Primary diversity mechanisms of the antibody synthesis in humans, mice and chickens

Nikolai Hecker

August 25, 2008

Contents

The current models for the evolution of antibody diversity do not reflect sufficiently the complexity of the associated mechanisms. The diversity of antibodies is essential for the survival of vertebrates against the huge amount of pathogens in the natural environment. The improvement of these models is a crucial step to understand the key factors that govern the evolution of antibody diversity. We analyze the V gene based primary diversity mechanisms in humans, mice and chickens. Our investigations provide the characteristic features of two mechanisms which help to create more suitable models for the evolution of antibodies.

1 Introduction

Among all organisms from the tree of life, vertebrates have the most highly developed immune system which is able to adapt to a rapidly changing and evolving environment of pathogens. This might be one of the reasons for the dominant position of vertebrates in nature. One of its major components is the adaptive immune system which is found in most higher vertebrates such as mammals, birds or bony fishes. However, even jawless fish lack an adaptive immune system. The antibody response is part of the adaptive immune system. Antibodies are necessary to protect the organism from pathogens. There are more pathogens than one genome could encode. Hence, vertebrates have evolved complex mechanisms to create diverse antibodies [Janeway et al., 2005, 6].

Since the immune system plays a crucial role for the survival of individual organisms as well as entire species, it is a milestone in the evolution of life. For this reason, it is of particular importance to understand evolutionary processes which have led to the development of the immune system. Thus, we create evolutionary models which reveal insights into these processes. Especially the evolution of antibodies is essential for most vertebrates. Forrest and Oprea have invented a model to evaluate an evolutionary favorable number of antibody genes [Oprea and Forrest, 1998]. This model does not seem to be sufficient because the evolution of antibodies involves highly complex mechanisms. For this reason, our aim is to identify the key characteristics of these processes which can actually be found in reality. These properties help to create a more suitable model for the antibody evolution.

Our investigations focus on three species that use two different mechanisms to achieve primary antibody diversity. Humans and mice use the same mechanism. We analyze both of them to compare differences of species with a close evolutionary relationship (figure 1). The third species we study is the chicken, since they use a very different mechanism to achieve primary antibody diversity. Janeway and others have proposed a factor called combinatorial diversity to describe the primary antibody diversity in humans and mice [Janeway et al., 2005, 6]. Our analysis suggests a new notion of the combinatorial diversity, and reveals important characteristics of the primary diversity mechanism in chickens.

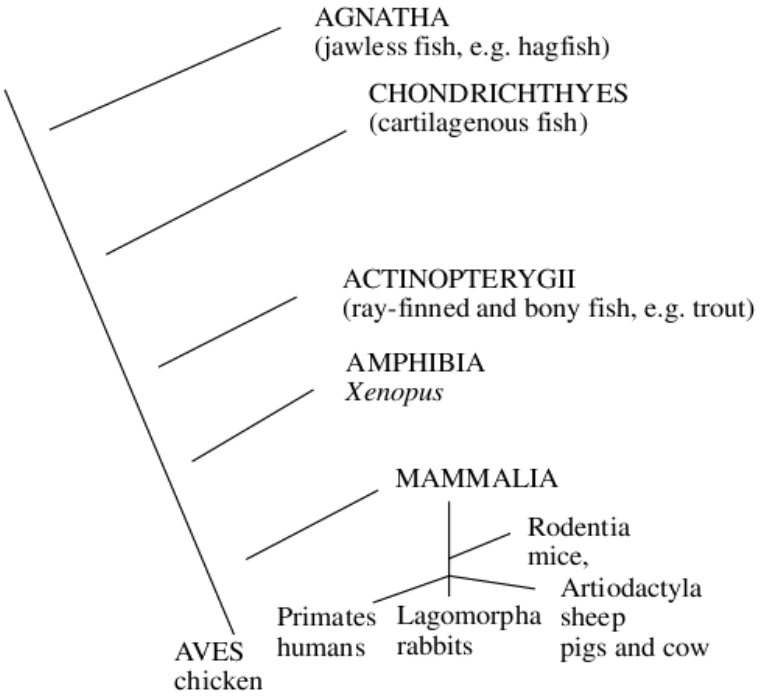


Figure 1: Sporadic tree of vertebrates. [M. Diaz et al., 2001]

2 Background

Antibodies are an essential part of the adaptive immunesystem. The natural environment contains a huge amount of pathogens. The amount of pathogens is even changing and increasing. An antibody recognizes specific pathogens. In order to survive in a natural environment an organism must produce more different antibodies than one genome could encode. Vertebrates with an adaptive immune system use highly complex mechanisms to create diverse antibodies. Humans and mice use a process called V(D)J recombination to achieve primary antibody diversity. In contrast to humans or mice, chickens use gene conversion. In this section we present important features of antibodies, V(D)J recombination and gene conversion.

2.1 Antibody structure and development

Antibodies are also known as immunoglubulins. They are expressed by B cells which are part of the adaptive immune system. Some important functions of antibodies are the neutralization of toxins and the labeling of foreign cells or particles or even worm larvas. Other cells of the immune system recognize the labeled foreign bodies and eliminate them. The primary development of B cells takes place in the central lymphoid tissues such as the bone marrow in humans or the yolk sac in chickens. Stem cells grow to mature B cells in the primary lymphoid tissues. Once they have developed to mature B cells they migrate to the peripheral lymphoid tissues (e.g. spleen). In the peripheral lymphoid tissues B cells are exposed to antigens and complete their development. A specific part of an antibody called antigen binding site binds antigens [Janeway et al., 2005, 6].
An antibody always consists of two identical light chains and two identical heavy chains (figure 2). Each of these chains is subdivided into variable and constant regions. A light chain has a single variable and a single constant region. Heavy chains have a single variable region and three or four constant regions according to their function [Alberts et al., 2002, 4]. The light chain and heavy chain variable region each contribute to the antigen binding site [Berg et al., 2007, 6]. Since an antibody consists of two identical heavy chains and two identical light chains it has two identical antigen binding sites. Antigens can be small particles or parts of proteins expressed on cell surfaces or on virus membranes. If one considers the huge amount of pathogens with different surfaces, it is obvious that antibodies must be very diverse. This diversity is created by several mechanisms. One of them is the combination of different light chains and heavy chains. Complex mechanisms, which change the antibody gene sequence, create diverse heavy

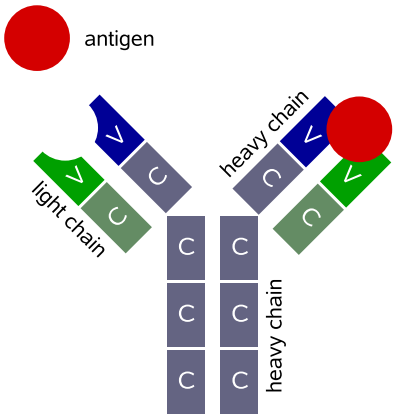


Figure 2: Antibody structure and antigen interaction. The Heavy chain and the light chain each provide one variable region which contributes to the antigen binding site. Variable regions are labeled with a white 'V', constant regions with a white 'C'.

and light chains [Janeway et al., 2005, 6].

2.2 Germline elements of antibody coding genes

There are at least four important elements in an antibody encoding gene sequence. These elements are called V,D,J and C segments. C segments are constant segments which are similar in all antibodies. The C segments build up the constant regions of an antibody. V,D and J segments encode the variable region. V segments, which are also referred to as V gene segments or V genes, contribute to most of the amino acid sequence of this region. V genes are responsible for the variability of antibody germline sequences. Thus, they are called V genes. In addition, D segments also contribute to the variablity of antibodies. They are called D genes because they are diversity increasing segments. D segments are only observed in heavy chain sequences. J segments are located between D and C segments in heavy chains and between V and C segments in light chains. They are called J segments because they help to join the other segments. In contrast to the dogma of molecular genetics, genomic antibody sequences undergo changes during the development of a cell. These changes of the germline sequence are mainly caused by two to three mechanisms. In all vertebrates with an adaptive immune system the sequences are altered by V(D)J recombination and somatic hypermutation. In addition, birds and several mammals such as rabbits, pigs or cows use gene conversion to change antibody

genes [Janeway et al., 2005, 6].

2.3 V(D)J recombination

During the process of V(D)J recombination V, (D) and J segments are rearranged in the genomic sequence. Figure 3 illustrates this process in heavy chains. First any of several D segments recombines with any of several J segments. D and J segments, which are located between the two recombining segments, are excised. Second the rearranged DJ segment is joined by any of several V segments. Again, segments, which are located between the joining V and DJ segment, are excised. After transcription all segments, which are not joined except the C segments, are spliced out. This process is similar in light chains. In contrast to the heavy chains, only a single rearrangement occurs in light chain sequences, namely V-J joining. The number of possibilities to combine different heavy and light chains, which are created by V(D)J recombination, is called combinatorial diversity [Janeway et al., 2005, 6].

2.4 Gene conversion and somatic hypermutation

After V(D)J recombination has occurred somatic hypermutation is observed in all vertebrates with an adaptive immune system. Somatic hypermutation means the accumulation of point mutations which are biased to transitions. At the same time when somatic hypermutation happen gene conversion takes place in several vertebrates. There is evidence, which suggests that gene conversion is coupled to somatic hypermutation. During the gene conversion process a pool of pseudogenes is used to overwrite parts of the functional V gene (figure 4). Unlike, the functional V gene pseudogenes don't have a functional promotor and no leader exon. Pseudogenes also lack recombination signal sequences (RSS). Thus, they are not involved in V(D)J recombination [Janeway et al., 2005, 6].

The processes mentioned above are used to create diverse light and heavy chains. For this reason, these mechanisms are essential to achieve diversity in antibody synthesis. Somatic hypermutation can be seen as a process which induces further changes in primarily altered sequences. Thus, somatic hypermutation can be referred to as secondary diversity mechanisms. Consequently, the primary diversity mechanisms are V(D)J recombination and gene conversion. Both humans and mice are representatives of V(D)J recombination while chickens use gene conversion to achieve primary diversity.

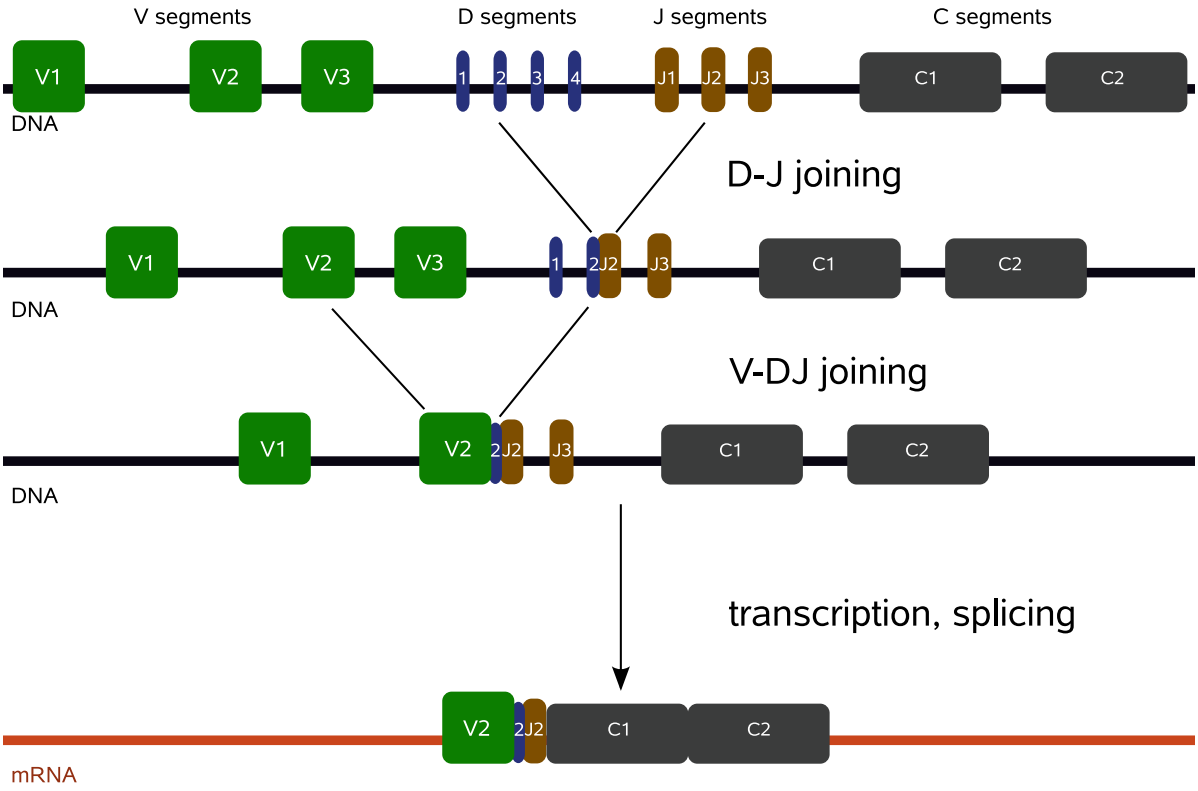


Figure 3: VDJ recombination. Firstly, any diversity (D) segment joins any joining (J) segment. Segments between the two joining segments are excised. Secondly, the DJ segment is joined by any variable (V) segment. After transcription not joined segments, except constant (C) segments, are spliced out.

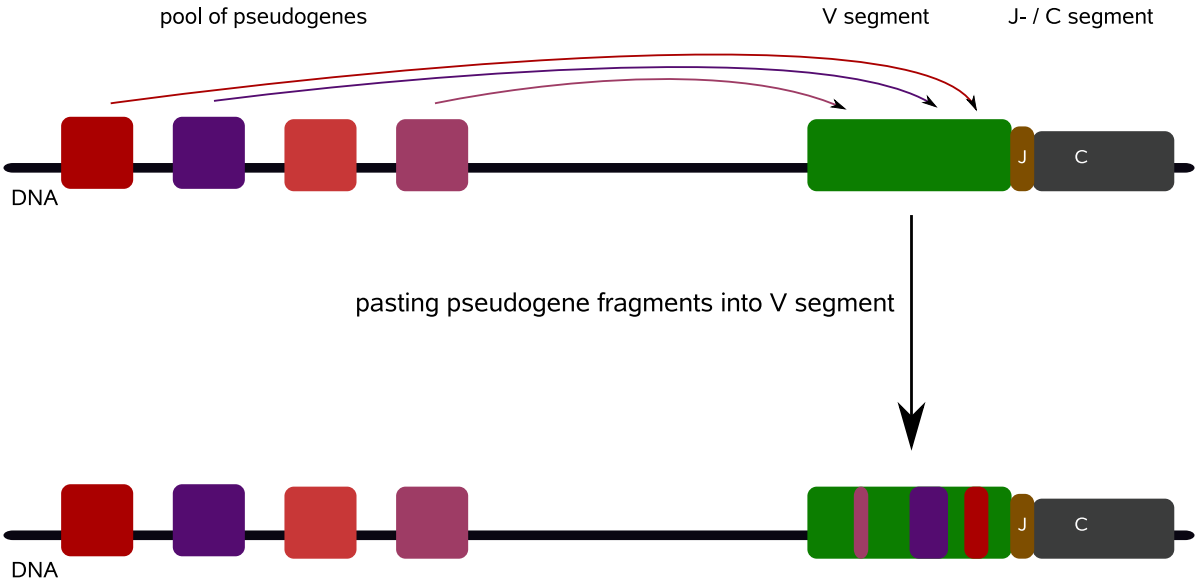


Figure 4: Gene conversion. A pool of pseudogenes is used to overwrite parts of the sequence of the functional V gene.

2.5 Germline structure and gene conversion in chickens

Chickens have a single heavy chain and a single light chain locus. Each of the loci contain a single functional V gene. Thus, chicken use gene conversion instead of V(D)J recombination to achieve primary diversity. About 80 pseudogenes are estimated for the heavy chain locus. 25 pseudogenes are identified at the light chain locus of chickens. Since most pseudogenes of the heavy chain locus are not determined, only the light chain locus is part of our investigations. Chickens have a specialized organ called bursa of fabricius. Before hatching, the bursa of fabricius is the organ where gene conversion takes place. At four to six month after hatching the bursa of fabricius involutes and gene conversion takes place in the spleen. In contrast to the spleen, somatic hypermutation occurs at low rate in the bursa of fabricius [Arakawa and Buerstedde, 2004]. For this reason, V gene sequences at bursal stage are most appropriate for our investigations of gene conversion in section 4.

2.6 Germline structure and V(D)J recombination in humans and mice

Humans and mice both use no gene conversion but V(D)J recombination to achieve primary diversity. Light chain sequences are located in two different loci in both humans and mice. The two loci are referred to as lambda locus and kappa locus. However, both have a single heavy chain locus. In terms of the V(D)J recombination our analysis considers only the diversity of V genes. The number of V genes is highly different in humans and mice. The kappa and heavy chain locus of mice consist of many V genes. 97 V genes are located at the kappa locus. The heavy chain locus of mice might contain 208 V genes. 75 of these are provisional V genes. Therefore the heavy chain locus contains at least 133 different V genes. All 208 V genes are considered in our analysis. The human heavy chain locus contains 48 and kappa locus contains 40 functional V genes. Only eight V genes were located at the lambda locus of mice. It is important that this data is considering laboratory mouse strains. The number of V genes, especially of the lambda locus, probably differs from the number of V genes in wildtype mice (Genbank accession (gb): NG_004051). The human lambda locus contains 33 V genes.

3 Combinatorial diversity in humans and mice

In this section we study the combinatorial diversity of V genes in humans and mice. These genes build the antigen binding sites which define the specificity of the antibodies. Our analysis is based on a new method of gene clustering. In order to demonstrate that this method is suitable for our analysis, we show that it produces valid results, when it is applied to the genes of the human lambda chain for which an established clustering is known. Our aim is to divide the V genes of humans and mice into groups, and analyze the evolutionary relationship between both species. Thus, we use our method to cluster the genes of human and mouse in separate sets, and all genes in one set. In addition, our algorithm provides an estimation of the diversity of the compared sequences. We suggest a new notion of the combinatorial diversity of V genes based on this estimation.

3.1 Database sources for V genes

We used two databases to obtain the sequences for our analysis: NCBI Genbank and IMGT. IMGT (ImMunoGenetics) is a database specialized on immunogenetics of vertebrates. This database provides sequences of immunoglobulins, T cell receptors, major histocompatibility complexes, immunoglobulin superfamilies, major histocompatibility complex superfamilies and related proteins [Lefranc et al., 2005], [IMGT/home]. IMGT has been created by Paul Lefranc in 1989. NCBI (National Center for Biotechnology Information) Genbank is a database which has an annotated collection of all publicly available DNA sequences [Benson et al., 2008], [Genbank/home]. NCBI Genbank has been created from 1979 to 1982, and is funded by the National Institute of Health (NIH).

IMGT was used to obtain all functional V genes of both humans and mice. The sequences were downloaded as FASTA files [IMGT/GENE-DB]. From these FASTA files we extracted reference sequences which are annotated with '*01'. Some V genes are listed on the web interface of IMGT but were missing inside the FASTA files. We downloaded these sequences manually from NCBI Genbank. All sequences were obtained in July 2008. Sequences, which were identical to or substrings of other sequences, were excluded. The mouse heavy chain V gene IGHV1-17-1 was removed because it is annotated as pseudogene in NCBI Genbank. The following sequences were downloaded manually from NCBI Genbank, and were used for the further analysis: IGKV1-33 (gb: M64856), IGKV1-39 (gb: X59312), IGKV2-28 (gb: X12691), IGKV2-40 (gb: X59311), IGHV1-62-1-pr and IGHV1-62-3-pr (gb: NG_005838).

3.2 Subgroup definition of V gene sequences

Immunogenetics, the genetic foundations of immunology, have been studied for almost three decades. Homologies between V genes have been analyzed over many years. These genes have been divided into families, groups and subgroups according to their sequence similarities. Our analysis follows the 'IMGT unique numbering for all IG V-REGIONs' [Lefranc et al., 2003]. Firstly, all genes are categorized by gene type and locus (e.g. IGLV for the immunoglobulin lambda chain V genes). Secondly, they are divided into subgroups (e.g. IGLV3). All members of a subgroup share at least 75% sequence identity. Thirdly, subgroups consist of different genes which are annotated with an unique number (e.g. IGLV3-11). Finally, these genes are divided in alleles which are annotated with an asterisk and a number (e.g. IGLV3-11*01). Alleles are supposed to have originated from the same gene. The reference sequences are the alleles which are annotated with '*01'. These sequences have been located at the genome reference assemblies.
All sequences, which we consider in our analysis, are V genes. Thus, we use the following abbreviations: 'L' for the lambda chain V genes, 'K' for the kappa chain V genes and 'H' for the heavy chain V genes. The associated number represents the subgroup (e.g. L3 is used instead of IGLV3). In order to distinguish the human genes from the mouse genes we annotate the mouse genes with 'm' and the human genes with 'h' (e.g. mL3 for the mouse lambda chain V gene subgroup 3). Sequences, which are annotated with 'S' (e.g. H8S), are provisional genes. The assignment of these provisional genes to the subgroups must be considered preliminarily. We use the abbreviations 'VL' for all lambda chain V genes, 'VK' for all kappa chain V genes and 'VH' for all heavy chain V genes in some figures and tables.

3.3 Heterogeneity calculation

The heterogeneity between two sequences is determined by the distance at the nucleotide level. Our algorithm is based on the best local alignment and the Hamming distance of two sequences. Firstly, the sequences are aligned by the best local alignment using the Smith-Waterman algorithm [Smith and Waterman, 1981]. The applied scores for nucleotide match, nucleotide mismatch, gap opening and gap extension are set to +5, -4, -12 and -4, respectively. Identities and gaps are counted inside the alignment (figure 5). The score is only used to determine the range of the best local alignment and does not contribute to the calculated distance. Secondly, the Hamming distance of the overlapping sequences before and after the alignment is computed (figure 6). The

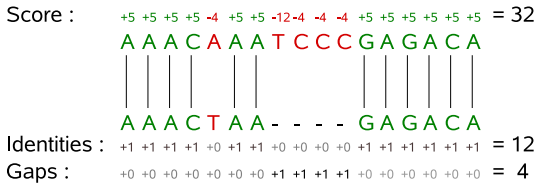


Figure 5: Identity and gap counting inside the best local alignment. This figure shows a fictional example of the best local alignment between two sequences. The bases to the left and to the right outside the alignment are not shown. The matching bases are colored green. The mismatches and the gaps are colored red.

distance between two sequences is calculated as follows:

$$\text{distance} \;=\; 1 - \frac{\text{identities} + \text{left overlap} + \text{right overlap}}{\text{length of smaller sequence} + \text{gaps}}$$

left overlap = overlap length before alignment – Hamming distance of overlap before alignment
right overlap = overlap length after alignment – Hamming distance of overlap after alignment

3.4 TSP gene clustering

Using the procedure described in the last section, we calculated the pairwise distances between all considered V genes. Each gene of this set has 433 distances to the other genes. This is a lot of data but it does not provide much information. The data must be reduced to the essential information. V genes, which are more homologous, have shorter distances to the each other genes. Thus, a clustering algorithm can identify the subgroups of homologous genes.

Two methods have been established to cluster genes: k-means and hierarchic clustering [Deonier et al., 2005, 1]. Since the k-means operates on vector space, it cannot be directly applied to our problem. The agglomerative hierarchic clustering is less demanding. However, in this work we use a new approach which is based on the traveling salesman problem (TSP). Like agglomerative hierarchic clustering, it only needs the pairwise distances between the input points. This method has recently been used to cluster gene expression data [Climer and Zhang, 2005]. However, it is a novelty approach to cluster the genes themselves in this manner.

