Analysis of the Human T Cell Receptor α/δ Locus

New Approaches to Mapping and Sequencing

Thesis by

Cecilie Boysen

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1996

(Submitted May 30, 1996)

Acknowledgment

First, and most of all, I want to thank Søren for putting up with me during my graduate work. Living in three different places at the same time is not easy: Søren in Denmark, I in Seattle, and both of us in Pasadena, where we would stay a couple of months every now and then. We are very much looking forward to a new life together with little Junior, just one job each, and all in the same location.

Second, I want to thank Lee Hood, my advisor. I'm not sure, I'll ever forgive Lee for moving to Seattle. That being said, Lee made it possible for me to continue living in Pasadena part time, while doing my work in Seattle. If I had nowhere else to sleep while in Seattle, I could always stay with Valerie and him. Lee always had five minutes in the early morning hours to discuss your project, and if that was not enough, you could call him or come by his house anytime. Or if you could catch up with him on his morning run by Lake Washington then there was a chance to chat. Lee was always a great inspiration with his big views and new ideas. And thank you, Lee, for working so hard to help me the last couple of months.

I want to thank Mel Simon for providing me with lab and office space, when I was at Caltech, and for making sure I was not completely lost.

I would like to thank Eric Davidson for letting me use his sequencer and PCR machine anytime, and him and Ellen Rothenberg for always trying to encourage me to come and talk science. I should have used that opportunity more.

I have worked with so many over the past five years, and I want to thank all of you for useful discussion about my project, help with laboratory work, and for frequent chats. In particular Lee Rowen, who taught me about sequencing, especially the problems. Lee R. and I spent hours together discussing work, helping each other with experiments, or sometimes just to chat. Kai Wang, taught me about mapping, did not mind helping one in making clone libraries, and he showed me how to make pot stickers. Inyoul Lee in sharing

some of this project and its problems with me, and giving me a ride home every now and then. Debbie Nickerson for teaching me various lab techniques the first few years. Hiroaki Shizuya and Pat Charmley for useful discussions on BAC procedures and polymorphisms studies, respectively. The sequencing crew in Seattle, especially Steve Swartzell for helping me with a lot of the more tricky procedures. The computer group in Seattle for providing a lot of great programs, which made this work possible, and for trying to explain Unix to me. David Mathog at Caltech for setting me up to do computer work, whenever I was at Caltech. Tawny Biddulph for being patient with me, when I was desperate and needed to fax or talk to Lee, no matter where in the world he was. I also want to thank everybody in Seattle and at Caltech for always having a smile ready.

Finally, I want to thank Tim Hunkapiller for letting me stay in his house in Seattle and giving me rides to or from work or the airport.

Abstract

The human T cell receptor (TCR) α/δ locus has been mapped and sequenced. This region occupies roughly one megabase (Mb) of DNA or equivalent to one three thousandth of the entire human genome, the longest continuous piece of human DNA yet sequenced. The sequence has provided new insights into the complex organization, structure and evolution of two intermingled multigene families (α and δ), and will hopefully in the future help answer interesting questions concerning the complex expression patterns of TCR α and δ chains and about possible associations between specific polymorphisms in the TCR α/δ locus and susceptibility to autoimmune diseases. Comparison to cDNA data has provided information about expression of each of the TCR elements and about the striking diversification in the third hypervariable or junctional region. The sequence has contributed a glimpse of closely associated genomic DNA, in that the sequences surrounding the TCR locus, include the defender against death gene as well as five olfactory receptor genes. The sequence also harbors many other stretches of DNA, highly similar to previously identified genes, although in most cases, these have been found to be nonfunctional due to one or a few mutations. Comparison of 130 kilobases (kb) in the 3' region of the human sequence with its murine counterpart, suggests this region is highly conserved. The same 3' region has also been found to be limited in the concentration of genome wide repeats compared to the remainder of the locus. Furthermore, it contains a substantially reduced frequency of DNA variations compared to the rest of the locus. Apart from DNA variations in noncoding sequence, polymorphisms have also been identified in the coding regions of the TCR variable (V) gene segments, where, if they lead to amino acid changes, may alter the function of the TCR.

During the physical clone mapping and sequencing, new strategies were tested using primarily bacterial artificial chromosome (BAC) clones. These clones proved to be much more reliable and stable than clones currently employed in the human genome project

(e.g., cosmids and yeast artificial chromosomes, YACs). BAC inserts can be sequenced completely by the high redundancy shotgun approach. Their insert size, stability, and capacity to be easily sequenced suggests that BAC clones are excellent mapping and sequencing reagents. The ends of BAC clone inserts can be sequenced directly. This has led to the proposal of a new strategy for obtaining the entire DNA sequence of the human genome without physical mapping.

Table of Contents

Introduction

The major goals for the human genome project are to create detailed genetic and physical maps for the three billion basepairs (bps) of human DNA, and ultimately to determine the complete DNA sequence. This project is one of the largest ever undertaken in biology and involves many laboratories from all over the world. The results from this project will have an enormous impact in biology and medicine in the coming years. Every gene and regulatory element will be sequenced. Many sequences responsible for chromosomal structure and function will be identified. Genes predisposing to genetic diseases can be located and identified. This will lead to diagnostic tools and eventually cures or prevention of the disease. Evolution within a species as well as across species can be followed, since along with the human genome, the genomes from five model organisms will also be sequenced: E.coli, yeast, nematode, Drosophila, and mouse. This will also facilitate the identification of functions for many human genes, since they will have homologues in one or more of the model organisms, where experiments are easier and faster to perform.

One of the interesting challenges in biology is the analysis of complex systems. Examples include: interactions between cells in the nervous system; the well coordinated expression of a single kind amongst hundreds of olfactory receptors to choose from in each cell of the olfactory system, or the complicated processes occurring during an immune response. To learn more about these complex systems, one can study each involved component separately, which will give very detailed information about that component. However, to understand the coordination of the whole complex, one has to look at how the system functions.

This thesis focuses on a subsystem of the human immune system: the TCR α/δ locus. This multigene family plays a major role in the immune response. Analysis of the DNA sequence of this region will provide insights into the organization, structure and

evolution of this gene locus, as well as information on the details of the TCR elements that will facilitate understanding complex issues like regulation of their expression patterns. Knowledge of its sequence will provide powerful tools to explore how it as a system responds to various signals such as immunity, tolerization, and development. Furthermore, the entire sequence of this locus will provide information about specific chromosomal structures found in genomic DNA, and give an idea about the nature of noncoding DNA.

The T cell receptor

T cell receptor gene structure and organization:

T cells play a major role in the immune defense, where they are involved in the destruction of foreign invaders like bacteria or viruses. T cells recognize fragments from these foreign substances via their T cell receptors. The mechanism involved is outlined in Figure 1. The TCR recognizes small peptides from antigen. These peptides are presented to the T cell by receptors encoded by the major histocompatibility complex (MHC) on antigen presenting cells. Thus the TCR recognizes not only the peptide, but the MHC molecule as well (Davis and Bjorkman, 1988). Each individual T cell generally expresses only one kind of TCR. Considering the large number of different peptides that could be presented, and the necessary specificity of each TCR, millions of different TCRs are required. There are two types of heterodimeric TCRs, $\alpha\beta$ and $\gamma\delta$. Each of the four TCR subunits is divided into two domains, the outer variable (V) domain with the antigen/MHC recognition site and the constant (C) domain, which attaches the receptor to the cell membrane and transfers the signal from the V domain to the interior of the cell (Figure 1). The V domain is encoded for by two or more distinct germline gene segments: V, (diversity (D)), and joining (J) elements (Marrack and Kappler, 1990). Each V gene segment includes a promoter, a smaller first exon, which encodes the majority of the signal sequence, an intron, a second exon encoding the majority of the V domain, and finally

recombinational signals (Figure 2). For the α and γ V domain, one of many V gene segments rearranges during T cell development in the thymus to one of many J gene segments (Figure 2). The β and δ genes employ a third gene segment, the D element, thus joining V to D to J. These rearranged gene segments are contiguous and encode the V gene. The C domains are encoded by the C genes. RNA splicing joins the V and C genes to generate TCR mRNA.

Hence a multiplicity of V, (D), and J gene segments can be joined in a combinatorial fashion to generate considerable diversity. An even higher level of diversity is generated in this rearrangement. During the joining of the different gene segments, nucleotides are deleted from the ends of the gene segments, and other random nucleotides (N) are inserted (Lieber, 1992). This can generate tremendous diversity in the junctional region, which is thought to be the structure making contact to the antigenic peptide. Further TCR diversity is generated by the combinatorial joining of α and β chains (or γ and δ chains). One more way the TCR repertoire is expanded is through allelic variations of gene segments leading to amino acid changes in the TCR.

The β and γ loci are encoded by two distinct gene families. The δ region, however, is inserted between the Vα and Jα gene segments (Figure 2). The Vα and Vδ gene segments are found interspersed with one another. The majority of these rearrange to Jα gene segments forming TCR α chains, whereas three of them primarily rearrange to Dδ gene segments generating TCR δ chains. Some of the V gene segments have been found in connection with both Cα and Cδ. This phenomenon generates several interesting questions.

αβ or γδ T cell receptor lineages:

One of the major questions in T cell development, is the choice between αβ and γδ commitment and how this choice is regulated. The immature T cell from the bone marrow enters the thymus where it goes through a complicated developmental process, before it

either dies, as it happens to the majority of the T cells, or matures and leaves the thymus for the periphery (reviewed by Rothenberg, 1992). Amongst the many processes in T cell development, are the rearrangement and expression of the different TCR gene segments. On a timescale, TCRβ, δ, and γ rearrangements take place in a more immature T cell compared to the TCRα rearrangement (Pearse et al., 1989 and Held et al., 1990). Whereas productive γ and β rearrangement seems to block further rearrangement at their own loci, allelic exclusion (Borgulya et al., 1992), they do not block rearrangement at any of the other loci. In αβ T cells, the δ locus has in the majority of cases been deleted from both chromosomes due to Vα-Jα joining, effectively eliminating the genes for the γδ lineage.

Several, not mutually exclusive, hypotheses have been set forward to explain the choice between the two lineages. One hypothesis suggests that lineage commitment takes place prior to the rearrangement processes. It has been suggested that the T cells rearrange their genes either in the α or the δ loci, but not in both (Winoto and Baltimore, 1989). This could be a function of specific transcriptional regulation, and relies on the notion that transcriptional activity at a specific gene segment will render this V gene segment susceptible to rearrangement by the recombinase activity (Yancopoulos and Alt, 1985, and Schlissel and Baltimore, 1989). Several enhancers and silencers have been found that in combination with specific transcription factors or relying on specific promoters for the different classes of V gene segments could lead to transcription and thus possible rearrangement of a specific class of V gene segments (Winoto and Baltimore, 1989, and Lauzurica and Krangel, 1994a,b). De Villartay et al., 1988, suggest that specific elements (δrec and ψJα) located on opposite sides of the δ region could effectively delete the δ region if activated, thus ensuring the αβ lineage.

However, whereas these mechanisms may be functional, they do not exclude δ rearrangement from occurring even if the T cell was precommitted to the αβ lineage (Thompson et al., 1990). Many of the TCRαβ bearing cells have previously undergone rearrangement at the δ locus as suggested by studies of extrachromosomal excision

products of the δ locus (Takeshita et al., 1989) or studies involving transgenic mice expressing αβ genes (Nakajima et al., 1995). It is also possible, that rearrangement may occur at either loci in a single cell, and that the rearrangement process determines the lineage outcome. One explanation suggests, that if the TCR γ and δ gene segments, which rearrange earlier than the α gene segments, rearranged to form a functional TCR, this would signal the cell to commit to the γδ lineage, and prevent rearrangement at the α locus. As discussed below, this still does not explain, why the Vδ gene segments that are found interspersed with the Vα gene segments would be able to rearrange earlier than the Vα gene segments, and why they would rearrange to Dδ and not to Jα gene segments.

Intermingled Vα and Vδ gene segments in human:

In mouse, most of the Vδ gene segments have been found isolated at the 3' end of the V gene segments (Wang et al., 1994) (Figure 3). It is thus possible that enhancers acting over limited distances would be involved in regulatory expression and/or rearrangement of these V gene segments. However, the same does not hold true in human, where at least one V gene segment (Vδ1) found almost exclusively associated with δ chains, has been located far from the δ region in the midst of many Vα gene segments (Satyanarayana et al., 1988, Hata, et al., 1989, and Ibberson et al., 1995). Two other human Vδ gene segments have been found exclusively in δ chains, Vδ2 and Vδ3. Vδ2 is found just 5' to the Dδ's, at a position equivalent to Vδ1 of the mouse. The specific location and specific function of human Vδ2 (see below) are similar to the ones of mouse Vδ1, and it is therefore of interest, that they show very little sequence homology (Clark et al., 1995). The human Vδ3 element is found at the 3' side of the Cδ gene in an inverted orientation. The mouse Vδ5 element is in a similar location, and these two gene segments are clearly orthologs. Besides these three Vδ gene segments approximately five other V gene segments have been found to rearrange to both α and δ elements (Takihara et al., 1989, and Migone et al, 1995). The accurate locations of these five Vαδ gene segments

have not been determined, but they appear interspersed within the Vα gene segments. What are the factors that determine whether a V gene segment rearranges to the Dδ or Jα elements?.

Most of the V to C association studies described above have been done using cDNAs from T cells from either the thymus or the periphery. One could argue, that the studies involving circulating T cells, would not necessarily reflect all of the possible rearrangement patterns, both because of the selection process they have experienced in the thymus and because of possible antigen driven clonal expansion. Most of the T cells die in the thymus either because they possess TCRs that will not recognize their own MHC molecules and therefore would be of no value, or because they recognize their own MHC possibly with some endogenous antigenic peptide so strongly, that if not destroyed in the thymus they could cause autoimmune disease. Only T cells with an intermediate affinity for self MHC/peptide complexes will be positively selected and leave the thymus. If all the V gene segments had an equal chance for recombining to either Dδ or Jα, but no positive selection occurred for the majority of the V elements in combination with the Cδ element, one could explain why so few V gene segments were found in association with the δ elements in cDNA from circulating T cells. However, one should still observe them in the thymus, since here, the majority of T cells with recombined gene elements have not yet been through the selective processes. Thus thymus derived T cells should better reflect the rearrangement potentials. However, it does not appear to be selection that is responsible for the fact, that only very few V gene segments are found associated with the Dδ element. What has been observed in the cDNA studies does in fact appear to reflect the rearrangement of these V gene segments as shown by Migone et al., 1995. They looked at eighty γδ T cells, characterizing their rearranged V gene segments on both their productive and non-productive alleles, arguing that rearrangements on the non-productive allele should not have been subjected to selection or pairing with the γ chain. The same small set of V gene segments was found on either allele, indicating that other mechanisms are involved in

the preferential usage of only a restricted set of V gene segments in δ rearrangements. Hence perhaps differences in promoter structures of these V gene segments in combination with specific transcriptional factors and enhancers and/or silencers would promote preferential rearrangement of certain V gene segments to form δ chains. Alternatively, differences in the recombinational signals in either the V gene segments or in the Dδ versus Jα gene segments could restrict the possible recombination partners. Both of these possibilities have been investigated here by obtaining the sequence for both promoters and recombinational signals for all V gene segments.

