the test using the hypergeometric distribution due to the fact that it contains genes from the significant subgroups. Therefore, in the refinement procedure starting from the lowest possible level of taxonomy, for each group containing significant subgroups all detected and changed genes that belong to the subgroups were removed and only the remaining genes were tested in the test using the hypergeometric distribution. If the group lost its significance, it was removed from further analysis; otherwise the groups higher in the taxonomy were further tested in the same procedure. An example of the algorithm is given in Figure 3.4.

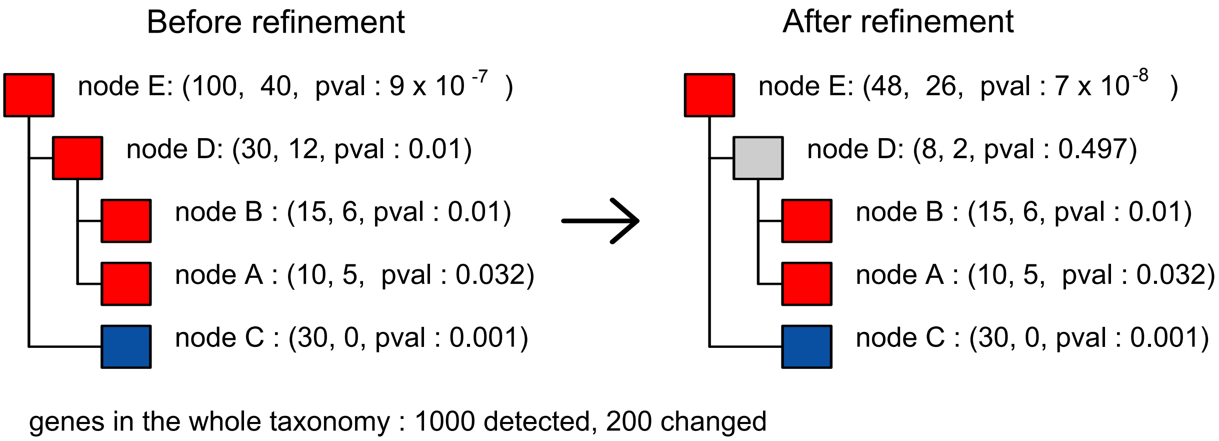

Before refinement
node E: (100, 40, pval : 9 x 10 -7)
node D: (30, 12, pval : 0.01)
node B : (15, 6, pval : 0.01)
node A : (10, 5, pval : 0.032)
node C : (30, 0, pval : 0.001)
After refinement
node E: (48, 26, pval : 7 x 10-8)
node D: (8, 2, pval : 0.497)
node B : (15, 6, pval : 0.01)
node A : (10, 5, pval : 0.032)
node C : (30, 0, pval : 0.001)
genes in the whole taxonomy : 1000 detected, 200 changed

Figure 3.4. Refinement algorithm. In the example, the tree *Before refinement* contains the groups significant in the hypergeometric test, used in the refinement. In the according taxonomy, 1000 detected and 200 changed genes can be found. The numbers attached to each node A – E are the corresponding numbers of detected and changed genes and the according *P*-value from the HG test. Conserved groups are marked blue, changed groups are marked red. To refine this list, one proceeds the following: First node D is tested for significance, as it only contains sub nodes on the last significance level. In this imaginary example, his sub nodes A and B share 3 detected and 1 changed gene. Together they therefore contain 22 detected and 10 changed genes. Without the genes from its sub nodes, node D contains 8 detected and 2 changed genes. The *P*-value for this remaining numbers of genes is 0.497. Therefore, this node is used in the next recursion of the algorithm. In the example it is marked grey in the tree *After refinement* This set consists now out of the nodes A, B, C and E. Now node E is tested. The number of the detected genes in its sub nodes is 52 and the corresponding number of changed genes is 10, as group C does not share any genes with group A and B. Therefore group E contains 48 remaining detected and 26 remaining changed genes. The *P*-value for these numbers of genes is $7 \cdot 10^{-8}$. Therefore the node is not removed in the refinement. As all groups have been tested here, the refinement algorithm ends.

Note that for each group there is a *P*-value for change and conservation before and after refinement, the latter one testing the significance of the remaining genes. For the significance after refinement a significance level of 5 % was chosen.

3.3. Functional profiling of genes differentially expressed among different brain regions

3.3.1. Annotation process

Table 3.1. Detected and changed probesets / unigenes after different steps of the annotation process

	detected	**changed**
probesets	19466	3817
unigenes	13591	3057
GO annotated (ann.) unigenes	3633	885
unigenes ann. in *molecular function*	2919	730
unigenes ann. in *cellular component*	2219	505
unigenes ann. in *biological process*	2897	710

The algorithms described were applied to the set of expressed and differentially expressed genes among cingulate cortex, Broca´s area, cerebellum and the caudate nucleus. In the first part of the analysis, these genes were annotated in the Gene Ontology.

A summary of the annotation process is shown in Table 3.1. The 19466 expressed probesets in the data set correspond to 13591 unigenes. This corresponds to 69.8 % of the probesets. The 3817 probesets differentially expressed between the four brain regions map to 3057 unigenes, which corresponds to 80.1 % of the probesets. 22.5 % of the 13591 detected unigenes are classified as differentially expressed. 26.7 % of the detected and 28.9 % of the differentially expressed unigenes are annotated in the Gene Ontology database.

Table 3.2. Distribution of annotated unigenes among functional groups

groups in *molecular function*	1299
groups in *cellular component*	282
groups in *biological process*	1105

Shown are the number of groups in each taxonomy into which the detected unigenes cluster

In Table 3.2. the number of groups in each taxonomy are shown into which these unigenes cluster. Though the numbers of annotated unigenes in each taxonomy are comparable, the unigenes are grouped into about four times more groups in the *function* and *process* taxonomy than in the *component* taxonomy.

3.3.2. Tests for the significance of the distribution of differentially expressed genes across functional categories

After annotating all genes in the Gene Ontology database, the overall significance of the data was examined. The data set shows a significantly high Chi square distance in the taxonomies *molecular function*, *cellular component* and *biological process*. In all cases the *P*-value is < 0.0001. This means, that the difference between the distribution of differentially expressed genes across functional categories and the distribution of detected genes is significantly higher than the difference between a distribution of randomly sampled and detected genes.

Table 3. 3. Significant groups at a 5 % significance level the test using the hypergeometric distribution

	groups changed in their expression profile			groups conserved in their expression profile		
taxonomy	# in data set	random mean	*P*-value	# in data set	random mean	*P*-value
molecular function	58	16.0	< 0.0001	28	6.7	0.0002
cellular component	6	4.8	0.358	26	2.7	< 0.0001
biological process	40	16.7	0.0032	30	8.4	0.0004

in data set : number of significant groups in the data set
random mean : mean number of significant groups among all random sets
P-value : *P*-value for the significance of the number of significant groups in the data set, given as the proportion of random sets with a number of significant groups that is higher or equal than the one given by the data set (< 0.0001, if no random set shows more significant groups than the data set)

Table 3.3. shows that at a significance level of 5 % an excess of significant groups can be found in the data set compared to the mean of significant groups found in the 10.000 random sets. Except to the significantly changed groups in the taxonomy *cellular component* no more than 3.2 % of all random sets show a number of significant groups higher than the one in the data set in all three taxonomies concerning groups changed as conserved in their expression profile. Yet in the taxonomy *cellular component* at a significance level of 1 % and 0.1 % more significantly changed groups than expected by random can be found. A table showing the number of false-positive groups at other significance levels is given in the appendix (Table A.1.).

Given the chosen significance levels, the lowest rate of false-positive groups can be found in the taxonomy *cellular component*. In this case the rate of false-positive groups is 10.4 %. The highest rate of false-positive groups is 41.8 % for changed groups in *biological process*. Many groups in the three taxonomies are significant even after a Bonferroni correction on a 5% level (see appendix Table B1.-3.)

3.3.3. Summary of the significant groups

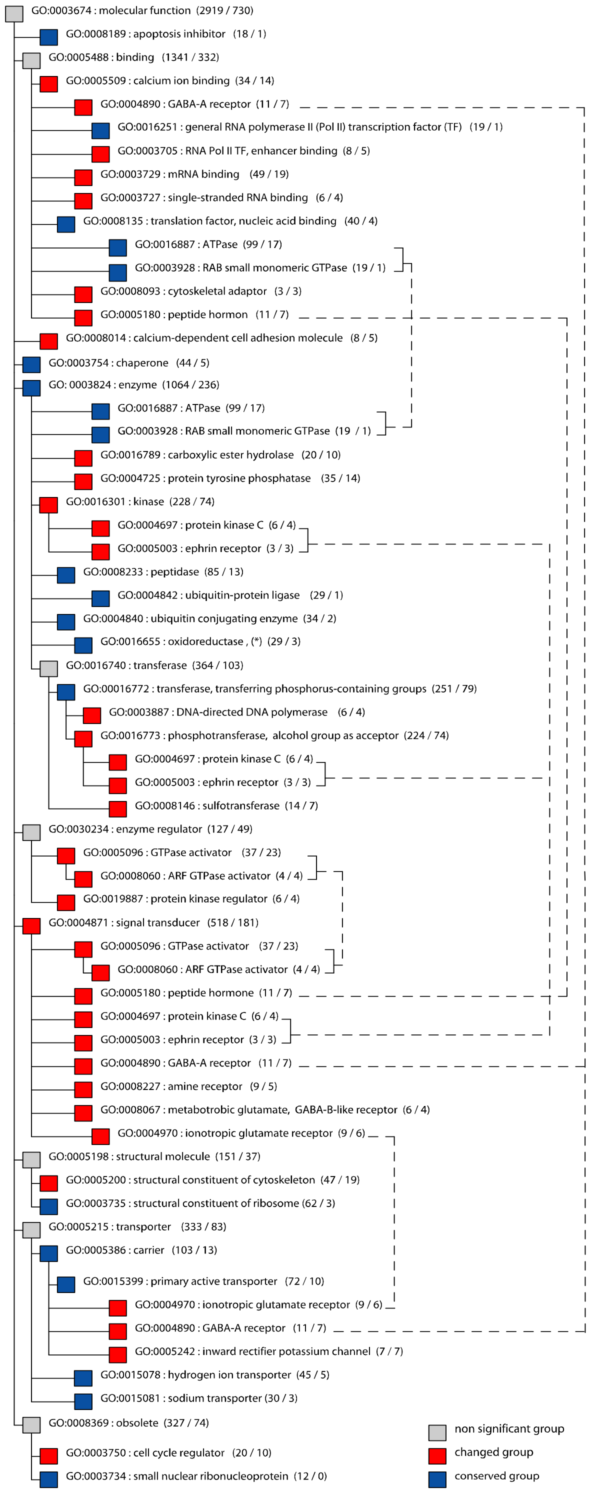
After the rate of false-positive groups had been estimated and the groups in the HG test had been refined, the tree of significant groups in each taxonomy was plotted to visualize the clustering of significant groups. Note, that there is a *P*-value for change and conservation of the groups before and after refinement. In Figure 3.5. the tree of significant groups after refinement is shown. A table of the significant groups with both *P*-values before and after refinement is given in the appendix (see Table B.1-.3).

As the *P*-value before refinement gives the significance of a group, unrelated to which subgroups it contains, this *P*-value is given in the following.

3.3.3.1. *Mecular function* taxonomy

As can be seen from Figure 3.5., using the test for the hypergeometic distribution after refinement 26 groups can be found to have changed in their expression profile between the brain regions and 18 show a conserved expression profile between the regions at a 5% level. A summary of these groups is given in the following.

The largest changed group in the taxonomy is *signal transducer* with 518 genes. It is also the group with the lowest *P*-value (1.45×10^{-8}). Only significantly changed groups cluster under *signal transducer*. Among these 10 groups are the groups *peptide hormone, amine receptor, GABA receptors, protein kinase C and GTPase activator. GTPase activator* is the second most significant group with a *P*-value of 1.6×10^{-6}. Another group, which does not cluster under *signal transducer*, but is known to be involved in signal transduction is *RAB small monomeric GTPase*. This group of genes shows a uniform expression between the brain regions.