The solution of a traveling salesman problem is the shortest circular tour between all nodes of a graph (figure 7). Each node is only touched once. Our algorithm, which we have discussed in section 3.3, creates a complete undirected graph. The edge weights

Figure 6: Algorithm to determine the distance between two sequences. The best local alignment is shown in red. The overlapping regions of the sequences are shown in blue-green.

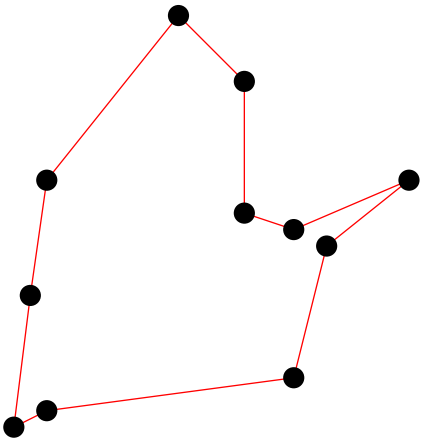


Figure 7: Geographical Traveling Salesman Problem (TSP). Eleven nodes, which are colored black, were chosen at random 2D-coordinates. The TSP instance was solved by the Concorde TSP solver which uses the NEOS server. The red lines indicate the shortest tour.

Figure 8: Example of Clusters. This figure shows fictional points which form two clusters.

are given by the computed gene distances. Each gene is connected to all other genes, since all distances are computed. In other words, we obtained a TSP instance. The solution of this TSP instance provides a minimal overall gene distance. The genes, which have close distances to each other, form a cluster (figure 8). Naturally, the genes outside this cluster have longer distances to the genes inside this cluster. Thus, it is not favorable to gain a short TSP tour, if we leave and visit the same clusters several times. Consequently, it is probable that the genes of a cluster will be sequentially ordered within a TSP tour. Thus, the solution of a TSP instance can be used for gene clustering. We use the publicly available TSP solver 'Concorde' to solve the TSP instances [Concorde]. The Concorde solver runs on the NEOS server at Arizona State University [Czyzyk et. al, 1998], [Gropp and More, 1997], [Dolan, 2001]. The Concorde TSP solver was written by D. Applegate, R. E. Bixby, V. Chvatal, and W. J. Cook and has been implemented by Hans Mittelmann. Once a TSP tour has been obtained, we can define a gene cluster as a series of genes on the tour which are connected by edges with a distance below a given threshold.

3.5 Validation of TSP gene clustering

We clustered the human lambda V genes to ensure that TSP clustering is a valid method. The resulting clusters should be equal to the established subgroups. In section 3.2 it was mentioned that these subgroups have been established in the literature. We use the subgroups as a reference clustering. All genes of one subgroup share at least 75% sequence identity. For this reason, we set the cluster threshold to 25% distance. Figure 9 shows the clustered TSP tour of the human lambda V genes. Each cluster contains the

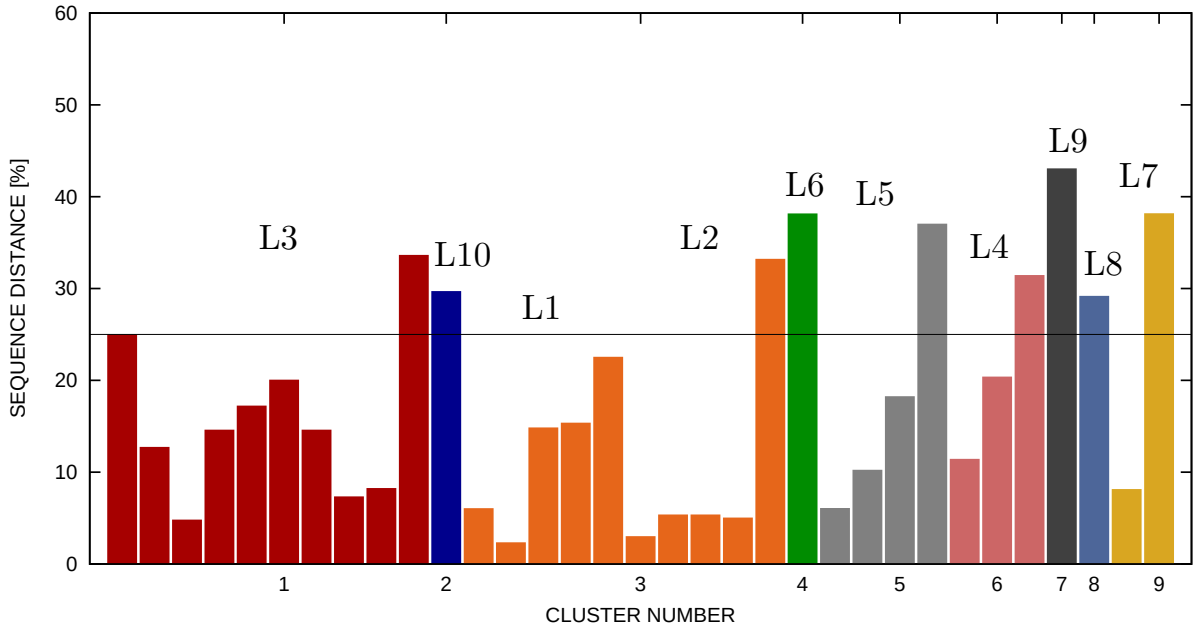
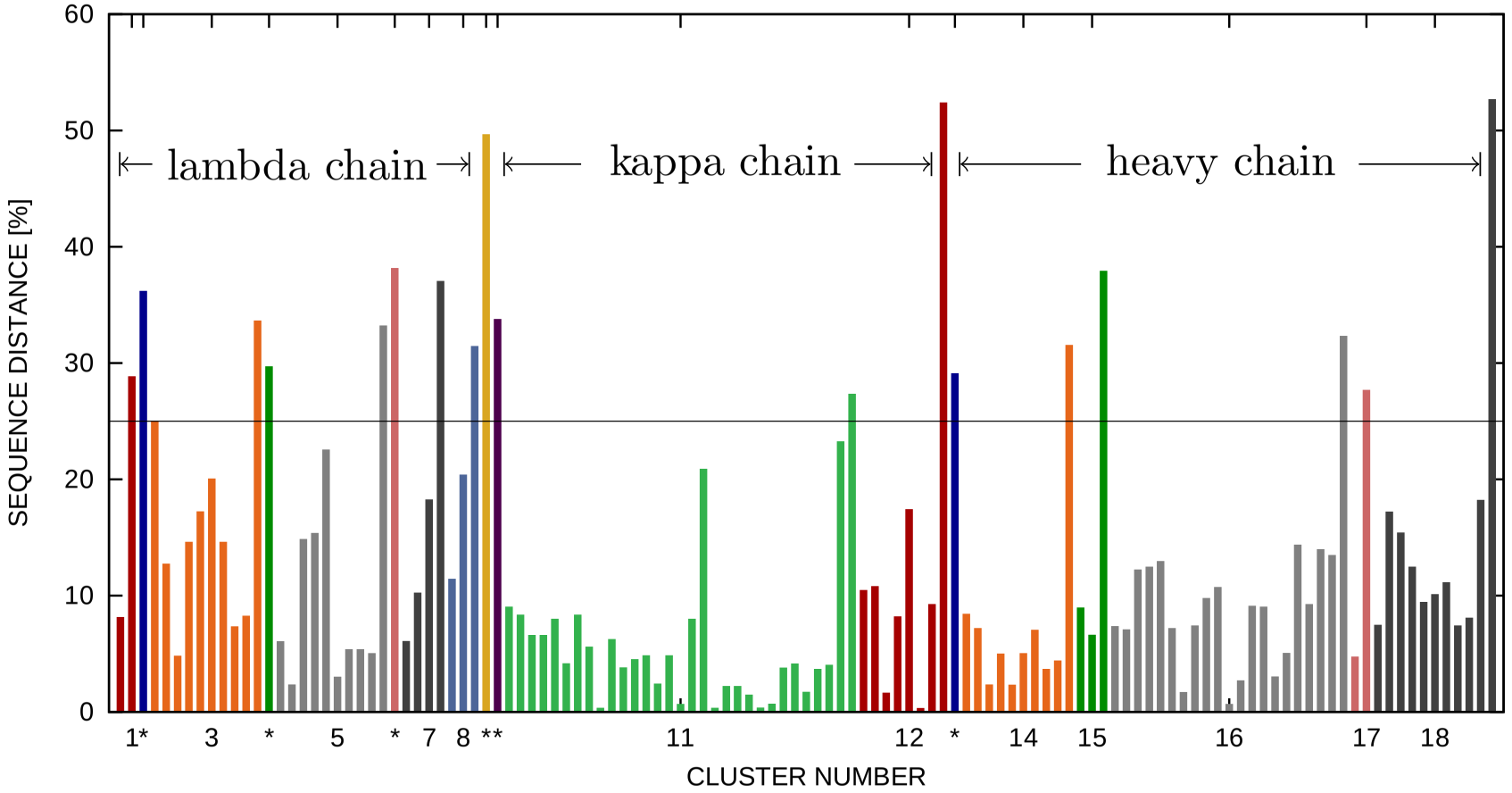


Figure 9: Comparison between the human lambda V genes. The TSP instance was solved by the Concorde TSP solver which uses the NEOS Server. The threshold for the clusters is set to 25%. The subgroups, which belong to a clusters, are labeled with black letters.

genes of a single subgroup except cluster 3. The genes of the subgroup L1 and L2 are found in the cluster 3. However, two distinct subclusters can be observed in the cluster 3. The first subcluster contains all members of the subgroup L1 and no other genes. All genes of the subgroup L2 and only sequences of L2 belong to the second subcluster. The subgroup definition does not imply that the sequences of two different clusters cannot be homologous. If we consider the human lambda locus, all computed clusters are equal to the established subgroups.
If we cluster the set of all human V genes, the genes are divided into three major clusters which reflect the lambda, the kappa and the heavy chain loci. The major cluster, which consists of all heavy chain genes (cluster 13 to 18), is separated from the major cluster which contains all kappa genes (cluster 10 to 12) and from the major cluster which contains all lambda genes (cluster 1 to 9). The cluster boundary between the kappa and heavy chain loci is about 52.4%. The boundary between the heavy chain and the lambda cluster has a value of 52.7%. The kappa and the lambda locus are separated by a boundary of 49.7% distance. All cluster boundaries within a major cluster have values less than 40%.

3.6 Homologies between human and mouse V genes

Both plots, that we have discussed above, suggest that the TSP clustering yields valid results. For this reason, we used our algorithm to analyze the homologies between the genes of one species and between the genes of both species. In other words, we clustered all human and all mouse V genes in separate sets and all sequences in one set. In additional experiments, we clustered the following sets: the human lambda and the mouse lambda V genes in one set, the human kappa and the mouse kappa V genes in one set, the human heavy chain and the mouse heavy chain V genes in one set, the human lambda V genes, the human kappa V genes, the human heavy chain V genes, the mouse lambda V genes, the mouse kappa V genes, the mouse heavy chain V genes. The results of our analysis are listed in the appendix (table 4 - table 16 and figure 22 - figure 29). The threshold for the clusters is set to 25%, as mentioned above.
Eleven clusters can be found in each the combined set of all kappa and the combined set of heavy chain of both human and mouse genes (table 1). The lambda genes of both species can be grouped into twelve clusters. The average number of genes inside a cluster of lambda genes is small compared to the other loci. Thus, we can assume that the lambda genes are more heterogenous than the genes of the kappa and the heavy chain locus. The heavy chain genes of each species appear to be more homologous. Figure 11



∗ cluster number can not be shown due to close distance on graphical display

Figure 10: Comparison between all human V genes. The TSP instance was solved by the Concorde TSP solver which uses the NEOS Server. The threshold for clusters is set to 25% The major clusters of heavy chain, of lambda chain and of kappa chain genes are labeled.

number of clusters

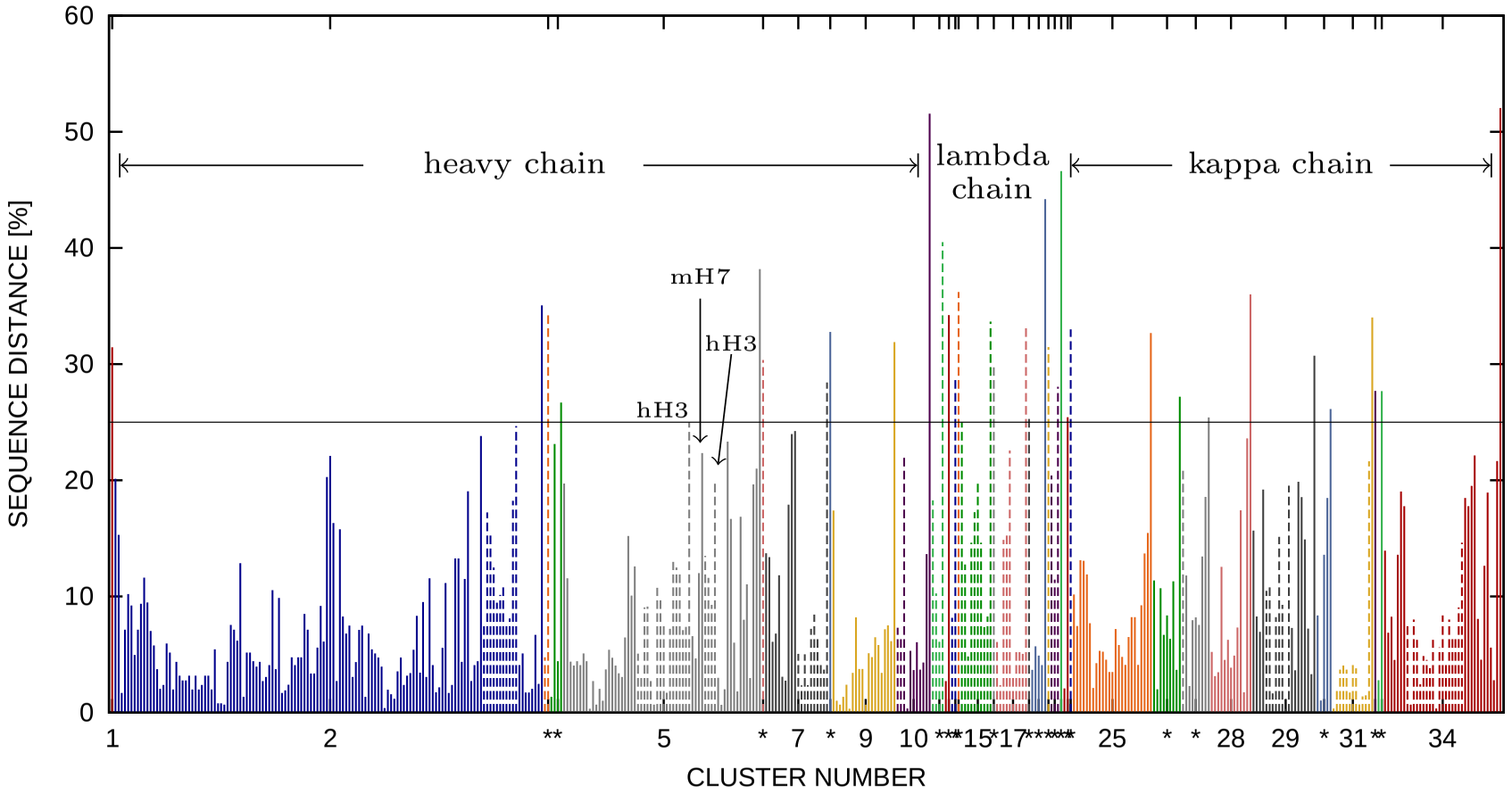
locus	human	mouse	human & mouse
all	18	15	34
VL	9	3	12
VK	3	9	11
VH	6	11	11

average number of genes per cluster

all	6.7 (±7.9)	12.5 (±22.1)	12.8 (±24.5)
VL	3.7 (±3.5)	2.7 (±1.7)	3.4 (±3.2)
VK	13.3 (±12.8)	10.8 (±8.9)	12.5 (±9.0)
VH	8.0 (±7.0)	18.9 (±31.2)	23.3 (±37.8)

Table 1: Number of clusters. Shows the number of clusters and the average number of genes inside one cluster of each loci for each species and of the combined sets of human and mouse genes.

shows the clustered TSP tour of all V genes of both humans and mice. Distinct divisions between the subgroups remain. Some of the human and the mouse V genes can be found in the same clusters. Clusters, which contain one subcluster of the human genes and one or more subclusters of the mouse genes, can be observed, namely cluster 2, 4, 5, 7, 10, 27, 31 and 34. However, frequently altering subclusters of the human and the mouse genes cannot be observed within a cluster. In addition, no genes of a species are scattered through the genes of the other species. This suggests heterogeneity between the human and the mouse V genes. The human subgroup H3 is an exceptional case. hH3 belongs to the cluster 5. The subcluster, which contains all sequences of hH3, is interspersed with a subcluster which consists of the sequences of the mouse subgroup H7. This does not necessarily mean that the sequences of the subgroup hH3 are very homologous to the sequences of the subgroup mH7. The sequences of mH7 and the sequences of hH3 are divided by relatively high boundaries, if we consider boundaries inside a cluster. To visit a subcluster twice, leads to a shorter TSP tour in this case.
Figure 12 shows the mean distances of the TSP tours for each locus and for each species. If sequences are very homologous, the minimum average distance from one gene to the next gene should be small. More heterogenous sets are likely to have a higher average distance. The TSP tour length divided by the number of the compared sequences gives a general idea about the homology. In fact, this is further evidence for our suggestion that



∗ the cluster number cannot be shown due to close distance on the graphical display

Figure 11: Comparison of all V genes of humans and mice. The TSP tour was computed by the Concorde TSP solver which uses the NEOS Server. The threshold for the clusters is set to 25%. The major clusters of the heavy chain, of the lambda chain and of the kappa chain genes are labeled. The dotted lines refer to the human genes, the full lines refer to the mouse genes. Additionally, the subclusters, which contain the genes of the human subgroup H3 (hH3) and the mouse subgroup H7 (mH7), are marked.

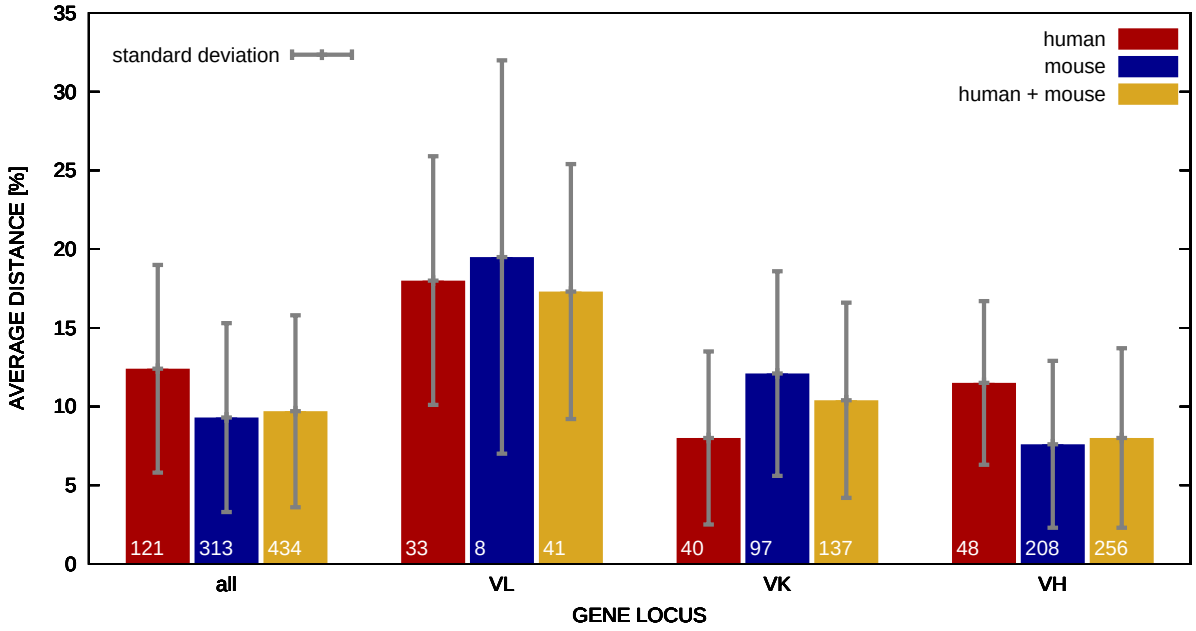


Figure 12: Average gene distance in a clustered TSP tour. The errorbars indicate the standard deviation of the mean distance. The number of sequences, which were used for the TSP instance, is labeled with a white number.

the sequences of the lambda locus are much more heterogenous than the sequences of the two other loci. The human kappa V genes are probably more homologous than the mouse kappa or the human heavy chain V genes. The mouse heavy chain genes appear to be more heterogen than the human heavy or than the mouse kappa V genes. Again, the computed data suggests that sequences of the heavy chain locus of each humans and mice are homologous.

Table 2 shows that the TSP tour length of the combined set of the human and the mouse genes is almost equal to the sum of the lengths of the separate TSP tours of each species. This suggests high a heterogeneity. If the human and the mouse V genes were very homologous, the TSP tour length of the combined set would only differ slightly from the length of the larger TSP tour of the separate sets. If the genes were identical, the TSP tour would not increase. The human and the mouse subcluster have long distances to another. Although the genes of the two species appear to be heterogenous all genes of a locus of both humans and mice belong to the same major cluster. This suggests at least little homologies between the two species.

TSP tour length

locus	human	mouse	human & mouse	human + mouse
all	15.0	29.3	42.2	44.3
VL	5.9	1.6	7.1	7.5
VK	5.9	11.8	14.3	15.0
VH	5.5	15.8	20.6	21.3

Number of V genes

all	121	313	434	
VL	33	8	41	
VK	40	97	137	
VH	48	208	256	

average TSP tour distance

all	12.4 (±6.6)	9.3 (±6.0)	9.7 (±6.1)	
VL	18.0 (±7.9)	19.5 (±12.5)	17.3 (±8.1)	
VK	8.0 (±5.5)	12.1 (±6.5)	10.4 (±6.2)	
VH	11.5 (±5.2)	7.6 (±5.3)	8.0 (±5.7)	

Table 2: TSP tour lengths and number of V genes. This table shows the length of the TSP tour, the number of V gene sequences and the average distance inside a TSP tour of each loci for each species. 'human & mouse' refers to the human and the mouse V genes combined in one set. 'human + mouse' refers to the sum of the set which contains only the human V genes and the set which contains only the mouse V genes.

3.7 Calculation of the V gene based combinatorial diversity in humans and mice

Janeway et al. have estimated the combinatorial diversity in humans to be 1.9 million. This value is the product of the number of different light chains of both loci and the number of different heavy chains which can be achieved by V(D)J recombination (see section 2). For each locus the number of different chains is the product of the number of each gene segment type, namely V(×D)×J [Janeway et al., 2005, 6]. Since our analysis considers only the V genes, we are not able to calculate the combinatorial diversity this way. However, we can calculate the combinatorial diversity of V genes based on the principle of Janeway et al. Both light chains in humans contain 73 V genes and the heavy chain contains 48 V genes. The resulting combinatorial V gene based diversity is 3504. Both light chains of mice contain 105 V genes. 208 V genes are identified at the mouse heavy chain locus. Hence, the combinatorial V gene based diversity is 21 840. The value calculated for mice is about 6.2 times as high as the value for humans. A disadvantage of the combinatorial diversity as proposed by Janeway et al. is that the resulting antibodies could have almost identical sequences. We suggest a new notion at least for the combinatorial diversity of V genes. This new notion describes the number and the homology of V genes. It is based on the TSP tour length of a clustered set of genes.