Waves of γδ T cells during fetal development:

Another scientific challenge emerges from the differential expression patterns of the Vδ gene segments during fetal development of the thymus. In the mouse, combinations of different Vγ gene segments with one particular Vδ gene segment (mVδ1), are expressed at different stages during fetal development. The earliest expression of TCR can be detected at day 14 and is a combination of mVγ3 with mVδ1, whereas a day or two later the most dominant TCR is mVγ4 with mVδ1, etc. (reviewed in Allison and Havran, 1991). These particular populations seem to disappear from the thymus later in life, where more diverse usage of the different Vγ and Vδ gene segments is found. Interestingly, these same early T cells seem to home to different tissues depending on their expressed Vγ gene segment. They exhibit almost no junctional diversity correlating with the low expression of terminal deoxynucleotidyl transferase (TdT), and initially the mVδ1 gene segment is found to rearrange to only one specific Dδ gene segment, whereas adult TCRδ chains are found to routinely utilize two or even three Dδ gene segments. Almost the same phenomenon has been found in the developing thymus of human (Krangel et al., 1990, van der Stoep et al., 1990, and McVay et al., 1991). In human fetal thymus the earliest detected Vδ gene segment to be expressed on T cells is Vδ2. Later in life, the Vδ2 gene segment is rarely expressed, and Vδ1 and to a less extent Vδ3 become the predominant V gene segments

expressed in δ chains in the thymus. Occasionally, one of the few Vαδ gene segments is expressed. Furthermore, the Vδ2 element rearranges to one particular Dδ gene segment with little diversity generated in the junctional region. In contrast, the Vδ1 and Vδ3 chains generally include two or all three of the Dδ gene segments, and extensive nucleotide nibbling and N nucleotide insertion occur. Figure 3 shows a comparison of the 3' end of the mouse and human TCR α/δ loci, indicating the mVδ1 and hVδ2 as well as mVδ5 and hVδ3 occupy similar positions 3' to the other V genes. The fact that mVδ1 and hVδ2 are isolated from the other V gene segments could account for their unique expression patterns. A comparison of the DNA sequence in these regions from both mouse and human might reveal conserved (presumably regulatory) regions that could play a role in the selection of these particular V gene segments for early expression in the fetal thymus.

Location dependent recombination of Vα gene segments with Jα gene segments:

It has been shown in mice, that several successive recombination steps can take place involving Vα gene segments to Jα gene segments, and that these occur in a nonrandom fashion (Takeshita et al., 1989, Roth et al., 1991, Thompson et al., 1991, and Petrie et al., 1993). The tendency of these multiple rearrangements is more pronounced in the adult than the fetal stage. The Vα gene segments found proximal to the δ region were found to rearrange more frequently to the most upstream Jα gene segments, whereas the more 5' Vα gene segments were found to recombine with Jα elements further 3'. Examination of the deleted DNA in these later rearranging steps suggested that the locus had rearranged earlier, and that a subsequent rearrangement had taken place. This could be due to a nonproductive rearrangement leading to a V gene incapable of producing a functional chain. It is possible that expression of a functional α chain is necessary for providing a signal to stop rearrangement. If this is the case, a non functional rearrangement would not provide this signal. Presumably further rearrangements could take place to

attempt to make a functional receptor. However, some successfully rearranged Vα-Jα gene segments producing a functional α chain expressed on the cell surface, failed to prevent further rearrangement at this locus. It was therefore suggested that DNA rearrangement does not cease until positive selection has occurred. This view is consistent with high RAG expression in immature T cells which is downregulated after positive selection has taken place (Petrie et al., 1993). Accordingly, this phenomenon would provide the T cell with several chances for positive selection.

It has been difficult to study this mechanism in human, because the relative locations of the Vα gene segments are known only for a few of the V gene segments. However, with the detailed map and the final sequence, described in this thesis, one should be in a position to investigate this problem. Apart from the location of the Vα and Jα gene segments other factors need to be considered (e.g., distinct types of DNA rearrangement signals).

Evolution of the TCR α/δ locus:

TCR genes are found in many different species, including fish and birds (Rast et al., 1995). By comparing the organization of TCR loci across different species one can get insights into the evolution of these complex gene families. As an example, the organization of the chicken TCRβ region appears to consist of perhaps ten V gene segments, each with associated J and C elements (Kai Wang, personal communication). The γ locus in mouse contains several VJC clusters (Vernooij et al., 1993). It is possible that an ancestral TCR unit was composed of a single V, a single J, and a single C gene segment, and over time evolved via duplications either of the whole unit itself or of only the V and J gene segments. Whether the D gene segment was present in this proposed original TCR unit is unknown. It could have been deleted from the duplications / translocations forming α and γ loci, or alternatively inserted to generate more extensive diversity in the β and δ loci.

Approximately 100 kb of the mouse and human TCR α/δ loci have previously been sequenced (Koop et al., 1992, and Koop et al, 1994). This region extends from the Cδ to the Cα region encompassing all the Ja gene segments. Comparison between the two species revealed an extraordinarily high sequence similarity (70%) between the two species over the entire 100 kb. The coding region comprise only five percent of this sequence. The species comparison identified several previously unknown Jα elements in both species that had not been identified by cDNA analysis. Furthermore, sequences with high similarity are likely to have important biological functions, and thus comparison of sequences between species, can reveal new genes, regulatory elements, or sequences related to chromosomal structures or functions.

The sequence of the entire mouse TCR α/δ locus will soon be available, and the comparison of the human sequence described in this thesis with its mouse counterpart should illuminate several interesting issues. A few selected regions of the mouse α/δ region have been sequenced already, and are here compared to the human sequence. One of the interesting questions is how much further upstream does the extreme sequence conservation extend and what function does it serve?

Besides research into the organization and evolution across species, evolution within the species itself can be investigated. Duplications, gene conversions etc., can readily be detected once the entire sequence is known. The 45 known TCRα/δ gene segments have been divided into 35 subfamilies based on 75% or more sequence similarity of the members (Arden et al., 1995). The existence of multimembered subfamilies suggests that gene duplication has occurred. The entire sequence, including all V gene segments, pseudogenes and relics of V gene segments, is necessary to analyze this evolutionary history in depth.

Polymorphisms in TCR elements:

MHC molecules are among the most polymorphic structures known (Klein et al., 1983). A polymorphism is a DNA variation that exist in more than one percent of the population. Virtually every unrelated individual has MHC alleles distinct from all others. This diversity of MHC alleles probably arises from selection for the ability of the human population to bind and present many different kinds of peptide antigens, so as to ensure that there are always a few in the human population who can respond to new infectious agents. Even if MHC molecules bind peptide with little specificity, the few MHC genes in any one individual, will not be able to bind all peptides. TCR genes need not be as polymorphic as their MHC counterparts, because each individual can generate an enormously diverse TCR repertoire. Several studies have investigated the extent of polymorphisms in or around either the V or C elements, particularly for the TCR β locus (Robinson et al., 1987, Grier et al., 1990, Posnett, 1990, Rowen et al. 1996). However, most of the earlier studies where based on restriction fragment length polymorphisms, and thus did not indicate whether the variations were found in the gene segments themselves nor did they characterize the nature of these polymorphisms. In the last couple of years, studies focusing on polymorphisms in the V gene segments have indicated a high level of V polymorphisms (Cornelis et al., 1993, Moss et al., 1993, Reyburn et al., 1993, and Charmley et al., 1994). In these studies not all of the polymorphisms could be detected, even if they were present in the DNA analyzed, and furthermore only randomly selected V gene segments were investigated. To determine the level of variations found in the V gene segments of the α/δ locus, I analyzed the majority of the Vα/δ gene segments by direct sequence analysis.

These polymorphisms can be used in the study of possible associations of specific TCR gene segments and susceptibility to autoimmune diseases. Autoimmune diseases are believed to be a result of an individual's immune system attacking one's self. Susceptibility to certain autoimmune diseases has been correlated with specific MHC alleles (Oksenberg et al., 1988, Todd et al., 1988, and Sinha et al., 1990), whereas autoimmune

disease correlations to TCR elements have been controversial (Hashimoto et al., 1992, Hillert and Olerup, 1992, and Steinman et al., 1992). In a mouse model of multiple sclerosis, it was shown that disease induction was correlated with specific MHC alleles as appears to be the case in human. It was furthermore found that transferring T cells from sick donors into healthy mice of the same MHC type would transfer the disease, thus indicating that T cells can initiate the disease (Hood et al., 1989). T cells in the diseased mice were found to utilize only a few types of TCRs with limited V gene segment usage. Antibodies against these specific TCRs could either reverse the disease in the sick mice or prevent the induction of disease in susceptible mice (Zaller et al., 1990).

Attempts to determine whether associations exist between certain TCR polymorphisms and susceptibility to autoimmune diseases in humans have been hampered by the fact that very few genetic markers are available in the V gene segments regions. Moreover, in the human TCR β locus there appear to be multiple hotspots of recombination, isolating the V element population into small clusters or islands (Seboun et al., 1993). Hence a multiplicity of genetic markers are necessary, one for each island, if one is to rigorously look for V polymorphism associations with disease susceptibility. We want to identify a series of genetic markers across the TCR α/δ locus.

Multiple new markers generated across the TCR α/δ region should help in these association studies. With the entire sequence of this region, twenty or more useful microsatellites evenly spanning the entire locus can be identified and tested for polymorphisms. Hopefully each of these will lie in linkage disequilibrium to the adjacent genetic markers on either side.

Getting the DNA sequence for 1.07 Mb

Mapping:

Traditionally in the human genome project the first step in large scale sequencing is to create a low resolution physical map with large inserts of human DNA, typically yeast

artificial chromosome (YAC) clones (Chumakov et al., 1995, and Doggett et al., 1995). These are generally mapped in relation to one another by sequence tagged site (STS) content mapping. STS mapping identifies a unique sequence that is amplifiable by the polymerase chain reaction (PCR) using two specific primers. YAC clones, however, have many defects, which complicate the mapping process (Green et al., 1991). First, they are often chimeric, that is during the cloning process two or more pieces of DNA from two different chromosomal locations are co-ligated before integration into the YAC vector. This artifact generates confusion in mapping, especially in regions with low coverage of YAC clones. In these cases, where for example only one or two YACs extend across a region, the faithfulness of genomic representation must be carefully checked by other methods for each YAC. Many YACs rearrange during growth, and contain deletions which can be hard to detect, unless the coverage of YAC clones is high. Depending on the YAC library, the chimera rate varies, but in typical libraries it is approximately 40-50%. Another inconvenience is that YAC DNA cannot easily be separated from host (yeast) DNA.

I initiated this project attempting to use YAC clones to map the TCR region. However, a bacterial artificial chromosome (BAC) clone library was being constructed employing a new cloning vector containing large insert DNA (Shizuya et al., 1992). The initial studies concerning the stability of the BAC clones were promising. So I decided to map the TCR α/δ locus using the new BAC clone library. The clone insert sizes averaged 140 kb, which is smaller than YACs (100-1000 kb), but considerable larger than cosmids (35-40 kb), Table 1. I had several concerns about BAC clones. Would the distribution of BACs be random along the chromosomes? Would they be stable and faithfully represent the genomic DNA, and not be chimeric? If they were ideal for mapping purposes, they had other advantages compared to YACs. Large amounts of DNA can be prepared free of host (E.coli) chromosomal DNA. Furthermore, they might serve as good reagents for the sequencing step in this project.

Sequencing:

In the genome project, the most commonly used sequencing substrate has been the cosmid. Cosmids are sequenced by one of several methods. In random or shotgun sequencing, the cosmid is randomly cleaved into smaller fragments of 1-2 kb. These fragments are then inserted into an M13 or pUC-vector, and DNA prepared from 800-1000 of these clones is sequenced. These sequences are then assembled computationally into one or more sequence strings (contigs). This approach generates a redundancy of around 6-8 fold, that is, any stretch of DNA have been sequenced on average in 6-8 individual sequences. This usually generates one or a few contigs with gaps to close with more directed methods. This method is the most widely used today, because of its simplicity and high accuracy in the final consensus sequence (because of the high redundancy).

To obtain a high resolution or sequence ready cosmid map from the mapped YACs, however, is no small effort. The YAC clones can be used in one of two ways: they can be subcloned into cosmids, and cosmids specific for human DNA selected from the majority of yeast DNA containing cosmids. Alternatively, YAC clones can be used to group cosmids made from total genomic DNA or from specific chromosomes into smaller regions. This can be done by determining the cosmids content of STSs, which have previously been mapped to specific YAC clones. Once these cosmids are obtained, they need to be mapped relative to one another, and a set of cosmids spanning the YAC insert with minimal overlap is chosen for sequencing.

The work in going from YACs to cosmids is labor intensive, and many cosmids, up to 40%, contain different defects as mentioned above for the YAC clones (Lee Rowen, personal communication). To test whether BAC clones could be used in sequencing, I first subcloned one of the bigger BACs into cosmids to use in a low redundancy sequencing project (Roach et al., 1995). However, it would be a big improvement, if one could avoid this subcloning step. Subcloning YACs or BACs into cosmids and then mapping the

cosmids constitute an enormous bottleneck in sequencing efforts. So I tested the possibility of sequencing BACs directly using the random shotgun approach.

In the process of mapping and sequencing BAC clones, I also developed a method for obtaining the sequence information from the ends of the BAC inserts by direct sequencing on total BAC DNA. This and many other advantages of BAC clones has lead to a new proposal for sequencing the entire human genome (Venter et al., 1996). It circumvents the majority of the mapping, and is essentially based on simple, automatable sequencing procedures alone.

The next five chapters

Chapter 2: Here I describe in detail the mapping of the TCR α/δ locus using YAC, BAC, P1-based artificial chromosome (PAC), and cosmid clones. All the known Vα/δ gene segments are located onto these clones, and most of them have been ordered relative to each other. The region has also been characterized with respect to rare restriction enzyme sites.

Chapter 3: The DNA sequence of the TCR α/δ region is investigated in this chapter. All V gene segments, including pseudogenes, have been characterized with respect to their structure (promoter, exon-introns, recombinational signals, amino acid sequences) and compared against the large repertoire of α/δ cDNAs. The sequence has been examined for homologies with previously identified genes, proteins or ESTs, open reading frames, genome wide repeats, microsatellite-sequences, etc. Two shorter regions have been compared to their mouse counterparts.

Chapter 4: Sequence variations were identified in the TCR α/δ V gene segments. A direct sequencing method is outlined. The rate of variations and whether the variations would lead to amino acid changes is considered.

Chapter 5: This chapter summarizes all the excellent features of the BAC clones as mapping and sequencing reagents. It introduces a new simple approach based on BAC clones to obtain the DNA sequence of the entire human genome.