GO:0003674 : molecular function (2919 / 730)
GO:0008189 : apoptosis inhibitor (18 / 1)
GO:0005488 : binding (1341 / 332)
GO:0005509 : calcium ion binding (34 / 14)
GO:0004890 : GABA-A receptor (11 / 7)
GO:0016251 : general RNA polymerase II (Pol II) transcription factor (TF) (19 / 1)
GO:0003705 : RNA Pol II TF, enhancer binding (8 / 5)
GO:0003729 : mRNA binding (49 / 19)
GO:0003727 : single-stranded RNA binding (6 / 4)
GO:0008135 : translation factor, nucleic acid binding (40 / 4)
GO:0016887 : ATPase (99 / 17)
GO:0003928 : RAB small monomeric GTPase (19 / 1)
GO:0008093 : cytoskeletal adaptor (3 / 3)
GO:0005180 : peptide hormon (11 / 7)
GO:0008014 : calcium-dependent cell adhesion molecule (8 / 5)
GO:0003754 : chaperone (44 / 5)
GO: 0003824 : enzyme (1064 / 236)
GO:0016887 : ATPase (99 / 17)
GO:0003928 : RAB small monomeric GTPase (19 / 1)
GO:0016789 : carboxylic ester hydrolase (20 / 10)
GO:0004725 : protein tyrosine phosphatase (35 / 14)
GO:0016301 : kinase (228 / 74)
GO:0004697 : protein kinase C (6 / 4)
GO:0005003 : ephrin receptor (3 / 3)
GO:0008233 : peptidase (85 / 13)
GO:0004842 : ubiquitin-protein ligase (29 / 1)
GO:0004840 : ubiquitin conjugating enzyme (34 / 2)
GO:0016655 : oxidoreductase , (*) (29 / 3)
GO:0016740 : transferase (364 / 103)
GO:00016772 : transferase, transferring phosphorus-containing groups (251 / 79)
GO:0003887 : DNA-directed DNA polymerase (6 / 4)
GO:0016773 : phosphotransferase, alcohol group as acceptor (224 / 74)
GO:0004697 : protein kinase C (6 / 4)
GO:0005003 : ephrin receptor (3 / 3)
GO:0008146 : sulfotransferase (14 / 7)
GO:0030234 : enzyme regulator (127 / 49)
GO:0005096 : GTPase activator (37 / 23)
GO:0008060 : ARF GTPase activator (4 / 4)
GO:0019887 : protein kinase regulator (6 / 4)
GO:0004871 : signal transducer (518 / 181)
GO:0005096 : GTPase activator (37 / 23)
GO:0008060 : ARF GTPase activator (4 / 4)
GO:0005180 : peptide hormone (11 / 7)
GO:0004697 : protein kinase C (6 / 4)
GO:0005003 : ephrin receptor (3 / 3)
GO:0004890 : GABA-A receptor (11 / 7)
GO:0008227 : amine receptor (9 / 5)
GO:0008067 : metabotrobic glutamate, GABA-B-like receptor (6 / 4)
GO:0004970 : ionotropic glutamate receptor (9 / 6)
GO:0005198 : structural molecule (151 / 37)
GO:0005200 : structural constituent of cytoskeleton (47 / 19)
GO:0003735 : structural constituent of ribosome (62 / 3)
GO:0005215 : transporter (333 / 83)
GO:0005386 : carrier (103 / 13)
GO:0015399 : primary active transporter (72 / 10)
GO:0004970 : ionotropic glutamate receptor (9 / 6)
GO:0004890 : GABA-A receptor (11 / 7)
GO:0005242 : inward rectifier potassium channel (7 / 7)
GO:0015078 : hydrogen ion transporter (45 / 5)
GO:0015081 : sodium transporter (30 / 3)
GO:0008369 : obsolete (327 / 74)
GO:0003750 : cell cycle regulator (20 / 10)
GO:0003734 : small nuclear ribonucleoprotein (12 / 0)
non significant group
changed group
conserved group

(*) : acting on NADH or NADPH, quinone or similar compound as acceptor

Figure 3.5. Significant groups in the *function* taxonomy after refinement. Given in parenthesis are the numbers of detected and changed genes in a group. Non significant groups are shown to structure the figure. If a group is categorised several times at different locations of the tree these locations are connected with a dashed line.

The second largest group, with an excess of genes that are differentially regulated between the four brain regions is *kinase* with 228 genes. Additionally all 224 genes in the changed group *phosphotransferase, alcohol group as acceptor* are kinase. Kinase and phosphotransferase both play an eminent role in activating proteins in signal cascades and are therefore involved in signal transduction.

The largest significantly conserved group is *enzyme* (1064 genes). It contains a mixture of conserved and changed groups, significant in the HG test, as enzymatic functions are related to many different biological phenomena.

The second largest group that is conserved after refinement is *transferase, transferring phosphorus containing groups*. It contains 251 genes. This group is a parent group of the changed groups *phosphotransferase, alcohol group as acceptor* and *DNA-directed DNA polymerase*. This group is the only case, where a group is changed before, but conserved after refinement. This means, that its remaining 21 detected and 1 changed genes after refinement are significantly conserved in their expression profile, whereas the whole group, including significant subgroups is changed.

Another large group that shows conserved expression between the brain regions is *peptidase*. Notably, related to this group are the conserved categories *ubiquitin conjugating enzyme* and *ubiquitin-protein ligase*, as ubiquitin targeting of proteins is a signal for degradation. These two groups share 21 genes out of 34 and 29 genes respectively.

Two related groups that are both conserved are *hydrogen ion transporter* (45 genes) and *sodium ion transporter* (30 genes). These two conserved groups share 27 genes among each other and almost all genes in both groups have mitochondrial localization. This localization is shared with the conserved group *oxidoreductase,(NADH to ubiquinone)*.

The most significant conserved group is *structural constituent of ribosome* ($P = 2.46 \times 10^{-7}$).

Another large group, under which many significant groups cluster is *binding*. It contains groups conserved and changed between the brain regions. It is a heterogeneous group containing significant sub groups involved in many, presumably biologically non related functions. It contains for example the groups *calcium ion binding*, *translation factor,* *nucleic acid binding* and *cytoskeletal adaptor*, which share no biological meaningful relation among each other.

3.3.3.2. *Cellular component* taxonomy

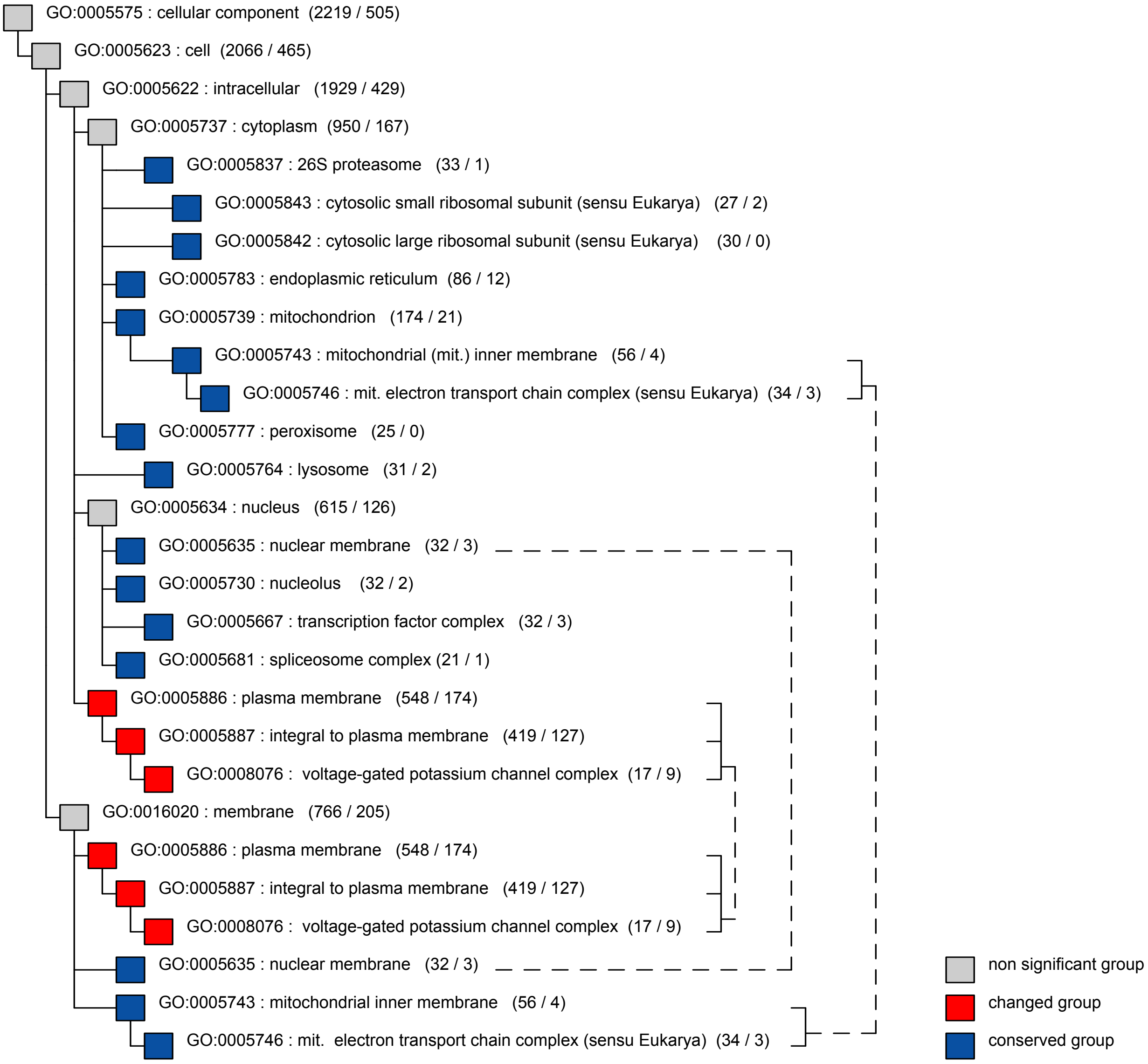

GO:0005575 : cellular component (2219 / 505)
GO:0005623 : cell (2066 / 465)
GO:0005622 : intracellular (1929 / 429)
GO:0005737 : cytoplasm (950 / 167)
GO:0005837 : 26S proteasome (33 / 1)
GO:0005843 : cytosolic small ribosomal subunit (sensu Eukarya) (27 / 2)
GO:0005842 : cytosolic large ribosomal subunit (sensu Eukarya) (30 / 0)
GO:0005783 : endoplasmic reticulum (86 / 12)
GO:0005739 : mitochondrion (174 / 21)
GO:0005743 : mitochondrial (mit.) inner membrane (56 / 4)
GO:0005746 : mit. electron transport chain complex (sensu Eukarya) (34 / 3)
GO:0005777 : peroxisome (25 / 0)
GO:0005764 : lysosome (31 / 2)
GO:0005634 : nucleus (615 / 126)
GO:0005635 : nuclear membrane (32 / 3)
GO:0005730 : nucleolus (32 / 2)
GO:0005667 : transcription factor complex (32 / 3)
GO:0005681 : spliceosome complex (21 / 1)
GO:0005886 : plasma membrane (548 / 174)
GO:0005887 : integral to plasma membrane (419 / 127)
GO:0008076 : voltage-gated potassium channel complex (17 / 9)
GO:0016020 : membrane (766 / 205)
GO:0005886 : plasma membrane (548 / 174)
GO:0005887 : integral to plasma membrane (419 / 127)
GO:0008076 : voltage-gated potassium channel complex (17 / 9)
GO:0005635 : nuclear membrane (32 / 3)
GO:0005743 : mitochondrial inner membrane (56 / 4)
GO:0005746 : mit. electron transport chain complex (sensu Eukarya) (34 / 3)
non significant group
changed group
conserved group

Figure 3.6. Significant groups in the *component* taxonomy after refinement. Given in parenthesis are the numbers of detected and changed genes in a group. Non significant groups are shown to structure the figure. If a group is categorised several times at different locations of the tree these locations are connected with a dashed line.

Figure 3.6. shows the hierarchical structure of the significant groups in the test using the hypergeometric distribution after refinement. 3 significantly changed and 13 significantly conserved groups can be identified in this test. A summary of these groups is given in the following.

The largest changed group is *plasma membrane*, containing 548 genes and the two other significantly changed groups - *integral to plasma membrane* and *voltage-gated potassium channel complex,* but no conserved group. Therefore all groups of genes that are differently expressed between the four brain regions have their localization in the plasma membrane. Plasma membrane is also the changed group with the lowest *P*-value (1.13×10^{-8}).

A large significant group that contains only conserved groups is *cytoplasma* with 950 genes. 8 significant sub groups of cytoplasma are conserved. They are located in vacuoles (*peroxisome, lysosome*), the cytosol (*26S proteasome, cytosolic large ribosomal subunit (sensu Eukarya), cytosolic small ribosomal subunit (sensu Eukarya)*), mitochondria (*mitochondrial inner membrane, mitochondrial electron transport chain complex*) and the *endoplasmic reticulum*. Mitochondrion is also the largest significantly conserved group after refinement.

Another major compartiment that contains only conserved groups is the *nucleus* with 615 genes, including four significant groups. Among these groups of genes, that are expressed on the same level throughout the brain regions are *transcription factor complex, spliceosome complex* and *nucleolus*, the region, where specifically tRNAs are transcribed.

3.3.3.3. *Biological process* taxonomy

Figure 3.7. shows the hierarchical structure of significant groups in the biological process taxonomy. After refinement 19 changed and 13 conserved groups can be found comparing the different brain regions.

Hereby the largest significantly changed group is *signal transduction* with 641 genes, containing 6 groups of genes with different expression among the brain regions, but none of the conserved groups. Several of these groups are involved in *G-protein signaling* and *glutamate signaling*. *Signal transduction* is also the second most significantly changed group in the HG test with a *P*-value of 1.55×10^{-8}. *Signal transduction* is a subgroup of *cell communication*. Notably another significantly changed group involved in neuronal functions, included in this category is *synaptic transmission*. The only conserved group clustering under *cell communication* is *humoral immune response*.

The third largest significantly changed group is *neurogenesis*. It contains 168 genes. It is the most significant changed group in the test using the hypergeometric distribution with a *P*-value of 1.46×10^{-10}. The changed groups *central nervous development* and *brain development* are, in this hierarchical order sub groups of *neurogenesis*.

The largest significantly conserved group is, with 605 genes, *protein metabolism*.

The genes in this group are regulated on the same level among the brain regions and it contains two conserved groups, one involved in the biosynthesis (*protein biosynthesis*) and one in the catabolism of proteins (*ubiquitin-dependent protein catabolism*). *Protein biosynthsis* is also the second largest significantly conserved group.