The TSP tour length shows a relation between the number of genes and the homology of those genes. As general rule many sequences, which are very heterogenous, lead to a long TSP tour. A few sequences, which share high similarities, lead to a short TSP tour. For this reason, the TSP tour length could be a useful factor to describe the diversity of genes. We have computed these values for the loci of humans and the loci of mice (table 2). The TSP tours of the mouse kappa locus and the heavy chain locus are relatively long. In contrast to these TSP tours, the tour of the mouse lambda chain is very short. The three human loci have also short TSP tours with respect to the mouse kappa and to the mouse heavy chain locus. We should remember that the mouse loci have more V genes than the human loci. The lambda locus of mice is an exceptional case. However, as mentioned in section 2, it is possible that the wildtype mice carry more lambda V genes. But, it is not the number of genes alone that matters. Many additional genes, which are very homologous to the other genes, lead to little increase in the TSP tour length. In other words, the TSP tour length is insensitive to duplicated sequences which show very few mutations. An approach

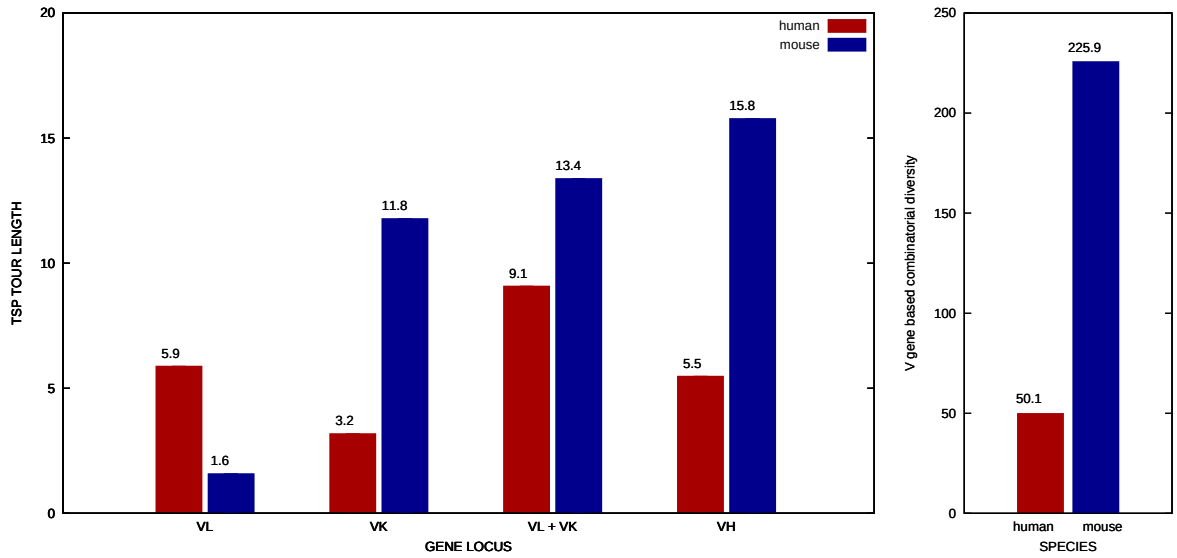


Figure 13: TSP tour lengths and V gene based combinatorial diversity. The length of the TSP tour for each locus and the V gene based combinatorial diversity of each species was computed. The black numbers indicate the value of the TSP tour length and the V gene based combinatorial diversity.

for the V gene based combinatorial diversity of humans or mice is product of the sum of the TSP tour lengths of both light chains and the TSP tour length of the heavy chain. Thus, the V gene based combinatorial diversity can be measured as follows:

$$\text{diversity} = \text{light chain diversity} \times \text{heavy chain diversity}$$
$$\Downarrow$$
$$\text{diversity} = (\text{TSP tour length VL} + \text{TSP tour length VK}) \times \text{TSP tour length VH}$$

Figure 13 shows the TSP tour length of each loci, the sum of the TSP tour lengths of both light chains and the V gene based combinatorial diversity of each species. The V gene based combinatorial diversity of mice is 4.5 times as high as the one of humans. We have calculated a ratio of 6.2 based on the principle of Janeway et al. In other words, the ratio is about 38% higher, if we do not take sequence homologies into account.

4 Pseudogene based analysis of gene conversion in chickens

We have analyzed the combinatorial diversity of V genes in humans and mice in section 3. It is less obvious to define a similar factor for the chickens. Chickens use gene conversion instead of V(D)J recombination to achieve primary diversity. During gene conversion the functional V gene is overwritten with fragments from pseudogenes. There is evidence that gene conversion preferentially involves homologous sequences [Arakawa and Buerstedde, 2004]. For this reason, we use TSP clustering (which we have validated in section 3) to determine the homologies between the pseudogenes and the functional V gene. Since most pseudogenes of the chicken heavy chain have not been located yet, we focus on the light chain. In order to analyze the characteristic features of gene conversion, we introduce a new algorithm. The algorithm identifies the pseudogenes which could have contributed to the gene converted sequence. Our investigations yield a preliminary estimation of the average number and length of gene conversion events per sequence. These results make it possible to compare the primary diversity mechanisms of humans, mice and chickens. We discuss this in section 5.

4.1 Database sources for pseudogenes, the functional V gene and ESTs

We obtained the sequences of the 25 pseudogenes and the single functional V gene of the chicken light from NCBI Genbank (gb: AH002536). In our analysis we annotate pseudogenes with 'ψ' and the according gene number. The functional V gene is referred to as 'L1'. Two expressed sequence tag (EST) libraries of chickens at bursal stage were downloaded from NCBI UniGene [UniGene/home]. NCBI UniGene is a transcriptome database. Transcribed sequences, which appear to originate from the same locus, are combined as a data set entry in UniGene. The two selected libraries 'riken1' (gb: AJ447823) and 'dkfz426' (gb: AJ394299) both contain ESTs of two to three weeks old lymphocytes which were obtained from the bursa of fabricius of chickens of the inbred CB strain. An EST is a subsequence of a mRNA sequence. The transcribed sequence is amplified with special primers and copied from RNA to DNA (cDNA). DT40subNB is a third bursal EST library for chickens which contains sequences of lymphoma cells. In other words, DT40 is a cancer cell line. DT40subNB is not considered in our investigations because cancer cells often show abnormal properties. Thus, results from

the analysis of this library might not be valid. ESTs are annotated with the associated clone number. As mentioned in section 2, bursal stage ESTs are most appropriate for the study of gene conversion. In the bursal stage gene conversion takes place at high rate and somatic hypermutation at low rate [Arakawa and Buerstedde, 2004]. For this reason, bursal sequences should mainly be affected by gene conversion. The use of ESTs should be valid because the V gene is expressed as a whole exon. There should be no change in the exon sequence after transcription. Thus, the bursal ESTs should not differ significantly from the genomic V gene sequence after gene conversion has occurred. All sequences, which we use in our analysis, were downloaded in July 2008.

4.2 Algorithm for the determination of pseudogene fragments within gene converted sequences

The following algorithm determines a set of pseudogene fragments which have most probably participated in the gene conversion process. Firstly, the original location of the V gene is determined. The position of the V gene shows the range where gene conversion should have occurred. The functional V gene has a length of 277 nucleotides. Thus, the important range for the investigation is about 280 nucleotides long. The range is determined by the best local alignment between L1 and the EST using the Smith-Waterman algorithm. The scores for nucleotide match, nucleotide mismatch, gap opening and gap extension were set to +5, -4, -12 and -4, respectively. Secondly, the largest pseudogene fragments are determined for each base of the EST sequence (figure 14). In other words, a substring starting from each base of the EST is extended while this substring matches a partial sequence of at least one pseudogene or L1. It is ensured that no fragment is a substring of another already determined fragment. A threshold is applied for the minimum length of a fragment. Figure 15 illustrates how our algorithm works.

It is possible that more than one pseudogene is assigned to a fragment. If this happens, all pseudogenes are selected. If one or more pseudogenes and L1 are assigned to the same fragment, only the reference sequence L1 is selected. A L1 fragment indicates that there is no change in the germline sequence. However, there are pseudogenes which are partially identical to the reference sequence. Since gene conversion might happen with pseudogenes which are in a different orientation with respect to L1, the reciprocal sequence of each pseudogene is considered, too [McCormack et al., 1993].

```
1  declare end_local
2
3  for i = first base to last base
4
5          end_local = max(i + threshold_length, end_local)
6          fragment = base i to end_local
7
8          while fragment is substring of pseudogene or of L1
9                  if fragment found
10                         end_local++
11         end while
12
13         if fragment found
14                 save fragment
15 end for
```

Figure 14: Pseudocode for the algorithm to determine the most probable fragments which overwrote parts of the V gene sequence.

4.3 Similarities between pseudogenes and the functional V gene

Other investigations suggest that gene conversion preferentially occurs with pseudogenes, which are homologous to the functional V gene [Arakawa and Buerstedde, 2004]. For this reason, we analyze homologies between the functional V gene and the pseudogenes. We compare the pseudogenes and the functional V gene in one set. In addition, we compare the pseudogenes without including L1. For this analysis we use TSP clustering (see section 3). Each sequence is compared as its actual sequence and as its reciprocal sequence. The threshold for the clusters is set to 25%. The chicken light chain pseudogenes and the functional V gene can be divided into eight clusters (figure 16). Two major clusters with boundaries above 50% can be identified. The first major cluster consists of cluster 1 to cluster 3. Cluster 4 to cluster 8 belong to the second major cluster. L1 is part of cluster 5. If the pseudogenes are clustered without including L1, a few pseudogenes change their cluster (figure 17). This cluster movement is an interesting observation. All pseudogenes, which participate in this event, are members of cluster 5 and cluster 7. This movement indicates a high similarity between the sequences of both clusters. The pseudogenes ψ4, ψ8, ψ12 and ψ17 on one side change their position around ψ22 with ψ19 on the other side. ψ22 is the only sequence which belongs to cluster 1. The pseudogenes of cluster 5 and 7 probably belong to the same cluster. This cluster

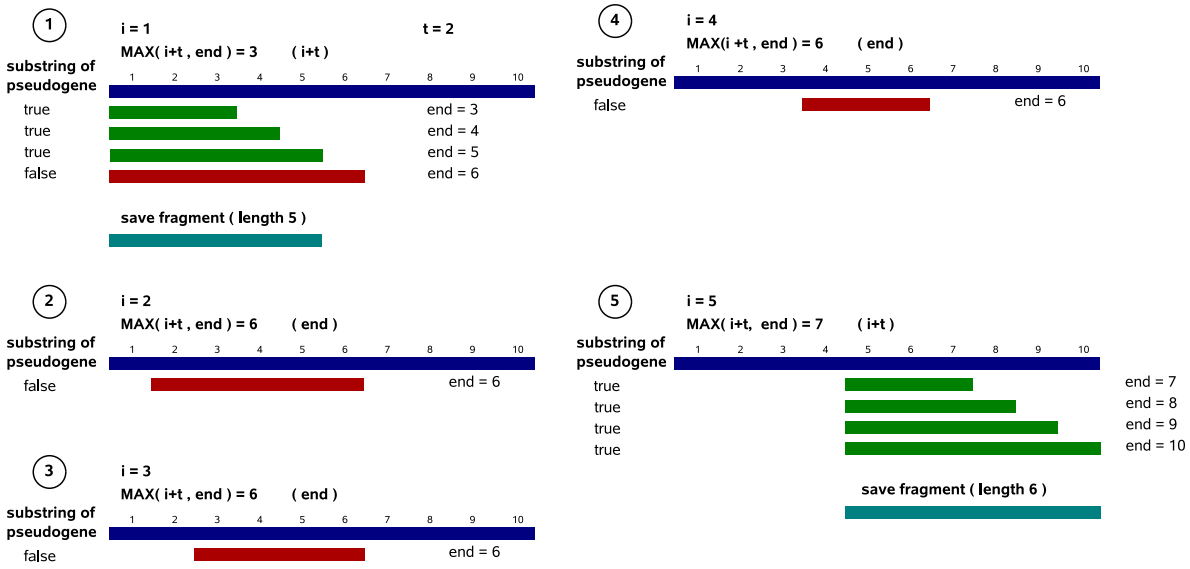


Figure 15: Illustration of the algorithm to determine pseudogene fragments. The fragment sequence is extended, if it is a substring of at least one pseudogene. The two resulting fragments of this example are overlapping at the fifth base on the EST. The blue line refers to the EST. A red, green or bluegreen line refers to the fragment, depending on if it is extended, not extended, or saved. 't' is the threshold for the length. 'i' is the current starting base. 'end' is the end base of the substring on the EST.

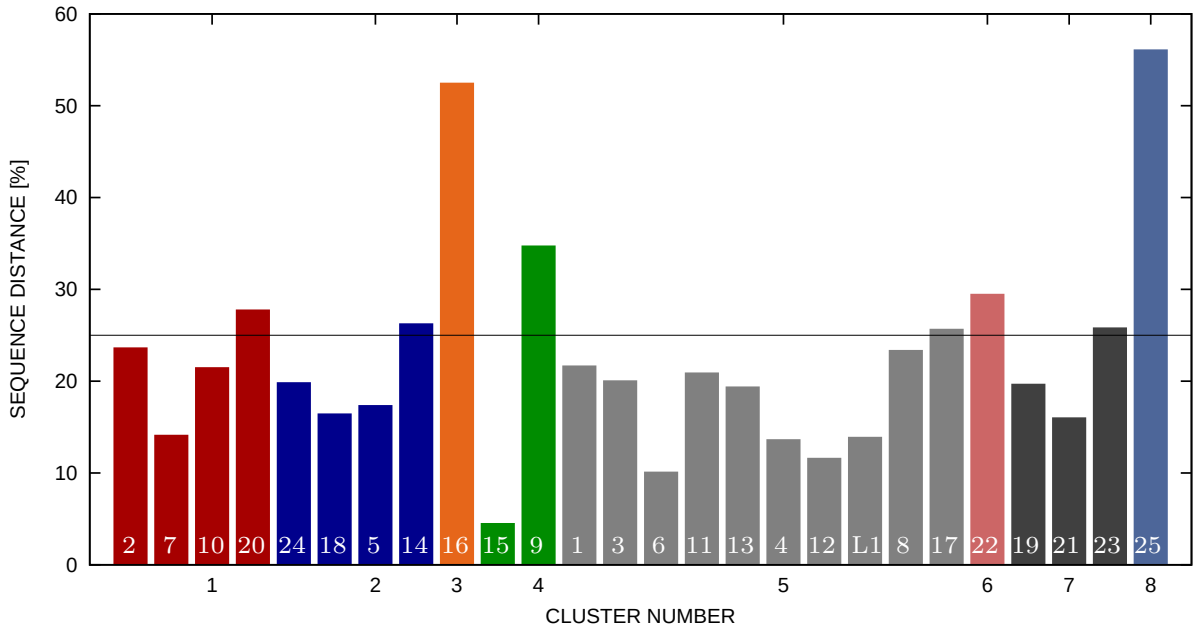


Figure 16: Comparison of pseudogenes and L1. The pseudogenes and L1 are clustered using TSP clustering. The pseudogenes and L1 are labeled with white letters (e.g. 1 refers to ψ1).

is interspersed with ψ22. ψ22 seems to be most similar to pseudogenes of cluster 5 and cluster 7. In order to obtain a shorter TSP tour, the sequences are separated in cluster 5 and cluster 7. The movement of the sequences between the two clusters suggests that L1 is most similar to the pseudogenes of cluster 5 and of cluster 7. In other words, ψ1, ψ3, ψ4, ψ6, ψ8, ψ11, ψ12, ψ13, ψ17, ψ19, ψ21 and ψ23 are the pseudogenes which are most homologous to L1.

4.4 Determination and interpretation of pseudogene fragments within antibody sequences

We have implemented a program, which uses the algorithm discussed above, to create LaTeX Tikz pictures. The fragments, which are inside the range of investigation, or which overlap inside this range, are considered in the output. All clones from the riken1 library showed relevant results, namely clone `29i11r1, 23b5r1, 2a17r1 and 17m18r1`. Additionally, eight clones from the dkfz426 library were useful for the analysis, namely `13c5r1, 30f11r1, 5g3r1, 15f1r1, 1c4r1, 25o13r1, 30n12r1 and 8c9r1`. The analyzed ESTs have about 500 to 600 nucleotides. The pseudogenes are supposed to be

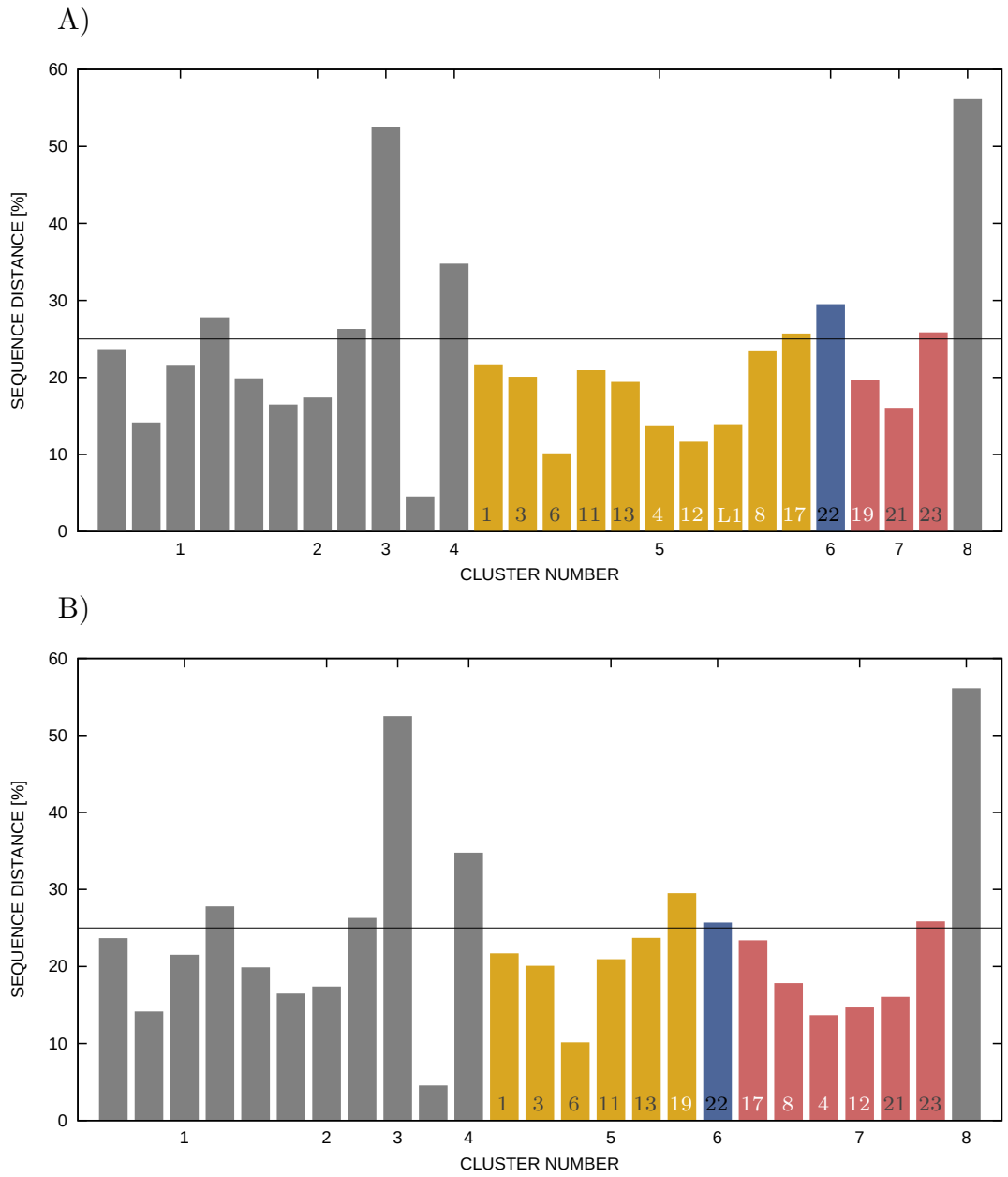


Figure 17: Comparison of pseudogenes with and without including L1. The pseudogenes of the chicken light chain are clustered including L1 (A) and without including L1 (B) using TSP clustering. The pseudogenes, which move between clusters, and L1 are labeled with white letters(e.g. 1 refers to ψ1). Other pseudogenes of the clusters of interest are labeled with black or gray numbers.

transposable elements which have evolved from the same functional genes due to duplication [Janeway et al., 2005, 6]. Consequently, their sequences are similar. Fragments with a length of less than ten nucleotides can be assigned to almost every pseudogene. Thus, we set the threshold length to 10. The regions, which cannot be assigned, are colored gray and annotated unknown. Pseudogenes appear to overwrite the sequence of the functional V gene, but also of the other pseudogene fragments. Thus, even a threshold length of 10 makes it difficult to interpret the created figures with certainty. For this reason, it is helpful to use a higher threshold to identify major components. Afterwards a lower threshold helps to assign additional fragments. Figure 18 shows the gene conversion plots of clone `5g3r1` from the dkfz426 database on threshold 20 and threshold 10. The pseudogene sequences are scaled 1:2 with respect to the partial clone sequence. The range of the best local alignment between L1 and clone `5g3r1` is shown at the center of each plot.

The major contributions to the gene converted sequence are shown in image A of the figure 18. ψ8, ψ4 and ψ23 are assigned to the largest determined fragments. Thus, they probably contribute to the sequence. Some parts of the sequence cannot be assigned to pseudogene fragments, if we use a threshold length of 20 nucleotides. Image B reveals additional fragments which help to assign the unknown parts of the sequence. One should notice that the distance of all fragments of ψ4 is equal on the pseudogene and on the clone sequence. In other words, they are in the same frame on both sequences. This suggests that ψ4 has pasted all fragments in a single event. The sequence of this whole fragment is interspersed with other fragments. The gap between the 77 and the 46 bases long fragment has probably been caused by a gene conversion event which involves ψ8. It is important to note that the fragments of ψ4 and ψ8 have long overlapping regions. These pseudogenes both belong to the same cluster (figure 16). The distance between these pseudogenes is 17.9%. Thus, they are homologous to each other. It can be assumed that ψ4 overwrote 168 bases of the original V gene in a single event. For similar reasons, it is probable that ψ13 pasted both fragments in a single event. ψ19 can be excluded because the range of this fragment can also be assigned to fragments of ψ4 and ψ17. These fragments are longer than the fragment of ψ19. If a fragment is longer, the probability is lower that it can be found on a pseudogene by chance. The 12 bases long fragment can be assigned to either ψ17 or ψ22. Since some pseudogenes share high similarities, identical parts of the sequences can frequently be observed.