Chapter 6: A procedure to obtain sequence information from ends of BAC or PAC inserts is described. This procedure has simplified many of the steps involved in mapping.

It should be mentioned that different nomenclatures have been used for the V gene segments. Chapter 2 and 4 use the old nomenclature (Arden et al., 1995), whereas Chapter 3 introduces a new nomenclature based on the 5' to 3' order of the V gene segments. A conversion table is found in Chapter 3.

References:

Allison, J. P., and Havran, W. L.: The immunobiology of T cells with invariant γδ antigen receptors. Annu Rev Immunol 9: 679-705, 1991.

Arden, B., Clark, S.P., Kabelitz, D., and Mak, T.W.: Human T-cell receptor variable gene segment families. Immunogenetics 42: 455-500, 1995.

Borgulya, P., Kishi, H., Uematsu, Y., and Von Boehmer, H.: Exclusion and inclusion of α and β T cell receptor alleles. Cell 69: 529-537, 1992.

Charmley, P., Nickerson, D., and Hood, L.: Polymorphism detection and sequence analysis of human T-cell receptor Vα-chain-encoding gene segments. Immunogenetics 39: 138-145, 1994.

Chumakov, I. M., Rigault, P., Le-Gall, I., Bellann'e-Chantelot, C. Billault, A., Guillou, S., Soularue, P., Guasconi, G., Poullier, E., Gros-I., et al.: A YAC contig map of the human genome. Nature 377 (6547 Suppl): 175-297, 1995.

Clark, S.P., Arden, B., Kabelitz, D., and Mak, T.W.: Comparison of human and mouse T-cell receptor variable gene segment subfamilies. Immunogentics 42: 531-540, 1995.

Cornelis, F., Pile, K., Loveridge, J., Moss, P., Harding, R., Julier, C., and Bell, J.: Systematic study of human αβ T cell receptor V segments shows allelic variations resulting in a large number of distinct T cell receptor haplotypes. Eur. J. Immunol. 23: 1277-1283, 1993.

Davis, M. M., and Bjorkman, P. J.: T-cell antigen receptor genes and T-cell recognition. Nature 334: 395-402, 1988.

De Villartay, J.P, Hockett, R.D., Copran, D., Korsmeyer, S.J., and Cohen, D.I.: Deletion of the human T-cell receptor δ-gene by a site-specifc recombination. Nature 335: 170-174, 1988.

Doggett, N. A., Goodwin, L. A., Tesmer, J. G., Meincke, L. J., Bruce, D. C., Clark, L. M., Altherr, M. R., Ford, A. A., Chi, H. C., Marrone, B. L., et al.: An integrated physical map of human chromosome 16. Nature 377 (6547 Suppl): 335-65, 1995.

Green, E. D., Riethman, H. C., Dutchik, J. E., Olson, M. V.: Detection and characterization of chimeric yeast artificial-chromosome clones. Genomics 11: 658-69, 1991.

Grier, A. H., Mitchell, M. P., and Robinson, M. A.: Polymorphism in human T cell receptor alpha chain variable region genes. Expl. Clin. Immunogenetics 7: 34-42, 1990.

Hashimoto, L. L., Mak, T. W., and Ebers, G. C.: T cell receptor α chain polymorphisms in multiple sclerosis. J. Neuroimmunol. 40: 41-48, 1992.

Hata, S., Clabby, M., Devlin, P., Spits, H., De Vries, J. E., and Krangel, M. S. Diversity and organization of human T cell receptor δ variable gene segments. J Exp Med 169: 41-57, 1989.

Held, W., Mueller, C., and MacDonald, H.R.: Expression of T cell receptor genes in the thymus: localization of transcripts in situ and comparison of mature and immature subsets. Eur. J. Immunol. 20: 2133-2136, 1990.

Hillert, J. and Olerup, O.: Germ-line polymorphism of TCR genes and disease susceptibility- fact or hypothesis? Immunol. Today 13: 47-49, 1992.

Hood, L., Kumar, V., Osman, G., Beall., S.S., Gomez, C., Funkhouser, W., Kono, D.H., Nickerson, D., Zaller, D.M., and Urban, J.L.: Autoimmune disease and T-cell immunologic recognition. Cold Spring Harbor Symp. Quant. Biol. LIV: 859-874, 1989.

Ibberson, M. R., Copier, J. P., and So, A. K. Genomic organization of the human T-cell receptor variable α (TCRAV) cluster. Genomics 28: 131-139, 1995.

Klein, J., Figueroa, F, and Nagy, Z.A.: Genetics of the major histocompatibility complex - the final act. Ann. Rev. Immunol. 1: 119-142, 1983.

Koop,B. F., Wilson, R.K., Wang, K., Vernooij, B., Zaller, D., Kuo, C.L., Seto, D., Toda, M., and Hood, L.: Organization, structure and function of 95 kb of DNA spanning the murine T-cell receptor Cα/Cδ region. Genomics 13: 1209-1230, 1992.

Koop, B. F., Rowen, L., Wang, K., Kuo, C. L., Seto, D., Lenstra, J. A., Howard, S., Shan, W., Deshpande, P., and Hood, L.: The human T-cell receptor TCRAC/TCRDC (Cα/Cδ) region: Organization, sequence, and evolution of 97.6 kb of DNA. Genomics 19: 478-493, 1994.

Krangel, M.S., Yssel., H., Brocklehurst, C., and Spits, H.: A distinct wave of human T cell receptor γ/δ lymphocytes in the early fetal thymus: Evidence for controlled gene rearrangement and cytokine production. J. Exp. Med. 172: 847-859, 1990.

Lauzurica, P and Krangel, M.S.: Enhancer-dependent and in-dependent steps in the rearrangement of a human T cell receptor δ transgene. J. Exp. Med. 179: 43-55, 1994a.

Lauzurica, P and Krangel, M.S.: Temporal and lineage-specifc control of the T cell receptor α/δ gene rearrangement by T cell receptor a and d enhancers. J. Exp. Med. 1913-1921, 1994b.

Lieber, M. R.: The mechanism of V(D)J recombination: A balance of diversity, specificity, and stability. Cell 70: 873-876, 1992.

Marrack, P. and Kappler, J. W.: The T cell receptors. Chem. Immunol. 49: 69-81, 1990.

McVay, L., Carding, S.R., Bottomly, K., and Hayday, A.C.: Regulated expression and structure of T cell receptor γ/δ transcripts in human thymic ontogeny. EMBO 10: 83-91, 1991.

Migone, N., Padovan, S., Zappador, C., Giachino, C., Bottaro, M., Matullo, G., Carbonara, C., De Libero, G., and Casorati, G.: Restriction of the T-cell receptor V delta gene repertoire is due to preferential rearrangement and is independent of antigen selection. Immunogenetics 42: 323-332, 1995.

Moss, P. A. H., Rosenberg, W. M. C., Zintzaras, E., and Bell, J. I.: Characterization of the human T cell receptor α-chain repertoire and demonstration of a genetic influence on Vα usage. Eur. J. Immunol 23: 1153-1159, 1993.

Nakajima, P.B., Menetski, J.P., Roth, D.B., Gellert, M., and Bosma, M.J.: V-D-J rearrangements at the T cell receptor δ locus in mouse thymocytes of the αβ lineage. Immunity 3: 609-621, 1995.

Oksenberg, J. R., Gaiser, C. N., Cavalli-Sforza, L. L., and Steinman, L.: Polymorphic markers of human T-cell receptor alpha and beta genes. Family studies and comparison of frequencies in healthy individuals and patients with multiple sclerosis and myasthenia gravis. Human Immunol. 22: 111-121, 1988.

Pearse, M., Wu, L., Egerton, M., Wilson, A., Shortman, K., and Scollay, R.: A murine early thymocyte developmental sequence is marked by transient expression of the interleukin 2 receptor. Proc. Natl. Acad. Sci. USA 86:1614-1618, 1989.

Petrie, H.T., Livak, F., Schatz, D.G., Strasser, A., Crispe, I.N., and Shortman, K.: Multiple rearrangements in T cell receptor α chain genes maximize the production of useful thymocytes. J. Exp. Med. 178: 615-622, 1993.

Posnett, D. N.: Allelic variations of human TCR V gene products. Immunol. Today 11: 368-373, 1990.

Rast, J.P., Haire, R.N., Litman, R.T., Pross, S., and Litman, G.W.: Identification and characterization of T-cell antigen receptor-related genes in phylogenetically diverse vertebrate species. Immunogenetics 42: 204-212, 1995.

Reyburn, H., Cornélis, F., Russell, V., Harding, R., Moss, P., and Bell, J.: Allelic polymorphism of human T-cell receptor V alpha gene segments. Immunogenetics 38: 287-291, 1993.

Roach, J.C., Boysen, C., Wang, K., and Hood, L.: Pairwise end sequencing: a unified approach to genomic mapping and sequencing. Genomics 26: 345-353, 1995.

Robinson, M. A. and Kindt, T. J.: Genetic recombination within the human T-cell receptor α-chain gene complex. Proc Natl Acad Sci USA 84: 9089-9093, 1987.

Roth, M.E., Holman, P.O., and Kranz, D.M.: Nonrandom use of Jα gene segments: Influence of Vα and Jα gene location. J. Immunol. 147: 1075-1081, 1991.

Rothenberg, E. V.: The development of functionally responsive T cells. Advances Immunol. 51: 85-214, 1992.

Rowen, L., Koop, B.F., and Hood, L.: The complete 685 kilobase DNA sequence of the human beta T cell receptor locus. Science *(in press)*, 1996.

Satyanarayana, K., Hata, S., Devlin, P., Roncarolo, M. G., De Vries, J. E., Spits, H., Strominger, J. L., and Krangel, M. S. Genomic organization of the human T-cell antigen-receptor α/δ locus. Proc Natl Acad Sci USA 85: 8166-8170, 1988.

Schlissel, M. S., and Baltimore, D.: Activation of immunoglobulin kappa gene rearrangement correlates with induction of germline kappa gene transcription. Cell 58: 1001-1007, 1989.

Seboun, E., Houghton, L., Hatem Jr., S.J., Lincoln, R., and Hauser, S.L.: Unusual organization of the human T-cell receptor β-chain gene complex is linked to recombination hotspots. Proc Natl Acad Sci USA 90: 5026-5029, 1993.

Shizuya, H., Birren, B., Kim, U.-J., Mancino, V., Slepak, T., Tachiiri, Y, and Simon, M.: Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proc Natl Acad Sci USA 89: 8794-8797, 1992.

Sinha, A. A., Lopez, M. T., and McDevitt, H. O.: Autoimmune diseases: The failure of self tolerance. Science 248: 1380-1388, 1990.

Steinman, L., Oksenberg, J. R., and Bernard, C. C. A.: Association of susceptibility to multiple sclerosis with TCR genes. Immunol. Today 13: 49-51, 1992.

van der Stoep, N., de Krijger, R., Bruining, J., Koning, F., and van der Elsen, P.: Analysis of early fetal T-cell receptor δ chains in humans. Immunogenetics 32:331-336, 1990.

Takeshita, S., Toda, M., and Yamagishi, H.: Excision products of the T cell receptor gene support a prgrassive rearrangement model of the α/δ locus. EMBO 8: 3261-3270, 1989.

Takihara, Y, Reimann, J., Michalopoulos, E., Ciccone, E., Moretta, L., and Mak, T.W.: Diversity and structure of human T cell receptor δ chain genes in pheripheral blood γ/δ-bearing T lymphocytes. J. Exp. Med. 169: 393-405, 1989.

Thompson, S.D., Pelkonen, j., and Hurwitz, J.L.: Concomitant T-cell receptor α and δ gene rearrangements in individuals T-cell precursors. Proc. Natl. Acad. Sci. USA 87: 5583-5586, 1990.

Thompson, S.D., Manzo, A.R., Pelkonen, J., Larche, M., and Hurwitz, J.: Developmental T cell receptor gene rearrangements: releatedness of the α/β and γ/δ T cell precursor. Eur. J. Immunol. 21: 1939-1950, 1991.

Todd, J.A., Acha-Orbea, H., Bell, J. I., Chao, N., Fronek, Z., Jacob, C. O., McDermott, M., Sinha, A. A., Timmerman, L., Steinmann, L., McDevitt, H. O.: A molecular basis for MHC class II-associated autoimmunity. Science 240: 1003-1009, 1988.

Venter, J.C., Smith, H.O., and Hood, L.: A new cooperative strategy for sequencing the human and other genomes. Submitted, 1996.

Wang. K., Klotz, J. L., Kiser, G. Bristol, G., Hays, E., Lai, E., Gese, E., Kronenberg, M., and Hood, L.: Organization of the V gene segments in mouse T-cell antigen receptor α/δ locus. Genomics 20, 419-428, 1994.

Winoto, A. and Baltimore, D.: Developmental regulation of the TCRαδ locus. Cold Spring Harbor Symp. Quant. Biol. LIV: 87-92, 1989.

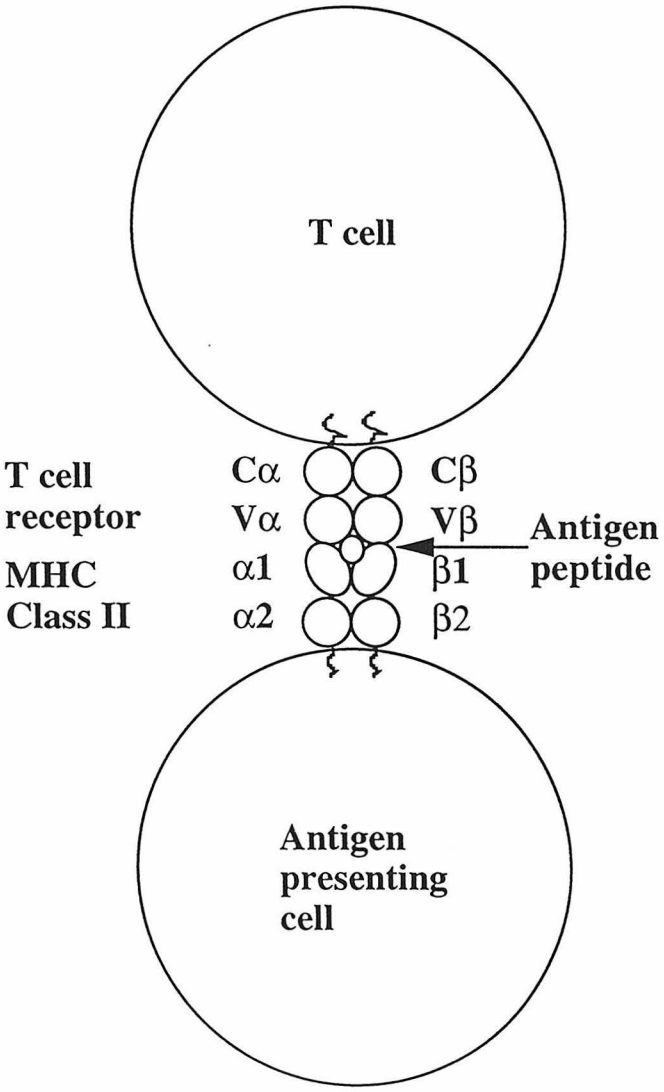
Yancopoulus, G., and Alt, F.: Developmentally controlled and tissue-specific expression of unrearranged V_H gene segments. Cell 40: 271- 281, 1985.