The third largest group, with genes expressed on the same level among the brain regions is *regulation of cell cycle*. Two other conserved groups involved in the cell cycle are *mitosis* and *positive regulation of cell proliferation*.

Two related groups that are equally expressed between the brain regions *are energy derivation by oxidation of organic compounds* and *oxidative phosphorylation, (NADH to ubiquinone)*. Both are involved in energy derivation.

All of the conserved groups mentioned above are subgroups of *cell growth and/or maintenance*, which includes more than two third of all genes in the taxonomy. It contains all groups conserved in their expression profile except one, but only four changed groups.

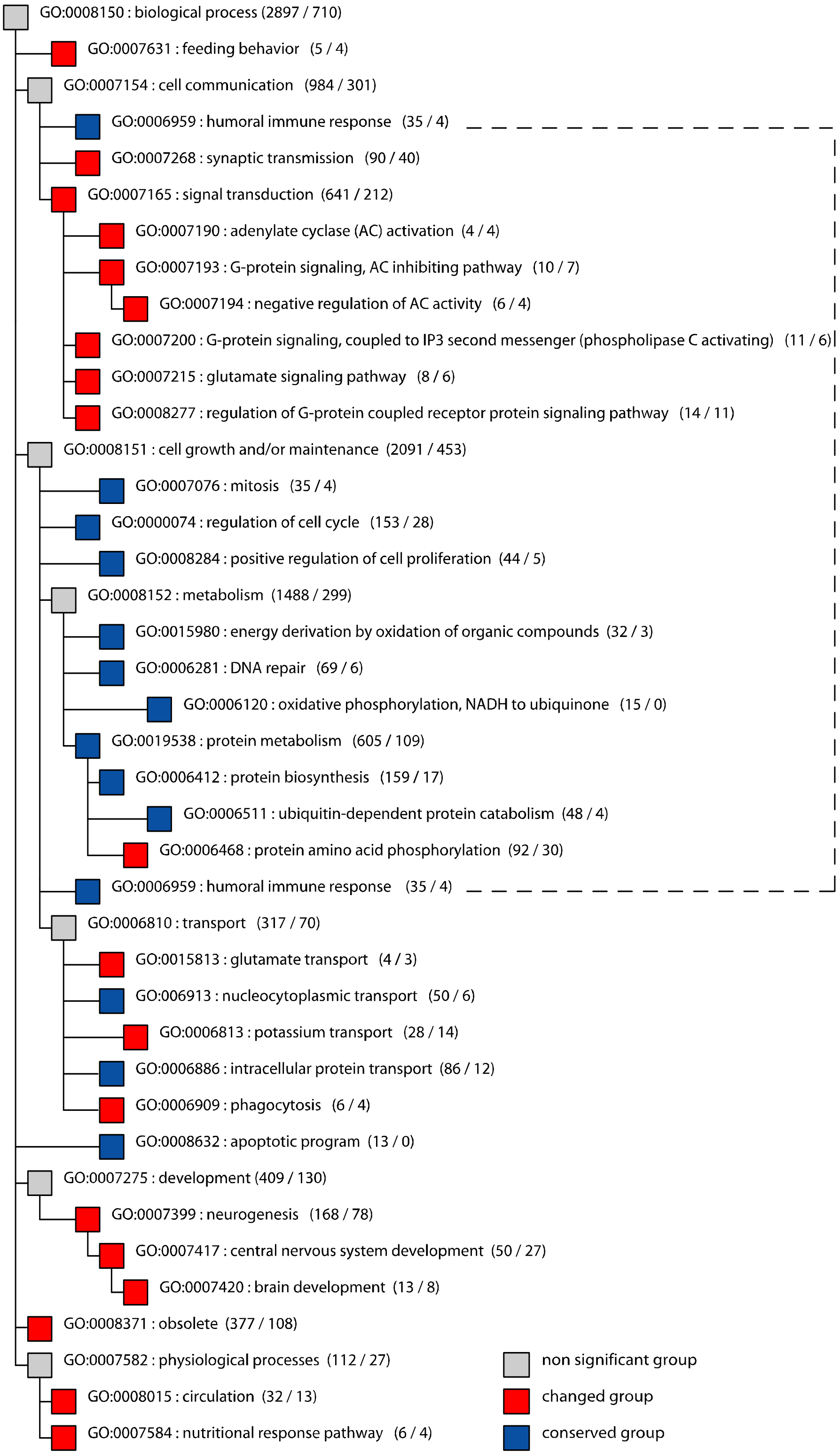

GO:0008150 : biological process (2897 / 710)
GO:0007631 : feeding behavior (5 / 4)
GO:0007154 : cell communication (984 / 301)
GO:0006959 : humoral immune response (35 / 4)
GO:0007268 : synaptic transmission (90 / 40)
GO:0007165 : signal transduction (641 / 212)
GO:0007190 : adenylate cyclase (AC) activation (4 / 4)
GO:0007193 : G-protein signaling, AC inhibiting pathway (10 / 7)
GO:0007194 : negative regulation of AC activity (6 / 4)
GO:0007200 : G-protein signaling, coupled to IP3 second messenger (phospholipase C activating) (11 / 6)
GO:0007215 : glutamate signaling pathway (8 / 6)
GO:0008277 : regulation of G-protein coupled receptor protein signaling pathway (14 / 11)
GO:0008151 : cell growth and/or maintenance (2091 / 453)
GO:0007076 : mitosis (35 / 4)
GO:0000074 : regulation of cell cycle (153 / 28)
GO:0008284 : positive regulation of cell proliferation (44 / 5)
GO:0008152 : metabolism (1488 / 299)
GO:0015980 : energy derivation by oxidation of organic compounds (32 / 3)
GO:0006281 : DNA repair (69 / 6)
GO:0006120 : oxidative phosphorylation, NADH to ubiquinone (15 / 0)
GO:0019538 : protein metabolism (605 / 109)
GO:0006412 : protein biosynthesis (159 / 17)
GO:0006511 : ubiquitin-dependent protein catabolism (48 / 4)
GO:0006468 : protein amino acid phosphorylation (92 / 30)
GO:0006959 : humoral immune response (35 / 4)
GO:0006810 : transport (317 / 70)
GO:0015813 : glutamate transport (4 / 3)
GO:006913 : nucleocytoplasmic transport (50 / 6)
GO:0006813 : potassium transport (28 / 14)
GO:0006886 : intracellular protein transport (86 / 12)
GO:0006909 : phagocytosis (6 / 4)
GO:0008632 : apoptotic program (13 / 0)
GO:0007275 : development (409 / 130)
GO:0007399 : neurogenesis (168 / 78)
GO:0007417 : central nervous system development (50 / 27)
GO:0007420 : brain development (13 / 8)
GO:0008371 : obsolete (377 / 108)
GO:0007582 : physiological processes (112 / 27)
GO:0008015 : circulation (32 / 13)
GO:0007584 : nutritional response pathway (6 / 4)
non significant group
changed group
conserved group

Figure 3.7. Significant groups in the *process* taxonomy after refinement. Given in parenthesis are the numbers of detected and changed genes in a group. Non significant groups are shown to structure the Figure. If a group is categorised several times at different locations of the tree these locations are connected with a dashed line.

3.3.3.4. Comparison between the significant groups in all three taxonomies

The major changed group in the taxonomy *molecular function* is, with 518 genes *signal transducer*, which shares 323 genes with the largest changed group in the taxonomy *biological process*, *signal transduction*, containing 641 genes.

About 38 % of the genes in the group *signal transducer* are annotated in the category *plasma membrane*, which contains itself 548 genes. Only about 16 % of these signal transducers are known to be localized in the cytoplasma, which as a group contains 950 genes and 8 % in *nucleus*, which contains 615 genes.

Among the conserved groups in the taxonomy *molecular function*, the genes in the groups *hydrogen-* and *sodium transporter* are to more than 70 % localized in the mitochondrial inner membrane, which is a conserved group in the taxonomy *cellular component*. Also among the genes in the group *oxidoreductase, acting on NADH or NADPH, quinone or similar compound as acceptor* 28 out of 29 are localized in the mitochondrial inner membrane. Additionally 15 of the sodium transporter genes constitute the group *oxidative phosphorylation, (NADH to ubiquinone)*, which is a significantly conserved group in the taxonomy *biological process*.

Also in the conserved group *ubiquitin-dependent protein catabolism* 40 genes out of 48 are contained in one of the three groups in *molecular function*, known to be involved in protein catabolism (*peptidase, ubiquitin protein ligase, ubiquitin conjugating enzyme*).

Finally the group *structural constituent of ribosome* shares 50 out of 62 of its genes with the two conserved groups in *cellular component* that constitute the small and large ribosomal subunit. It also shares 56 of its genes with the group *protein biosynthesis* in *biological process*. These genes constitute about one third of the genes in this group.

3.4. Functional profiling of genes differentially expressed in the human and chimpanzee brain

In the same way as the expressed and differentially expressed genes among four different brain regions, the expressed and changed genes in the comparison of human and chimpanzee brain were analysed. Though sequence controlled (masked) data was available, also the unmasked data was analysed, to get an insight in how far the sequence differences affect the measurements and the profiling.

3.4.1. Annotation process

Table 3.4. Detected and changed probesets after different steps of the annotation process – brain masked and unmasked data

	detected before masking	detected after masking	changed before masking	changed after masking
probesets	19466	18522	2577	2014
unigenes (UGs)	13591	13033	2155	1716
GO annotated (ann.) unigenes	3633	3492	575	451
UGs ann. in *function*	2919	2807	479	367
UGs ann. in *component*	2219	2122	362	272
UGs ann. in *process*	2897	2789	455	360

Table 3.4. shows a summary of the annotation process. The set of detected genes before masking is the one used also in the comparison between brain regions. For the 18522 detected probesets after masking 13033 unigene cluster can be found. After masking the resulting detected set of unigenes is about 4.1 % smaller than before. In the unmasked and masked data set 26.7 and respectively 26.8 % of all expressed unigenes have at least one annotation in the Gene Ontology.

Table 3.5. Distribution of annotated unigenes among functional groups

	brain unmasked data	brain masked data
groups in *molecular function*	1299	1280
groups in *cellular component*	282	280
groups in *biological process*	1105	1097

Shown are the number of groups in the *function, component* and *process* taxonomy into which the detected unigenes cluster.

After masking the changed set of unigenes is 20.4 % smaller than the corresponding unmasked data set. In the unmasked and masked data set 26.7 and 26.3 % of all differentially regulated unigenes are annotated in the Gene Ontology.

Table 3.5. shows the number of groups in each taxonomy, among which the detected genes are distributed. These numbers do not change significantly after masking.

3.4.2. Tests for the significance of the distribution of differentially expressed genes across functional categories

In both the unmasked and masked data set the overall distribution of differentially expressed genes among the functional categories differs significantly from the distribution of detected genes in the *process* taxonomy (*P* = 0.0027 and *P* = 0.0003). Yet, for both the unmasked and masked data set, this distribution of changed genes across functional categories is not significant in the *function* (*P* = 0.698, *P* = 0.168) and *component* taxonomy (*P* = 0.440 and *P* = 0.386).

Table 3.6. significant groups at a 5% level – *biological process* taxonomy

	groups changed in their expression profile			groups conserved in their expression profile		
biological process	**# in data set**	**random mean**	***P*-value**	**# in data set**	**random mean**	***P*-value**
unmasked data set	35	17.4	0.015	5	6.5	0.65
masked data set	39	19.7	0.015	13	5.3	0.040

# in data set : number of significant groups in the data set
random mean : mean number of significant groups among all random sets
P-value : *P*-value for the significance of the number of significant groups in the data set, given as the proportion of random sets with a number of significant groups that is higher or equal than the one given by the data set (< 0.0001, if no random set shows more significant groups than the data set)

The average number of significant groups from all random sets gives and estimate of the number of false-positive significant groups at a certain significance level. Compared to this number about twice as many significantly changed groups can be found at a 5% significance level in both the masked and unmasked data set (see Table 3.6.). In concordance with the results of the Chi square test, in no other taxonomy a significantly high number of changed groups can be found in both the masked and unmasked data set. This number of significantly conserved groups does not exceed the expected number of false-positive groups in the unmasked data set. In contrast at a significance level of 5% more than expected significantly conserved groups can be found after masking.

In this data set, furthermore four significantly conserved groups can be found at a significance level of 1% in the *function* taxonomy, whereas the number of false-positive groups is 0.53. (see Table A.2.). Except to this, in no other taxonomy a significantly high number of conserved groups can be found in both the masked and unmasked data set at any significance-cutoff examined. No significant group is identified using a Bonferroni correction at a 5 % significance level. The rate of false-positive groups ranges from 40.8 to 50.5 % at the chosen cutoffs. It is lower or equal at lower significance cutoffs (see Table A.2.).

3.4.3. Summary of the significant groups

The hierarchical structure of the significant groups after refinement has been plotted in the following figures. (see Figure 3.8. / 9.) Tables of the significant groups in the *biological process* taxonomy that include *P*-values before and after refinement are shown in the appendix (see Table C.1. / 2.). In the following part, the *P*-values before refinement will be given.

3.4.3.1. *Biological process* taxonomy - unmasked data set

12 changed groups are found in the HG test after refinement in the unmasked data. At no significance level, the observed number of conserved groups in the data set exceeded significantly the number of false-positive groups. Therefore only the changed groups, significant in the HG test on a 5% level have been used for the refinement.