The clone sequence `8c9r1` originated from five paste events. One of those events includes a reciprocal sequence of a pseudogene. The fragment of ψ10 can be excluded because

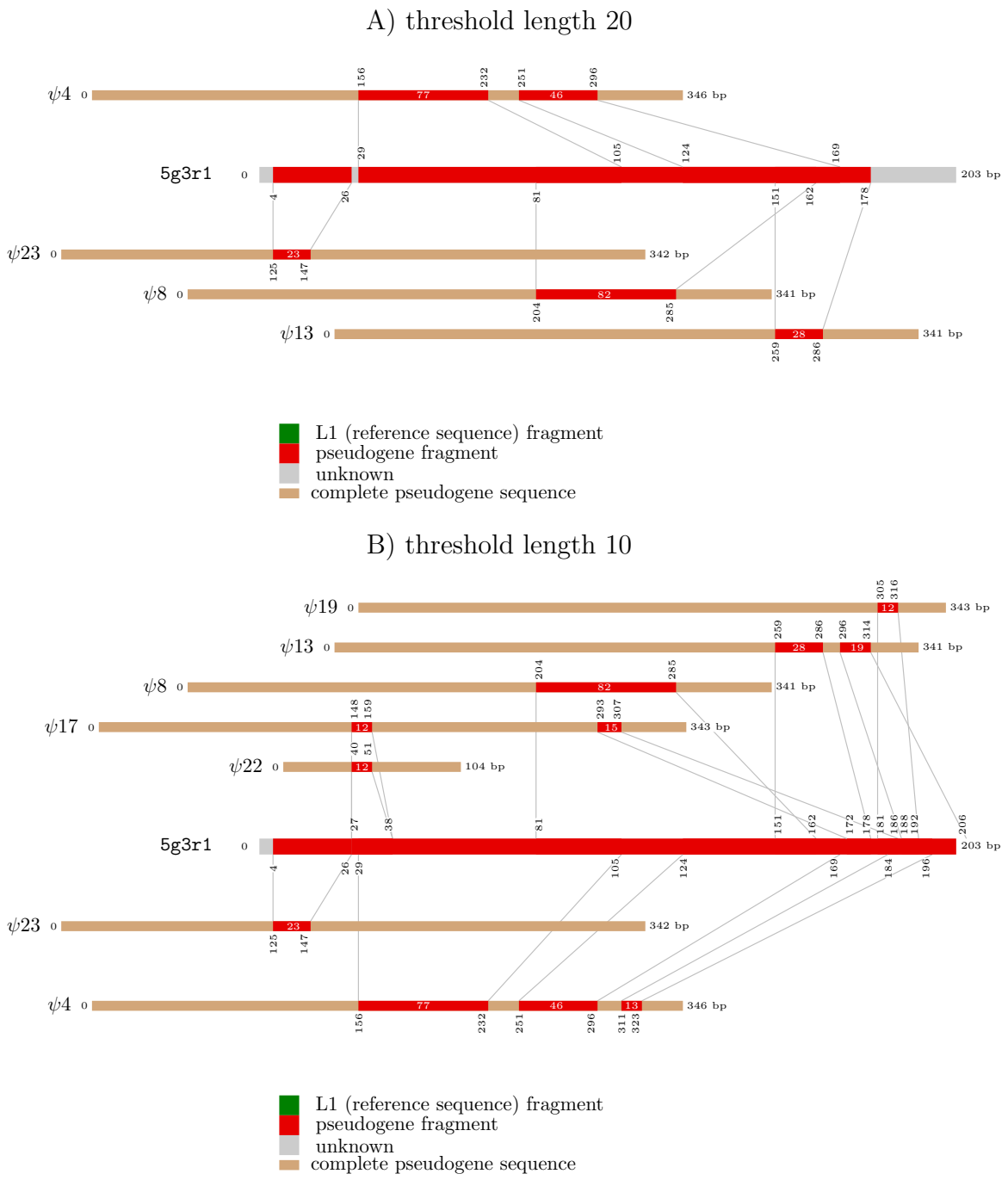


Figure 18: Gene conversion - clone `5g3r1` shown on two different thresholds. White numbers indicate the length of the pseudogene fragments. The range of pseudogene fragments is shown in red on the partial sequence of the clone and on the whole pseudogene sequence. Black numbers refer to start and end positions.

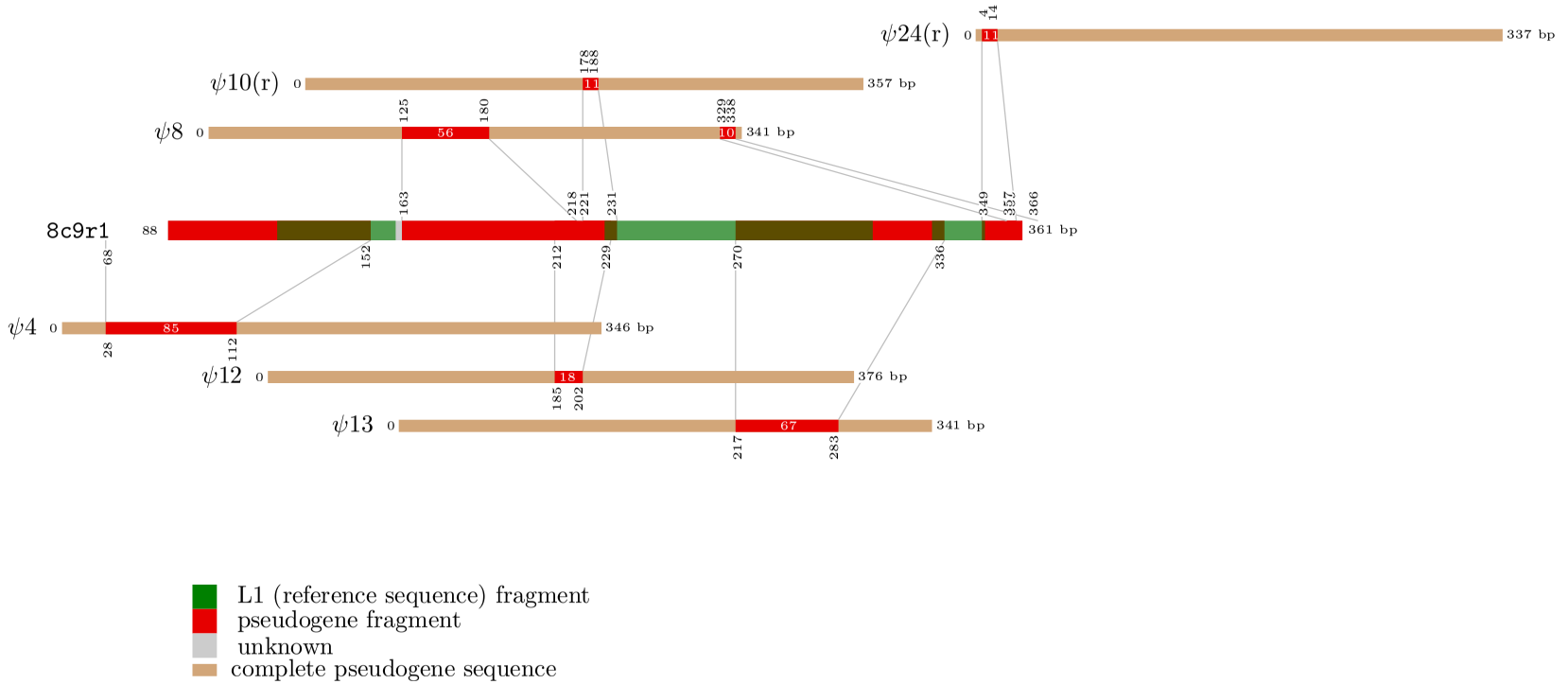


Figure 19: Gene conversion - clone `8c9r1` shown on threshold 10. White numbers indicate the length of the pseudogene fragments. The range of pseudogene fragments is shown in red on the partial sequence of the clone and on the whole pseudogene sequence. Black numbers refer to start and end positions. '(r)' indicates that the reciprocal sequence of the pseudogene is matching the fragment.

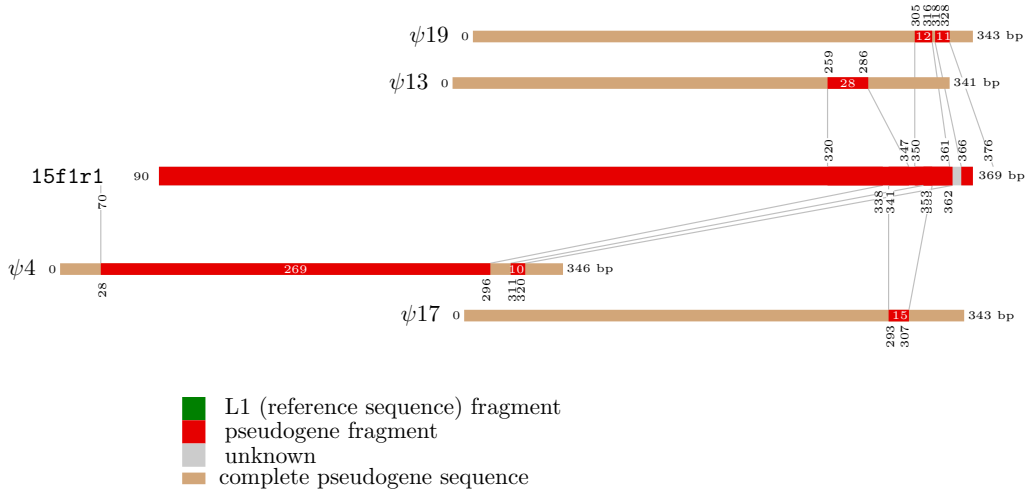


Figure 20: Gene conversion - clone `15f1r1` shown on threshold 10. White numbers indicate the length of pseudogene fragments. The range of pseudogene fragments is shown in red at the partial sequence of the clone and at the whole pseudogene sequence. Black numbers indicate start and end positions.

this part of the sequence can be assigned to the reference sequence L1 and a fragment of ψ12. The reciprocal sequence of ψ24 probably overwrote 11 nucleotides of the original sequence. This is likely because ψ24 has a different orientation with respect to L1 [McCormack et al., 1993]. We can also observe that the sequence of the original gene is not always completely overwritten.

Some sequences show much larger paste events. Clone `15f1r1` from the dkfz426 database shows a huge paste event (figure 20). ψ4 seems to have overwritten 269 bases in a single event. Since the functional V gene has length of 277 nucleotides, it has been almost completely overwritten. A three bases long fragment cannot be assigned. Unknown regions of a similar length can be found in other clones, too (see appendix figure 31 - figure 39). These regions suggest, that gene conversion could involve fragments which are smaller than ten base pairs.

The plots of the other clones and the interpretation of the fragments are listed in the appendix (figure 31 - figure 39 and table 19 - table 21).

It can be observed that the inserted fragments mostly originate from the center of the pseudogene sequences. Only a few fragments were found which have a start range on a pseudogene lower than nucleotide 27. If we consider the functional V gene, our analysis does not reveal any bias with respect to the range. The pseudogene fragments are

pseudogene	average length	events
ψ4	67.0 (±18.0)	22
ψ8	39.4 (±29.4)	15
ψ12	78.4 (±60.4)	10
ψ11	47.9 (±25.9)	9
ψ13	36.0 (±31.0)	8
ψ17	25.8 (±10.8)	6
ψ6	67.5 (±43.5)	4
ψ19	15.2 (±4.2)	4
ψ3	49.0 (±12.0)	3
ψ1	42.0 (±4.0)	3
ψ23	53.3 (±30.3)	3
ψ25	12.0 (±0.0)	2
ψ9	10.0 (±0.0)	2
ψ21	31.0 (±0.0)	1
ψ22	12.0 (±0)	1
ψ24(r)	10.0 (±0.0)	1

Table 3: Pseudogene contributions to all ESTs. In the right column the number of events, which occurred in all 12 ESTs, is listed. The average length of paste events was calculated for each pseudogene which contributes to at least one of the ESTs. '(r)' indicates that the fragments were assigned to the reciprocal sequence of the pseudogene. The pseudogenes are ordered first by the number of events, second by the average length of a gene conversion event.

found in the range of the functional V gene and beyond. The average number and the average lengths of the determined fragments are shown on table 3. The results of our analysis show that the number and length of gene conversion events depends on the pseudogene. ψ4, ψ8, ψ12, ψ13, ψ17 and ψ6 show the biggest contributions to the analyzed sequences. All of these pseudogenes belong to cluster 5 (figure 16). L1 is found in the same cluster. Only four of the 16 involved pseudogenes do not belong to cluster 5 or cluster 7, namely ψ23, ψ25, ψ9 and ψ22. As mentioned above, cluster 5 and cluster 7 contain the sequences which seem to be most homologous to L1. Indeed, these results suggest that gene conversion preferentially occurs with more homologous sequences. In general, long overlaps between the sequences of two fragments can be observed. It is probable that the presence of homologous sequences is necessary for gene conversion. In conclusion, the length of inserted fragments varies from about ten to over 200 nucleotides in the analyzed sequences. The results for clone `15f1r1` suggest that it possible to

overwrite the whole functional V gene in a single event. On average 48.3 ($\pm$42.2) bases of the functional V gene are overwritten per gene conversion event. At bursal stage on average 7.5 ($\pm$1.6) events can be observed during the gene conversion process of a light chain sequence. Homologous sequences seem to be preferentially involved in gene conversion. However, to obtain statistically relevant results, more sequences should be analyzed in future research.

5 Discussion

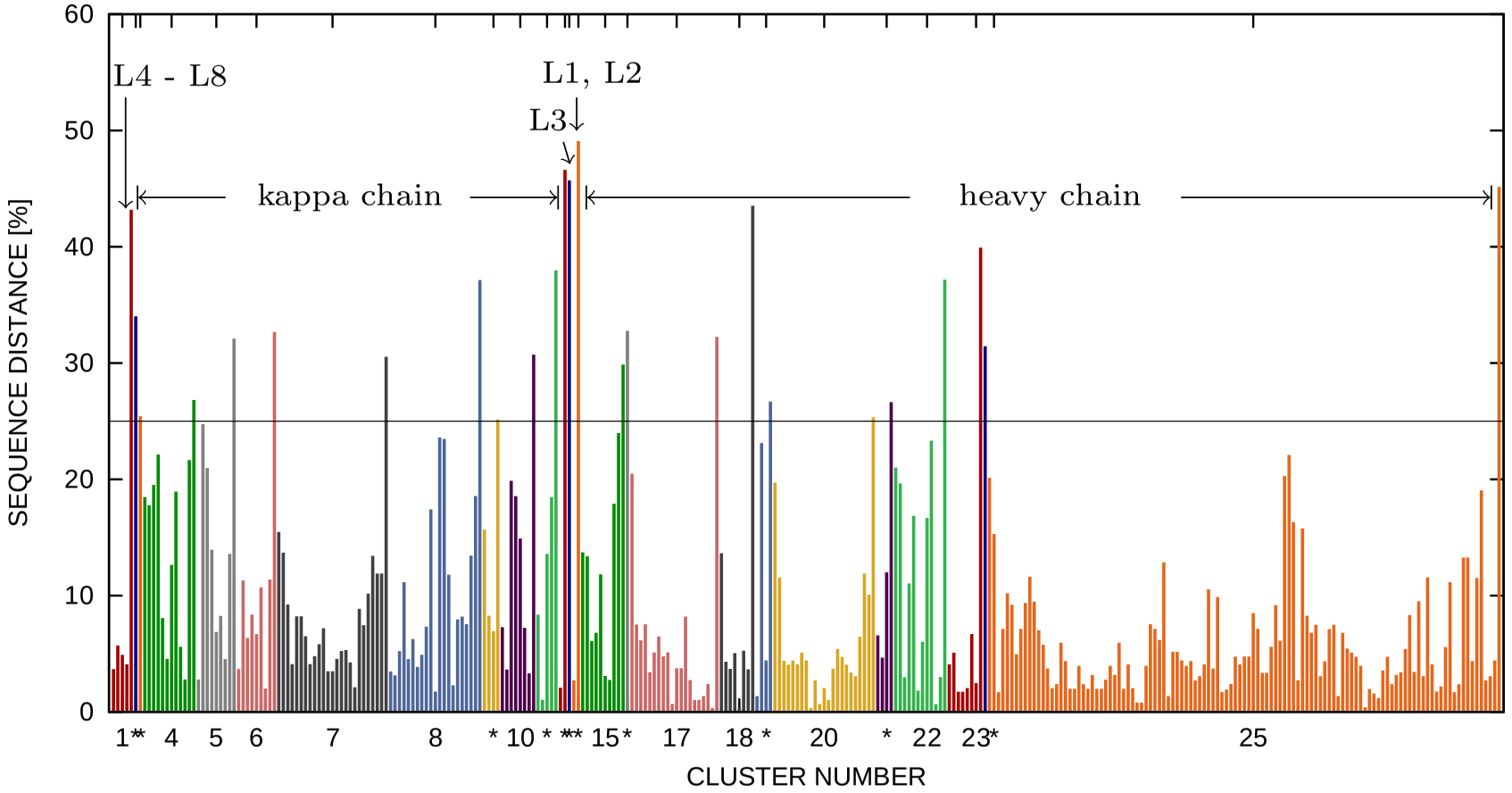
In this section we evaluate the results which we have described in the last two sections. We start with a discussion of the advantages of TSP clustering compared to other clustering methods. Next, we turn our attention to the results of our gene conversion algorithm. With the help of these conclusions, we compare the primary diversity mechanisms of humans, mice and chickens. Finally, we evaluate the implications of our results to computational modeling of the immune system and indicate possibilities for future research.

5.1 Advantages of TSP clustering

TSP clustering provides good estimations for the heterogeneity of all sets of the genes that we have compared. There are only a few V genes which are less suitable for TSP clustering, namely the mouse lambda genes. The mouse lambda V genes are very few sequences. Our analysis in section 3.6 reveals that the genes of the lambda loci are most heterogenous. If we compare all mouse V genes, the lambda genes belong to two different major clusters (figure 21). L4 to L8 belong to the first major cluster. L1 to L3 form the second major cluster. L1 and L2 are members of the same class. The second class contains only L3. The last class consits of L4 to L8. These classes can be distinguished between the sequences of the other loci. As mentioned in section 2, the inbred labortory mouse strains lack several functional V lambda genes. For this reason, the mouse lambda genes should be treated as a special case. However, we can suggest that L1, L2 and L3 are more homologous to each other than to the genes which build the second lambda gene cluster. If all human and mouse genes are clustered in one set, all genes are found in clusters which are according to their loci.

However, TSP clustering has advantages compared to other clustering algorithms. In contrast to k-means or QT clustering, TSP clustering does not use cluster centers and needs no metric. For this reason, TSP clustering can directly be applied to gene comparisons. Since TSP clustering uses no cluster centers, it should provide a better relationship between all genes of a set than k-means.

We have suggested that the TSP tour length could be used to determine a new notion for the combinatorial diversity, since it provides information about the number and the homology of the compared sequences (section 3.7). The minimum spanning tree could provide similar information. However, TSP clustering should be more suitable for this approach because every gene is treated equally. Since the number of edges of each node



∗ the cluster number can not be shown due to close distance on the graphical display

Figure 21: Comparison between all mouse V genes. The TSP instance was solved by the Concorde TSP solver which uses the NEOS Server. The threshold for the clusters is set to 25 % distance. The ranges of the heavy and the kappa chain cluster are marked. In addition, the clusters, which contain lambda genes, are labeled.

inside a minimum spanning tree might vary, it does not provide a good relationship between all genes. One gene might be connected to many others. Other genes might have a single edge. Consequently, some genes have bigger contributions than others. Every gene should contribute in the same way, if we want to determine a good relationship between the genes. If TSP clustering is applied to the genes, every gene has exactly two edges. Thus, every gene contributes in the same way.
For similar reasons, TSP clustering could be a more suitable approach to determine the diversity of genes than hierarchic clustering. The agglomerative hierarchic clustering shows a good relationships between clusters and needs no vector space. However, the relationship between all genes is less obvious than in a clustered TSP tour of genes. The clusters can be connected at different hierarchic levels. If average linking clustering is used, the mean distance the genes of a cluster is calculated to determine homologous clusters. TSP clustering uses no mean distance but always exact values. If the minimum or the maximum distance of the elements of each cluster is used to determine homologous clusters (single linkage or complete linkage clustering), the genes of a cluster do not contribute in the same way. Thus, TSP clustering should reveal a better relationship between all compared genes.
Another possibility to determine the diversity of the genes is to calculate the product of the mean distance between all genes and the number of genes. However, the mean distance is likely to have a big deviation. The TSP tour length includes no deviation. For this reason, the TSP tour length should provide better information about the diversity of the compared genes.

5.2 Interpretation of gene converted sequences

Gene conversion is a highly complex process. The analysis of a gene converted sequence yields not always all pseudogenes which participated in the process. Our analysis of the twelve ESTs suggests that gene conversion can occur with fragments that are smaller than ten nucleotides. The fragments of this size can be assigned to many different pseudogenes. In addition, we have observed that some pseudogenes are identical in their partial sequence. For this reason, the origin of some fragments cannot be identified unequivocally. If many gene conversion events have occurred in the same range, it is difficult, in some cases impossible, to interpret this region correctly. The algorithm, we have discussed in section 4, leads to very useful results. Almost all regions of the analyzed sequences have been identified as pseudogene fragments or germline sequence. Only a few small fragments could not be assigned to neither pseudogene nor reference

sequence in some clones. Besides a small fragment size, somatic hypermutation or an error prone overwrite process are possible explanations for these fragments. However, inside and between larger fragments is no evidence that either somatic hypermutation has occurred, or that the paste process is error prone. It is also possible that the original V gene sequence of the clones differs from the reference sequence. However, the most probable explanation seems to be that fragments can be smaller than ten nucleotides. Other research suggests that fragments with a length of eight bases can be observed [Arakawa and Buerstedde, 2004].
We are not able to apply meaningful investigations to the chicken heavy chain. Furthermore, only twelve sequences of the selected EST libraries are suitable for the analysis of the chicken light chain. Thus, the resulting data is statistically not relevant. It is not possible to calculate a useful value for the resulting V gene and pseudogene based diversity in chickens. Nevertheless, we were able to calculate preliminary estimations for the average number and and average length of paste events per sequence in the chicken light chain.

5.3 Comparability of diversity mechanisms between humans, mice and chickens

Humans, mice and chickens differ in the primary diversity mechanisms. Humans and mice both use V(D)J-recombination. In contrast to humans and mice, chickens use gene conversion to achieve the primary antibody diversity. All three of them combine light chains and heavy chains to achieve diversity. Humans and mice each have two gene loci for the light chains. Chickens have a single light chain locus. All of them have a single heavy chain locus. The primary diversity in humans and mice is created by the combination of several functional V genes. Unlike humans and mice, chickens just have a single functional V gene in each of the two loci. Chickens use gene conversion which involves a pool of pseudogenes to achieve the primary diversity.
Although humans and mice use the same mechanism mice have much more functional V genes. 75 of the mouse heavy chain V genes are provisional genes. Even without the provisional genes, the mouse heavy chain carries 133 V genes. In other words, about three times as much genes as the human heavy chain. The lambda chain of both human and mice contains the most heterogenous V genes. The mouse lambda chain leads to little contribution to the combinatorial diversity. The lack of several lambda V genes in labortory mice could be the reason for this. In contrast to mice, the lambda and

the heavy chain locus have the biggest contributions to the V gene based combinatorial diversity in humans. However, mice achieve most of their combinatorial diversity, if they combine kappa chains with heavy chains.
Mice have much more V genes than humans and achieve a higher combinatorial diversity of V genes than humans. Probably, not all V genes are used at the same frequency. Additionally, there are some combinations of heavy and light chains which are not stable [Janeway et al., 2005, 6]. Thus, they do not contribute to the antibody diversity in a natural environment. These facts must be taken into account for both humans and mice. Nevertheless, there must be reasons why mice have much more V genes than humans. In the evolution it is usual unfavorable to carry more genes than it is necessary. Our investigations suggest that the human and the mouse V genes are heterogenous. This suggests a more independent evolution of the V genes of both species than one might have expected. Thus, an independent evolution is a possible explanation. Another explanation is that humans use other diversity mechanisms more efficient than mice to compensate for the missing V genes. It is possible that the human evolution was not under the same selectional pressure as the evolution of mice. An aspect might be that humans always had access to better sanitation. Thus, humans might need less potential for antibody diversity. However humans, mice and domicile chickens share a similar habitat. Thus, the selectional pressure should be relatively similar for all three them.
Chickens use a different way to achieve diversity. The results of our analysis of twelve ESTs suggest that 16 of the 25 pseudogenes are involved in the gene conversion process of the chicken light chain. We have computed that fragments with an average size of about 50 nucleotides are copied more than seven times over the functional gene during the gene conversion process of a sequence. Since the functional V gene has 277 bases, it can be completely substituted by pseudogene fragments. Furthermore, our investigations reveal that large fragments are able to overwrite the whole V gene sequence in a single event. In addition, pseudogenes appear to overwrite the pseudogene fragments which have already been copied over the functional V gene. Consequently, even if less sequences are involved, gene conversion could be able to create a diversity potential which is higher than the one achieved by V(D)J recombination. However, humans and mice do not use gene conversion. Humans and mice are able to create a good primary diversity potential with V(D)J recombination which has been evolutionary established. Additional processes such as junctional diversity play an important role for the antibody diversity of humans and mice. These process are sufficient to recognize diverse antigens. There are other mammals such as rabbits, which have several functional V genes, and use both

mechanisms gene conversion and V(D)J recombination [Janeway et al., 2005, 6]. In addition, somatic hypermutation appears to play a crucial role for the antibody specificity in humans and mice. After a B cell is induced by antigens somatic hypermutations can lead to a 1000-fold increase in antigen affinity in humans [Berg et al., 2007, 6].