Zaller, D.M., Osman, G., Kanagawa, O., and Hood, L.: Prevention and treatment of murine experimental allergic encephalomyelitis with T cell recepotr Vβ-specific antibodies. J. Exp. Med. 171: 1943-1955, 1990.

Table 1. Insert sizes of clones most commonly used in the human genome project.

Clone	Insert size
YAC	100-1000 kb
BAC	50-300 kb
PAC	50-300 kb
P-1	80-90 kb
Cosmid	35-40 kb
Plasmid	<15 kb
M13	<4 kb

Figure 1. Schematic of antigen recognition by the T cell receptor. Antigen is processed to small peptides inside the antigen presenting cell. These peptides are presented on the surface by MHC molecules. This dual structure is recognized by the V domains of the heterodimeric TCR. A signal is send to the interior of the T cell through the C domain, which anchors the TCR to the cell.



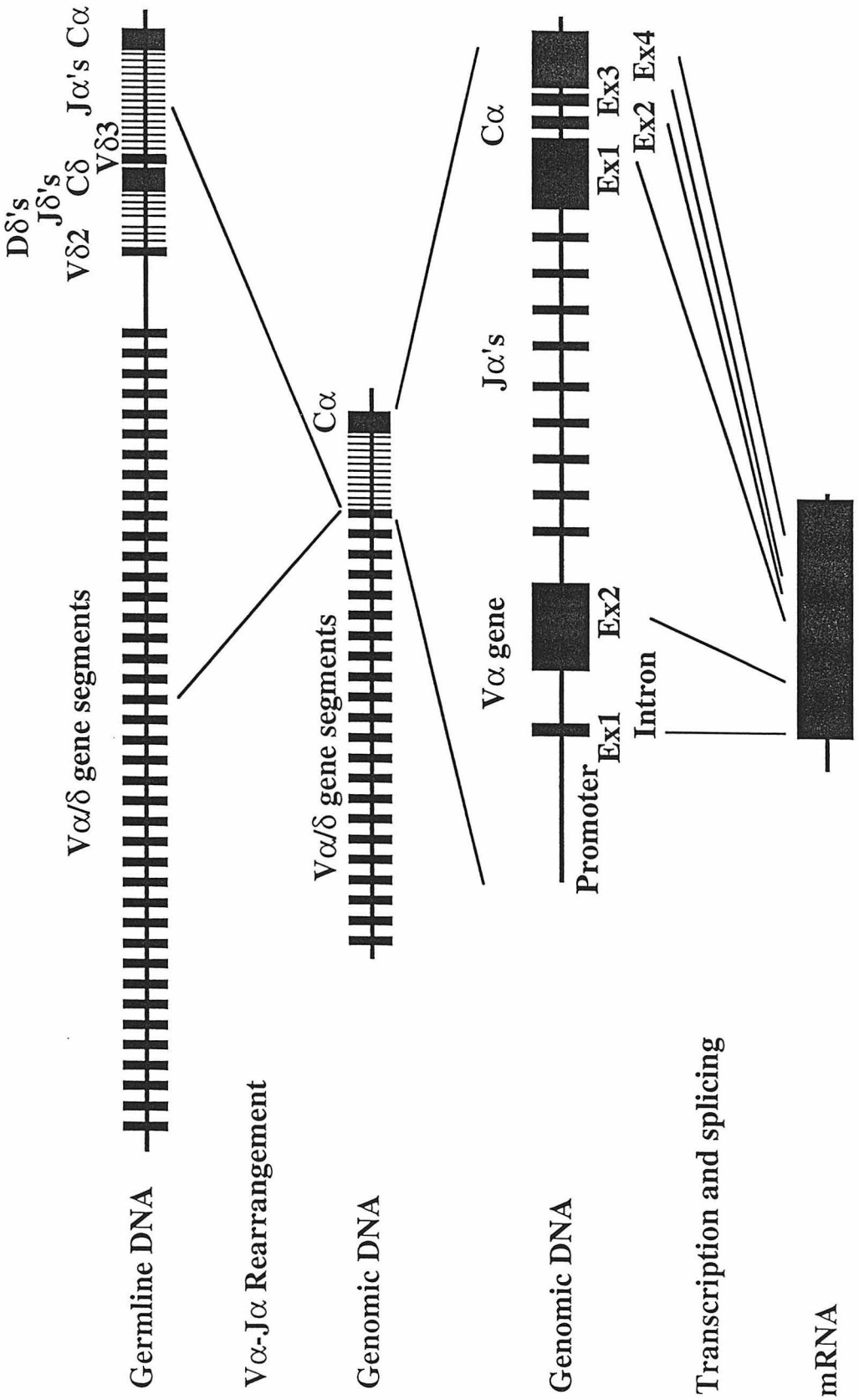


Figure 2. Organization of the human TCR α/δ locus. DNA rearrangement of a Vα gene segment to a Jα gene segment is shown, followed by transcription and splicing of the exons and the Cα gene to generate mRNA.

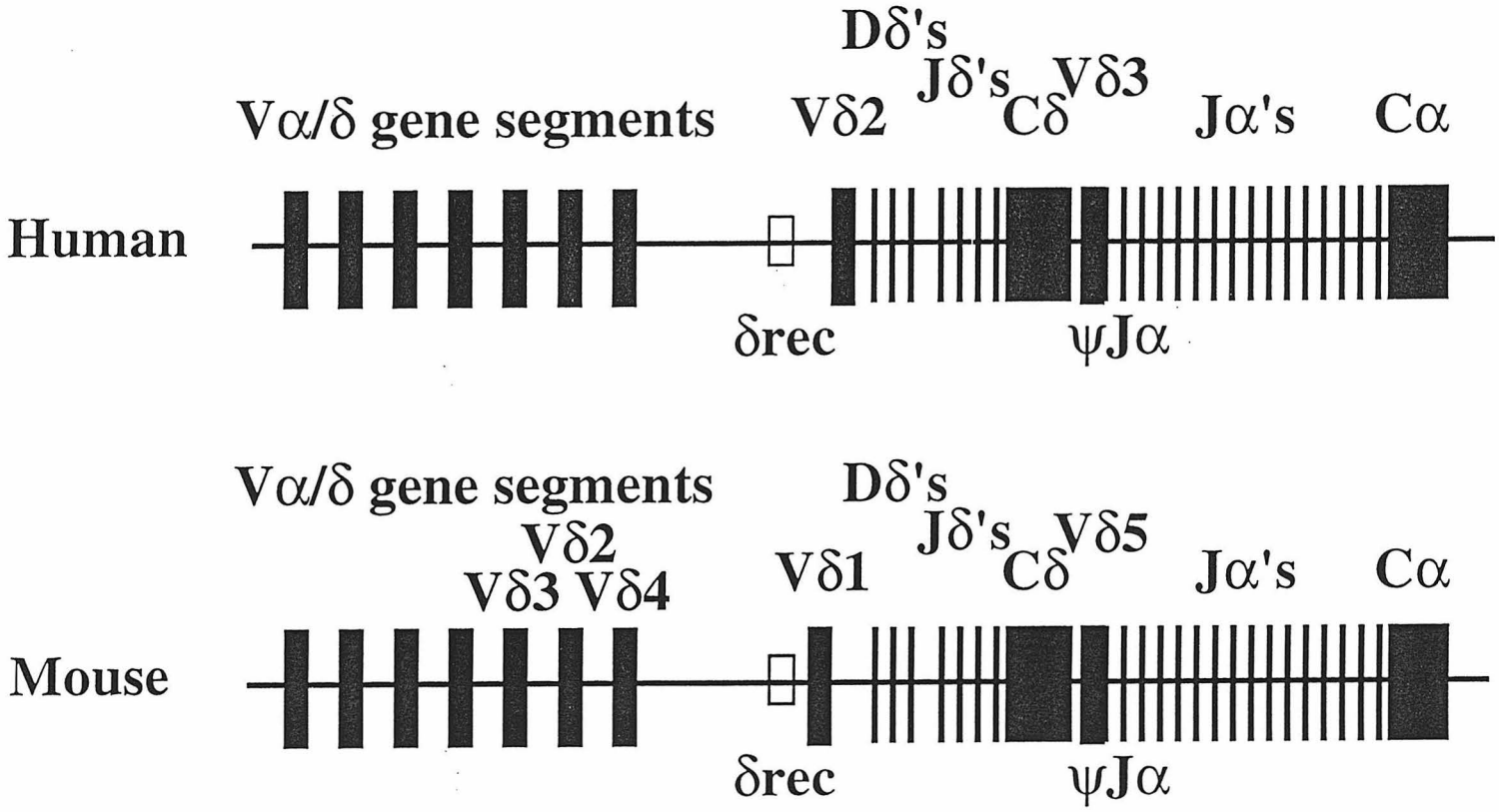


Figure 3. Comparison of the 3' end of the human and mouse TCR α/δ loci.

Organization of the human T cell receptor α/δ variable gene segments determined by bacterial, P1-derived, and yeast artificial chromosome, and cosmid mapping

Running title: Organization of the human T cell receptor α/δ locus

Cecilie Boysen†, Kai Wang≠‡, Ung-Jin Kim†, Melvin I. Simon†, Leroy Hood≠

† Division of Biology 147-75, California Institute of Technology, Pasadena, California 91125

≠ Department of Molecular Biotechnology 357730, University of Washington, Seattle, Washington 98195

‡ Current address: Darwin Molecular Corp., 1631 220th St. SE, Bothell, WA 98021

Correspondence to Leroy Hood

ABSTRACT

A physical map of the human α/δ T cell receptor locus, spanning ~one megabase (Mb), has been constructed from yeast artificial chromosome (YAC), bacterial artificial chromosome (BAC), P1-derived artificial chromosome (PAC), and cosmid inserts. Fifty-one variable (V) gene segments have been mapped with respect to one another and with respect to several rare cutting restriction enzymes by hybridization and PCR analyses. The 3' region including the δrec rearrangement element, the diversity (D), joining (J), and constant (C) elements have been organized. BACs appear to be excellent mapping reagents.

INTRODUCTION

T cells play a major role in the immune response against bacterial or viral infections. The T cell's specificity is determined by its T cell receptor (TCR), which recognizes antigenic peptide embeded in major histocompatibility complex (MHC) class I or class II molecules on the surface of antigen presenting cells (Davis and Bjorkman, 1988). Mammals have two types of heterodimeric T cell receptors, α/β and γ/δ (Marrack and Kappler, 1990). The TCR polypeptides are encoded by distinct gene families, α/δ, β and γ. The α and δ polypeptides are divided into a variable (antigen recognition) and constant (attached to the cell surface) regions. The Vα region is encoded by a multiplicity of Vα and Jα gene segments--one each of which undergoes DNA rearrangement and joining during T cell differentiation to generate a Vα gene (Lieber, 1992). Likewise, the Vδ region is encoded by Vδ, Dδ, and Jδ gene segments that also rearrange during development to create a Vδ gene. The Cα and Cδ regions are encoded by distinct Cα and Cδ genes. The α and δ nuclear RNA transcripts are spliced, joining the V and C genes, to form the mature mRNA. The V gene segments can be divided into discrete subfamilies whose members share 75% or more homology. Thirty-five subfamilies have been identified in the human α/δ locus

each with one to five members (Arden et al., 1995). Some V gene segments appear to be associated only with the Cδ gene (Vδ1, Vδ2, and Vδ3), whereas five or six others may join with either C gene (the remainder appear only to join with the Cα gene). An element that joins to Jα gene segments at the DNA level and deletes all of the intervening DNA is designated δrec (De Villartay et al., 1988 and Hockett et al., 1988). Its physiological role, if any, is uncertain.

Several types of studies have contributed to our current understanding of the human α/δ T cell locus. (*i*) About 250 α or δ cDNA sequences have been determined. These data suggest that there are about 45 different Vα or Vδ gene segments, although they provide no positional information. (*ii*) Some chromosomal DNA clones have been sequenced. For example, the 97 kilobases (kb) of DNA encompassing the 3' end of the locus has been sequenced (Koop et al., 1994a) and reveals the following order of gene segments: 5' Jδ-Jδ-Cδ-Vδ3-$(Jα)_{61}$-Cα 3'. (*iii*) The relative order of V gene segments can be determined in homogeneous T cell lines or tumors by deletional analyses. Most T cells must rearrange both the maternal and paternal chromosomes to get a functional Vα or Vδ gene. Accordingly, in any individual T cell line or tumor, the 3' most rearranged V gene separates all other V genes into two classes: those that are 3' to the rearranged V gene are deleted in both chromosomes; and those that are 5' to the rearranged V gene are present on one or both chromosomal copies. If multiple T cells are analyzed, a deletional map of the analyzed gene segment order can be determined. Several deletional maps of differing resolution have been determined for the human α/δ locus (Wilson et al., 1988 and Ibberson et al., 1995). (*iv*) Cleavage by rare cutting restriction enzymes produce long DNA fragments that can be separated by pulsed field gel electrophoresis (PFGE) and analyzed with different V probes by Southern blot analyses (Griesser et al., 1988, Satyanarayana et al., 1988, Hockett et al., 1988, and Ibberson et al., 1995). These studies suggest that the human α/δ T cell receptor locus is approximately one Mb, and the relative order of some of the V gene segments can be identified. However, determination of the detailed organization of this

region is limited by the number of T cells analyzed and/or the number of V gene probes used in these studies.

As a prelude to the sequence analysis of the human α/δ T cell receptor locus, we have developed a detailed physical map of the α/δ locus employing YAC, BAC, PAC, and cosmid clones. In addition, we have developed a highly detailed map of the order of the V gene segments.

MATERIALS AND METHODS

DNA Sources

YAC clones were obtained from the St. Louis' human genomic DNA YAC library. This library was constructed from a lymphoblastoid cell line, CGM-1 (Brownstein et al., 1989). BAC clones were obtained from a human BAC library at California Institute of Technology. This library was developed from a normal human male fibroblast cell line, (ATCC: CRL 1905: CCD-978Sk) (Shizuya et al., 1992). This cell line was also used in PFGE analysis of human genomic DNA. The PAC library at Genome Systems, Inc. was constructed from a normal human male fibroblast cell line, HSF7 (Ioannou et al., 1994).

Screening of Genomic YAC, BAC, and PAC Libraries

Human TCRα/δ specific YAC clones were obtained from St. Louis by PCR screening using primer pairs specific for a few Vα and the Cδ gene segments (Table 1).