Hereby, the largest changed group is *lipid metabolism* containing 129 genes. It is also the second most significant group in this test with a *P*-value of 0.002. *Lipid catabolism* and *glycosphingolipid biosynthesis* are changed groups clustering under *lipid metabolism*. The second largest changed group with 13 genes is *transcription initiation from Pol II promoter*. It is a subgroup of *nucleobase, nucleoside, nucleotide and nucleic acid metabolism*. This group contains 623 genes and besides *transcription initiation from Pol II promoter*, the significantly changed group *GTP biosynthesis*.

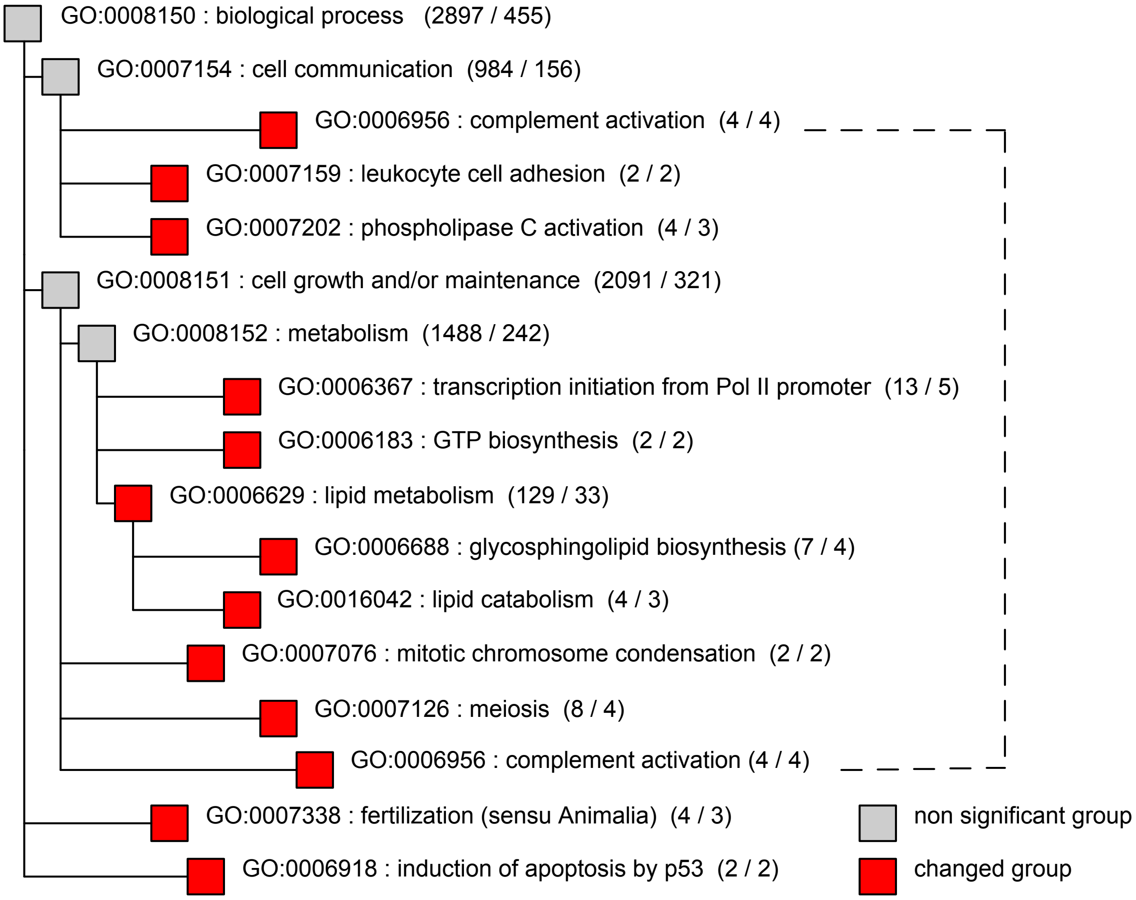

GO:0008150 : biological process (2897 / 455)
GO:0007154 : cell communication (984 / 156)
GO:0006956 : complement activation (4 / 4)
GO:0007159 : leukocyte cell adhesion (2 / 2)
GO:0007202 : phospholipase C activation (4 / 3)
GO:0008151 : cell growth and/or maintenance (2091 / 321)
GO:0008152 : metabolism (1488 / 242)
GO:0006367 : transcription initiation from Pol II promoter (13 / 5)
GO:0006183 : GTP biosynthesis (2 / 2)
GO:0006629 : lipid metabolism (129 / 33)
GO:0006688 : glycosphingolipid biosynthesis (7 / 4)
GO:0016042 : lipid catabolism (4 / 3)
GO:0007076 : mitotic chromosome condensation (2 / 2)
GO:0007126 : meiosis (8 / 4)
GO:0006956 : complement activation (4 / 4)
GO:0007338 : fertilization (sensu Animalia) (4 / 3)
GO:0006918 : induction of apoptosis by p53 (2 / 2)
non significant group
changed group

Figure 3.8. Significant groups in the *process* taxonomy after refinement. Given in parenthesis are the numbers of detected and changed genes in a group. Non significant groups are shown to structure the figure. If a group is categorised several times at different locations of the tree these locations are connected with a dashed line.

Another major group, under which a group, differentially expressed in a comparison of human and chimpanzee brain is matching, is *immune response* with 130 genes. It contains the changed subgroup *complement activation*. This group is the most significant changed group in the HG test with a *P*-value of 0.0006. Though not clustered under this category, *leukocyte cell adhesion*, another significantly changed group, is involved in immune response. Although the overall distribution in *molecular function* is not significant, the group *antiviral response protein*, also involved in immune response is significantly different expressed in the human and chimpanzee brain.

Furthermore the category *cell cyle*, though not significant itself contains two changed groups. These are *meiosis* and *mitotic chromosome condensation*. Another changed group, *induction of apoptosis by p53* plays a role in the cell cycle (Anderson 1997), but is not divided into this category by the GO consortium.

3.4.3.2. *Biological process* taxonomy - masked data set

15 changed and 5 conserved groups are significant in the test using the hypergeometric distribution after refinement.

The largest changed groups are *embryogenesis and morphogenesis* and *induction of apoptosis*, both containing 50 genes. The third largest group of genes differentially regulated in human and chimpanzee brains is *sexual reproduction*, with 37 genes.

A major group, under which only changed groups cluster is *cell cycle*. This group is significant in the tree test. It contains 236 genes and 3 differentially expressed groups. These are involved in meiotic and mitotic processes. Another group of genes that plays a role in the cell cycle, but is not classified in this category is *induction of apoptosis (Anderson 1997).*

Furthermore two changed groups are sub groups of *lipid metabolism*. It contains the significant sub groups *phosphatidyinositol biosynthesis* and *glycosphingolipid biosynthesis*. This group is the most significant changed group in the HG test with a pvalue of 0.00003. In the *glycosphingolipid biosynthesis* pathway 6 out of 7 genes are differentially regulated in human and chimpanzee brains. In the taxonomy *molecular function*, which is not significant overall, genes of the changed groups *sialyltransferase* and *mannosyltransferase* are involved in glycosphingolipid biosynthesis (Flitsch, Goodridge et al. 1994). Genes of the changed group *inositol/phosphatidylinositol kinase* are concerned with signal transduction involving phosphatidylinositols (Hong, Mikami et al. 2003) .

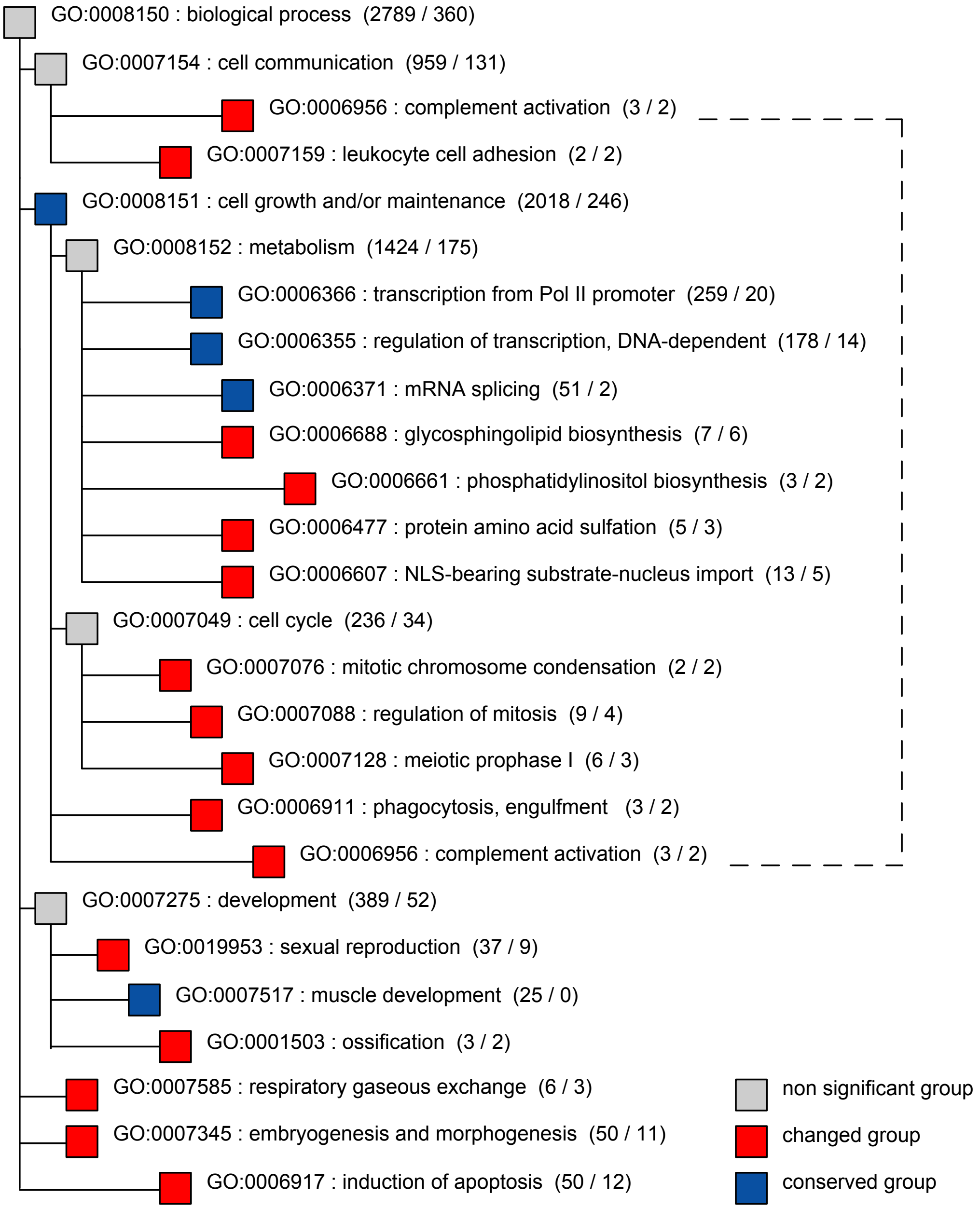

GO:0008150 : biological process (2789 / 360)
GO:0007154 : cell communication (959 / 131)
GO:0006956 : complement activation (3 / 2)
GO:0007159 : leukocyte cell adhesion (2 / 2)
GO:0008151 : cell growth and/or maintenance (2018 / 246)
GO:0008152 : metabolism (1424 / 175)
GO:0006366 : transcription from Pol II promoter (259 / 20)
GO:0006355 : regulation of transcription, DNA-dependent (178 / 14)
GO:0006371 : mRNA splicing (51 / 2)
GO:0006688 : glycosphingolipid biosynthesis (7 / 6)
GO:0006661 : phosphatidylinositol biosynthesis (3 / 2)
GO:0006477 : protein amino acid sulfation (5 / 3)
GO:0006607 : NLS-bearing substrate-nucleus import (13 / 5)
GO:0007049 : cell cycle (236 / 34)
GO:0007076 : mitotic chromosome condensation (2 / 2)
GO:0007088 : regulation of mitosis (9 / 4)
GO:0007128 : meiotic prophase I (6 / 3)
GO:0006911 : phagocytosis, engulfment (3 / 2)
GO:0006956 : complement activation (3 / 2)
GO:0007275 : development (389 / 52)
GO:0019953 : sexual reproduction (37 / 9)
GO:0007517 : muscle development (25 / 0)
GO:0001503 : ossification (3 / 2)
GO:0007585 : respiratory gaseous exchange (6 / 3)
GO:0007345 : embryogenesis and morphogenesis (50 / 11)
GO:0006917 : induction of apoptosis (50 / 12)
non significant group
changed group
conserved group

Figure 3.9. Significant groups in the *process* taxonomy after refinement. Given in parenthesis are the numbers of detected and changed genes in a group. Non significant groups are shown to structure the figure. If a group is categorized several times in different locations of the tree these locations are connected with a dashed line.

The largest conserved group in the comparison between human and chimpanzee brains is *cell growth and/or maintenance*, containing more than two third of all genes in this taxonomy. It contains 3 conserved and 9 changed groups.

The three conserved groups are all subgroups of *nucleobase, nucleoside, nucleotide and nucleic acid metabolism* (*n.n.n.n. metabolism*), which contains 594 genes and no further changed groups. The conserved groups are involved in transcription and splicing. *Transcription initiation from Pol II promoter*, a subgroup *of n.n.n.n. metabolism* is the most significant conserved group with a *P*-value of 0.004.