5.4 Implications to computational modeling

It is possible that a species with less genes is able to achieve a similar antibody diversity as a species with many V genes. Consequently, the size of the gene repertoire alone is not a sufficient feature for models of the antibody evolution. Sequence homologies must be taken into account, too. We have suggested that the TSP tour length could be used to compute the combinatorial diversity of V genes (section 3.7). The combinatorial diversity as proposed by Janeway et al. is not considering sequence homologies. Thus, this value is only representing the number of different antibodies but not their specificity. Our new notion of the combinatorial diversity of V genes takes sequence homologies into account. Thus, it can be used for an estimation of the primary diversity achieved by V(D)J recombination in models for the evolution of antibodies.

We have computed the values of the V gene based combinatorial diversity in humans and mice. Mice are able to achieve a V gene based combinatorial diversity which is more than four times as high as the one of humans. However, both species have survived in a similar environment. For this reason, additional features such as junctional diversity are important for the computational modeling of antibody evolution.

Our investigations concerning chickens reveal that a similar or even higher diversity can be achieved with a single functional gene, if gene conversion is used. For this reason, gene conversion should be a fundamental feature of models for the evolution of antibodies. We have calculated preliminary estimations for the average number and the average lengths of paste events per sequence. In order to obtain statistically relevant results, more sequences should be analyzed.

Rabbits and several other mammals achieve the primary diversity by both V(D)J recombination and gene conversion. Investigations of these species could reveal interesting properties of the evolutionary relationship of both mechanisms.

The process of somatic hypermutation is not important for the initial antigen recognition in humans and mice. However, somatic hypermutation is important to increase the the antigen affinity after an antigen has been recognized. Thus, somatic hypermutation might be essential to ensure the survival of an organism.

Further investigations of other species such as rabbits must be applied to future research. In addition to the features that we have mentioned in this section, the characteristic properties of junctional diversity and also of somatic hypermutation must be determined and integrated in the models for the antibody evolution.

6 Summary

We have used a new approach for gene clustering to analyze homologies between the V genes of humans and mice: TSP clustering. Since TSP clustering yields valid results for the human lambda genes, we have applied it to all V genes of both humans and mice. Although humans and mice have a close evolutionary relationship the results of our analysis reveal that the V genes of both species appear to be heterogenous. In addition, the kappa genes and the heavy chain genes of each species are more homologous than the lambda genes. We have discussed the advantages of TSP clustering compared to the k-means algorithm and to the hierarchic clustering. In addition, we have demonstrated that the TSP tour length could be used for a new notion of the combinatorial diversity of V genes which might be more suitable than the one proposed by Janeway et al.

For our investigations of the gene conversion process in chickens, we have proposed a new algorithm that identifies the involved pseudogenes. We have estimated preliminary values for the average length and the average number of paste events which have occurred during the gene conversion process in a sequence. Our results suggest that about 50 nucleotides are overwritten by each gene conversion event. On average 7.5 paste events per sequence are observed. We present evidence that gene conversion preferentially involves pseudogenes which are more homologous to the functional V gene. 16 of the 25 pseudogenes of the chicken light chain contribute to the analyzed sequences.

The gene conversion process appears to be able to achieve a higher primary diversity potential with less sequences than the V(D)J recombination. Humans and mice use additional mechanisms like junctional diversity to compensate for this.

Our analysis suggests that the number of genes but also the homology of those genes are important for the models of the evolution of antibodies. Our notion for the combinatorial diversity of V genes describes both of these features. Additionally, the gene conversion process must be a integrated in these models. The characteristic properties of other mechanisms and of other species must be determined in future research and afterwards be applied to the evolutionary models for the development of the immune system.

7 Acknowledgments

Thanks to Nico Dehnert for proof-reading. A special thanks to Johannes Textor for his great support. It has to be mentioned that it was basically his idea to use TSP clustering for the analysis.

References

[Janeway et al., 2005, 6] C. A. Janeway Jr., P. Travers and M. Walport, M. J. Shlomchik. 2005. *Immunobiology: the immune system in health and disease.* Garland Science Publishing

[Alberts et al., 2002, 4] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter. 2002. *Molecular biology of the cell.* Garland Science, a member of the Taylor & Francis Group

[Berg et al., 2007, 6] J. M. Berg, J. L. Tymoczko and L. Stryer. 2007. *Biochemistry.* W. H. Freeman and Company

[Deonier et al., 2005, 1] R. C. Deonier, S. Tavar, and M. S. Waterman. 2005. *Computational Genome Analysis: An Introduction.* Springer

[Oprea and Forrest, 1998] M. Oprea and S. Forrest. 1998. Simulated evolution of antibody gene libraries under pathogen selection. Paper presented at IEEE International Conference: Systems, Man, and Cybernetics, October 11-14, in San Diego, CA, USA

[M. Diaz et al., 2001] M. Diaz, M. F. Flajnik and N. Klinman. 2001. Evolution and the molecular basis of somatic hypermutation of antigen receptor genes. *Phil. Trans. R. Soc. Lond.* 356: 67-72

[Arakawa and Buerstedde, 2004] H. Arakawa and J-M. Buerstedde. 2004. Immunoglobulin Gene Conversion: Insights From Bursal B Cells and the DT40 Cell Line *Development Dynamics* 229: 458-464

[McCormack et al., 1993] W. T. McCormack, E. A. Hurley and C. B. Thompson. 1993. Germ Line Maintenance of the Pseudogene Donor Pool for Somatic Immunoglobulin Gene Conversion in Chickens *Molecular and Cellular Biology* 13, No.2: 821-830

[Smith and Waterman, 1981] T. F. Smith and M. S. Waterman. 1981. Identification of Common Molecular Subsequences *J . Mol. Biol.* 147: 195-197

[Climer and Zhang, 2005] S. Climer und W. Zhang. 2005. A traveling salesmans approach to clustering gene expression data. Technical Report WUCSE-2005-5, Washington, University in St. Louis

[Lefranc et al., 2005] M.-P. Lefranc, V. Giudicelli, Q. Kaas, E. Duprat, J. Jabado-Michaloud, D. Scaviner, C. Ginestoux, O. Clment, D. Chaume and G. Lefranc. 2005. IMGT, the international ImMunoGeneTics information system *Nucl. Acids Res.* 33: D593-D597. PMID: 15608269

[Giudicelli et al., 2005] V. Giudicelli, D. Chaume and M.-P.Lefranc. 2005. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucl. Acids Res.* 33: D256-D261. PMID: 15608191 pdf

[Lefranc et al., 2003] M.-P. Lefranc, C. Pommi, M. Ruiz, V. Giudicelli, E. Foulquier, L. Truong, V. Thouvenin-Contet and G. Lefranc. 2003. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.* 27: 55-77. PMID: 12477501 pdf

[Benson et al., 2008] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, D. L. Wheeler. 2008. GenBank. *Nucleic Acids Res.* 36(Database issue): D25-30

[Czyzyk et. al, 1998] J. Czyzyk, M. Mesnier, and J. Mor. 1998. The NEOS Server. *IEEE Journal on Computational Science and Engineering* 5: 68-75

[Gropp and More, 1997] W. Gropp and J. More. 1997. Optimization Environments and the NEOS Server. *Approximation Theory and Optimization* 167-182

[Dolan, 2001] E. Dolan. 2001. The NEOS Server 4.0 Administrative Guide. Technical Memorandum ANL/MCS-TM-250, Mathematics and Computer Science Division, Argonne National Laboratory

[IMGT/home] IMGT home page.
http://imgt.cines.fr

[IMGT/GENE-DB] IMGT/GENE-DB Query page.
http://imgt.cines.fr/IMGT_GENE-DB/GENElect?livret=0

[Genbank/home] NCBI Genbank home page.
http://www.ncbi.nlm.nih.gov/Genbank/

[UniGene/home] NCBI UniGene home page.
http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene

[Concorde] NEOS Server: the Concorde TSP solver.
http://www-neos.mcs.anl.gov/neos/solvers/co:concorde/TSP.html

8 Appendix

number	gene subgroups	elements	mean distance	cluster distance
1	mH15	1	0.0 (± 0.0)	31.5
2	mH1 mH1S mH14 hH1 hH7 mH9 mH9S	134	6.4 (± 5.1)	35.1
3	hH5	2	4.8 (± 0.0)	34.5
4	mH11 mH4	4	9.6 (± 9.6)	26.7
5	mH5 mH5S hH3 mH7 mH10 mH10S mH6 mH6S mH13	62	8.4 (± 6.3)	38.2
6	hH6	1	0.0 (± 0.0)	30.4
7	mH3 mH3S mH12 hH4	20	8.9 (± 6.7)	28.4
8	mH12	1	0.0 (± 0.0)	32.8
9	mH2 mH2S	20	4.7 (± 3.8)	31.9
10	hH2 mH8S mH8	11	7.3 (± 5.9)	51.6
11	hL5	4	11.5 (± 5.1)	40.5
12	mL2 mL1	2	2.7 (± 0.0)	34.2
13	hL7	2	8.2 (± 0.0)	28.9
14	hL8	1	0.0 (± 0.0)	36.2
15	hL3	10	13.9 (± 6.1)	33.7
16	hL10	1	0.0 (± 0.0)	29.7
17	hL1 hL2	10	8.9 (± 6.6)	33.2

Table 4: Comparison between all V genes of humans and mice - 1.

number	gene subgroups	elements	mean distance	cluster distance
18	hL6	1	0.0 (± 0.0)	25.3
19	mL4 mL5 mL7 mL8 mL6	5	4.6 (± 0.8)	44.2
20	hL9	1	0.0 (± 0.0)	31.5
21	hL4	3	15.9 (± 4.5)	28.1
22	mL3	1	0.0 (± 0.0)	46.6
23	mK17	2	2.1 (± 0.0)	25.4
24	hK5	1	0.0 (± 0.0)	33.0
25	mK4	25	7.5 (± 3.6)	32.7
26	mK3	9	7.6 (± 3.3)	27.2
27	hK4 mK8	9	11.3 (± 5.8)	25.4
28	mK6 mK7	13	7.8 (± 6.4)	36.0
29	mK2 hK2 mK1	20	10.5 (± 6.1)	30.7
30	mK5	5	10.4 (± 6.5)	26.1
31	hK3 mK18	13	4.1 (± 5.5)	34.0
32	mK16	1	0.0 (± 0.0)	27.7
33	mK13	2	2.8 (± 0.0)	27.7
34	mK12 hK1 mK11 mK14 mK9 mK10 mK19	37	9.6 (± 6.2)	52.0
	overall	434	7.7 (± 5.5)	33.3 (± 6.8)
	all sequences	434	9.7 (± 6.1)	0.0 (± 0.0)

Table 5: Comparison between all V genes of humans and mice - 2. The human genes are colored gray. The mean distance from one element to the next is calcualted within a cluster. The standard deviation is shown inside parenthesis. Cluster distance means the distance from the last element of one cluster to the first element of the next cluster. The overall mean distance is the weighted mean value of the mean distances inside of all clusters containing more than one element. The overall cluster distance is the unweighted mean value of all cluster distances. All sequences: mean distance of all sequences including cluster bounds is computed.

number	gene subgroups	elements	mean distance	cluster distance
1	hL8	1	0.0 (± 0.0)	28.9
2	hL7	2	8.2 (± 0.0)	34.2
3	mL1 mL2	2	2.7 (± 0.0)	40.5
4	hL5	4	11.5 (± 5.1)	37.1
5	hL4	3	15.9 (± 4.5)	29.0
6	mL3	1	0.0 (± 0.0)	35.0
7	hL9	1	0.0 (± 0.0)	44.2
8	mL6 mL8 mL7 mL5 mL4	5	4.6 (± 0.8)	25.3
9	hL6	1	0.0 (± 0.0)	33.2
10	hL2 hL1	10	8.9 (± 6.6)	29.7
11	hL10	1	0.0 (± 0.0)	33.7
12	hL3	10	13.9 (± 6.1)	36.2
	overall	41	10.4 (± 5.4)	33.9 (± 5.1)
	all sequences	41	17.3 (± 8.1)	0.0 (± 0.0)

Table 6: Comparison between human and mouse VL genes. Human genes are colored gray. The mean distance from one element to the next is calcualted within a cluster. The standard deviation is shown inside parenthesis. Cluster distance means the distance from the last element of one cluster to the first element of the next cluster. The overall mean distance is the weighted mean value of the mean distances inside of all clusters containing more than one element. The overall cluster distance is the unweighted mean value of all cluster distances. All sequences: mean distance of all sequences including cluster bounds is computed.

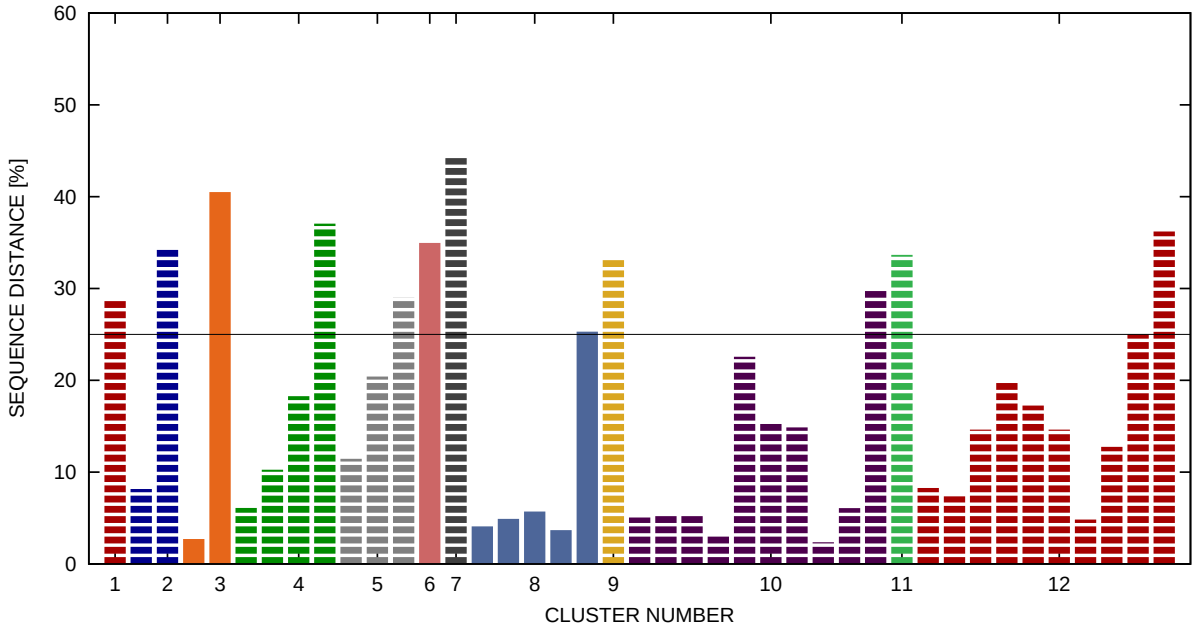
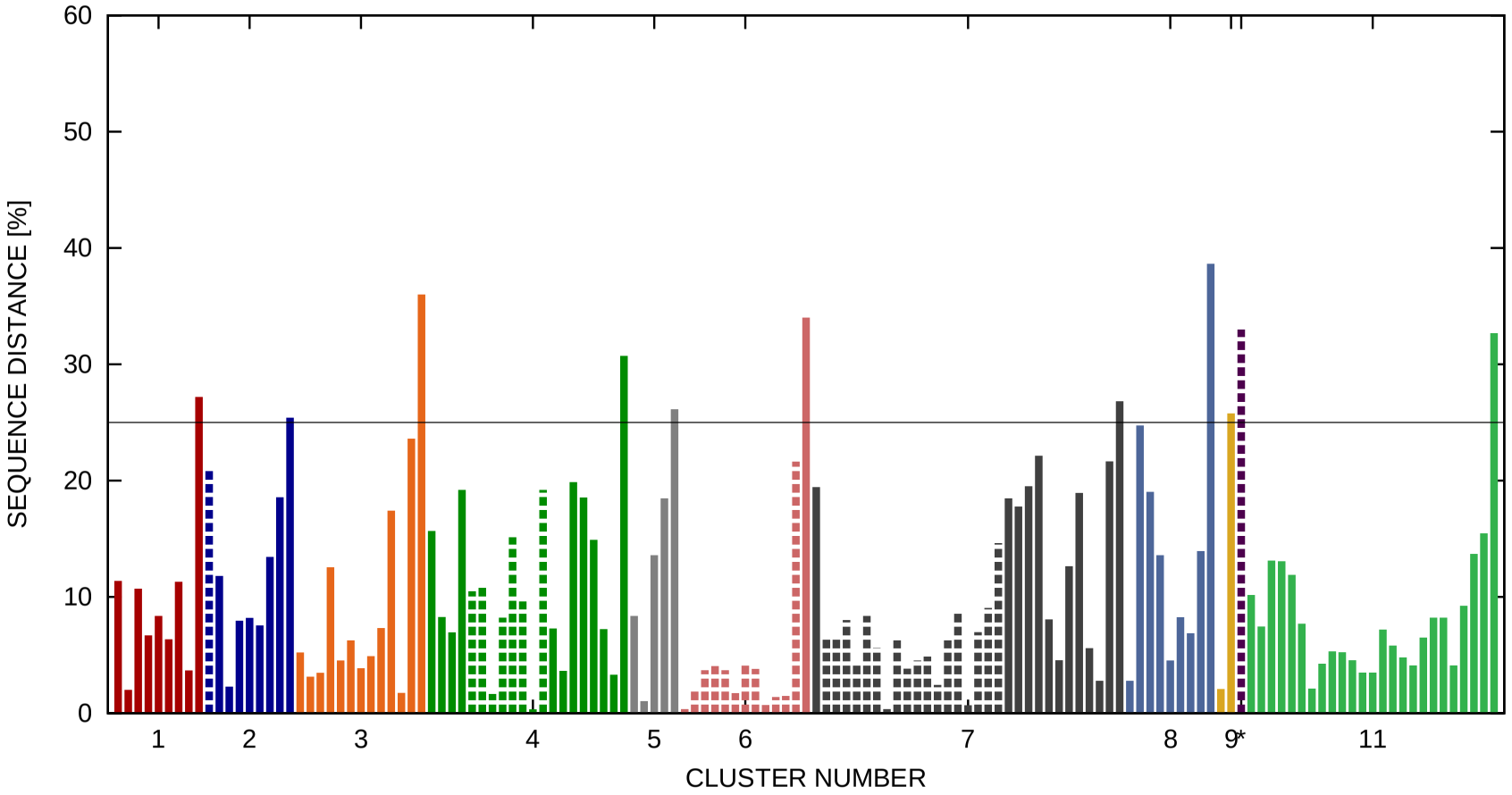


Figure 22: Comparison between human and mouse VL genes. The TSP instance was solved by the Concorde TSP solver which uses the NEOS Server. The threshold for the clusters is set to 25%. The dotted lines refer to the human genes. The full lines refer to the mouse genes.

number	gene subgroups	elements	mean distance	cluster distance
1	mK3	9	7.6 (± 3.3)	27.2
2	hK4 mK8	9	11.3 (± 5.8)	25.4
3	mK6 mK7	13	7.8 (± 6.4)	36.0
4	mK2 hK2 mK1	20	10.5 (± 6.1)	30.7
5	mK5	5	10.4 (± 6.5)	26.1
6	hK3 mK18	13	4.1 (± 5.5)	34.0
7	mK16 hK1 mK11 mK14 mK9 mK10 mK19	31	9.3 (± 6.4)	26.8
8	mK13 mK12	9	11.7 (± 7.1)	38.6
9	mK17	2	2.1 (± 0.0)	25.8
10	hK5	1	0.0 (± 0.0)	33.0
11	mK4	25	7.5 (± 3.6)	32.7
	overall	137	8.7 (± 5.7)	30.6 (± 4.4)
	all sequences	137	10.4 (± 6.2)	0.0 (± 0.0)

Table 7: Comparison between human and mouse VK genes. Human genes are colored gray. The mean distance from one element to the next is calcualted within a cluster. The standard deviation is shown inside parenthesis. Cluster distance means the distance from the last element of one cluster to the first element of the next cluster. The overall mean distance is the weighted mean value of the mean distances inside of all clusters containing more than one element. The overall cluster distance is the unweighted mean value of all cluster distances. All sequences: mean distance of all sequences including cluster bounds is computed.