To obtain specific BACs we used PCR amplified Vα/δ gene segments as probes. These were labeled with P-32 using a random labeling approach (T7 QuickPrime, Pharmacia, or Multiprime DNA Labeling System, Amersham) and hybridized overnight at 65°C to the BAC library membranes in SET (0.6 M NaCl, 0.02 M EDTA, 0.2 M Tris-HCl [pH 8.0], 2% SDS, and 0.1% pyrophosphate). The membranes were washed 10 minutes in 1 x SSC + 0.1% SDS, followed by 2-3 washes in 0.1 x SSC + 0.1% SDS at 65°C for

10-20 minutes each. Positive clones were identified after exposure at -70°C to Kodak X-AR film with intensifying screen overnight or longer. Specific PCR-probes were also made for the ends of different clones, and the PCR product labeled and used as above. Whole cosmids and BACs were also used as probes in hybridization to the BAC library. Cosmid and BAC DNAs were digested with NotI to separate the vector from the inserts, and run on a PFGE (see below). The inserts were cut out of the gel and the DNA extracted from the agarose using beads (Sephaglas BandPrep, Pharmacia or Qiaex, Qiagen). When using P-32 labeled cosmids or BACs as probes, cold vector DNA, human Cot-1 or total placental DNA, and total E.*coli* DNA were used to suppress hybridization of repeat sequences and contaminating vector and E.*coli* DNA.

The PAC library was screened as described above for the BAC library using specific V gene segments or PCR products generated from ends of BACs.

DNA Preparation

Total human genomic DNA from the same cell line used to make the BAC library was prepared in low melting point (LMP) agarose. Cells were washed twice in phosphate buffered saline (PBS) and resuspended to 10^8 cells/ml in PBS. The cells were then warmed to 37°C before they were mixed with an equal volume of melted 1% LMP-agarose and poured into molds. The solidified plugs were incubated overnight at 50°C in a solution of 0.5 M EDTA [pH 9.0], 1% Sarcosyl, and Proteinase K (0.5 mg/ml). This step was repeated once for one more overnight incubation. The plugs were then rinsed and stored in 0.5 M EDTA.

YAC DNA was prepared in LMP agarose. Yeast cells were lysed for 30 minutes at 37°C in SCE (1 M sorbitol, 0.1 M trisodium citrate [pH 7.0], and 50 mM EDTA [pH 8.0]) and Yeast Lytic Enzyme. The cell lysates were then mixed with an equal volume of 1% LMP-agarose in 125 mM EDTA [pH 8.0] at 37°C. The cell suspension was poured into molds and allowed to solidify. The DNA plugs were placed in a buffer (0.5 M EDTA [pH

9.0], 10 mM Tris-HCl [pH 8.0], and 50 mM DTT) at 37°C overnight, followed by washing in 0.5 M EDTA and 10 mM Tris-HCl. The plugs were then incubated at 50°C overnight in (0.5 M EDTA [pH 9.0], 10 mM Tris-HCl [pH 8.0], 1% Sarcosyl, and Proteinase K, 1 mg/ml). Finally, the blocks were treated with 100 µg/ml RNase A in TE (10 mM Tris-HCl, 1 mM EDTA), before storing at 4°C in 0.5 M EDTA, 10 mM Tris-HCl.

BAC, PAC and cosmid DNA was prepared using standard alkaline lysis procedures (Sambrook et al., 1989). Minipreparations were made either by hand without organic extractions or by an automated minipreparation machine, Autogen 740, Integrated Separation Systems.

Construction of Cosmid Libraries from YACs

Before partial digestion with SauIIIA (New England Biolabs), ten 1 mm slices of YAC DNA in LMP agarose were dialyzed in double distilled (dd) H_20 overnight. The plugs were melted and SauIIIA buffer added to 1x according to manufacturers instructions. The tubes were equilibrated to 37°C before addition of 1 Unit β-agarase I (New England Biolabs) and varying amounts of SauIIIA (0.01 - 0.1 Units / 200 µl reaction). Incubation was continued for 30 minutes after which the reactions were terminated by addition of EDTA to a final concentration of 50 mM. The DNA was pooled and loaded onto a gradient of 10 - 50 % sucrose in TE + 1 M NaCl. The gradients were spun in an ultracentrifuge for 20 hours at 25.000 rpm. Half ml fractions were collected and tested on a 0.3 % agarose gel for size. The 40-50 kb fractions were precipitated, and resuspended in 20 µl of TE. The size selected DNA were used in ligation to the BamHI digested, dephosphorylated cosmid vector, pWE15A (Lai et al, 1991). The ligation mix was packaged and transformed into DH5αMCR cells (Gibco, BRL) using Stratagenes Gigapack II Gold Packaging Extract. The cells were spread on nylon membranes on agar plates containing ampicillin, and the colonies grown overnight, before replicas were taken. The replicas were used to make membranes for library screening. Cosmid clones with human insert were initially

identified by hybridization to a P-32 labeled Alu probe, and later by probes made from V gene segments or cosmid ends.

Southern Blots and Hybridizations

DNA was digested with various restriction enzymes according to the suggestions of the manufacturer. The DNA was run in a 0.8% agarose gel, and transferred to nylon membranes by capillary action using 0.4 N NaOH. The membranes were rinsed twice in 2 X SSC before use. PFGE gels were irradiated at 254 nm ultraviolet light for 45 seconds in the presence of ethidium bromide before blotting. The blots were prehybridized in hybridization solution (50% formamide, 5 X SSC, 0.02 M sodium phosphate [pH 6.7], 100 μg/ml denatured salmon sperm DNA, 1% SDS, 0.5% nonfat dry milk, and 10% dextran sulfate) at least half an hour prior to hybridization. As above, probes were labeled and hybridized at 65°C overnight followed by washing, although the washing conditions would vary in their concentration of SSC from 0.1 X SSC to 2 X SSC, dependent on the desired stringency.

End Sequencing of YAC, BAC, and Cosmid inserts

Sequence from the ends of YAC and BAC inserts were obtained using the vectorette or bubble technique as described in Riley et al., 1990, or by sequencing Alu-vector PCR products (Nelson et al., 1991). Cosmid, and later BAC ends were obtained by direct sequencing, either radioactive or fluorescent analyses (Boysen et al., 1996a).

Pulsed Field Gel Electrophoresis

Large DNA molecules were separated in 1% agarose in 0.5 X TBE at 14°C using a variety of devices for PFGE, either homemade or from Biorad. Voltage applied was 6 V/cm, switch times and total time depended on the sizes separated (Birren and Lai, 1993).

RESULTS

YAC Map

Specific PCR assays for the Vα9, Vα12, Vα13, and Cδ elements (Table 1) were used to screen the St. Louis YAC library. Nine YAC clones were obtained, ranging in size from 180 kb to 350 kb as determined by PFGE and hybridization using pUC19 DNA as probe (data not shown). One of the YAC clones contained two YACs. Southern blot analyses of YAC DNAs digested with several different restriction enzymes (EcoRI, BamHI, HindIII, and PstI) and probed with Vα/δ gene segments were used together with specific PCR assays to determine the V gene segments present in each YAC clone. Five of the YAC clones were chimeric based on their paucity of V gene segments in conjunction with their total length. The ends of the remaining three YAC clones were sequenced. Specific PCR assays were generated and used to amplify a human chromosome specific/hamster hybrid panel to ensure that the ends of the YACs both were on chromosome 14. These three YAC clones appear to faithfully represent the genomic DNA. Two of these, YAC234D11G1 and YAC116B220D9, make a contig of about 400 kb covering more than half of the V gene segments (Figure 1). The third clone, YAC190A232G1, which was obtained using the primers specific for the Cδ gene, covers the most 3' V gene segments through to the Cδ gene (Figure 1).

Cosmid Map

To identify more precisely the location of the V gene segments and to provide substrates for DNA sequencing, the three faithful YAC clones were partially digested with SauIIIA and subcloned into cosmid vectors. Human specific cosmid clones were identified by screening the YAC-cosmid libraries with an Alu specific probe under conditions of low stringency. Cosmid contig maps of each YAC clone were constructed based on restriction enzyme analyses and V gene segment hybridization (Figure 1). Cosmid clones were digested with several different restriction enzymes and the sizes of the resulting fragments

compared. Southern blots were made from the same gels and used in hybridization with an Alu-repeat specific probe, or probes specific for V gene segments (e.g., Figure 2). Initial cosmid overlaps were determined from restriction enzyme analyses and V probe hybridizations. Further contig building was done using STSs for V gene segments (Figure 3). Finally, end-sequences were obtained from the cosmids at the ends of contigs, and used either to make new STSs or as probes to check for overlap between contigs. This approach resulted in cosmid contigs spanning all three YACs (Figure 1).

BAC and PAC Clone Map of the α/δ locus

The human BAC library at Caltech was initially screened when it included 2.5-fold coverage of the human genome. It was screened with three V probes missing from the YAC clones (Vα7.1, Vα16, and Vα24). Five BAC clones were identified (the 5' most clones in Figure 1 with numbers less than 600). Subsequently, the BAC library was screened with cosmids spaced every 50 kb from the cosmid map described above. Several additional clones were obtained (clone numbers under 600). When the Caltech BAC library reached a 3.7-fold coverage, it was screened a last time using some of the BAC clones obtained previously for the 5' and 3' ends under conditions where the hybridization of repeat sequences was suppressed with human placental or Cot-1 DNA. The BAC clones were analyzed by restriction enzyme (HindIII, EcoRI, and BamHI or PstI) mapping (Figure 4), hybridization against individual V probes or whole BAC DNA probes (Figure 5), and STS content mapping. End sequences from BAC clones were obtained using either the vectorette technique (Riley et al., 1990), the Alu-vector technique (Nelson et al, 1991), or in some cases by sequencing the BAC DNA directly with chain terminators (Boysen et al., 1996a). All of the known V gene segments could be mapped to these BAC clones. One gap remained at the 5' end of the locus between the BAC135 and the BAC10 clones (Figure 1). A PAC library was screened with probes of Vα gene segments located close to the gap and probes generated from the ends of these BACs. Several positive PAC clones

were obtained, one of which (PAC161) closed the gap. We also obtained PAC230 to cover the region around a very unstable cosmid (cosmid 9.8 in Figure 1).

Pulsed Field Gel Map of the α/δ locus

Human genomic DNA as well as YAC, BAC, and PAC DNAs were digested with the rare cutting enzymes NotI, SalI, SfiI, and BssHII. The resulting fragments were separated by PFGE. The sizes of the BAC and PAC fragments were determined by comparison with standards on ethidium bromide stained gels. The BAC and PAC insert lengths ranged from 85 to 240 kb, as determined after excision from the vector by NotI. Fragment sizes from human genomic and YAC DNAs had to be determined after the generation of Southern blots and hybridization with appropriate probes (Figure 6). The rare restriction enzyme sites shown in Figure 1 are identified from the BAC and YAC data Predictions from these data for the fragment sizes for human genomic DNA did not always match the actual length obtained directly from human genomic blots. This is probably due to methylation of CpG dinucleotides because some of the longer genomic fragments could be accounted for by addition of two or more of the fragment sizes determined by BAC and YAC DNAs.

Approximate sizes of the rare restriction site fragments are given in Figure 1. As shown, all of the V gene segments can be found on one of four contiguous SfiI fragments of 500 kb, 190 kb, 180, and 175 kb, 5' to 3'. The enzyme BssHII gave inconclusive results for the four 5' most BAC clones where it seems to cut several times. For that reason, it has not been included in Figure 1. However, one BssHII site was unambiguously determined to be right next to the Sfi I-3 site, although the orientation of these sites has not been determined since the fragment sizes are too similar.

Location of Human Vα/δ Gene Segments

Most of the V gene segments could be uniquely located relative to one another by inspection of their hybridization patterns (Figure 2) or via V-specific PCR analysis of the different BAC, PAC, and cosmid clones (Figure 3). In Figure 1 where the V subfamily members have been uniquely identified by specific PCR analysis or DNA sequencing, the subfamily number is given (e.g., 14.1, 14.2, etc.). When the subfamily sites but not specific numbers are identified (e.g., hybridization with one Vα1 member and washing under low stringency conditions), the subfamily number is followed by a letter (e.g., 1a, 1b, 1c, etc.) starting from the 3' end. Only three Vδ gene segments (Vδ1-Vδ3) are given on the map. The other known Vδ gene segments (Vδ4, Vδ5, Vδ6, Vδ7, and Vδ8) have also been identified as Vα gene segments (Vα6, Vα21, Vα17, Vα28, and Vα14.1, respectively), and are indicated with stars in Figure 1. Thus, most V gene segments capable of rearranging to Dδ gene segments are scattered among the Vα gene segments and are concentrated toward the 3' end of the V gene segment cluster (Figure 1). The Vδ2 and Vδ3 gene segments lie 3' to all other V gene segments and 3' to the Cδ gene, respectively. The deletional element, δrec, lies between the SalI-4 site and the Vδ2 gene segment (Figure 1).

Where detailed cosmid maps are available, almost all of the V gene segments could be ordered relative to each other. The few V gene segments that could not be ordered are aligned in vertical arrays. In regions with BAC and PAC clone coverage alone, the ordering of the gene segments relies on overlapping BAC and/or PAC clones.

To locate the V gene segments with respect to the rare restriction enzyme sites, the V gene segments were used as probes in hybridizations with the PFGE blots. These analyses suggest Vα9 is just 5' to the SfiI-2 site; Vα3 and Vα12 lie between the SfiI-2 and SalI-3 sites, and Vα30 lies just 3' to the SalI-3 site (Figures 1 and 6). The Vα14.2, Vα14.1, Vα27, Vα31, and Vα19 gene segments were all found 5' to the SfiI-4 and SalI-4

sites, whereas the Vδ2 element was found 3' to these two restriction enzyme sites, and thus is the 3' most V gene segment upstream of the Dδ, Jδ and Cδ elements.

DISCUSSION

Complete physical map of α/δ locus.

The BAC, cosmid, and PAC clones appear to provide a complete physical map of the α/δ locus. All known V gene segments have been placed on the map, either by hybridization or PCR analyses, as have the δrec, Cδ, and Cα elements. The clones in this map overlap completely from 5' to 3'. Size estimates from the restriction maps employing rare cutting enzymes (NotI, SalI, and SfiI) and a summing of the clone insert sizes suggest that the α/δ locus is approximately 1 Mb in length.

BAC clones appear to be excellent mapping reagents.

The 17 BAC clones selected from 2.4- to 3.7-fold human library appear to span the 1 Mb locus, but for a single gap (covered by a PAC clone). Detailed restriction map analyses and end sequencing suggest that BAC clones are rarely chimeric or rearranged (e.g., incur deletions) (Boysen et al., 1996b). Furthermore, it appears that BAC clones can be sequenced directly by shotgun analyses (C. Boysen, unpublished).

All known V gene segments have been mapped across the α/δ locus.