A large category which includes both changed and conserved groups is *metabolism*. It contains 1424 genes and contains 3 conserved and 4 changed groups. The three conserved groups are the subgroups of *n.n.n.n. metabolism*. Two of the changed groups are groups involved in lipid metabolism.

Another large category under which both changed and conserved groups cluster is development. It contains 389 genes and is not significant itself. The changed groups are *sexual reproduction* and *ossification*. The conserved group is *muscle development*.

3.4.3.3. comparison between the masked and unmasked data set

In the *process* taxonomy 17 and 4 significant groups, respectively are shared between the masked and unmasked data set before and after refinement.

After refinement, most of the changed groups in the masked and unmasked data set fall into the 3 categories *immune response*, *cell cycle* and *lipid metabolism*. Groups not included in these major categories in the masked data are *protein amino acid sulfation*, *NLS-bearing substrate-nucleus import*, *respiratory gaseous exchange*, *ossification*, *embryogenesis and morphogenesis* and *sexual reproduction.* None of these groups are among the significantly changed groups in the unmasked data set. Still, *fertilization (sensu Animalia)*, a significantly changed subgroup in the unmasked data is a subgroup of *sexual reproduction*.

Whereas 2 significantly changed groups are subgroups of *n.n.n.n. metabolism* in the unmasked data set, only conserved groups cluster under this category in the masked data set. Still, the 15 detected genes contained in the changed groups in the unmasked data set make up a proportion of 2,4 % of all genes in this category. In contrast in the masked data set the 343 detected genes of the conserved group in this category constitute 57.7 % of all genes in the *n.n.n.n. metabolism* group.

Whereas the overall number of changed genes in the process taxonomy is reduced by 20.8 % from 455 to 360, this number of genes in the group immune response is reduced by 32 % from 25 to 17 genes.

4. Discussion

4.1. Algorithms used in the functional profiling

The algorithms applied to the microarray data sets in order to get an insight of the underlying biological phenomena involved in the gene expression comparisons studied were the annotation procedure, the tests for the overall significance of the distribution of genes across functional categories and the refinement procedure. A discussion of the advantages and drawbacks of these methods will be given in the following.

4.1.1. Annotation process

In the annotation process, the probesets are matched to their corresponding unigenes and than further linked to their corresponding GO functions, resulting in a list of detected and differentially expressed genes in each functional category of a taxonomy.

The most important result the annotation process is that only about a quarter of the unigenes have an annotation in the database. This result is in concordance, with the finding that probably also about one forth of all human genes are annotated in the Gene Ontology. The power to detect a functional category with significantly more or less differentially expressed genes than expected is limited by the percentage of annotated genes available. While this percentage may be enough to detect a significant change or conservation in a larger category, it may not be enough to detect change or conservation in categories which contain at this state of available annotation relatively few genes. However, with the rapidly increasing knowledge about the human genome more data will be present for follow-up studies with upcoming versions of the database.

4.1.2. Statistical tests for the significance of the overall distribution of changed genes into functional categories

With the Chi square test and the test for the rate of false-positive groups, the overall significance of the data is examined. These two tests are ordered hierarchically. Only if the overall distribution of changed genes across functional categories in a taxonomy is significant, the rate of false-positive groups will be estimated at different significance levels.

In all three data sets, the results of both tests were in concordance. If a taxonomy was significant in the Chi square test also the number of changed or conserved significant groups exceeded the number of false-positive groups significantly.

Whereas in both tests all three taxonomies were significant in the comparison between different brain regions, only the *process* taxonomy was significant in the comparison between human and chimpanzee brain in both the sequence controlled and uncontrolled data. Several reasons may account for this.

One explanation is that there is no significance in the *component* and *function* taxonomies that concern cellular localization and biochemical activity of genes. This means that, when genes are clustered according to these categories, no more significantly changed or conserved groups than expected can be found. This means that genes in different cellular localizations and with different biochemical activity can vary in their expression between humans and chimpanzee. This difference is of no apparent functional importance, so that this variance in expression is nonselective or neutral.

However other explanations can not be excluded. Notably, additionally to the experimental variance that affects both the expression measurements in the in-brain and the human chimpanzee comparison, the measurements in the latter comparison are also affected by biological variance and – though a masking algorithm was applied - variance due to sequence differences. This additional variance might make the data less cleaner than the data from the in-brain comparison and might reduce the significance of the distribution into categories due to random noise. Another indication that this additional variance alloyed the analysis is that after controlling for sequence differences both an excess of changed and conserved groups was observed in the process taxonomy, while without masking only more changed groups than expected were found.

At the moment it is not possible to distinguish between the two possible explanations. Further indication, which of these two explanations is true might be given, when the sequence data for the whole chimpanzee genome is available and all oligonucleotides that contain sequence differences can be removed from the analysis.

If the significance of the data in the functional profiling was reduced due to variance and noise, it is obviously crucial for the results of the analysis, how many false-positive and false-negative differentially expressed genes are contained in the data analysed.

Notably, if the overall distribution of changed genes across functional categories is significant, in almost all data sets examined, the rate of false-positive groups gets lower, if a lower significance cutoff is chosen in the HG test (see appendix Table A.1 / 2.). This way the selection for the cutoff presents a dilemma. At a higher cutoff, more information is present as more significant groups are contained, but the results are less certain as the false-positive rate is higher. At a lower significance level, the results are more certain, but less significant groups are observed and the number of significant groups can vary more, as the numbers of significant groups are much smaller. The selection criterion, chosing the number of significant groups at a cutoff where this number is significantly higher than in the random sets, is therefore thought to represent a good compromise. Hereby the rate of false-positive groups ranged from 10 to 42 % in the in-brain comparison and from 40.8 to 51 % in the human-chimpanzee comparison. However, as the information for other significance levels is available, it can be taken into consideration (see Table A1. / 2.).

Summarizing these results, the two statistical tests proposed thoroughly examine the significance of the whole data set and if the significance is due to significantly more conserved or changed genes. Furthermore a selection criterion can be applied, that results in a low number of false-positive groups in the analysis.

4.1.3. Refinement

The aim of the refinement is to remove significant groups, whose significance is due to a significant subgroup. If this is the case, such a higher node does not provide any further information about a functional change or conservation, as the significant categories further downstream provide more specific information. Such a group can therefore be removed.

If differentially expressed genes were randomly chosen, like in a random set, one would also expect the significantly changed and conserved groups from such a set to be randomly distributed across a taxonomy. Therefore, if in the analysis of a data set many groups, all significant in the same direction cluster under a major category, a major change or conservation in this group is more likely.

Still it is very difficult to estimate the significance of such a finding, as size of the groups, clustering of the significant groups and overall number of groups under a certain category throughout the whole tree would have to be taken into account to get an estimate of the significance of this correlation.

What complicates the analysis of this correlation is that the number of changed groups under a certain major category might be artificially blown up by single significant subgroups on a low hierarchical level of a taxonomy. In such a large category a major change or conservation is then presumed, where there is none.

A good example, where this effect can be seen is the category *nucleotide metabolism* in the sequence uncontrolled comparison between chimpanzee and human brain. Though not significant itself, it contains eight significant groups before refinement. A subgroup of seven of these groups is *GTP biosynthesis*. In the refinement procedure, these seven groups are removed and only the group *GTP biosynthesis* is significant in the category *nucleotide metabolism*. If the list was not refined, one might assume a major difference between humans and chimpanzees in the expression of genes involved *nucleotide metabolism*. This presumed evidence that major changes occurred in this category is falsified, when the list of significant groups is reduced in the refinement.

Therefore, the result of the refinement is a more comprehensive list of the significant groups, as it is smaller and contains only the essential informative nodes and gives a better picture of the distribution of the significantly changed and conserved groups.

In all data sets analysed, more than half of the significant groups were removed in the refinement algorithm. Yet, the *P*-value after refinement is only used, to answer the question, if a node carries additional information, whereas the *P*-value before refinement gives an estimate of the significant change or conservation from all genes of a functional group and is therefore the more important one to answer biological meaningful questions.

4.1.4. Summary

A number of tools is available that categorizes genes in different databases and tests each category, if significantly more or less genes in the condition under study have accumulated in such a group (Luyf, de Gast et al. 2002; Doniger, Salomonis et al. 2003; Draghici, Khatri et al. 2003). The drawback of all these methods and of the one presented is that only a small percentage of the available genes are annotated in the databases. Another disadvantage is that mostly a large number of categories is tested, so that in a data set that contains a small number of genes to test probably the number of significant groups does not substantially exceed the number of false-positive groups.

However this problem can be circumvented, if driven by a certain hypothesis one examines only a small number of selected categories which one thinks are involved in the condition under study.

The advantage of these approaches is that the biological phenomena concerned with any large scale analysis can be – as far as knowledge about these processes is available – immediately determined and do not have to be traced down in months spent with search through the literature.

The novelty of this tool is the resampling method and the two tests for the overall significance of the data based on this resampling approach. Therefore this is the first approach that rigorously examines if the result of the categorization is significant at all and not a product of the multiple testing. Another novelty is the reduction of the list of significant groups into a more comprehensive one in the subsequent refinement, after which a clearer picture of the underlying biological processes involved in the study is given.

4.2. Functional groups of genes differentially expressed among different brain regions

When the genes expressed and differentially expressed among different brain regions are analysed, the overall distribution of changed genes is significant in all three taxonomies of the Gene Ontology.

Hereby the largest changed groups after refinement in both the *function* and *process* taxonomy are involved in signal transduction. Both these groups have one of the lowest *P*-values in their taxonomy and are still significant after a Bonferroni correction at a 5% significance level. Additionally more than a half of the significantly changed groups in both taxonomies are concerned with or related to signal transduction. Among the signal transducers, for which the localization is known, a large proportion is located in the outer membrane, whereas a much smaller part is located within the cell (see Chapter 3.2.3.4.).

Therefore most of the differences in the expression profile between different brain regions concern signal transducers located in the plasma membrane. Corresponding genes heterogeneously expressed between the observed regions range from peptide hormones, a variety of receptors, like glutamate, GABA and amine receptors, genes concerned with G-protein coupled signaling to protein kinase.

These genes are presumably involved in neuronal functions (Kandel 2000) and it is possible, that the distribution of signal transducers in the membrane involved in neuronal activities plays a crucial role for the identity and specific function of a certain brain region. These results correspond to previous findings, where it was found, that different cortical regions exhibit a specific distribution of glutamate and GABA receptors (Zilles, Palomero-Gallagher et al. 2002).

Other major groups of genes heterogeneously expressed between the brain regions involved in neuronal functions are *synaptic transmission* and *neurogenesis*. These two groups in the *process* taxonomy also have one of the lowest *P*-values in this taxonomy and are significant after a Bonferroni correction at a 5 % level.

Additionally, after refinement, two significant sub groups, *central nervous system development* and *brain development* belong to *neurogenesis*. Two explanations may account for the fact that these differences are observed in the adult brain. First it might be that these genes have a different function in the adult brain, second it might be that the differential expression among the brain regions is related to the different expression of these genes during development. This would mean that the developmental program that shapes the identity of a brain region during development, is involved in maintaining and coding the specific function of a region (Zhang, Hume et al. 2002). It may be possible that the types of receptors, neurons and patterns of connections developed during embryogenesis, the signature of a region to carry out a certain function, is maintained through the same program in the adult brain. (Oberto, Tolosano et al. 1998; Kandel 2000) Such a function in development and maintainance has been proposed for two of the eight changed genes in the group *brain development*, Slit1 and Six3 (Halder, Callaerts et al. 1998; Itoh, Miyabayashi et al. 1998).

The largest conserved group in the *process* taxonomy is protein metabolism. It is the second most significant conserved group and the corresponding *P*-value is still significant after a Bonferroni correction at a 5% significance level. Conserved sub groups are involved both in protein biosynthesis and catabolism.

One third of the genes in protein biosynthesis are ribosomal genes. The groups involved in protein catabolism*, ubiquitin-dependent protein catabolism* shares a large proportion of its genes with related groups in the *function* taxonomy. Therefore, it can be assumed that ribosomal genes and other genes involved in protein biosynthesis and genes involved in ubiquitin dependent catabolism are expressed on the same level throughout the brain.

Furthermore, the most significant conserved group in the cellular component taxonomy is mitochondrion. It is also significant after a Bonferroni correction at a 5% level. Conserved groups located in the mitochondrial inner membrane, involved in energy derivation can be found in the *process* and *function* taxonomy.

Furthermore, in the *process* taxonomy, three groups of genes involved in the cell cycle are homogenously expressed throughout the brain. As neurons are supposed to divide rarely, this group of genes might be detectable due to the proportion of glia cells in the samples, which are known to divide more frequently (Korr, Schultze et al. 1975).

Summarizing these findings groups involved in protein metabolism, energy derivation and cell division, all involved in housekeeping functions and cell maintenance do not differ in expression throughout the brain, so that it seems to be important for the integrity of the brain to keep up these functions on a constant rate among all regions.

4.3. Functional profiling of genes differentially expressed in human and chimpanzee brain

The second data set annotated and analysed were the detected and changed genes in the comparison of the human and chimpanzee brain. Hereby it is apparent from the results of the annotation procedure that after controlling for sequence differences the number of detected unigenes is 4.1 % smaller than before masking, while the number of changed unigenes is reduced by 20.4 %.