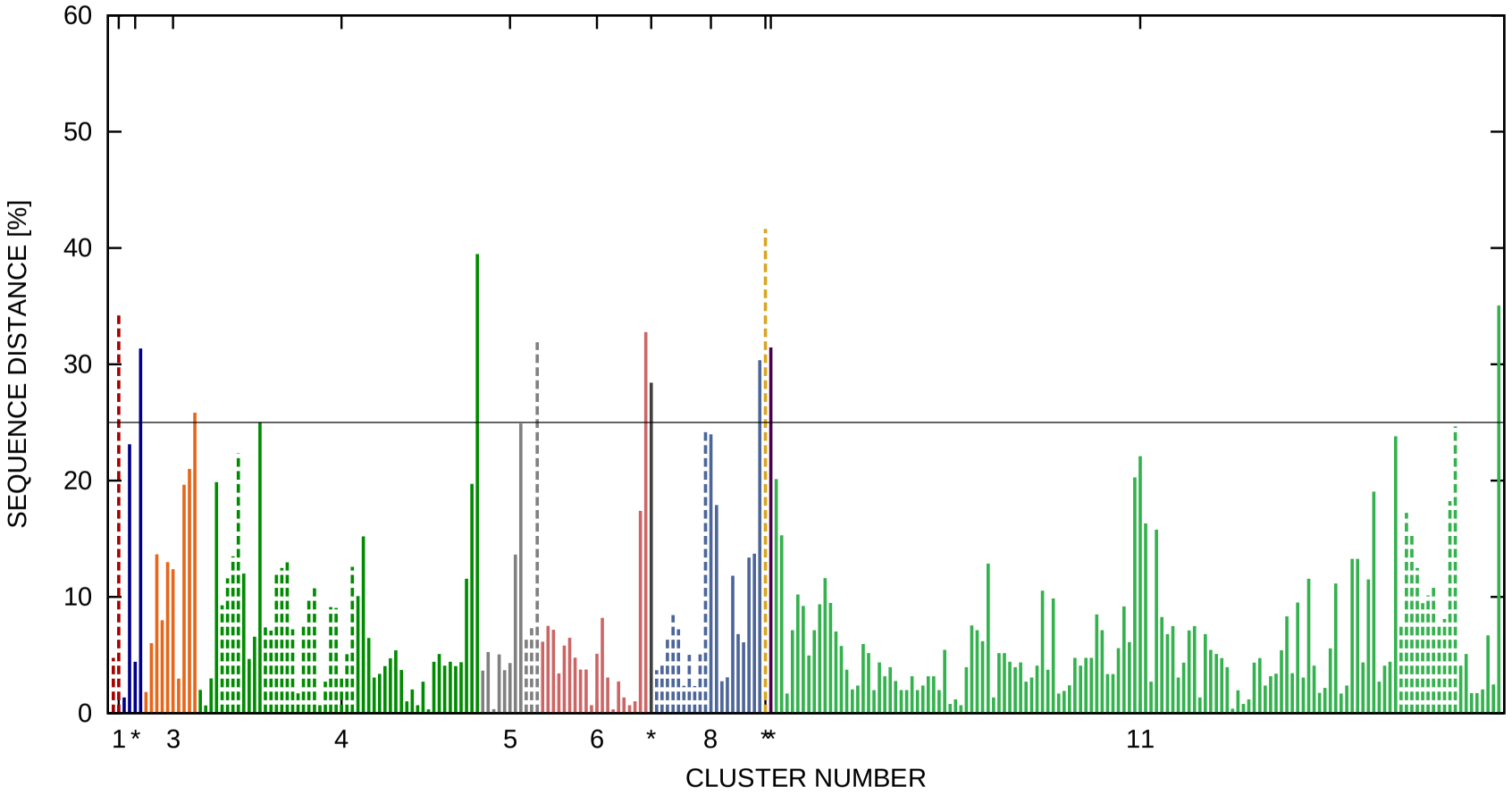


∗ cluster number can not be shown due to close distance on graphical display

Figure 23: Comparison between human and mouse VK genes. The TSP instance was solved by the Concorde TSP solver which uses the NEOS Server. The threshold for the clusters is set to 25%. The dotted lines refer to the human genes. The full lines refer to the mouse genes.

number	gene subgroups	elements	mean distance	cluster distance
1	hH5	2	4.8 (± 0.0)	34.5
2	mH11 mH4	4	9.6 (± 9.6)	31.4
3	mH6S mH6 mH13	10	10.9 (± 6.4)	25.8
4	mH10 mH10S hH3 mH7 mH5 mH5S	52	7.5 (± 5.7)	39.5
5	mH8 mH8S hH2	11	7.5 (± 6.7)	31.9
6	mH2 mH2S	20	4.7 (± 3.8)	32.8
7	mH12	1	0.0 (± 0.0)	28.4
8	hH4 mH12 mH3 mH3S	20	8.9 (± 6.7)	30.4
9	hH6	1	0.0 (± 0.0)	41.6
10	mH15	1	0.0 (± 0.0)	31.5
11	mH1 mH1S mH14 hH1 hH7 mH9 mH9S	134	6.4 (± 5.1)	35.1
	overall	256	6.9 (± 5.5)	33.0 (± 4.4)
	all sequences	256	8.0 (± 5.7)	0.0 (± 0.0)

Table 8: Comparison between human and mouse heavy chain V genes. Human genes are colored gray. The mean distance from one element to the next is calcualted within a cluster. The standard deviation is shown inside parenthesis. Cluster distance means the distance from the last element of one cluster to the first element of the next cluster. The overall mean distance is the weighted mean value of the mean distances inside of all clusters containing more than one element. The overall cluster distance is the unweighted mean value of all cluster distances. All sequences: mean distance of all sequences including cluster bounds is computed.



∗ cluster number can not be shown due to close distance on graphical display

Figure 24: Comparison between human and mouse heavy chain V genes. The TSP instance was solved by the Concorde TSP solver which uses the NEOS Server. The threshold for the clusters is set to 25%. The dotted lines refer to the human genes. The full lines refer to the mouse genes.

number	gene subgroups	elements	mean distance	cluster distance
1	L7	2	8.2 (± 0.0)	28.9
2	L8	1	0.0 (± 0.0)	36.2
3	L3	10	13.9 (± 6.1)	33.7
4	L10	1	0.0 (± 0.0)	29.7
5	L1 L2	10	8.9 (± 6.6)	33.2
6	L6	1	0.0 (± 0.0)	38.2
7	L5	4	11.5 (± 5.1)	37.1
8	L4	3	15.9 (± 4.5)	31.5
9	L9	1	0.0 (± 0.0)	49.7
10	K5	1	0.0 (± 0.0)	33.8
11	K1 K3 K4	31	5.4 (± 5.2)	27.4
12	K2	8	8.3 (± 5.4)	52.4
13	H6	1	0.0 (± 0.0)	29.1
14	H4	10	5.1 (± 2.0)	31.6
15	H2	3	7.8 (± 1.2)	37.9
16	H3	21	8.5 (± 4.1)	32.3
17	H5	2	4.8 (± 0.0)	27.7
18	H1 H7	11	11.7 (± 3.8)	52.7
	overall	121	8.3 (± 4.8)	35.7 (± 7.8)
	all sequences	121	12.4 (± 6.6)	0.0 (± 0.0)

Table 9: Comparison between all human V genes. The mean distance from one element to the next is calcualted within a cluster. The standard deviation is shown inside parenthesis. Cluster distance means the distance from the last element of one cluster to the first element of the next cluster. The overall mean distance is the weighted mean value of the mean distances inside of all clusters containing more than one element. The overall cluster distance is the unweighted mean value of all cluster distances. All sequences: mean distance of all sequences including cluster bounds is computed.

number	gene subgroups	elements	mean distance	cluster distance
1	L3	10	13.9 (± 6.1)	33.7
2	L10	1	0.0 (± 0.0)	29.7
3	L1 L2	10	8.9 (± 6.6)	33.2
4	L6	1	0.0 (± 0.0)	38.2
5	L5	4	11.5 (± 5.1)	37.1
6	L4	3	15.9 (± 4.5)	31.5
7	L9	1	0.0 (± 0.0)	43.1
8	L8	1	0.0 (± 0.0)	29.2
9	L7	2	8.2 (± 0.0)	38.2
	overall	33	11.6 (± 5.9)	34.9 (± 4.3)
	all sequences	33	18.0 (± 7.9)	0.0 (± 0.0)

Table 10: Comparison between human VL genes. The mean distance from one element to the next is calcualted within a cluster. The standard deviation is shown inside parenthesis. Cluster distance means the distance from the last element of one cluster to the first element of the next cluster. The overall mean distance is the weighted mean value of the mean distances inside of all clusters containing more than one element. The overall cluster distance is the unweighted mean value of all cluster distances. All sequences: mean distance of all sequences including cluster bounds is computed.

number	gene subgroups	elements	mean distance	cluster distance
1	K2	8	8.0 (± 4.9)	42.5
2	K5	1	0.0 (± 0.0)	33.8
3	K1 K3 K4	31	5.4 (± 5.2)	27.4
	overall	40	5.9 (± 5.1)	34.5 (± 6.2)
	all sequences	40	8.0 (± 5.5)	0.0 (± 0.0)

Table 11: Comparison between human VK genes. The mean distance from one element to the next is calcualted within a cluster. The standard deviation is shown inside parenthesis. Cluster distance means the distance from the last element of one cluster to the first element of the next cluster. The overall mean distance is the weighted mean value of the mean distances inside of all clusters containing more than one element. The overall cluster distance is the unweighted mean value of all cluster distances. All sequences: mean distance of all sequences including cluster bounds is computed.

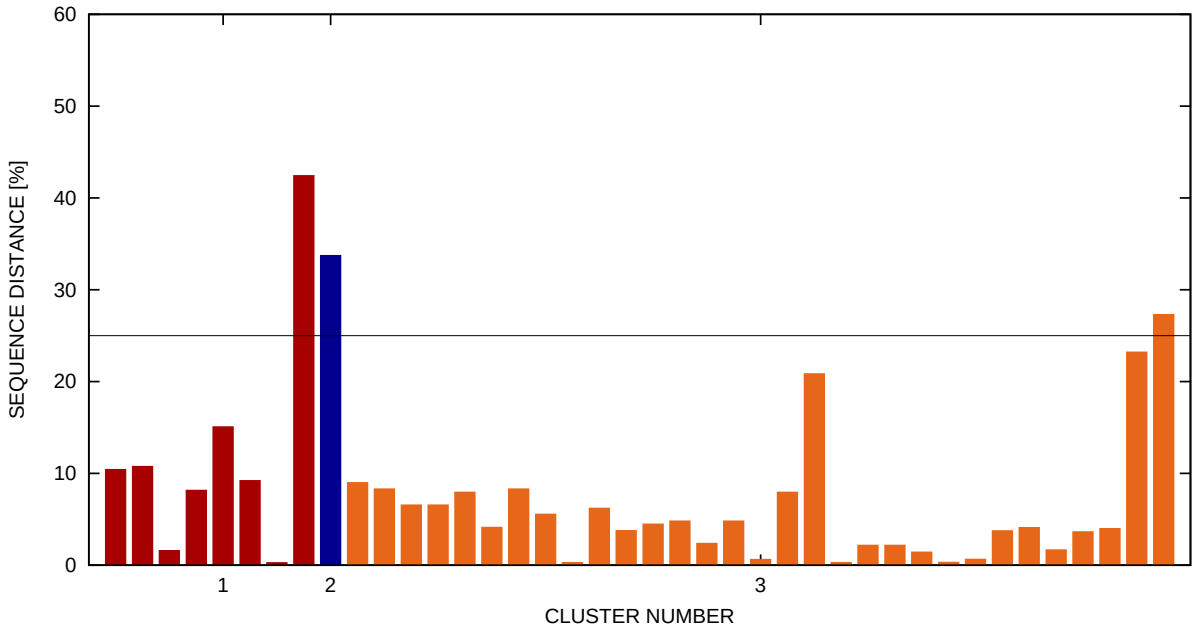


Figure 25: Comparison between human VK genes. The TSP instance was solved by the Concorde TSP solver which uses the NEOS Server. The threshold for the clusters is set to 25%. The dotted lines refer to the human genes. The full lines refer to the mouse genes.

number	gene subgroups	elements	mean distance	cluster distance
1	H6	1	0.0 (± 0.0)	36.5
2	H2	3	7.0 (± 0.3)	39.2
3	H3	21	8.5 (± 4.1)	35.0
4	H7 H1	11	11.7 (± 3.8)	27.7
5	H5	2	4.8 (± 0.0)	34.9
6	H4	10	4.8 (± 1.7)	29.7
	overall	48	8.3 (± 3.5)	33.8 (± 3.9)
	all sequences	48	11.5 (± 5.2)	0.0 (± 0.0)

Table 12: Comparison between human heavy chain V genes. The mean distance from one element to the next is calcualted within a cluster. The standard deviation is shown inside parenthesis. Cluster distance means the distance from the last element of one cluster to the first element of the next cluster. The overall mean distance is the weighted mean value of the mean distances inside of all clusters containing more than one element. The overall cluster distance is the unweighted mean value of all cluster distances. All sequences: mean distance of all sequences including cluster bounds is computed.

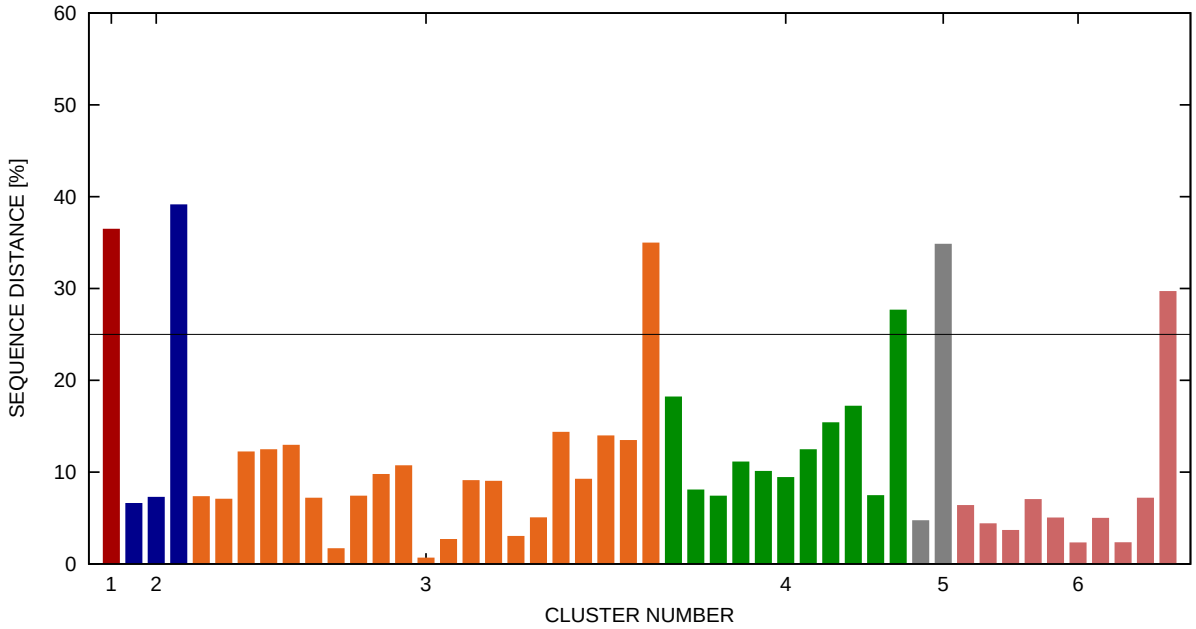


Figure 26: Comparison between human heavy chain V genes. The TSP instance was solved by the Concorde TSP solver which uses the NEOS Server. The threshold for the clusters is set to 25%. The dotted lines refer to the human genes. The full lines refer to the mouse genes.

number	gene subgroups	elements	mean distance	cluster distance
1	L4 L5 L7 L8 L6	5	4.6 (± 0.8)	43.2
2	K18	1	0.0 (± 0.0)	34.0
3	K16	1	0.0 (± 0.0)	25.4
4	K11 K14 K9 K10 K19	12	13.8 (± 7.0)	26.8
5	K13 K12	9	12.0 (± 7.3)	32.1
6	K3	9	7.6 (± 3.3)	32.7
7	K4	25	7.5 (± 3.6)	30.5
8	K6 K7 K8	21	9.3 (± 6.6)	37.1
9	K2	4	10.3 (± 3.8)	25.2
10	K1	8	10.7 (± 6.5)	30.7
11	K5	5	10.4 (± 6.5)	38.0
12	K17	2	2.1 (± 0.0)	46.6
13	L3	1	0.0 (± 0.0)	45.7
14	L1 L2	2	2.7 (± 0.0)	49.1
15	H3 H3S H12	10	11.1 (± 6.7)	29.9
16	H12	1	0.0 (± 0.0)	32.8
17	H2 H2S	20	4.8 (± 4.4)	32.2
18	H8 H8S	8	5.1 (± 3.8)	43.5
19	H11 H4	4	9.6 (± 9.6)	26.7
20	H5 H5S	23	5.3 (± 4.3)	25.3
21	H7	4	7.7 (± 3.1)	26.6
22	H13 H6S H6 H10 H10S	14	10.2 (± 7.9)	37.2
23	H9 H9S	8	3.4 (± 1.8)	39.9
24	H15	1	0.0 (± 0.0)	31.5
25	H1 H1S H14	115	5.8 (± 4.4)	45.1
	overall	313	7.1 (± 5.1)	34.7 (± 7.3)
	all sequences	313	9.3 (± 6.0)	0.0 (± 0.0)

Table 13: Comparison between all mouse V genes. The mean distance from one element to the next is calcualted within a cluster. The standard deviation is shown inside parenthesis. Cluster distance means the distance from the last element of one cluster to the first element of the next cluster. The overall mean distance is the weighted mean value of the mean distances inside of all clusters containing more than one element. The overall cluster distance is the unweighted mean value of all cluster distances. All sequences: mean distance of all sequences including cluster bounds is computed.

number	gene subgroups	elements	mean distance	cluster distance
1	L3	1	0.0 (± 0.0)	45.7
2	L1 L2	2	2.7 (± 0.0)	45.0
3	L6 L8 L4 L5 L7	5	4.3 (± 0.8)	45.2
	overall	8	4.0 (± 0.8)	45.3 (± 0.3)
	all sequences	8	19.5 (± 12.5)	0.0 (± 0.0)

Table 14: Comparison between mouse VL genes. The mean distance from one element to the next is calcualted within a cluster. The standard deviation is shown inside parenthesis. Cluster distance means the distance from the last element of one cluster to the first element of the next cluster. The overall mean distance is the weighted mean value of the mean distances inside of all clusters containing more than one element. The overall cluster distance is the unweighted mean value of all cluster distances. All sequences: mean distance of all sequences including cluster bounds is computed.

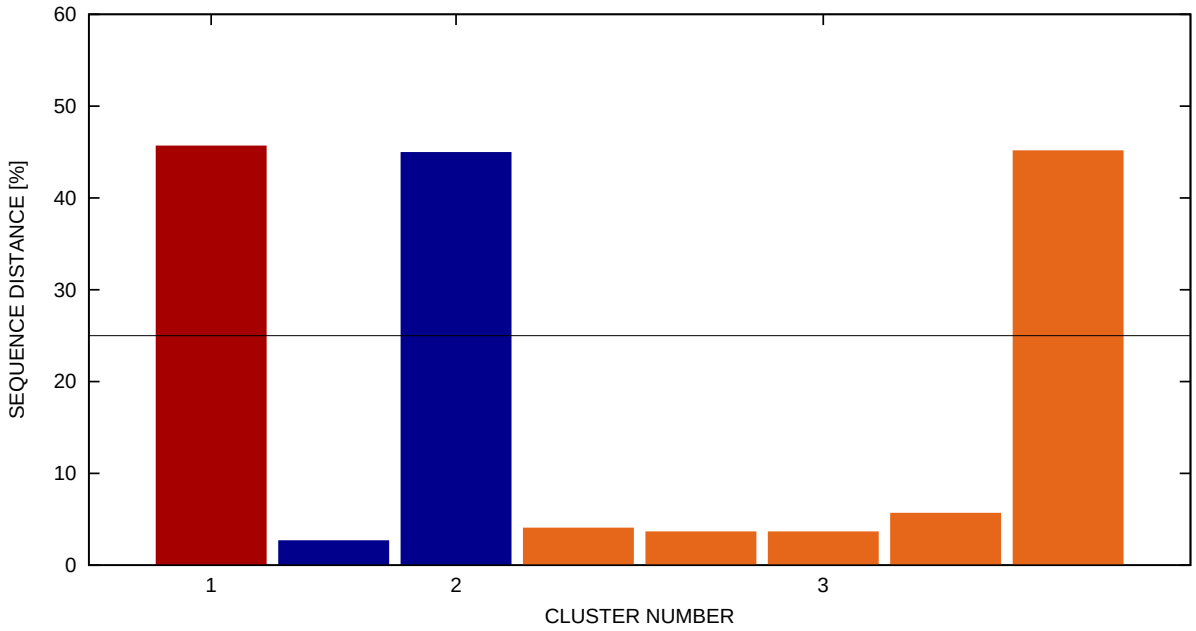
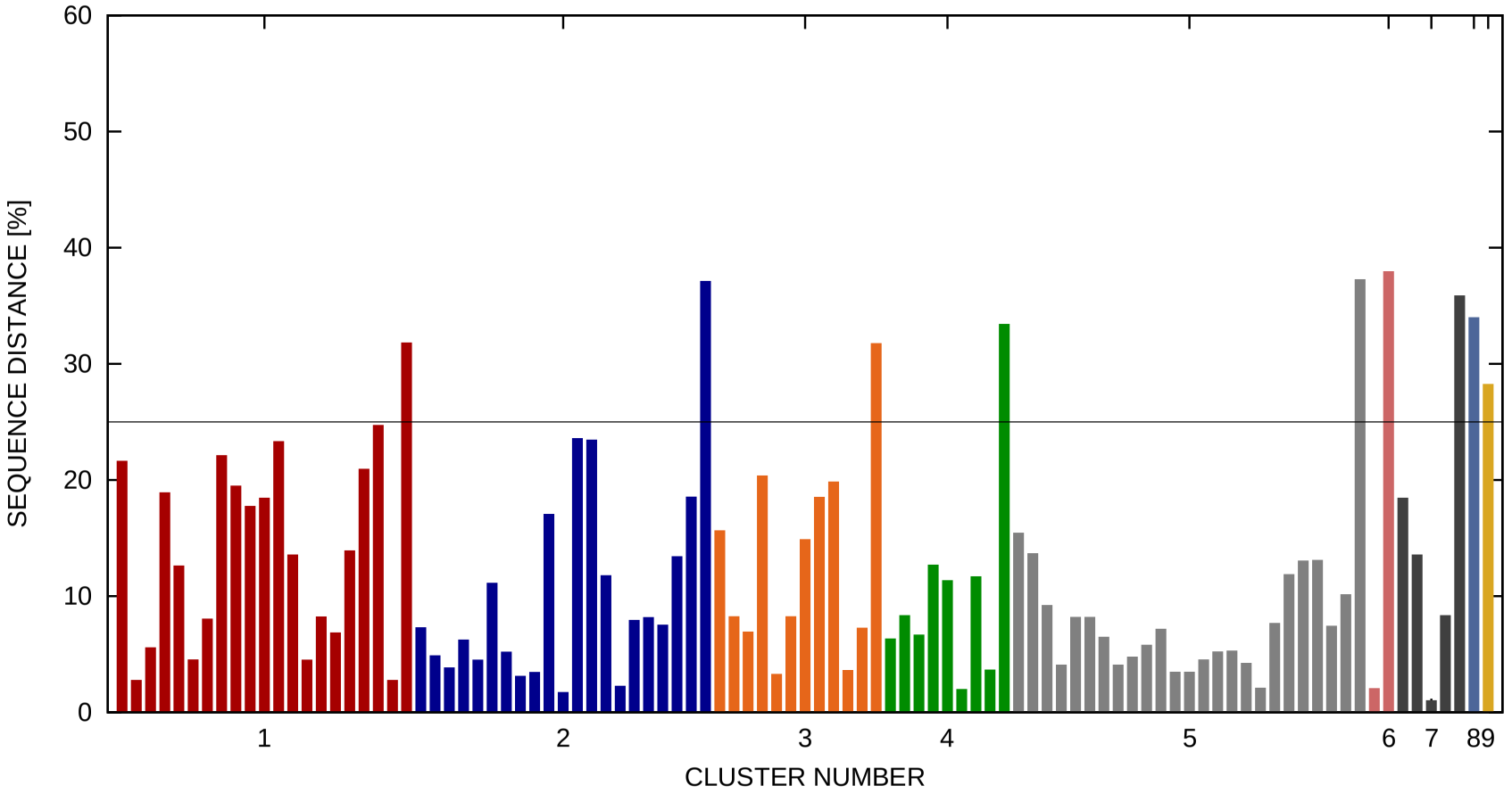


Figure 27: Comparison between mouse VL genes. The TSP instance was solved by the Concorde TSP solver which uses the NEOS Server. The threshold for the clusters is set to 25%. The dotted lines refer to the human genes. The full lines refer to the mouse genes.

number	gene subgroups	elements	mean distance	cluster distance
1	K19 K10 K9 K14 K11 K12 K13	21	13.6 (± 7.4)	31.8
2	K6 K7 K8	21	9.3 (± 6.6)	37.1
3	K2 K1	12	11.6 (± 6.1)	31.8
4	K3	9	7.9 (± 3.6)	33.4
5	K4	25	7.5 (± 3.6)	37.3
6	K17	2	2.1 (± 0.0)	38.0
7	K5	5	10.4 (± 6.5)	35.9
8	K18	1	0.0 (± 0.0)	34.0
9	K16	1	0.0 (± 0.0)	28.3
	overall	97	9.9 (± 5.8)	34.2 (± 3.0)
	all sequences	97	12.1 (± 6.5)	0.0 (± 0.0)

Table 15: Comparison between mouse VK genes. The mean distance from one element to the next is calcualted within a cluster. The standard deviation is shown inside parenthesis. Cluster distance means the distance from the last element of one cluster to the first element of the next cluster. The overall mean distance is the weighted mean value of the mean distances inside of all clusters containing more than one element. The overall cluster distance is the unweighted mean value of all cluster distances. All sequences: mean distance of all sequences including cluster bounds is computed.