The majority of the 51 V gene segments have been ordered with respect to one another, except for 16 V elements that are assigned to five bins with two to six members in each of these bins. The total of 51 V gene segments detected by hybridization corresponds well to the reported numbers of 45 cDNAs (Arden et al., 1995). Since not all of the V gene segments identified by hybridization have been sequenced, we can not be certain of precise correspondences. Additional V gene segments, particularly pseudogenes may emerge as

the locus is sequenced. For example, the human β T cell receptor locus, spanning ~700 kb, has 65 V gene segments, 19 of which are pseudogenes (Rowen et al., 1996). The V gene segments are distributed more or less evenly across the locus except at the 3' end where the Vδ2 gene segment lies 100 kb 3' to the Vα19 gene segment (Figure 1). The ~750 kb region between the Vα19 and Vα7.1 gene segments contains on average, a V gene segment every 15 kb. This is somewhat less densely packed than the human β (Rowen et al., 1996) or the murine α/δ (Wang et al., 1994) T cell receptor loci. For example, the mouse α/δ locus contains 86 V gene segments over approximately 900 kb (1 per ~10 kb) (Wang et al., 1994). This increased V density arises from the duplication (45 to 80 kb homology units) of more densely packed V gene segment regions. The low resolution physical map of the human α/δ locus suggests, that smaller regions at the 5' end including the Vα1, Vα2, and Vα8 subfamilies may have arisen as a result of duplication.

Seven of the 35 V subfamilies are multi-membered (Vα1, Vα2, Vα4, Vα7, Vα8, Vα14, and Vα22). For three of the two-membered subfamilies, Vα4, Vα7, and Vα14, their two members have been uniquely identified, and so have two of the three members of Vα8 (Figure 1). The Vα22 and Vα8 subfamilies show two and three members, respectively, by Southern blot analysis. However, only one Vα22 and two Vα8 members have been identified by cDNA sequence analysis. The additional member in each case could be a pseudogene, as sequence analysis has not been done. Likewise, for the Vα2 and Vα1 subfamilies, we have identified by Southern blot analysis four and eight members, respectively, while only three different Vα2 and five Vα1 cDNAs were reported earlier (Arden et al., 1995). Complete sequencing of this region will indicate if the extra V gene segments we detect by hybridization are functional.

The detailed locations of V gene segments reported here are, for the most part, in agreement with a recently published map of this region. Ibberson et al., 1995, described the rough organization of subfamilies 1-24 based on PFGE analysis and deletional mapping. Most of the single member families are positioned identically, although it is more

difficult to compare the multi-member families, in part because of the lower resolution of their map. The sizes of the SfiI fragments are approximately the same, and these are also in agreement with earlier PFGE studies of this region (Hockett et al., 1988, Satyanarayana et al., 1988, and Hata et al., 1989). Not all of the V gene segments have been located to the same fragments, nor are they in the same order. For example, the Vα4 members have been localized to different SfiI fragments in the two studies. In addition, the Vα10 and Vα3 members are located in different positions with respect to the other V gene segments. Detailed DNA sequence analyses now underway will resolve these discrepancies.

A comparison of the mouse and human α/δ loci reveals a striking similarity at the 3' ends of these families.

The 3' α/δ regions in both human and mouse are remarkably similar from the Cα gene through to the Vδ2 element, a region spanning approximately 130 kb of DNA. In the ~100 kb of DNA sequenced in both species encompassing the region from the Cδ to the Cα genes, a striking 71% homology is seen even if the coding regions only occupy 5% of the sequence. Each of the 61 Jα gene segments has an orthologue in the opposite species in the same 5' to 3' order (Koop and Hood, 1994). This DNA sequence similarity seems to extend 5' through the δ region (Rowen and Boysen, unpublished). In the human β T cell receptor locus, the 70 kb region between the 3' most V gene segment and the Dβ1 gene segment contains five tandemly arrayed trypsinogen genes (Rowen et al., 1996). It will be interesting to determine whether the 100 kb counterpart region in the human α/δ locus (e.g., between Vα19 and Vδ2) also contains non-T cell receptor genes.

The α and δ gene elements are intermingled.

Eight different V elements have been found associated with the Cδ gene in various cDNA clones (Migone et al., 1995). Five of these V gene segments can also be associated with the Cα gene. The Vδ1, Vδ2, and Vδ3 gene segments are always associated with the

Cδ gene. The Vδ1 element is approximately 350 kb 5' to the Cδ gene and five V elements associated with both the Cα and Cδ genes are scattered across much of the V element region (Figure 1). The Vδ2 and Vδ3 gene segments are the 3' most V elements, separated from the others by more than 100 kb.

Hence, most V elements associate only with the Cα gene; five associate with both Cα and Cδ; and three associate only with the Cδ gene. There are two general explanations for these selective associations. (*i*) Any V element may rearrange to either C gene; antigen driven selection may permit only the associations described above to be clonally expanded at the functional T cell level. (*ii*) The DNA rearrangement sequences lying to the 3' side of the V gene segments may limit the rearrangement process to the associations described above.

It is worth stressing again that the 3' region from the Vδ2 element to the Cα gene is highly conserved across the human and mouse evolutionary lines (Koop and Hood, 1994). In contrast, most of the mouse Vδ-specific elements are located at the 3' end of the V gene segment cluster (Wang et al., 1994). It is interesting that δrec typically rearranges to a pseudo J element just downstream from Vδ3. Accordingly, this rearrangement ensures that the Vδ2, Vδ3, all Dδ, all Jδ and the Cδ elements are deleted--thus permitting that chromosome to only express successfully rearranged Vα genes (de Villartay et al., 1988).

SUMMARY

A detailed physical map of the human α/δ locus extending over more than 1 Mb has permitted the identification and localization of all known V gene segments. These studies suggest that BAC clones are excellent physical mapping reagents. The complete DNA sequence analysis of the α/δ locus is now underway.

ACKNOWLEDGMENT

We would like to thank Deborah Nickerson for providing oligo nucleotides and Hillary Massa, Cynthia Freeman, and Barbara Trask for *in situ* hybridizations. This work was supported by a grant from the Department of Energy (DOE).

REFERENCES

Arden, B., Clark, S. P., Kabelitz, D., and Mak, T. W.: Human T-cell receptor variable gene segments families. Immunogenetics 42: 455-500, 1995.

Birren, B. and Lai, E.: Pulsed field gel electrophoresis: A practical guide. Academic Press, Inc., 1993.

Boysen, C., Simon, M. I., and Hood, L.: Fluorescent sequencing directly from bacterial or P1-derived artificial chromosomes. Submitted, 1996a.

Boysen, C., Simon, M. I., and Hood, L.: The use of bacterial artificial chromosomes (BACs) as mapping and sequencing reagents. Submitted, 1996b.

Brownstein, B. H., Silverman, G. A., Little, R. D., Burke, D. T., Korsmeyer, S. J., Schlessinger, D., and Olson, M. V.: Isolation of single-copy human genes from a library of yeast artificial chromosome clones. Science 244, 1348-1351, 1989.

Davis, M. M., and Bjorkman, P. J.: T-cell antigen receptor genes and T-cell recognition. Nature 334: 395-402, 1988.

De Villartay, J.-P., Hockett, R. D., Coran, D., Korsmeyer, S. J., and Cohen, D. I.: Deletion of the human T-cell receptor δ-gene by a site specific recombination. Nature 335: 170-174, 1988.

Griesser, H., Champagne E., Tkachuk D., Takihara, Y., Lalande, M., Baillie, E., Minden, M., and Mak, T.W.: The human T cell receptor α-δ locus: a physical map of the variable, joining and constant region genes. Eur J Immunol 18: 641-644, 1988.

Hata, S., Clabby, M., Devlin, P., Spits, H., De Vries, J. E., and Krangel, M. S.: Diversity and organization of human T cell receptor δ variable gene segments. J Exp Med 169: 41-57, 1989.

Hockett, R. D., De Villartay, J.-P., Pollock, K., Poplack, D. G., Cohen, D. I., and Korsmeyer, S. J.: Human T-cell antigen receptor (TCR) δ-chain locus and elements responsible for its deletion are within the TCR α-chain locus. Proc Natl Acad Sci USA 85: 9694-9698, 1988.

Ibberson, M. R., Copier, J. P., and So, A. K.: Genomic organization of the human T-cell receptor variable α (TCRAV) cluster. Genomics 28: 131-139, 1995.

Ioannou, P. A., Amemiya, C. T., Garnes, J., Kroisel, P. M., Shizuya, H., Chen, C., Batzer, M. A., and de Jong, P. J.: A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. Nature Genetics 6: 84-89, 1994.

Koop, B. F., Rowen, L., Wang, K., Kuo, C. L., Seto, D., Lenstra, J. A., Howard, S., Shan, W., Deshpande, P., and Hood, L.: The human T-cell receptor TCRAC/TCRDC (Cα/Cδ) region: Organization, sequence, and evolution of 97.6 kb of DNA. Genomics 19: 478-493, 1994a.

Koop, B. F., and Hood, L.: Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. Nature Genetics 7, 48-53, 1994.

Lai, E., Wang, K., Avdalovic, N., and Hood, L.: Rapid restriction map constructions using a modified PWE15 cosmid vector and a robotic workstation. Biotechniques 11: 212-217, 1991.

Lieber, M. R. The mechanism of V(D)J recombination.: A balance of diversity, specificity, and stability. Cell 70: 873-876, 1992.

Marrack, P. and Kappler, J. W. The T cell receptors.: Chem Immunol 49: 69-81, 1990.

Migone, N., Padovan, S., Zappador, C., Giachino, C, Bottaro, M., Matullo, G., Carbonara, C., De Libero, G., and Casorati, G.: Restriction of the T-cell receptor V delta gene repertoire is due to preferential rearrangement and is independent of antigen selection. Immunogenetics 42: 323-332, 1995.

Nelson D. L., Ballabio, A., Victoria, M. F., Pieretti, M., Bies, R. D., Gibbs, R. A., Maley, J. A., Chinault, A. C., Webster, T. D., and Caskey, C. T.: Alu-primed polymerase chain reaction for regional assignment of 110 yeast artificial chromosome clones from the human X chromosome: Identification of clones associated with a disease locus. Proc Natl Acad Sci USA 88: 6157-6161, 1991.

Riley, J., Butler, R., Ogilvie, D., Finniear, R., Jenner, D., Powell, S., Anand, R., Smith, J. C., and Markham, A. F.: A novel, rapid method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones. Nucleic Acids Res 18: 2887-2890, 1990.

Rowen, L., Koop, B. F., and Hood, L.: The complete 685 kb sequence of the human beta T cell receptor locus. Science (*in press*), 1996.

Sambrook, J., Fristsch, E. F., and Maniatis, T.: Molecular cloning: A laboratory manual. (Second Edition). Editors: N. Ford, C. Nolan, and M. Ferguson. Cold Spring Harbor Laboratory Press. Chapter 1, 1.25-1.28, 1989.

Satyanarayana, K., Hata, S., Devlin, P., Roncarolo, M. G., De Vries, J. E., Spits, H., Strominger, J. L., and Krangel, M. S.: Genomic organization of the human T-cell antigen-receptor α/δ locus. Proc Natl Acad Sci USA 85: 8166-8170, 1988.

Shizuya, H., Birren, B., Kim, U.-J., Mancino, V., Slepak, T., Tachiiri, Y, and Simon, M.: Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proc Natl Acad Sci USA 89:8794-8797, 1992.

Wilson, R. K., Lai, E., Concannon, P., Barth, R. K., and Hood, L. E.: Structure, organization and polymorphism of murine and human T-cell receptor α and β chain gene families. Immunol. Reviews 101, 149-171, 1988.

Wang. K., Klotz, J. L., Kiser, G. Bristol, G., Hays, E., Lai, E., Gese, E., Kronenberg, M., and Hood, L.: Organization of the V gene segments in mouse T-cell antigen receptor α/δ locus. Genomics 20, 419-428, 1994.

Table 1. Oligonucleotide sequences and annealing temperatures for specific amplification of TCR elements. Most oligonucleotide pairs were used for both specific PCR assays as well as to generate probes for use in hybridization.

Gene segment	5' primer	3' primer	Annealing temp
TCRDV101S1	AAGGTTACTCAAGCCCAGTC	CTGTAAGGCTGAAATGGTTAAG	60°C
TCRDV102S1	TTGGTGCCTGGACACCAAAC	GATGGTGCAAGTATCTTAAGTA	60°C
TCRDV103S1	CAGAGTTCCCCGGACCAGAC	CCTTACTGGAGAGATCACCA	60°C
TCRDC	AGCCTCATACCAAACCATCCG	CACTTCAAAGTCAGTGGAGTGCAC	60°C
DREC	TAAGATCCTCAAGGGTCGAG	TGTGCTGGCATCAGAGTGTG	60°C
TCRAV1S1	TCATACCAGTGCTGGGGA	GGCACAGAAGTACTCAGCT	58°C
TCRAV2S1	TGAAATCCTTGAGAGTTTTACTA	GAGAAACATACTGGCTGG	60°C
TCRAV2S2*	GGCATCTCTGTAGAAACATA	GTCTCTGATGAACAAGGAGAT	60°C
TCRAV3S1	CTGGGAGTGTCTTTGGTGATT	AAGGAACTGCTTTTCTTGGAAG	60°C
TCRAV4S1	TGAAGTTGGTGACAAGCATTACT	AGGTACTGGACTTTCTGT	60°C
TCRAV4S2*	GGCTGGTGGCAAGAGTAACTG	GCGTAGCGTGGGGCAGGA	60°C
TCRAV5S1	GGAGACGAATGGAGTCATCC	GGCTGTGATATGAAACAAACTC	58°C
TCRAV6S1	GTCACTTTCTAGCCTGCTGA	CAGTTGTGAAGCGGAGATGACA	60°C
TCRAV7S1	GGGGAGCTTTCCTTCTCTATG	ATCTGGAGCTCCTGTAGAAGG	60°C
TCRAV7S2*	GTGGGGAGTTTTCCTTCTT	TTCATCTGGAGCTCCTTCAA	58°C
TCRAV8S1	ACATCCATTCGAGCTGTATTTATAT	TGTGATCTGCAGGGAGAAATGT	60°C
TCRAV8S2*	ATTCGAGCTTTATTTATGTACTTGT	AATTTGCAGAGAGAGATGTTTCA	60°C
TCRAV9S1	ATGAAGCCCACCCTCATCT	TGAGTCTTCCTCTTGAGCAA	58°C
TCRAV10S1	GTCCTGAAATTCTCCGTGTCCA	AGTGATGTGGAGAGAACTGTC	60°C
TCRAV11S1	ATGGCTTTGCAGAGCACTCTGG	CTGGAGGATGAGCAGCGATG	60°C
TCRAV12S1	GCCAGCCTGTTGAGGGCAG	TACTGCTGAGTCCACGACT	60°C
TCRAV13S1	GGACCTCTGCTGGGGCTC	TGTGGTCTGGGAAGAGGAA	60°C
TCRAV14S1	TTCCTGTGGGCACTTGTGA	CACAGAAATACATCGCGGCA	60°C
TCRAV14S2*	GCTGTGGGCAGTCGTGG	GCACAGAAATACATCGCAGTG	60°C
TCRAV15S1	GAGGATGTGGAGCAGAGTT	CAGTCTGGGTGTCTGCAATG	60°C
TCRAV16S1	GCCTCTGCACCCATCTCGA	CAAAGCGGAGTCGCTCACAA	58°C
TCRAV17S1*	TAGTTCTGTGGCTTCAACTATG	AGTCTCCAGGCTGGGAATCCA	60°C
TCRAV18S1	GAATCCTTTGGCAGCCCCATTA	AAATAGCTGTAACCCTCCTTGG	60°C
TCRAV19S1	AGATCCGGCAATTTTTGTTGGCT	GGGAGCTGTGCTTTTCCTGTA	60°C
TCRAV20S1	GGCAAGTGGCGAGAGTGATC	AGTGTCGCTCAGGGAAACCCG	60°C
TCRAV21S1	GTGGAAGGACATGAATAAAGCAC	TCCAGGCTGGGAGGGCACA	60°C
TCRAV22S1	AGGCTTAGTATCTCTGATACTC	ACACCGCTGAGTCTGACACT	59°C
TCRAV23S1*	GTCTAAGTGACAGAAGGAATG	AATGTATAAAGTACTACGTCCTGA	60°C
TCRAV24S1*	GCATCTGACGACCTTCTTGGT	ATGTAGGAGGCTGAATCGCTGAG	60°C
TCRAV25S1	ATGCTCCTTGAACATTTATTA	GTAGATGCCTACATCACTAGG	60°C
TCRAV26S1	GAGACTGTTCTGCAAGTACTCCTA	AGGTAGATGCCTGCATGGCTGG	60°C
TCRAV27S1	ATGAAGAAGCTACTAGCAATG	GTAGGTGGCAGAGAGGTCATG	60°C
TCRAV28S1	ATGATGAAGTGTCCACAGGCT	GGTAGACGGCCGAGTCTCCGG	60°C
TCRAV29S1	ATGGAGACTCTCCTGAAAGTGC	AAGTAGGTTCCTGAGTAACTG	60°C
TCRAV30S1	AGAAGTGGCGCCTCTGAG	CACAGAGATAAGTGGCTGAG	60°C
TCRAV31S1	TAACAGTGATGCCCTCTG	CTGAGTCTGATGCCTGGACTG	60°C
TCRAV32S1	AACTTCGATGCACCTCTTTCC	GCTTCTTCACTTCTCCACTC	60°C
TCRAC	CTTGAAGCTGGGAGTGG	CTAAGAGAGCCGTACTGG	60°C