This may be due to a stricter selection criteria for change than for detection, through which a higher proportion of changed than detected genes is removed after masking. Another reason may be that a larger proportion of genes identified as differentially expressed genes, than genes expressed on the same level contained sequence differences. This would mean that a larger proportion of changed genes was classified erroneously as changed because they contained sequence differences.

In the Chi square test for the overall significance of the distribution of changed genes across functional categories, the taxonomy *biological process* is significant in both the masked and unmasked data. Yet in the data set, which is not controlled for sequence differences, only an excess of significantly changed groups can be found, while in the masked data set both more changed and conserved groups than expected can be found. This might be due to the fact that the data, in which it was not controlled for sequence differences contains more false-positive differentially expressed genes. Therefore the genes were classified wrongly and lead to less significant results. Also categories of genes containing relatively more sequence differences than others might have a higher probability to be classified as changed due to this effect on the measurements.

After refinement both the masked and unmasked data set contain significantly changed groups in the major categories *lipid metabolism, immune response* and *cell cycle*, which indicates that major changes on a gene expression level might have occurred in these categories between the two species.

In the category *lipid metabolism*, the most significant changed group in the masked data set is *glycosphingolipid biosynthesis*. Glycosphingolipids are supposed to be involved in transducing signals for apoptosis and are known to contain sialic acids. This is interesting, as one of the few biochemical differences known between humans and chimpanzees is the absence of a certain sialic acid (Neu5Ac) in humans which leads to the overexpression of its precursor (Neu5Gc) (Chou, Takematsu et al. 1998; Irie, Koyama et al. 1998; Chou, Hayakawa et al. 2002). Additionally groups involved in apoptosis can be found in the masked and unmasked data set.

The rate of apoptosis is crucial for the density of nerve cells in the adult brain (Kovac, Grammig et al. 2002). It can be speculated that a higher rate of apoptosis of brain cells during development and aging of the chimpanzee in comparison to humans contributes to a higher plasticity and density of nerve cells in the human brain. Supporting this view is the finding, that in contrary to other mammal brains, the human brain is still growing after birth. This might lead to a reduced rate of apoptosis of nerve cells in the human developing and adult brain. This is further supported by the finding that the inactivation of the specific hydroxylase that turns Neu5Gc into Neu5Ac occurred prior to brain expansion on the human lineage and a role in the human-specific brain development was speculated. (Chou, Hayakawa et al. 2002)

Another major category, containing only changed groups is *cell cycle*. The different expression of these genes in human and chimpanzee brain might be due to a different rate of division and turnover in glia cells, as neuronal cells are supposed to divide rarely in the adult brain. It might also be that the induction of apoptosis leads to differential expression of genes regulating the cell cycle in the human and chimpanzee brain.

Notably, in the category *cell cycle* one group in each data set is concerned with meiosis. It is not clear, why the genes involved in this function are expressed in the brain, as meiosis itself does not occur in the brain. It might be that these genes are involved in related mitotic processes.

Two other groups of genes, which are not expected to be expressed in the brain are concerned with sexual reproduction. Here it might be the case, that these groups are false-positive due to a relatively high proportion of sequence differences between the two species. Genes involved in sexual reproduction are supposed to evolve rapidly due to sexual selection (Johnson, Viggiano et al. 2001; Gage, Parker et al. 2002) and therefore are genes in this group are supposed to show a higher proportion of sequence differences than other genes.

Another group of genes, which is supposed to evolve rapidly is immune response, which is another major group, that contains in both data sets only changed groups. Pathogens are a major force of evolution and force hosts to adapt rapidly (Flajnik 1994). Therefore it might be that a large number of sequence differences in this group also account for the change in the gene expression profile between humans and chimpanzees.

Finally, the most interesting finding is that one of the largest changed groups in the human-chimpanzee comparison in the masked data set is *embryogenesis and morphogenesis*.

The, compared to chimpanzees enlarged size of the human brain and the more extensive folding of the human neocortex are probably due to different regulation of genes during the morphogenesis of the human brain. Additionally DSH and EYA1, two of the genes in this group differentially expressed between humans and chimpanzee are involved in neuroblast specification and eye development, respectively (Pizzuti, Amati et al. 1996; Abdelhak, Kalatzis et al. 1997). It might be that these genes differentially regulated in the human and chimpanzee brain are involved in the developmental processes that account for this difference between the human and chimpanzee brain. Again, it might be that the different expression in the adult brain of humans and chimpanzees result from this different expression during development.

5. Outlook

The questions about the function and uniqueness of the human mind will, to a degree, always be speculative and philosophical. Yet, as we know more and more about the underlying biological processes of brain functions, these have turned out to be questions, also to be studied by methods of psychology and molecular biology. The results of the analysis presented here, show the usefulness of these approaches.

From the results of the analysis, it can be seen that differences among brain regions correlate with expression differences involved in the neuronal functions synaptic transmission, signal transduction and neurogenesis.

Furthermore ít can be speculated that differences in the human and chimpanzee brain are related to different expression of genes regulating development during embryogenesis, changes in the expression of genes involved in the regulation of the cell cycle and genes involved in lipid metabolism.

Concerning the methodical side of this analysis it is clear that other tools are available for functional profiling for gene expression data. It is the novelty of this study, to examine the significance of the overall distribution of genes across functional category, therefore assessing, if the groups found to be significant are meaningful at all. Furthermore the list of significant groups is afterwards reduced to a more comprehensive one.

The aim in the future development of this tool will be to include other databases, with which one will be able to link the differences under study to related diseases or enzymatic pathways, as well as to include annotations for other available model organisms like mouse or yeast. With the advent of other large scale data sets for the human-chimpanzee comparison, like the chimpanzee genome, which will be available in a couple of month, it will also be possible to exhibit this tool to these kinds of comparisons.

It is an advantage of this method, that it can be applied to analyse data in any of these large-scale comparisons.

The arrival and analysis of these large scale experiments also marks a major change in biology itself. For long the enormous complexity and diversity of the living world has detracted biology from a mathematical exact description. With the recent advent of these large scale experiments, their statistical analysis and new mathematical theories that contribute increasingly to explain this data, practical and theoretical steps have been made towards this aim. Following the path which physics and chemistry have gone before, it is the challenge of our time that biology as a science changes its face from a generally descriptive to an exact and predictive one.

*"Who wants to read the book of nature, must know the language in which it is written.
This language is mathematics. "*
(Galileo Galilei, 1564-1642)

6. Bibliography

Abdelhak, S., V. Kalatzis, et al. (1997). "A human homologue of the Drosophila eyes absent gene underlies branchio-oto-renal (BOR) syndrome and identifies a novel gene family." Nat Genet **15**(2): 157-64.

Anderson, P. (1997). "Kinase cascades regulating entry into apoptosis." Microbiol Mol Biol Rev **61**(1): 33-46.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

Buxhoeveden, D. P., A. E. Switala, et al. (2001). "Lateralization of minicolumns in human planum temporale is absent in nonhuman primate cortex." Brain Behav Evol **57**(6): 349-58.

Carter, T. A., J. A. Del Rio, et al. (2001). "Chipping away at complex behavior: transcriptome/phenotype correlations in the mouse brain." Physiol Behav **73**(5): 849-57.

Chen, F.-C. and W.-H. Li (2001). "Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees." Am J Hum Genet **68**(2): 444-56.

Chou, H. H., T. Hayakawa, et al. (2002). "Inactivation of CMP-N-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution." Proc Natl Acad Sci U S A **99**(18): 11736-41.

Chou, H. H., H. Takematsu, et al. (1998). "A mutation in human CMP-sialic acid hydroxylase occurred after the Homo-Pan divergence." Proc Natl Acad Sci U S A **95**(20): 11751-6.

de Chaldee, M., M. C. Gaillard, et al. (2003). "Quantitative assessment of transcriptome differences between brain territories." Genome Res **13**(7): 1646-53.

Doniger, S. W., N. Salomonis, et al. (2003). "MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data." Genome Biol **4**(1): R7.

Draghici, S., P. Khatri, et al. (2003). "Global functional profiling of gene expression." Genomics **81**(2): 98-104.

Flajnik, M. F. (1994). "Advances in immunology." Bioessays **16**(9): 671-5.

Flitsch, S. L., D. M. Goodridge, et al. (1994). "The chemoenzymatic synthesis of neoglycolipids and lipid-linked oligosaccharides using glycosyltransferases." Bioorg Med Chem **2**(11): 1243-50.

Gage, M. J., G. A. Parker, et al. (2002). "Sexual selection and speciation in mammals, butterflies and spiders." Proc R Soc Lond B Biol Sci **269**(1507): 2309-16.

Halder, G., P. Callaerts, et al. (1998). "Eyeless initiates the expression of both sine oculis and eyes absent during Drosophila compound eye development." Development **125**(12): 2181-91.

Hong, Y. K., A. Mikami, et al. (2003). "A new class of mutations reveals a novel function for the original phosphatidylinositol 3-kinase binding site." Proc Natl Acad Sci U S A **100**(16): 9434-9.

Irie, A., S. Koyama, et al. (1998). "The molecular basis for the absence of N-glycolylneuraminic acid in humans." J Biol Chem **273**(25): 15866-71.

Itoh, A., T. Miyabayashi, et al. (1998). "Cloning and expressions of three mammalian homologues of Drosophila slit suggest possible roles for Slit in the formation and maintenance of the nervous system." Brain Res Mol Brain Res **62**(2): 175-86.

Johnson, M. E., L. Viggiano, et al. (2001). "Positive selection of a gene family during the emergence of humans and African apes." Nature **413**(6855): 514-9.

Kandel, E. R. (2000). Priciples of neural science. New York, McGraw-Hill.

Khaitovich, P., Muetzel, B. et al (2003). "Evolution of gene expression in the primate brain." submitted for publication.

Korr, H., B. Schultze, et al. (1975). "Autoradiographic investigations of glial proliferation in the brain of adult mice. II. Cycle time and mode of proliferation of neuroglia and endothelial cells." J Comp Neurol **160**(4): 477-90.

Kovac, A. D., J. Grammig, et al. (2002). "Comparison of neuronal density and subfield sizes in the hippocampus of CD95L-deficient (gld), CD95-deficient (lpr) and nondeficient mice." Eur J Neurosci **16**(1): 159-63.

Letinic, K. and P. Rakic (2001). "Telencephalic origin of human thalamic GABAergic neurons." Nat Neurosci **4**(9): 931-6.

Liu, G., A. E. Loraine, et al. (2003). "NetAffx: Affymetrix probesets and annotations." Nucleic Acids Res **31**(1): 82-6.

Lockhart, D. J., H. Dong, et al. (1996). "Expression monitoring by hybridization to high-density oligonucleotide arrays." Nat Biotechnol **14**(13): 1675-80.

Luyf, A. C., J. de Gast, et al. (2002). "Visualizing metabolic activity on a genome-wide scale." Bioinformatics **18**(6): 813-8.

Nadon, R. and J. Shoemaker (2002). "Statistical issues with microarrays: processing and analysis." Trends Genet **18**(5): 265-71.

Oberto, A., E. Tolosano, et al. (1998). "The murine Y1 receptor 5' upstream sequence directs cell-specific and developmentally regulated LacZ expression in transgenic mice CNS." Eur J Neurosci **10**(10): 3257-68.

Pizzuti, A., F. Amati, et al. (1996). "cDNA characterization and chromosomal mapping of two human homologues of the Drosophila dishevelled polarity gene." Hum Mol Genet **5**(7): 953-8.

Pruitt, K. D. and D. R. Maglott (2001). "RefSeq and LocusLink: NCBI gene-centered resources." Nucleic Acids Res **29**(1): 137-40.

Rilling, J. K. and T. R. Insel (1999). "The primate neocortex in comparative perspective using magnetic resonance imaging." J Hum Evol **37**(2): 191-223.

Sandberg, R., R. Yasuda, et al. (2000). "Regional and strain-specific gene expression mapping in the adult mouse brain." Proc Natl Acad Sci U S A **97**(20): 11038-43.

Semendeferi, K., A. Lu, et al. (2002). "Humans and great apes share a large frontal cortex." Nat Neurosci **5**(3): 272-6.

Tomasello, M. and J. Call (1997). Primate Cognition. Oxford, Oxford University Press.

Wheeler, D. L., D. M. Church, et al. (2003). "Database resources of the National Center for Biotechnology." Nucleic Acids Res **31**(1): 28-33.

Wildman, D. E., M. Uddin, et al. (2003). "Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: enlarging genus Homo." Proc Natl Acad Sci U S A **100**(12): 7181-8.

Wodicka, L., H. Dong, et al. (1997). "Genome-wide expression monitoring in Saccharomyces cerevisiae." Nat Biotechnol **15**(13): 1359-67.

Zhang, Y. H., K. Hume, et al. (2002). "Expression of the Ste20-like kinase SLK during embryonic development and in the murine adult central nervous system." Brain Res Dev Brain Res **139**(2): 205-15.

Zhao, N., H. Hashida, et al. (1995). "High-density cDNA filter analysis: a novel approach for large-scale, quantitative analysis of gene expression." Gene **156**(2): 207-13.

Zilles, K., N. Palomero-Gallagher, et al. (2002). "Architectonics of the human cerebral cortex and transmitter receptor fingerprints: reconciling functional neuroanatomy and neurochemistry." European Neuropsychopharmacology: The Journal Of The European College Of Neuropsychopharmacology **12**(6): 587-599.