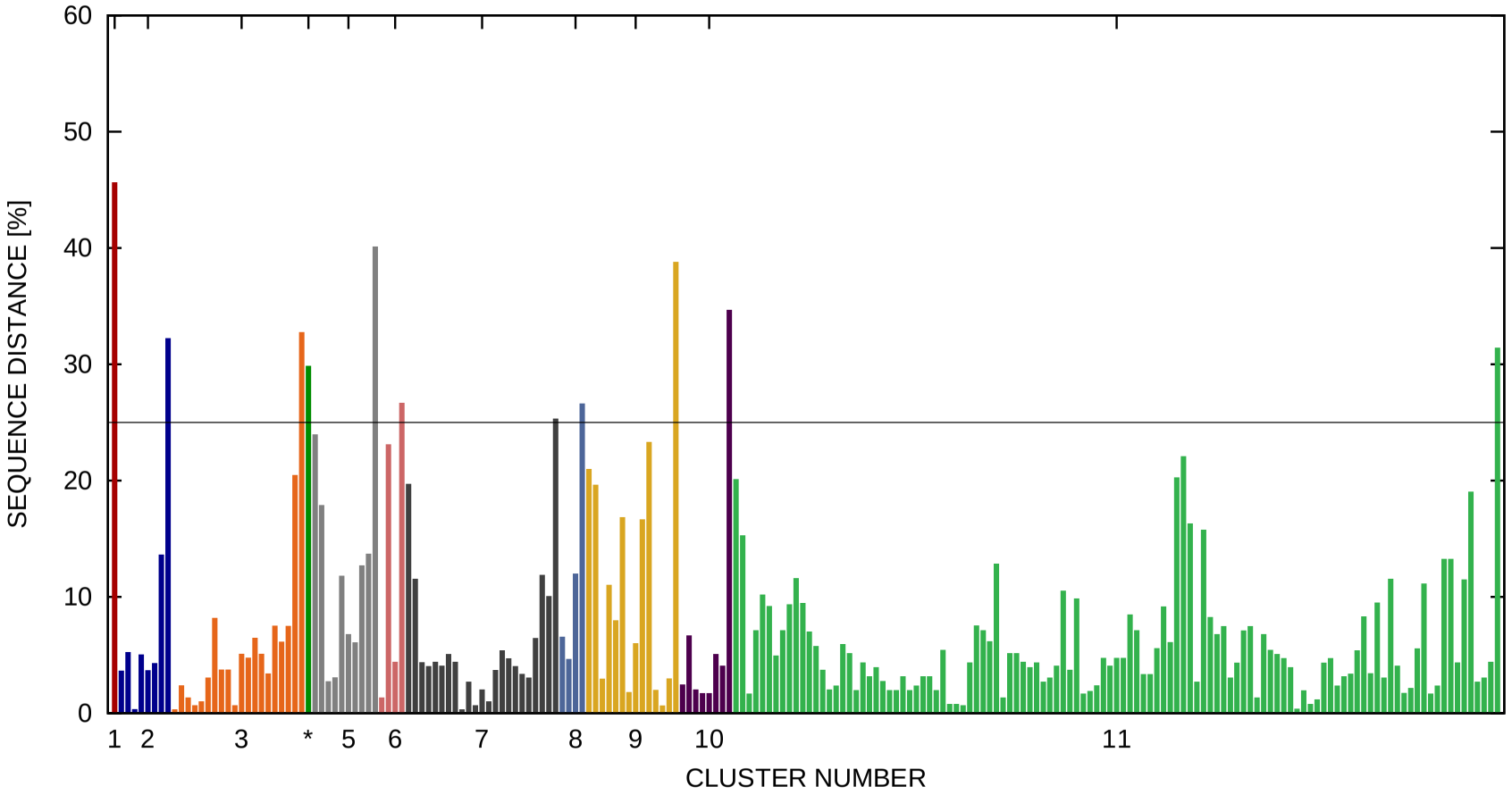


∗ cluster number can not be shown due to close distance on graphical display

Figure 28: Comparison between mouse VK genes. The TSP instance was solved by the Concorde TSP solver which uses the NEOS Server. The threshold for the clusters is set to 25%. The dotted lines refer to the human genes. The full lines refer to the mouse genes.

number	gene subgroups	elements	mean distance	cluster distance
1	H15	1	0.0 (± 0.0)	45.6
2	H8 H8S	8	5.1 (± 3.8)	32.2
3	H2 H2S	20	4.8 (± 4.4)	32.8
4	H12	1	0.0 (± 0.0)	29.9
5	H12 H3 H3S	10	11.0 (± 6.7)	40.1
6	H11 H4	4	9.6 (± 9.6)	26.7
7	H5 H5S	23	5.3 (± 4.3)	25.3
8	H7	4	7.7 (± 3.1)	26.6
9	H13 H6S H6 H10 H10S	14	10.2 (± 7.9)	38.8
10	H9 H9S	8	3.4 (± 1.8)	34.7
11	H1 H1S H14	115	5.8 (± 4.4)	31.4
	overall	208	6.2 (± 4.9)	33.1 (± 6.0)
	all sequences	208	7.6 (± 5.3)	0.0 (± 0.0)

Table 16: Comparison between mouse heavy chain V genes. The mean distance from one element to the next is calcualted within a cluster. The standard deviation is shown inside parenthesis. Cluster distance means the distance from the last element of one cluster to the first element of the next cluster. The overall mean distance is the weighted mean value of the mean distances inside of all clusters containing more than one element. The overall cluster distance is the unweighted mean value of all cluster distances. All sequences: mean distance of all sequences including cluster bounds is computed.



∗ cluster number can not be shown due to close distance on graphical display

Figure 29: Comparison between mouse heavy chain V genes. The TSP instance was solved by the Concorde TSP solver which uses the NEOS Server. The threshold for the clusters is set to 25%. The dotted lines refer to the human genes. The full lines refer to the mouse genes.

number	gene subgroups	elements	mean distance	cluster distance
1	ψ2 ψ7 ψ10 ψ20	4	19.8 (± 4.1)	27.8
2	ψ24 ψ18 ψ5 ψ14	4	17.9 (± 1.4)	26.3
3	ψ16	1	0.0 (± 0.0)	52.5
4	ψ15 ψ9	2	4.5 (± 0.0)	34.8
5	ψ1 ψ3 ψ6 ψ11 ψ13 ψ4 ψ12 L1 ψ8 ψ17	10	17.2 (± 4.6)	25.7
6	ψ22	1	0.0 (± 0.0)	29.5
7	ψ19 ψ21 ψ23	3	17.9 (± 1.8)	25.9
8	ψ25	1	0.0 (± 0.0)	56.1
	overall	26	17.1 (± 3.7)	34.8 (± 11.6)
	all sequences	26	22.6 (± 6.2)	0.0 (± 0.0)

Table 17: Comparison between the light chain V gene and light chain pseudogenes of chicken. The mean distance from one element to the next is calcualted within a cluster. The standard deviation is shown inside parenthesis. Cluster distance means the distance from the last element of one cluster to the first element of the next cluster. The overall mean distance is the weighted mean value of the mean distances inside of all clusters containing more than one element. The overall cluster distance is the unweighted mean value of all cluster distances. All sequences: mean distance of all sequences including cluster bounds is computed.

number	gene subgroups	elements	mean distance	cluster distance
1	ψ2 ψ7 ψ10 ψ20	4	19.8 (± 4.1)	27.8
2	ψ24 ψ18 ψ5 ψ14	4	17.9 (± 1.4)	26.3
3	ψ16	1	0.0 (± 0.0)	52.5
4	ψ15 ψ9	2	4.5 (± 0.0)	34.8
5	ψ1 ψ3 ψ6 ψ11 ψ13 ψ19	6	19.3 (± 4.7)	29.5
6	ψ22	1	0.0 (± 0.0)	25.7
7	ψ17 ψ8 ψ4 ψ12 ψ21 ψ23	6	17.1 (± 3.4)	25.9
8	ψ25	1	0.0 (± 0.0)	56.1
	overall	25	17.6 (± 3.7)	34.8 (± 11.6)
	all sequences	25	23.1 (± 6.2)	0.0 (± 0.0)

Table 18: Comparison between chicken light chain pseudogenes. The mean distance from one element to the next is calcualted within a cluster. The standard deviation is shown inside parenthesis. Cluster distance means the distance from the last element of one cluster to the first element of the next cluster. The overall mean distance is the weighted mean value of the mean distances inside of all clusters containing more than one element. The overall cluster distance is the unweighted mean value of all cluster distances. All sequences: mean distance of all sequences including cluster bounds is computed.

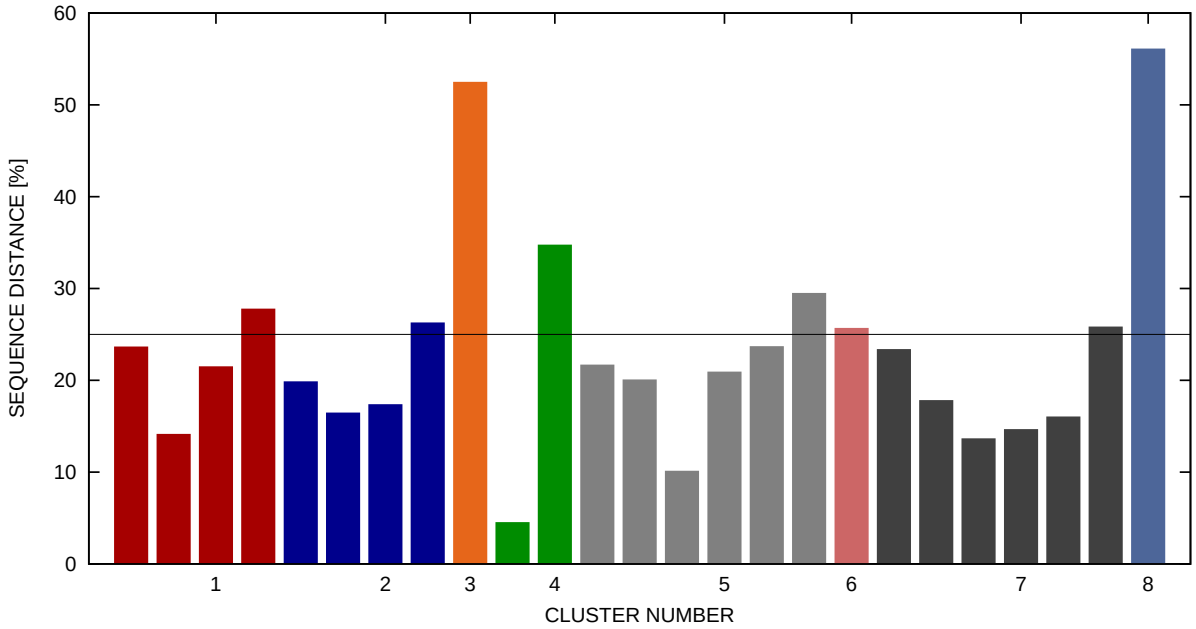


Figure 30: Comparison between chicken light chain pseudogenes. The TSP instance was solved by the Concorde TSP solver which uses the NEOS Server. The threshold for the clusters is set to 25%. The dotted lines refer to the human genes. The full lines refer to the mouse genes..

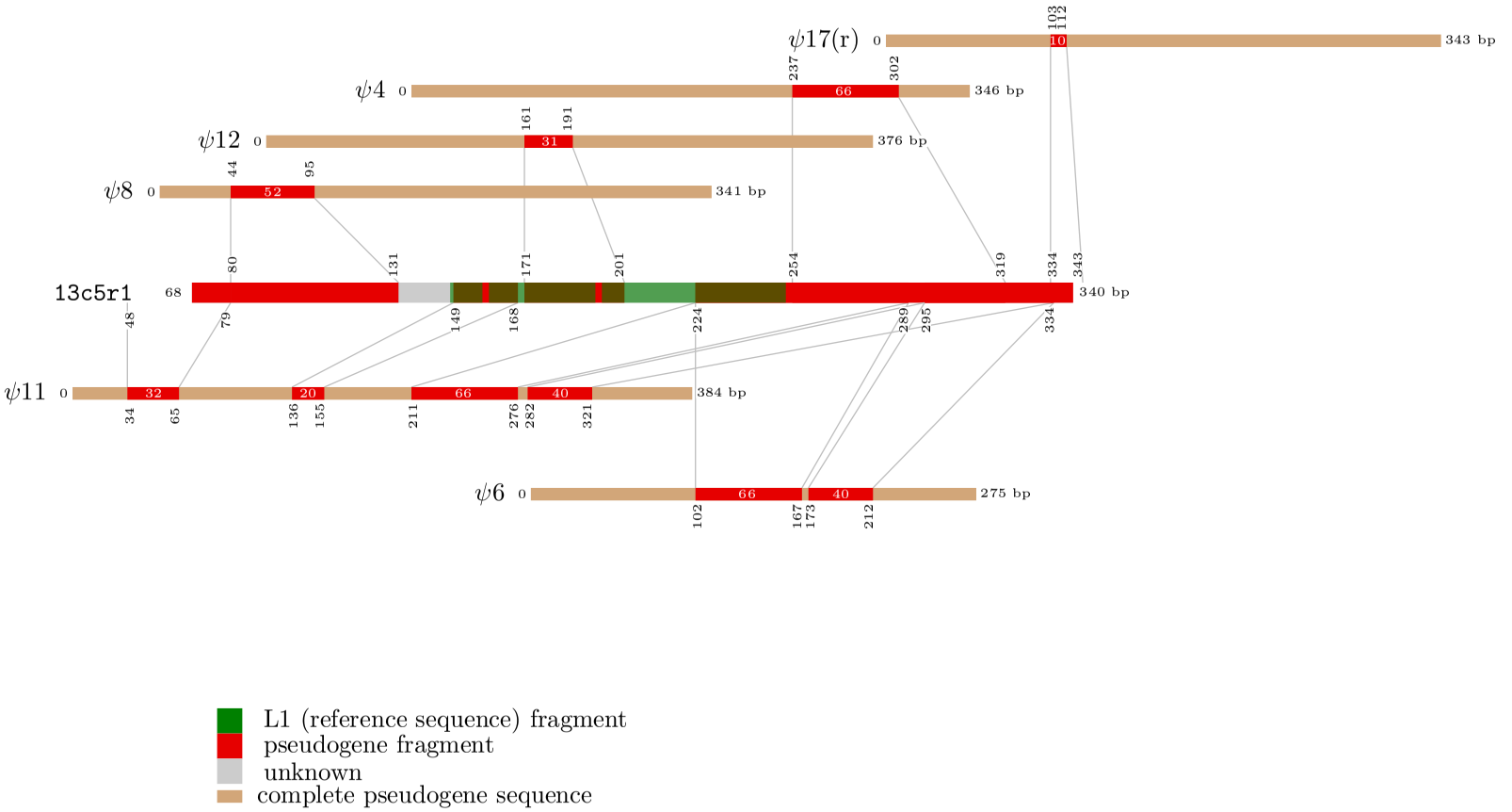


Figure 31: Gene conversion - clone `13c5r1` shown on threshold 10. White numbers indicate the length of pseudogene fragments. The range of pseudogene fragments is shown in red at the partial sequence of the clone and at the whole pseudogene sequence. Black numbers indicate the start and the end positions. '(r)' indicates that the reciprocal sequence of the pseudogene is matching the fragment.

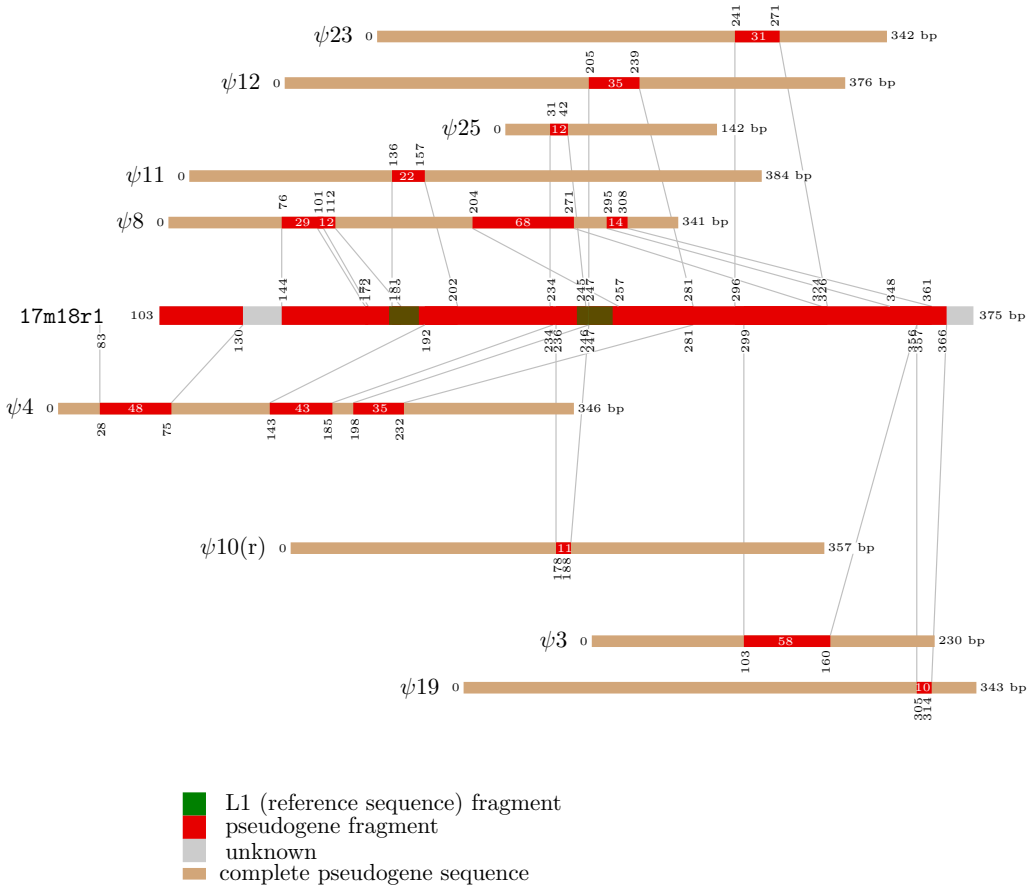


Figure 32: Gene conversion - clone `17m18r1` shown on threshold 10. White numbers indicate the length of pseudogene fragments. The range of pseudogene fragments is shown in red at the partial sequence of the clone and at the whole pseudogene sequence. Black numbers indicate the start and the end positions. '(r)' indicates that the reciprocal sequence of the pseudogene is matching the fragment.

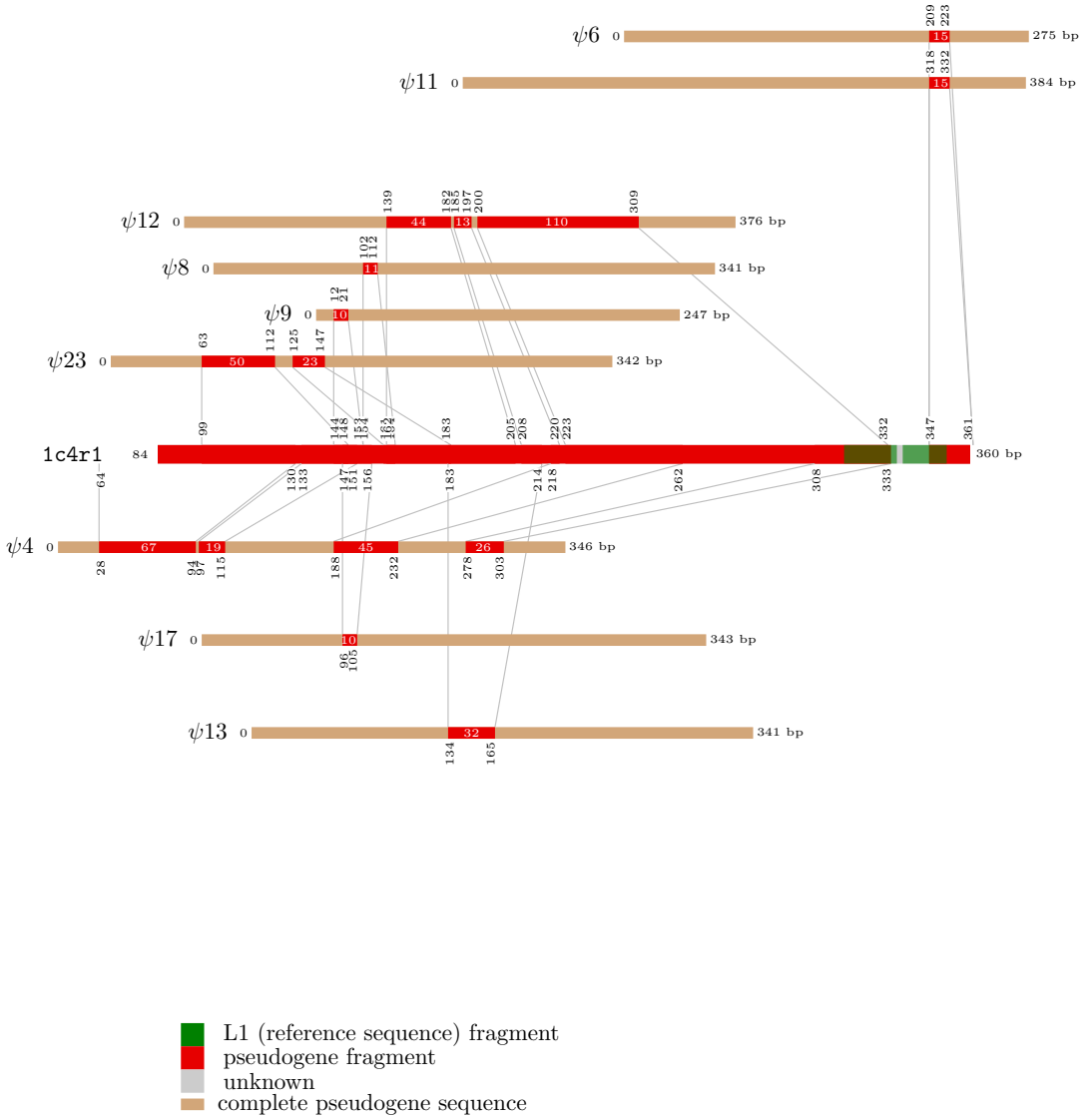


Figure 33: Gene conversion - clone `1c4r1` shown on threshold 10. White numbers indicate the length of pseudogene fragments. The range of pseudogene fragments is shown in red at the partial sequence of the clone and at the whole pseudogene sequence. Black numbers indicate the start and the end positions.