*) Oligonucleotides were only used in specific PCR assays.

FIGURE LEGENDS

Figure 1. Map of the human TCR α/δ locus. The TCR elements were located on the map via hybridization to Southern blots of restricted YAC, BAC and cosmid DNAs, or via specific PCR assays of YAC, BAC, PAC and cosmid DNAs (See Table 1). V gene segments in larger subfamilies which were only identified by hybridization have a letter affixed to their subfamily name. V gene segments without unique location have been written in a column with other V gene segments binned to that region. Rare restriction enzyme sites have been determined for NotI, SalI, and SfiI. V gene segments which have been found in cDNAs of δ chains are indicated by *.

Figure 2. Hybridization of V gene segment probe to cosmids. DNAs from cosmids developed from YAC234 were digested with EcoRI and the fragments separated in a 0.8 % agarose gel. Southern blots were made and hybridized to a Vα23 probe. A band were identified around 4.0 kb in several cosmids, indicating the location of this V gene segment.

Figure 3. STS content mapping of cosmids. Specific primers for Vα4.2 (Table 1) were used in amplification of DNA prepared from cosmids derived from YAC234. M: 123 bp marker. G: Genomic DNA.

Figure 4. Restriction mapping of BAC DNAs. DNA prepared from BACs were digested with HindIII and run on 0.8 % agarose gel. HindIII cuts the insert out, leaving a vector band of 6.8 kb for clones with numbers below 420, and 7.5 kb for clones with numbers above. M: 1 kb ladder.

Figure 5. BAC to BAC hybridization. BAC363 DNA was used as a probe in hybridization to a Southern blot of EcoRI digested BAC DNAs.

Figure 6. V gene segments ordered with respect to rare restriction enzymes. YAC DNAs were digested with NotI, SalI, SfiI, and BssHII, and run on PFGE. Southern blots were made and used in hybridization to specific V gene segment probes. YAC116 included three controls (÷) to check for degradation of DNA under different conditions without enzyme. Indicated sizes are estimated by comparison to the original ethidium bromide stained agarose gel that included DNA size markers (lambda ladder and yeast chromosomes).

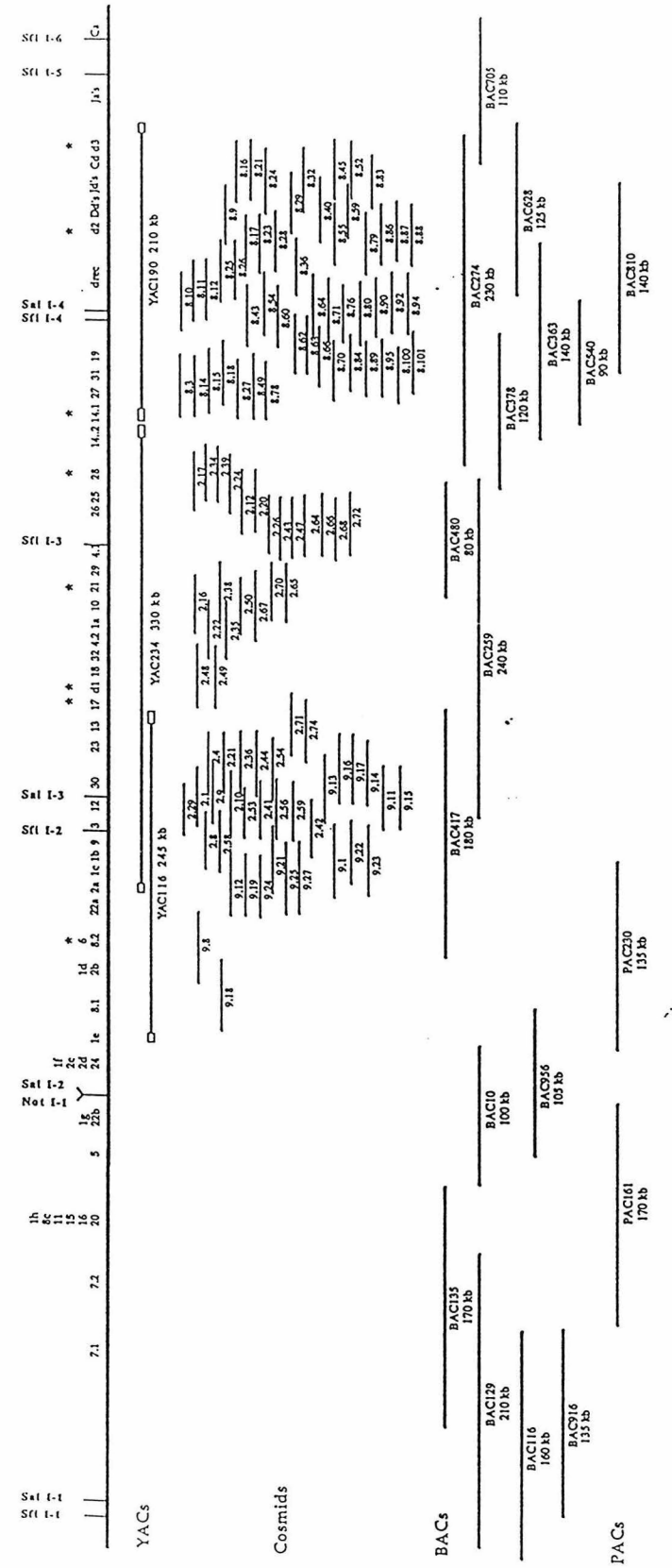

YACs
Cosmids
BACs
PACs
YAC116 245 kb
YAC234 330 kb
YAC190 210 kb
BAC417 180 kb
BAC480 80 kb
BAC259 240 kb
BAC274 230 kb
BAC278 120 kb
BAC628 125 kb
BAC705 110 kb
BAC363 140 kb
BAC540 90 kb
BAC810 140 kb
BAC135 170 kb
BAC129 210 kb
BAC116 160 kb
BAC916 135 kb
BAC10 100 kb
BAC956 105 kb
PAC161 170 kb
PAC230 135 kb
100 kb

Figure 1.

Figure 2.

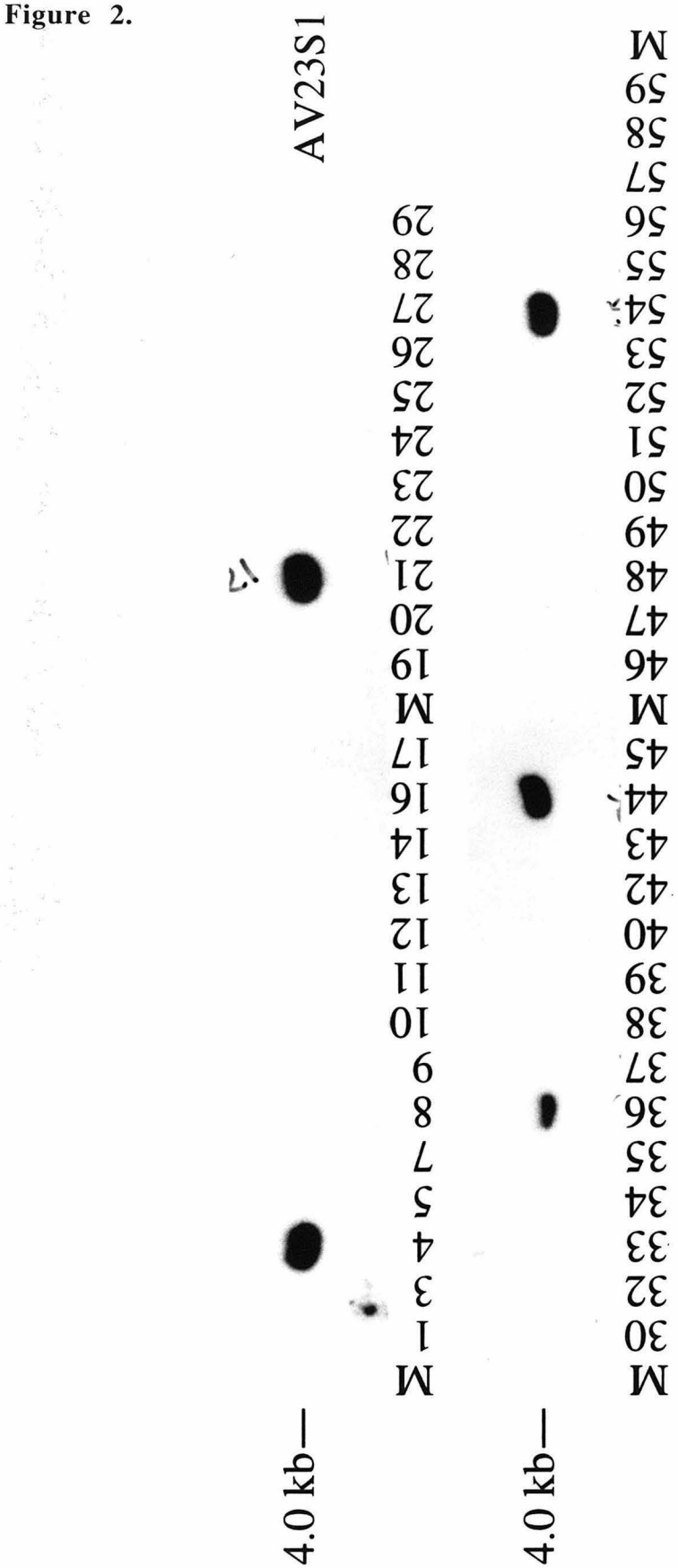

AV23S1
4.0 kb
4.0 kb
M 1 3 4 5 7 8 9 10 11 12 13 14 16 17 M 19 20 21 22 23 24 25 26 27 28 29
M 30 32 33 34 35 36 37 38 39 40 42 43 44 45 M 46 47 48 49 50 51 52 53 54 55 56 57 58 59 M

Figure 3.

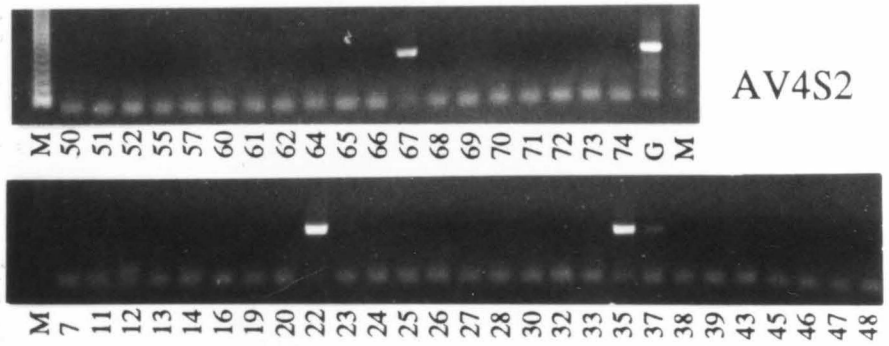

AV4S2
M 50 51 52 55 57 60 61 62 64 65 66 67 68 69 70 71 72 73 74 G M
M 7 11 12 13 14 16 19 20 22 23 24 25 26 27 28 30 32 33 35 37 38 39 43 45 46 47 48

Figure 4.

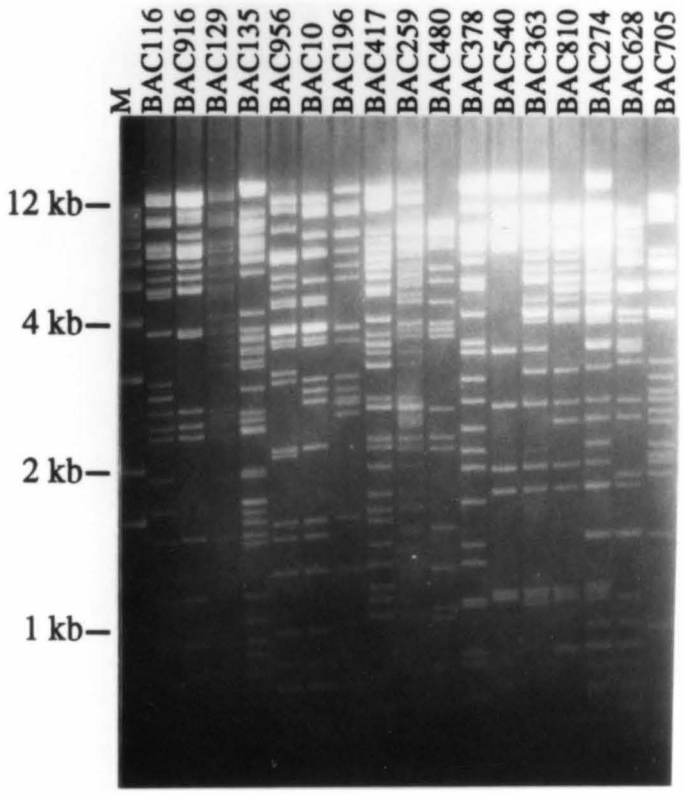

M
BAC116
BAC916
BAC129
BAC135
BAC956
BAC10
BAC196
BAC417
BAC259
BAC480
BAC378
BAC540
BAC363
BAC810
BAC274
BAC628
BAC705
12 kb
4 kb
2 kb
1 kb

Figure 5.