7. Appendix

7.1. Numbers of false-positive groups at different significance levels in the HG test

Table A.1. Significant groups at different significance level in the HG test / in-brain comparison

		groups changed in their expression profile			groups conserved in their expression profile		
taxonomy	**SL**	**# in data set**	**random mean**	***P*-value**	**# in data set**	**random mean**	***P*-value**
molecular function	*0.1*	93	40.4	< 0.0001	50	15.7	< 0.0001
	0.05	58	16.0	< 0.0001	28	6.7	0.0002
	0.01	35	2.1	< 0.0001	10	0.9	0.0009
	0.001	14	0.2	< 0.0001	2	0.1	0.0096
cellular component	**SL**	**# in data set**	**random mean**	***P*-value**	**# in data set**	**random mean**	***P*-value**
	0.1	21	12.4	0.0634	32	5.8	< 0.0001
	0.05	6	4.8	0.358	26	2.7	< 0.0001
	0.01	5	0.7	0.0144	14	0.4	< 0.0001
	0.001	4	0.1	0.0017	8	0.0	< 0.0001
biological process	**SL**	**# in data set**	**random mean**	***P*-value**	**# in data set**	**random mean**	***P*-value**
	0.1	67	42.0	0.0218	42	19.2	0.0022
	0.05	40	16.7	0.0032	30	8.4	0.0004
	0.01	26	2.4	< 0.0001	17	1.3	< 0.0001
	0.001	13	0.2	< 0.0001	7	0.1	0.0001

SL : significance level

in data set : number of significant groups in the data set

random mean : mean number of significant groups among all random sets

P-value : *P*-value for the significance of the number of significant groups in the data set, given as the proportion of random sets with a number of significant groups that is higher or equal than the one given by the data set (< 0.0001, if no random set shows more significant groups than the data set)

Table A.2. Significant groups at different significance level in the HG test / human-chimpanzee brain comparison

		groups changed in their expression profile			groups conserved in their expression profile		
data set	**SL**	**# in data set**	**random mean**	**P-value**	**# in data set**	**random mean**	**P-value**
unmasked	*0.1*	78	37.5	0.0007	15	14.7	0.480
	0.05	35	17.4	0.0147	5	6.5	0.648
	0.01	5	2.6	0.190	0	0.9	1
	0.001	2	0.2	0.0502	0	0.1	1
sequence controlled	**SL**	**# in data set**	**random mean**	**P-value**	**# in data set**	**random mean**	**P-value**
	0.1	59	36.0	0.0212	24	13.3	0.0523
	0.05	39	19.7	0.0151	13	5.3	0.0404
	0.01	13	2.8	0.0051	6	0.8	0.0122
	0.001	7	0.2	0.0009	2	0.1	0.0121

Legend: see Table A.1.

7.2. Significant groups in the HG test

Table B.1. Changed and conserved groups significant at a 5% level before and after refinement / *molecular function* – in-brain comparison

molecular function **(N = 2919, K = 730)**					
changed groups at a 5% significance level	**GO id**	**M**	**x**	***P*-value (a)**	***P*-value (b)**
signal transducer	GO:0004871	518	181	*1.45E-08*	0.0362066
GTPase activator	GO:0005096	37	23	*1.62E-06*	6.22E-05
small GTPase regulatory/interacting protein	GO:0005083	81	37	*3.32E-05*	0.188717
inward rectifier potassium channel	GO:0005242	7	7	*5.99E-05*	5.99E-05
extracellular ligand-gated ion channel	GO:0005230	26	16	*7.98E-05*	0.169366
enzyme activator	GO:0008047	61	28	0.00027427	0.754709
receptor	GO:0004872	257	88	0.000329	0.19412
receptor signaling protein	GO:0005057	156	58	0.00035439	0.152896
enzyme regulator	GO:0030234	127	49	0.00037862	0.442074
transmembrane receptor	GO:0004888	188	67	0.00051069	0.218391
alpha-type channel	GO:0015268	91	37	0.00061864	0.433665
ion channel	GO:0005216	88	36	0.00063747	0.461382
glutamate receptor	GO:0008066	15	10	0.00077185	1
channel/pore class transporter	GO:0015267	92	37	0.00079282	0.462624
cation channel	GO:0005261	63	27	0.00131349	0.270867
voltage-gated potassium channel	GO:0005249	16	10	0.00160052	0.399563
ligand-gated ion channel	GO:0015276	35	17	0.00206503	0.539332
voltage-gated ion channel	GO:0005244	30	15	0.00262686	0.195751
hormone	GO:0005179	12	8	0.00273256	0.250086
phosphotransferase, alcohol group as acceptor	GO:0016773	224	74	0.00306538	0.0202248
potassium channel	GO:0005267	25	13	0.00325114	0.2824
ARF GTPase activator	GO:0008060	4	4	0.00388752	0.00388752
kinase	GO:0016301	228	74	0.0051101	0.0302846
excitatory extracellular ligand-gated ion channel	GO:0005231	13	8	0.00556014	0.261791
inhibitory extracellular ligand-gated ion channel	GO:0005237	13	8	0.00556014	0.437693
GABA receptor	GO:0016917	13	8	0.00556014	0.437693
neurotransmitter receptor	GO:0030594	16	9	0.00733535	0.367368
neurotransmitter binding	GO:0042165	16	9	0.00733535	0.367368
GABA-A receptor	GO:0004890	11	7	0.00746964	0.00746964
peptide hormone	GO:0005180	11	7	0.00746964	0.00746964
*transferase, transferring phosphorus-containing groups **	GO:0016772	251	79	0.0092805	0.997684
ionotropic glutamate receptor	GO:0004970	9	6	0.00990564	0.00990564
glutamate-gated ion channel	GO:0005234	9	6	0.00990564	1
monocarboxylate channel	GO:0015256	9	6	0.00990564	1
glutamate channel	GO:0015259	9	6	0.00990564	1
cell cycle regulator	GO:0003750	20	10	0.0136098	0.0136098
carboxylic ester hydrolase	GO:0016789	20	10	0.0136098	0.0136098
protein kinase	GO:0004672	190	61	0.0137087	0.0740401
structural constituent of cytoskeleton	GO:0005200	47	19	0.0137706	0.0137706
ephrin receptor	GO:0005003	3	3	0.0155929	0.0155929
cytoskeletal adaptor	GO:0008093	3	3	0.0155929	0.0155929
receptor binding	GO:0005102	74	27	0.0174451	0.136145
mRNA binding	GO:0003729	49	19	0.0222276	0.0222276
G-protein coupled receptor	GO:0004930	46	18	0.0233218	0.366226
RNA polymerase II transcription factor, enhancer binding	GO:0003705	8	5	0.0271595	0.0271595
calcium-dependent cell adhesion molecule	GO:0008014	8	5	0.0271595	0.0271595
calcium ion binding	GO:0005509	34	14	0.0274367	0.0274367
chloride channel	GO:0005254	19	9	0.0283804	0.633453
protein tyrosine phosphatase	GO:0004725	35	14	0.0355395	0.0355395
transmembrane receptor protein tyrosine kinase	GO:0004714	35	14	0.0355395	0.152622

single-stranded RNA binding	GO:0003727	6	4	0.0374846	0.0374846
DNA-directed DNA polymerase	GO:0003887	6	4	0.0374846	0.0374846
protein kinase C	GO:0004697	6	4	0.0374846	0.0374846
metabotropic glutamate, GABA-B-like receptor	GO:0008067	6	4	0.0374846	0.0374846
protein kinase regulator	GO:0019887	6	4	0.0374846	0.0374846
phorbol ester receptor	GO:0001565	6	4	0.0374846	1
sulfotransferase	GO:0008146	14	7	0.0379686	0.0379686
amine receptor	GO:0008227	9	5	0.0487269	0.0487269
conserved groups at a 5 % significance level					
structural constituent of ribosome	GO:0003735	62	3	*2.52E-05*	2.52E-05
ligase	GO:0016874	54	4	0.00082042	0.0950857
acid-D-amino acid ligase	GO:0016881	32	1	0.00111776	0.421586
carrier	GO:0005386	103	13	0.00129695	0.0300252
ligase, forming carbon-nitrogen bonds	GO:0016879	37	2	0.0019777	0.366547
ubiquitin-protein ligase	GO:0004842	29	1	0.00244522	0.00244522
ubiquitin conjugating enzyme	GO:0004840	34	2	0.00404722	0.00404722
enzyme	GO:0003824	1064	236	0.004133	0.00044736
oxidoreductase, acting on NADH or NADPH	GO:0016651	37	3	0.00859876	0.0997014
cation transporter	GO:0008324	75	10	0.00955476	0.262266
small protein conjugating enzyme	GO:0008639	36	3	0.0106504	0.937521
primary active transporter	GO:0015399	72	10	0.0152965	0.0152965
translation factor, nucleic acid binding	GO:0008135	40	4	0.0154633	0.0154633
translation regulator	GO:0045182	40	4	0.0154633	1
hydrogen ion transporter	GO:0015078	45	5	0.0172361	0.0172361
peptidase	GO:0008233	85	13	0.0204381	0.0204381
chaperone	GO:0003754	44	5	0.0207118	0.0207118
monovalent inorganic cation transporter	GO:0015077	49	6	0.0220932	0.937521
RAB small monomeric GTPase	GO:0003928	19	1	0.030568	0.030568
general RNA polymerase II transcription factor	GO:0016251	19	1	0.030568	0.030568
small nuclear ribonucleoprotein	GO:0003734	12	0	0.0313946	0.0313946
ion transporter	GO:0015075	91	15	0.0331338	0.441024
sodium transporter	GO:0015081	30	3	0.0366646	0.0366646
apoptosis inhibitor	GO:0008189	18	1	0.0389722	0.0389722
apoptosis regulator	GO:0016329	18	1	0.0389722	1
hydrolase, acting on acid anhydrides	GO:0016817	181	35	0.0391808	0.701944
ATPase	GO:0016887	99	17	0.0394674	0.0394674
oxidoreductase, acting on NADH or NADPH, quinone or similar compound as acceptor	GO:0016655	29	3	0.0446456	0.0446456

N = number of genes annotated in the taxonomy
K = number of changed genes annotated in the taxonomy
M = number of detected genes in the functional group
x = number of changed genes in the functional group
P-value (a) = *P*-value in the HG test before refinement
P-value (b) = *P*-value in the HG test after refinement

The group *transferase, transferring phosphorus containing groups* is the only group, which is changed before, but conserved after refinement. It is therefore marked with a *. The *P*-value before refinement of groups that are still significant using a Bonferoni correction at a 5% significance level is given in italic. The *P*-value field of groups which are significant on a 5% level after refinement is filled grey.

Table B.2. Changed and conserved groups significant at a 1 % and 5% level before and after refinement /*cellular component* – in-brain comparison

cellular component **(N = 2219, K = 505)**					
changed groups at a 1% significance level	**GO id**	**M**	**x**	***P*-value (a)**	***P*-value (b)**
plasma membrane	GO:0005886	548	174	*1.13E-08*	0.00020032
integral to plasma membrane	GO:0005887	419	127	*4.15E-05*	0.00040461
voltage-gated potassium channel complex	GO:0008076	17	9	0.00629858	0.00629858
membrane	GO:0016020	766	205	0.00071277	0.945402
integral to membrane	GO:0016021	513	147	0.00022112	0.676881
changed groups at a 5% significance level					
cytoplasm	GO:0005737	950	167	*2.46E-07*	0.73512
ribonucleoprotein complex	GO:0030529	101	5	*6.48E-07*	0.0772683
cytosol	GO:0005829	139	11	*1.82E-06*	0.181711
ribosome	GO:0005840	76	4	*2.79E-05*	0.157522
cytosolic ribosome (sensu Eukarya)	GO:0005830	58	2	*4.30E-05*	0.77242
large ribosomal subunit	GO:0015934	34	0	*0.00014266*	0.355687
mitochondrion	GO:0005739	174	21	*0.00015186*	0.0142884
cytosolic large ribosomal subunit (sensu Eukarya)	GO:0005842	30	0	0.00040763	0.00040763
peroxisome	GO:0005777	25	0	0.0015097	0.0015097
mitochondrial inner membrane	GO:0005743	56	4	0.0017209	0.0250042
inner membrane	GO:0019866	56	4	0.0017209	1
26S proteasome	GO:0005837	33	1	0.00202354	0.00202354
mitochondrial membrane	GO:0005740	69	7	0.00526894	0.659267
endomembrane system	GO:0012505	60	6	0.00855764	0.0897553
nucleolus	GO:0005730	32	2	0.0133015	0.0133015
lysosome	GO:0005764	31	2	0.0163083	0.0163083
lytic vacuole	GO:0000323	31	2	0.0163083	1
vacuole	GO:0005773	31	2	0.0163083	1
endoplasmic reticulum	GO:0005783	86	12	0.0272612	0.0272612
spliceosome complex	GO:0005681	21	1	0.0311662	0.0311662
mitochondrial electron transport chain complex (sensu Eukarya)	GO:0005746	34	3	0.0317704	0.0317704
cytosolic small ribosomal subunit (sensu Eukarya)	GO:0005843	27	2	0.036152	0.036152
small ribosomal subunit	GO:0015935	27	2	0.036152	1
eukaryotic 48S initiation complex	GO:0016283	27	2	0.036152	1
nuclear membrane	GO:0005635	32	3	0.0453747	0.0453747
transcription factor complex	GO:0005667	32	3	0.0453747	0.0453747

Legend: see Table B.1.