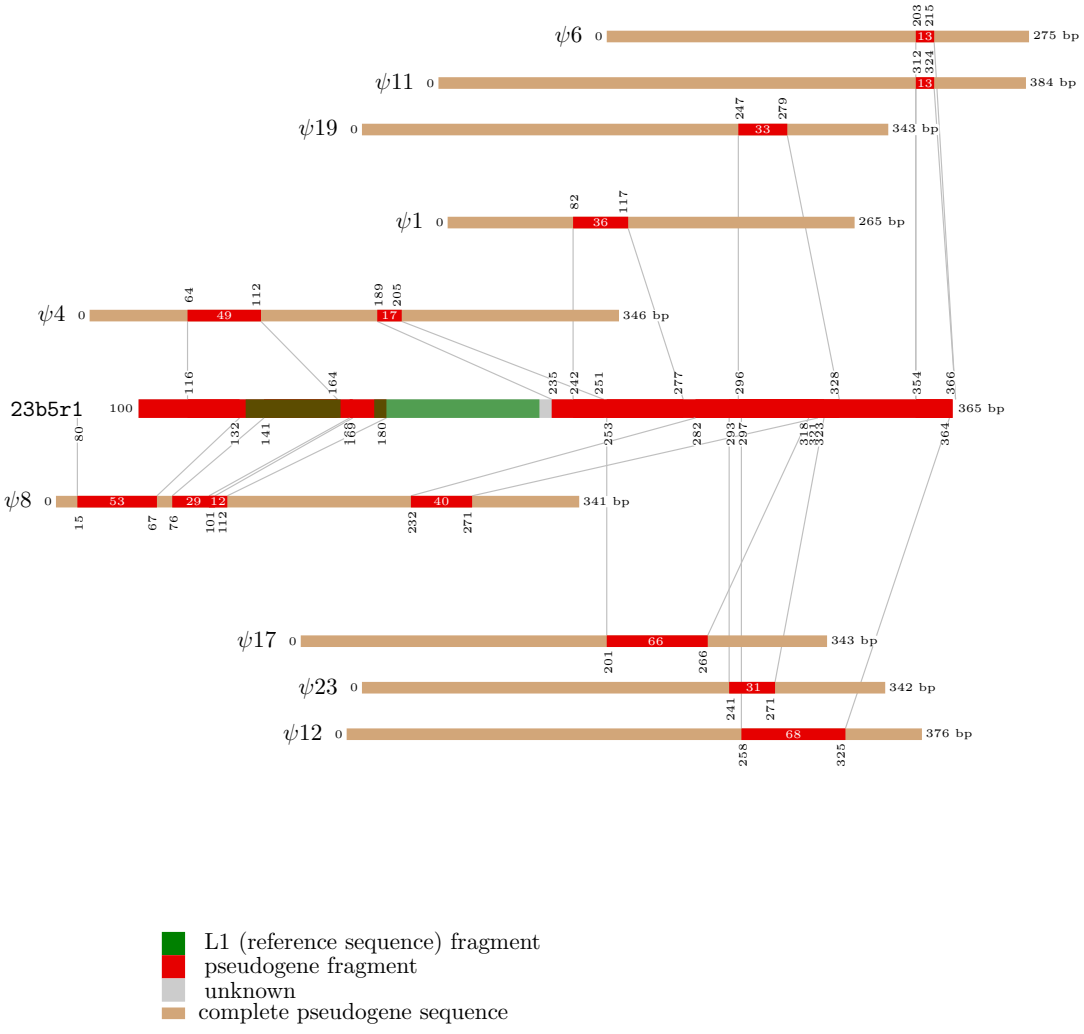


Figure 34: Gene conversion - clone `23b5r1` shown on threshold 10. White numbers indicate the length of pseudogene fragments. The range of pseudogene fragments is shown in red at the partial sequence of the clone and at the whole pseudogene sequence. Black numbers indicate the start and the end positions.

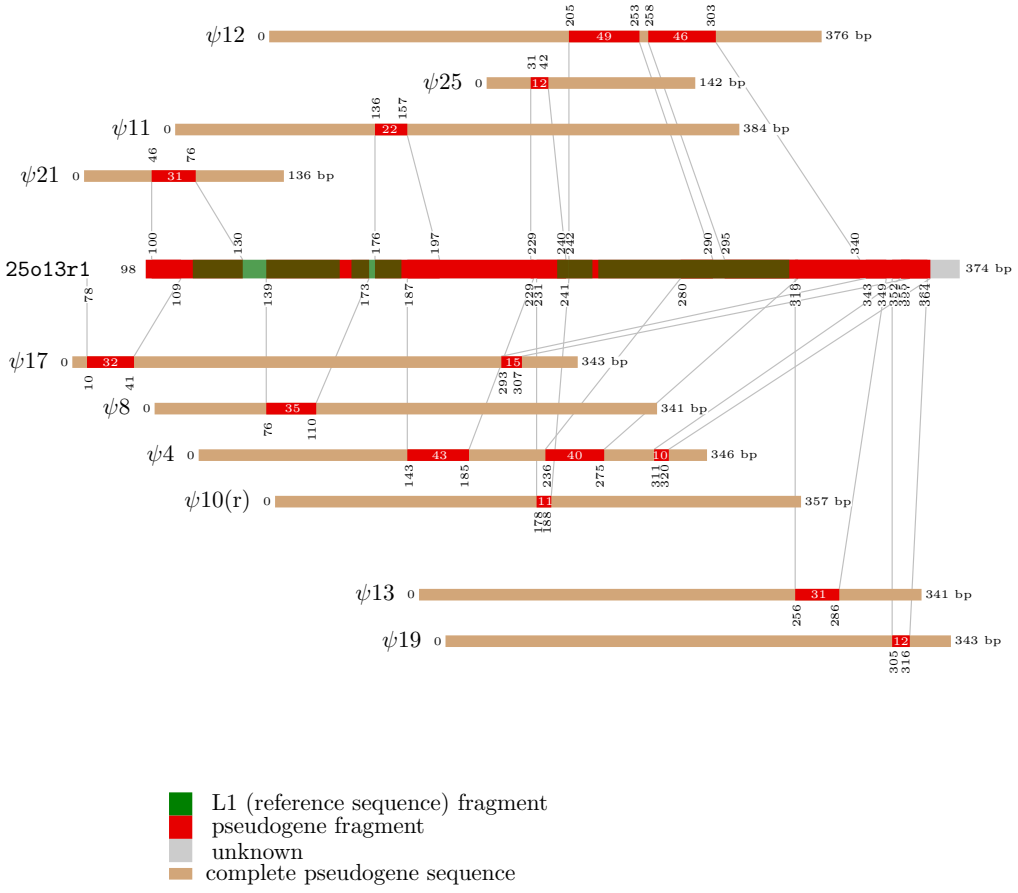


Figure 35: Gene conversion - clone `25o13r1` shown on threshold 10. White numbers indicate the length of pseudogene fragments. The range of pseudogene fragments is shown in red at the partial sequence of the clone and at the whole pseudogene sequence. Black numbers indicate the start and the end positions. '(r)' indicates that the reciprocal sequence of the pseudogene is matching the fragment.

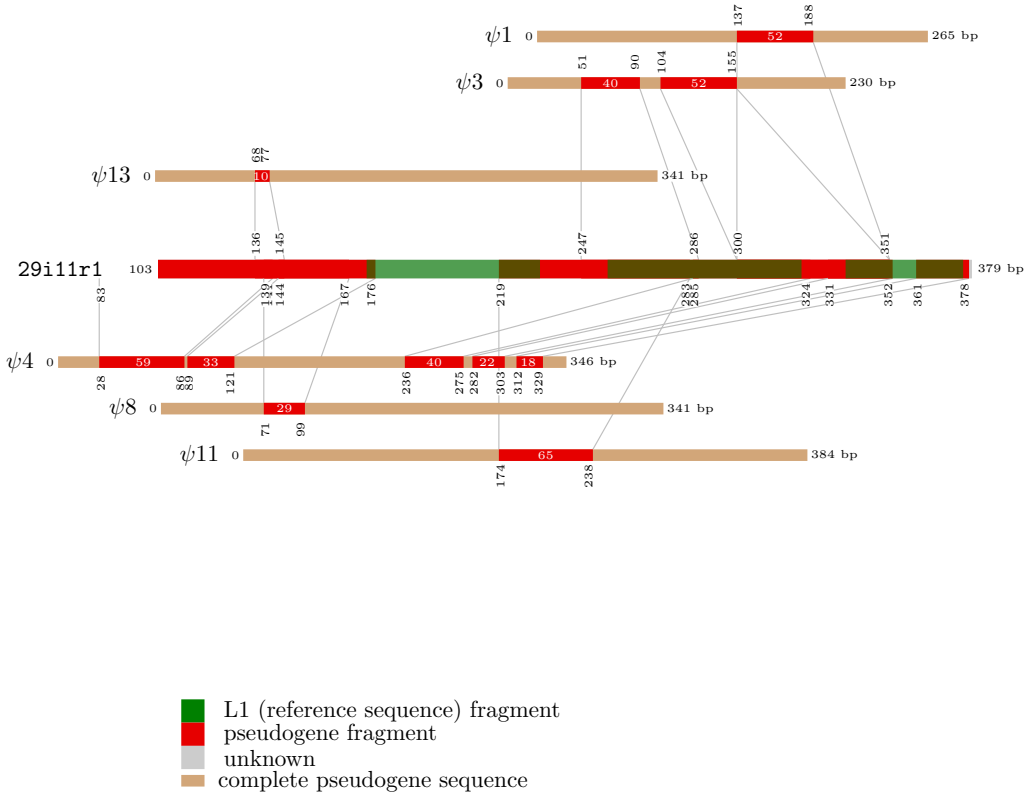


Figure 36: Gene conversion - clone `29i11r1` shown on threshold 10. White numbers indicate the length of pseudogene fragments. The range of pseudogene fragments is shown in red at the partial sequence of the clone and at the whole pseudogene sequence. Black numbers indicate the start and the end positions.

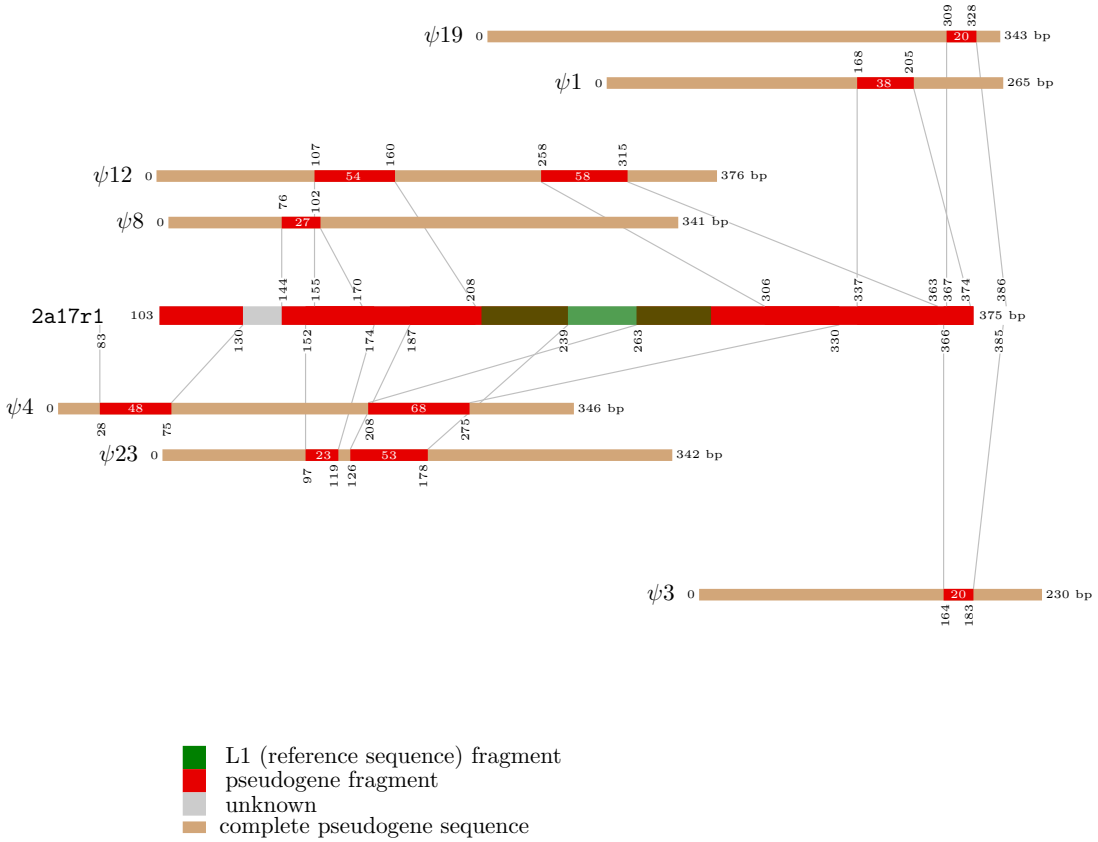


Figure 37: Gene conversion - clone `2a17r1` shown on threshold 10. White numbers indicate the length of pseudogene fragments. The range of pseudogene fragments is shown in red at the partial sequence of the clone and at the whole pseudogene sequence. Black numbers indicate the start and the end positions.

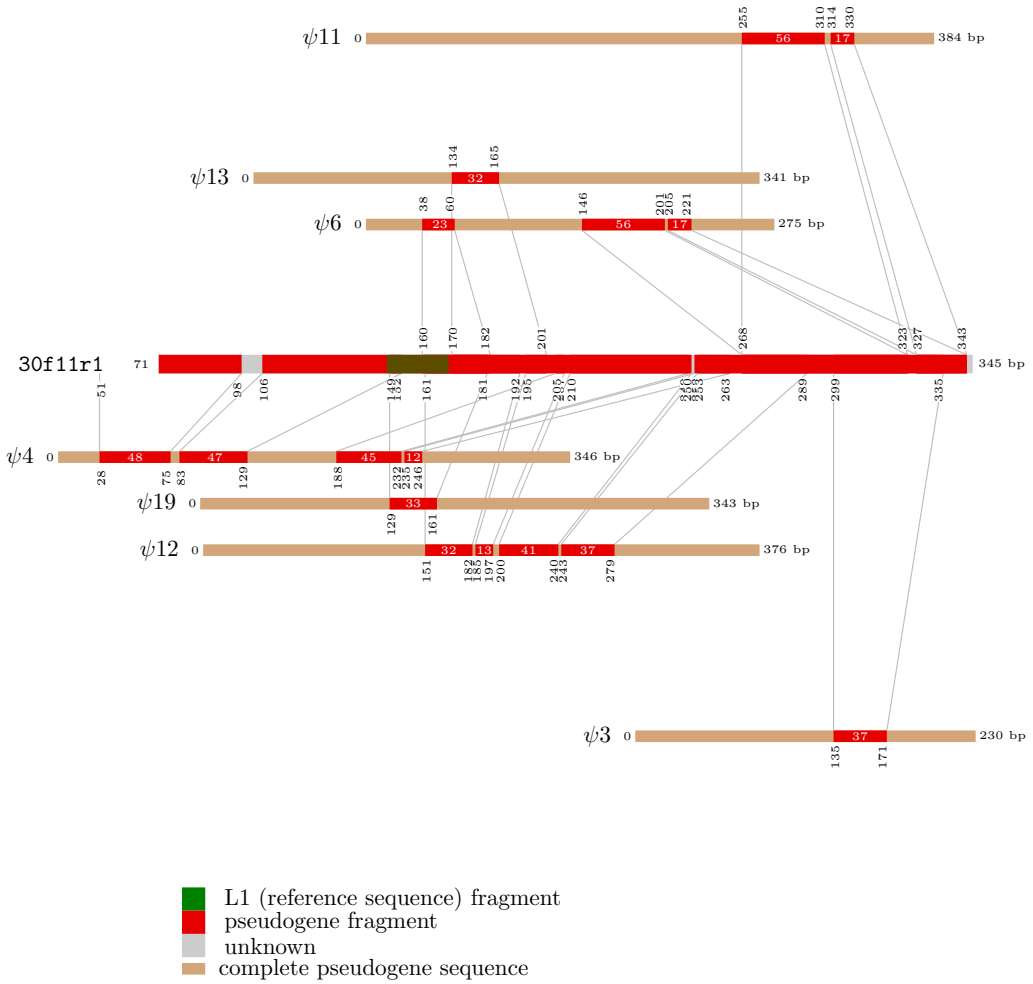


Figure 38: Gene conversion - clone `30f11r1` shown on threshold 10. White numbers indicate the length of pseudogene fragments. The range of pseudogene fragments is shown in red at the partial sequence of the clone and at the whole pseudogene sequence. Black numbers indicate the start and the end positions.

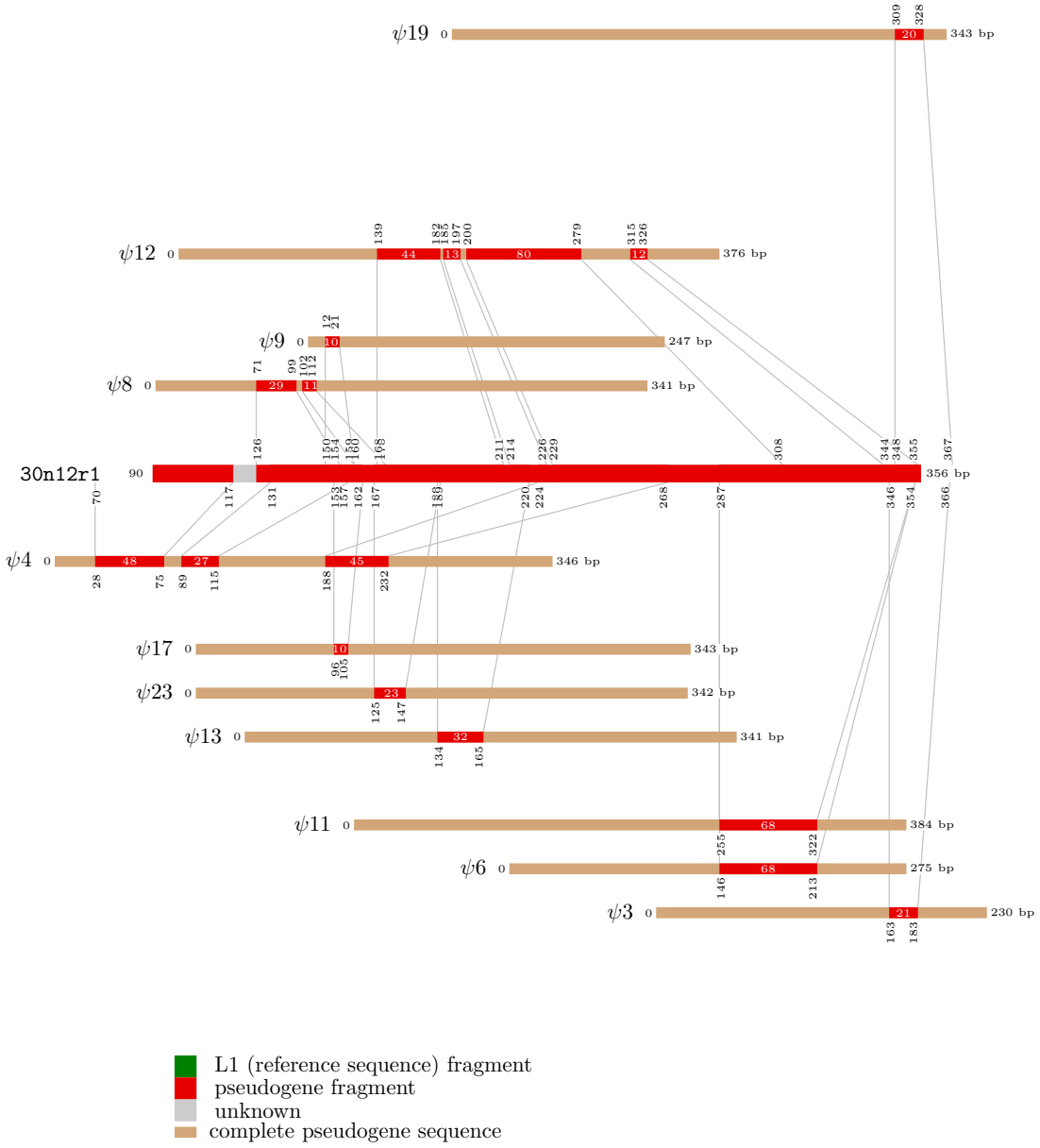


Figure 39: Gene conversion - clone `30n12r1` shown on threshold 10. White numbers indicate the length of pseudogene fragments. The range of pseudogene fragments is shown in red at the partial sequence of the clone and at the whole pseudogene sequence. Black numbers indicate the start and the end positions.

clone	event	length	pseudogene
29i11r1	1	94	ψ4
29i11r1	2	94	ψ4
29i11r1	3	65	ψ11
29i11r1	4	52	ψ1 ψ3
29i11r1	5	18	ψ4
29i11r1	6	10	ψ13
23b5r1	1	89	ψ8
23b5r1	2	69	ψ12
23b5r1	3	66	ψ17
23b5r1	4	49	ψ4
23b5r1	5	36	ψ1
23b5r1	6	17	ψ4
23b5r1	7	12	ψ8
2a17r1	1	68	ψ8
2a17r1	2	58	ψ12
2a17r1	3	54	ψ12
2a17r1	4	53	ψ23
2a17r1	5	48	ψ4
2a17r1	6	38	ψ1
2a17r1	7	27	ψ8
2a17r1	8	20	ψ19
17m18r1	1	90	ψ4
17m18r1	2	68	ψ8
17m18r1	3	58	ψ3
17m18r1	4	48	ψ4
17m18r1	5	29	ψ8
17m18r1	6	22	ψ11
17m18r1	7	12	ψ25
17m18r1	8	12	ψ8
17m18r1	9	10	ψ19

Table 19: Interpretation of the clones from the riken1-library. The lengths of the fragments, which were identified in the clone sequence, and the pseudogenes from which they originated, are listed in this table.

clone	event	length	pseudogene
1c4r1	1	173	ψ12
1c4r1	2	88	ψ4
1c4r1	3	84	ψ23
1c4r1	4	45	ψ4
1c4r1	5	32	ψ13
1c4r1	6	15	ψ6 ψ11
1c4r1	7	11	ψ8
1c4r1	8	10	ψ9
30n12r1	1	141	ψ12
30n12r1	2	68	ψ11 ψ6
30n12r1	3	48	ψ4
30n12r1	4	45	ψ4
30n12r1	5	32	ψ13
30n12r1	6	29	ψ8
30n12r1	7	20	ψ19
30n12r1	8	12	ψ12
30n12r1	9	11	ψ8
30n12r1	10	10	ψ9
5g3r1	1	168	ψ4
5g3r1	2	82	ψ8
5g3r1	3	56	ψ13
5g3r1	4	23	ψ23
5g3r1	5	15	ψ17
5g3r1	6	12	ψ17 ψ22
30f11r1	1	129	ψ12
30f11r1	2	76	ψ11 ψ6
30f11r1	3	48	ψ4
30f11r1	4	47	ψ4
30f11r1	5	45	ψ4
30f11r1	6	37	ψ3
30f11r1	7	32	ψ13

Table 20: Interpretation of the clones from the dkf426-library - 1. The lengths of the fragments, which were identified in the clone sequence, and the pseudogenes from which they originated, are listed in this table.

clone	event	length	pseudogene
13c5r1	1	111	ψ6 ψ11
13c5r1	2	66	ψ4
13c5r1	3	52	ψ8
13c5r1	4	32	ψ11
13c5r1	5	31	ψ12
13c5r1	6	20	ψ11
15f1r1	1	269	ψ4
15f1r1	2	28	ψ13
15f1r1	3	15	ψ17
15f1r1	4	11	ψ19
15f1r1	5	10	ψ4
25o13r1	1	99	ψ12
25o13r1	2	43	ψ4
25o13r1	3	35	ψ8
25o13r1	4	32	ψ17
25o13r1	5	31	ψ21
25o13r1	6	31	ψ13
25o13r1	7	22	ψ11
25o13r1	8	15	ψ17
25o13r1	9	12	ψ25
25o13r1	10	10	ψ4
8c9r1	1	85	ψ4
8c9r1	2	67	ψ13
8c9r1	3	56	ψ8
8c9r1	4	18	ψ12
8c9r1	5	10	ψ8
8c9r1	6	10	ψ24(r)

Table 21: Interpretation of the clones from the dkf426-library - 2. The lengths of the fragments, which were identified in the clone sequence, and the pseudogenes from which they originated, are listed in this table.