Table B.3. Changed and conserved groups significant at a 5% level before and after refinement / *biological process* – in-brain comparison

biological process **(N = 2897, K = 710)**					
changed groups at a 5% significance level	**GO id**	**M**	**x**	***P*-value (a)**	***P*-value (b)**
neurogenesis	GO:0007399	168	78	*1.46E-10*	4.12E-06
G-protein coupled receptor protein signaling pathway	GO:0007186	95	51	*4.99E-10*	0.0584234
signal transduction	GO:0007165	641	212	*1.55E-08*	0.00038303
cell communication	GO:0007154	984	301	*4.27E-08*	0.29515
cell surface receptor linked signal transduction	GO:0007166	195	78	*5.22E-07*	0.08005
organogenesis	GO:0009887	269	99	*1.54E-06*	0.841835
central nervous system development	GO:0007417	50	27	*6.03E-06*	0.00035568
morphogenesis	GO:0009653	287	101	*1.24E-05*	0.929546
cell-cell signaling	GO:0007267	162	63	*2.08E-05*	0.0914617
synaptic transmission	GO:0007268	90	40	*2.10E-05*	2.10E-05
transmission of nerve impulse	GO:0019226	91	40	*2.89E-05*	1
regulation of G-protein coupled receptor protein	GO:0008277	14	11	*3.10E-05*	3.10E-05

signaling pathway					
development	GO:0007275	409	130	0.00019201	0.8826
G-protein signaling, coupled to cAMP nucleotide second messenger	GO:0007188	18	11	0.00100255	1
cAMP-mediated signaling	GO:0019933	18	11	0.00100255	1
potassium transport	GO:0006813	28	14	0.00296799	0.00296799
G-protein signaling, adenylate cyclase inhibiting pathway	GO:0007193	10	7	0.00304947	0.0479408
adenylate cyclase activation	GO:0007190	4	4	0.00358478	0.00358478
glutamate signaling pathway	GO:0007215	8	6	0.00374384	0.00374384
intracellular signaling cascade	GO:0007242	142	49	0.00392198	0.10043
metal ion transport	GO:0030001	50	21	0.0045645	0.281103
brain development	GO:0007420	13	8	0.00485975	0.00485975
behavior	GO:0007610	25	12	0.00883438	0.0917664
G-protein signaling, coupled to cyclic nucleotide second messenger	GO:0007187	25	12	0.00883438	0.954896
second-messenger-mediated signaling	GO:0019932	25	12	0.00883438	0.954896
cyclic-nucleotide-mediated signaling	GO:0019935	25	12	0.00883438	0.954896
feeding behavior	GO:0007631	5	4	0.0144246	0.0144246
G-protein signaling, adenylate cyclase activating pathway	GO:0007189	5	4	0.0144246	1
obsolete	GO:0008371	377	108	0.0275023	0.0275023
G-protein signaling, coupled to IP3 second messenger (phospholipase C activating)	GO:0007200	11	6	0.0310057	0.0310057
circulation	GO:0008015	32	13	0.0316874	0.0316874
phagocytosis	GO:0006909	6	4	0.0349087	0.0349087
negative regulation of adenylate cyclase activity	GO:0007194	6	4	0.0349087	0.0349087
nutritional response pathway	GO:0007584	6	4	0.0349087	0.0349087
regulation of adenylate cyclase activity	GO:0045761	6	4	0.0349087	1
protein amino acid phosphorylation	GO:0006468	92	30	0.0463977	0.0463977
cation transport	GO:0006812	64	22	0.0472952	0.688043
glutamate transport	GO:0015813	4	3	0.0479408	0.0479408
acidic amino acid transport	GO:0015800	4	3	0.0479408	1
monovalent inorganic cation transport	GO:0015672	47	17	0.0483397	0.88021
conserved groups at a5% significance level					
metabolism	GO:0008152	1488	299	*8.56E-09*	0.161614
cell growth and/or maintenance	GO:0008151	2091	453	*1.13E-08*	0.647602
protein biosynthesis	GO:0006412	159	17	*5.62E-06*	5.62E-06
macromolecule biosynthesis	GO:0009059	159	17	*5.62E-06*	1
protein metabolism	GO:0019538	605	109	*1.22E-05*	0.00901204
biosynthesis	GO:0009058	220	30	*2.82E-05*	0.3389
DNA repair	GO:0006281	69	6	0.00064569	0.00064569
DNA metabolism	GO:0006259	144	21	0.0020698	0.219217
catabolism	GO:0009056	140	21	0.00356469	0.104533
ubiquitin-dependent protein catabolism	GO:0006511	48	4	0.00375834	0.00375834
protein-ligand dependent protein catabolism	GO:0019941	48	4	0.00375834	1
oxidative phosphorylation	GO:0006119	19	0	0.00469615	0.32457
proteolysis and peptidolysis	GO:0006508	97	13	0.0047417	0.202698
protein catabolism	GO:0030163	97	13	0.0047417	0.202698
macromolecule catabolism	GO:0009057	107	15	0.00503467	0.183872
energy pathways	GO:0006091	68	8	0.00674125	0.0926055
electron transport	GO:0006118	25	1	0.00788661	0.254803
intracellular protein transport	GO:0006886	86	12	0.011236	0.011236
oxidative phosphorylation, NADH to ubiquinone	GO:0006120	15	0	0.0145671	0.0145671
nucleocytoplasmic transport	GO:0006913	50	6	0.0223802	0.0223802
cytoplasmic transport	GO:0016482	50	6	0.0223802	1
positive regulation of cell proliferation	GO:0008284	44	5	0.0246048	0.0246048
apoptotic program	GO:0008632	13	0	0.0256386	0.0256386
cell cycle	GO:0007049	243	47	0.0280708	0.492885
energy derivation by oxidation of organic compounds	GO:0015980	32	3	0.0282586	0.0282586

nuclear division	GO:0000280	43	5	0.0293085	0.379143
response to biotic stimulus	GO:0009607	181	34	0.0365907	0.148096
regulation of cell cycle	GO:0000074	153	28	0.0381843	0.0381843
humoral immune response	GO:0006959	35	4	0.0459631	0.0459631
mitosis	GO:0007067	35	4	0.0459631	0.0459631

Legend: see Table B.1.

Table C.1. Changed and conserved groups significant at a 5% level before and after refinement / human chimpanzee comparison–unmasked/ *biological process*

biological process **(N = 2897, K = 455)**					
changed groups at a 5% significance level	**GO id**	**M**	**x**	***P*-value (a)**	***P*-value (b)**
complement activation	GO:0006956	4	4	0.00060174	0.00060174
humoral defense mechanism (sensu Vertebrata)	GO:0016064	4	4	0.00060174	1
sphingolipid metabolism	GO:0006665	13	7	0.00163106	0.0532076
lipid metabolism	GO:0006629	129	33	0.00208245	0.0401148
chromosome condensation	GO:0030261	3	3	0.00385274	0.157059
M phase	GO:0000279	57	16	0.0117766	0.192315
phospholipase C activation	GO:0007202	4	3	0.0136057	0.0136057
fertilization (sensu Animalia)	GO:0007338	4	3	0.0136057	0.0136057
lipid catabolism	GO:0016042	4	3	0.0136057	0.0136057
purine nucleotide biosynthesis	GO:0006164	4	3	0.0136057	0.289496
purine ribonucleotide biosynthesis	GO:0009152	4	3	0.0136057	0.289496
fertilization	GO:0009566	4	3	0.0136057	1
glycosphingolipid biosynthesis	GO:0006688	7	4	0.0141575	0.0141575
glycolipid biosynthesis	GO:0009247	7	4	0.0141575	1
sphingolipid biosynthesis	GO:0030148	7	4	0.0141575	1
membrane lipid metabolism	GO:0006643	31	10	0.0163358	0.162779
GTP biosynthesis	GO:0006183	2	2	0.0246218	0.0246218
induction of apoptosis by p53	GO:0006918	2	2	0.0246218	0.0246218
mitotic chromosome condensation	GO:0007076	2	2	0.0246218	0.0246218
leukocyte cell adhesion	GO:0007159	2	2	0.0246218	0.0246218
mitotic prophase	GO:0000088	2	2	0.0246218	1
nucleoside triphosphate biosynthesis	GO:0009142	2	2	0.0246218	1
purine nucleoside triphosphate biosynthesis	GO:0009145	2	2	0.0246218	1
ribonucleoside triphosphate biosynthesis	GO:0009201	2	2	0.0246218	1
purine ribonucleoside triphosphate biosynthesis	GO:0009206	2	2	0.0246218	1
GTP metabolism	GO:0046039	2	2	0.0246218	1
meiosis	GO:0007126	8	4	0.0248456	0.0248456
glycosphingolipid metabolism	GO:0006687	8	4	0.0248456	1
transcription initiation	GO:0006352	16	6	0.0286309	0.401164
membrane lipid biosynthesis	GO:0046467	16	6	0.0286309	0.424967
lipid biosynthesis	GO:0008610	29	9	0.0286728	0.255977
nuclear division	GO:0000280	43	12	0.0286755	0.418575
mitotic cell cycle	GO:0000278	117	26	0.0364557	0.0809244
humoral immune response	GO:0006959	35	10	0.0377673	0.358251
transcription initiation from Pol II promoter	GO:0006367	13	5	0.0404531	0.0404531

Legend: see Table B.1.

Table C.2. Changed and conserved groups significant at a 5% level before and after refinement / human chimpanzee comparison– sequence controlled */ biological process*

biological process (N = 2789, K = 360)					
changed groups at a 5% significance level	**GO id**	**M**	**x**	***P*-value (a)**	***P*-value (b)**
glycosphingolipid biosynthesis	GO:0006688	7	6	2.78E-05	2.78E-05
glycolipid biosynthesis	GO:0009247	7	6	2.78E-05	1
sphingolipid biosynthesis	GO:0030148	7	6	2.78E-05	1
glycosphingolipid metabolism	GO:0006687	8	6	9.91E-05	1
glycolipid metabolism	GO:0006664	9	6	0.00027	1
membrane lipid biosynthesis	GO:0046467	16	8	0.00036	1
sphingolipid metabolism	GO:0006665	13	7	0.00048	0.563959
chromosome condensation	GO:0030261	3	3	0.00214	0.129079
lipid biosynthesis	GO:0008610	29	10	0.00233	0.724813
sulfur metabolism	GO:0006790	13	6	0.00343	0.072677
M phase	GO:0000279	56	15	0.0037	0.390038
lipid metabolism	GO:0006629	122	26	0.00552	0.188389
sulfur amino acid metabolism	GO:0000096	11	5	0.00828	0.175408
membrane lipid metabolism	GO:0006643	30	9	0.01068	0.937602
cell-cell adhesion	GO:0016337	8	4	0.01248	0.175408
meiosis	GO:0007126	8	4	0.01248	0.241536
nuclear division	GO:0000280	42	11	0.01464	0.852603
mitotic chromosome condensation	GO:0007076	2	2	0.01662	0.016621
leukocyte cell adhesion	GO:0007159	2	2	0.01662	0.016621
mitotic prophase	GO:0000088	2	2	0.01662	1
protein amino acid sulfation	GO:0006477	5	3	0.01746	0.017456
NLS-bearing substrate-nucleus import	GO:0006607	13	5	0.01852	0.018522
regulation of mitosis	GO:0007088	9	4	0.02021	0.020205
induction of apoptosis	GO:0006917	50	12	0.02194	0.021944
induction of programmed cell death	GO:0012502	50	12	0.02194	1
meiotic prophase I	GO:0007128	6	3	0.0316	0.031604
respiratory gaseous exchange	GO:0007585	6	3	0.0316	0.031604
meiosis I	GO:0007127	6	3	0.0316	1
protein modification	GO:0006464	243	41	0.03692	0.087763
sexual reproduction	GO:0019953	37	9	0.04107	0.041071
reproduction	GO:0000003	37	9	0.04107	1
ossification	GO:0001503	3	2	0.04559	0.045593
phosphatidylinositol biosynthesis	GO:0006661	3	2	0.04559	0.045593
phagocytosis, engulfment	GO:0006911	3	2	0.04559	0.045593
complement activation	GO:0006956	3	2	0.04559	0.045593
humoral defense mechanism (sensu Vertebrata)	GO:0016064	3	2	0.04559	1
phosphatidylinositol metabolism	GO:0046488	3	2	0.04559	1
M phase of mitotic cell cycle	GO:0000087	38	9	0.04794	0.696761
embryogenesis and morphogenesis	GO:0007345	50	11	0.04965	0.049651
conserved groups at a 5% significance level					
transcription	GO:0006350	321	23	0.00038	0.085039
transcription, DNA-dependent	GO:0006351	311	22	0.00038	0.062398
nucleobase, nucleoside, nucleotide and nucleic acid metabolism	GO:0006139	594	55	0.00132	0.427367
transcription from Pol II promoter	GO:0006366	259	20	0.00407	0.004071
RNA metabolism	GO:0016070	125	7	0.00538	0.06935
RNA processing	GO:0006396	120	7	0.00835	0.102087
regulation of transcription, DNA-dependent	GO:0006355	178	14	0.02038	0.020376
regulation of transcription	GO:0045449	178	14	0.02038	1
mRNA splicing	GO:0006371	51	2	0.03068	0.030679
muscle development	GO:0007517	25	0	0.03108	0.031082
mRNA processing	GO:0006397	72	4	0.03447	0.478423
RNA splicing	GO:0008380	60	3	0.03861	0.672786
cell growth and/or maintenance	GO:0008151	2018	246	0.0399	0.030649

Legend: see Table B.1

Content