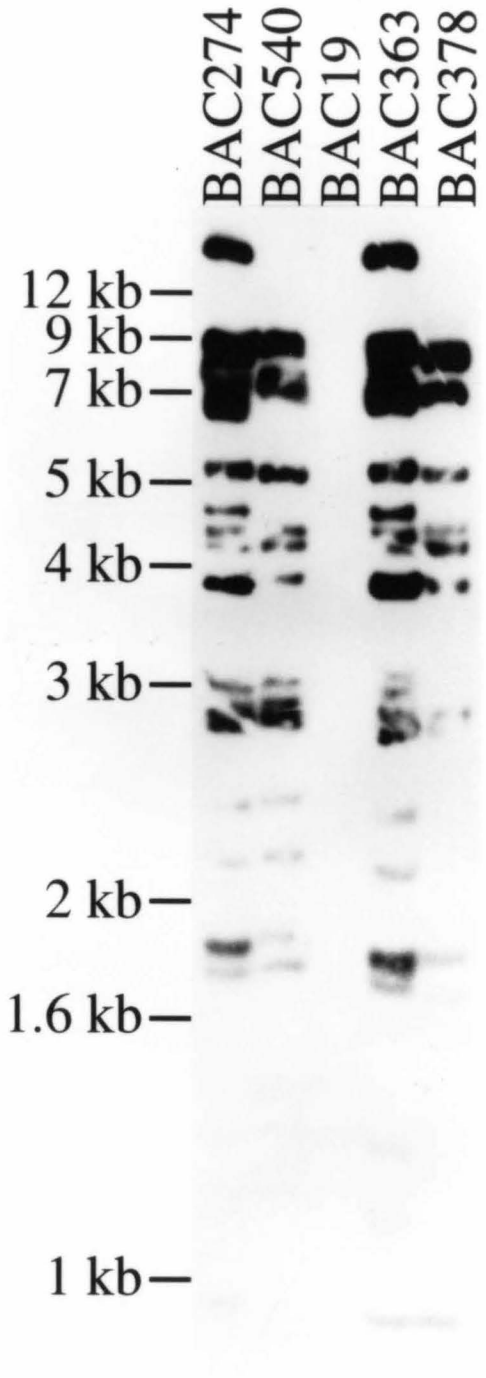
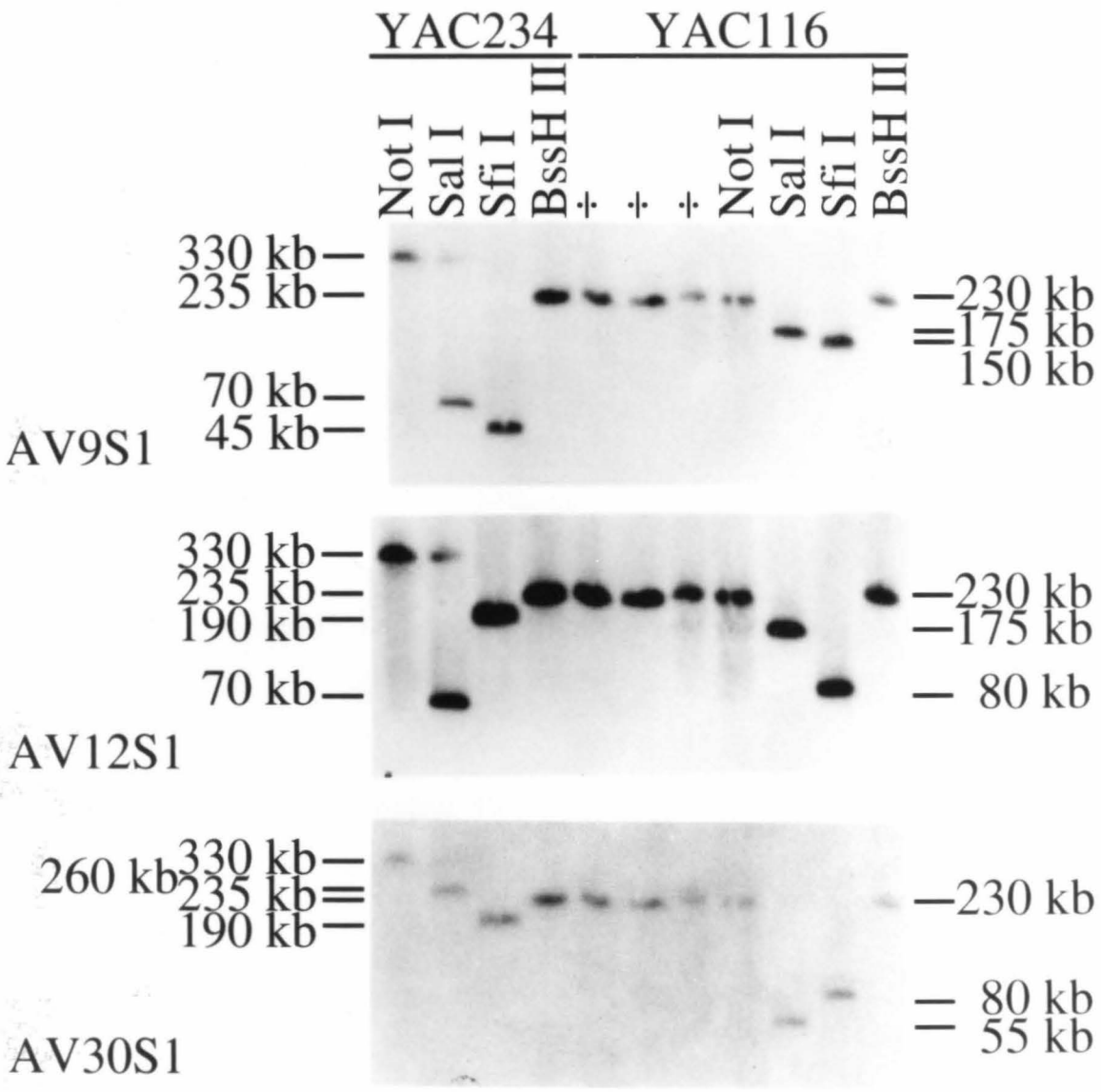

BAC274
BAC540
BAC19
BAC363
BAC378
12 kb
9 kb
7 kb
5 kb
4 kb
3 kb
2 kb
1.6 kb
1 kb

Figure 6.


YAC234
YAC116
Not I
Sal I
Sfi I
BssH II
÷
÷
÷
Not I
Sal I
Sfi I
BssH II
330 kb
235 kb
70 kb
45 kb
AV9S1
—230 kb
175 kb
150 kb
330 kb
235 kb
190 kb
70 kb
AV12S1
—230 kb
—175 kb
— 80 kb
260 kb
330 kb
235 kb
190 kb
AV30S1
—230 kb
— 80 kb
— 55 kb

DNA Sequence of the Human T cell receptor α/δ Locus: Biological Implications

Cecilie Boysen‡, Inyoul Lee†, Kai Wang≠, Arian Smit†, and Leroy Hood†*

‡ California Institute of Technology, Division of Biology, 147-75, Pasadena, California 91125

† University of Washington, Department of Molecular Biotechnology, Box 357730, Seattle, Washington 98195

≠ Darwin Molecular Corporation, 1631 - 220th Street Southeast, Bothell, Washington 98021

* Corresponding author

ABSTRACT

The sequence of the human α/δ T cell receptor locus, 1.07 megabases (Mb), has been determined. Fifty-seven variable gene elements, 48 of which appear functional, have been identified. Five olfactory receptor genes are intercalated with the 5' end of this locus and the highly conserved anti-apoptosis gene, defender against death (DAD), lies at the 3' end of this locus. More than 250 α/δ cDNA sequences have been compared against their chromosomal counterparts and the striking observation is that no pseudogenes are expressed, whereas most functional genes are. This suggests efficient degradation of pseudogene mRNAs. The α/δ locus is divided into three chromosomal domains by GC nucleotide content and the presence or absence of genome-wide L1 repetitive elements. The 3' domain correlates perfectly with a highly conserved 130 kb sequence (71%) across the human and mouse evolutionary lines.

INTRODUCTION

The human α/δ T cell receptor (TCR) locus is a complex multigene family spanning ~1 Mb on human chromosome 14. It encodes two distinct TCR polypeptides, α and δ, each contributing to one of the two types of heterodimeric T cell receptors--α/β and γ/δ. The γ/δ TCR is expressed early in T cell development and still has a somewhat uncertain function. The α and β families encode the classical TCR responsible for most T cell responses. The α and δ polypeptides are divided into variable (V) (antigen recognition) and constant (C) (fixation to the T cell membrane) regions. The V regions are encoded by a multiplicity of gene segments, Vα and joining (Jα) for α and Vδ, diversity (Dδ), and Jδ for δ, one each of which rearranges and joins together during T cell development to form Vα and Vδ genes, respectively (reviewed in Davis and Bjorkman, 1988). The DNA rearrangement process is mediated by short DNA signals of two types--hexamer/23 nucleotide spacer/nanomer joining to a hexamer/12 nucleotide spacer/nanomer--lying adjacent to the gene segments to be joined (Lieber, 1992). After transcription, the Vα and Vδ genes are joined to their respective C genes by RNA splicing to generate α or δ mRNAs. The rearrangement mechanism permits TCRs to be expressed in a quantized manner, with generally only one type of TCR per T cell. Hence, the recognition functions of the T cell are also quantized. The V gene segments have been divided into 35 subfamilies whose members exhibit 75% or more similarity at the DNA level and, accordingly, can be identified by DNA hybridization. Diversity is generated in the Vδ and Vα genes by three distinct mechanisms: (1) combinatorial joining of the gene segments; (2) diversity at the gene segment junctions created by exonuclease removal of bases from the ends of the gene segments; and (3) the addition of non-chromosomal

encoded nucleotides by the enzyme terminal deoxynucleotidyl transferase (TdT) to the ends of the gene segments prior to joining.

Three V element features are of special interest. (*i*) Some V gene segments appear to associate primarily with the Cδ gene (Vδ1, Vδ2, and Vδ3). Five others associate with both Cδ and Cα genes. The remainder of the V gene segments appear to associate only with Cα genes. The key question is whether these associations are mechanistically determined (e.g. three classes of DNA rearrangement signals) or determined by selection (e.g. antigen-driven or special developmental signals). (*ii*) The Vδ genes are expressed in successive developmental waves, Vδ2 first followed by Vδ1 and Vδ3 (van der Stoep et al., 1990, Krangel et al., 1990, and McVay et al., 1991). (*iii*) The Vδ genes are expressed in distinct tissues dependent on the developmental stage (Morita et al., 1994). Vδ1 expressing T cells are predominant in the adult thymus, whereas γδ T cells in adult blood predominantly express Vδ2 (Kabelitz, 1992, and Haas et al., 1993). An analysis of the regulatory elements of these genes may provide some insights into these features.

The gene products of the human α/δ locus have been studied by three general approaches. First, the sequence analyses of more than 250 α/δ cDNAs suggests that there are perhaps 45 V gene segments, although alleles, closely related gene duplications and sequencing errors cannot be readily distinguished. These fall into 35 V subfamilies (Arden et al., 1995). Second, physical mapping studies employing V gene segment hybridization, analysis of large DNA fragments produced by restriction digest with infrequent cutting enzymes and separated by pulsed field gel electrophoresis (Griesser et al., 1988, Satyanarayana et al., 1988, Hata et al., 1989, Ibberson et al., 1995), as well as hybridizations to deleted chromosome 14s in individual T cells (created because both maternal and paternal chromosome 14s generally undergo

rearrangement during T cell development)(Ibberson et al., 1995). Various sequence and mapping studies on smaller chromosomal fragments of this locus have also been carried out (Isobe et al., 1988, and Loh et al., 1988). These studies also suggested there were approximately 40-50 V gene segments and crude localization was possible. Finally, 97 kb of sequence at the 3' end of the human α/δ locus, spanning the Cδ to Cα regions, has been sequenced (Koop et al., 1994). These studies collectively suggest that with respect to its TCR elements the α/δ locus is organized as follows: (V1 . . . V45), Dδ1-3, Jδ1-4, Cδ, Vδ3, Jα1-61, Cα.

The complete sequence of the human α/δ locus has been determined, spanning 1.07 Mb of DNA. Forty-eight apparently functional and nine pseudo V gene segments have been identified. A most striking observation is that none of the pseudogenes are represented as cDNAs, whereas virtually all of the functional V elements are encoded as cDNAs--implying an efficient mechanism for degrading pseudogene mRNAs. Five olfactory receptor genes are intercalated among the 5' most V gene segments, whereas an anti-apoptosis gene, defender against death (DAD), appears 15 kb 3' to the Cα gene. These associations could have interesting biological implications. This locus is sharply divided into three domains by GC content and by the presence of high or low levels of the L1 genome-wide repeat element. One of these domains corresponds to a highly conserved 3' sequence region previously noted (Koop et al., 1994, Koop and Hood, 1994). The 3' ~130 kb of the mouse and human α/δ TCR families appears as highly conserved as are most coding regions (~71%) in spite of the fact that coding sequence comprises approximately 5% of this region. A comparison with the completely sequenced human TCRβ locus (Rowen et al., 1996) shows both striking similarities and differences. Knowledge of the complete human α/δ TCR

locus permits the initiation of powerful new systems approaches to how this locus functions in immunity.

RESULTS AND DISCUSSION

Mapping and Sequencing the α/δ locus

Yeast, bacterial, and P1-derived artificial chromosome (YAC, BAC, and PAC) clones were obtained across the α/δ locus from appropriate clone libraries using various V and C elements as probes. Three YACs, spanning the 3' end of the V portion of the locus, were subcloned into cosmids and these were mapped. Seventeen BAC clones were obtained spanning the entire locus, but for a single gap that was covered by a PAC clone. A second PAC clone was also isolated covering an unstable cosmid clone region. Eight cosmid, five BAC, and two PAC clones were sequenced using the random or shotgun strategy (C. Boysen, in preparation). Our estimated error rate is 1/5,000 base pairs. These studies suggest that BAC clones, in contrast to their YAC and cosmid counterparts, are highly stable and generally faithfully represent chromosomal DNA. Furthermore, BAC inserts can be readily sequenced by the random method. Hence, BAC clones appear to represent excellent mapping and sequencing reagents (Boysen et al, 1996a).

T Cell Receptor Elements

A schematic diagram of the α/δ locus is given in Figure 1. Fifty-seven V elements have been identified and nine of these appear to be pseudogenes by various criteria (Table 1). Twenty-five relics, highly mutated V elements, have also been identified. This locus also contains 3 Dδ, 4 Jδ, and 61 Jα gene

segments, as well as the Cδ and Cα genes. The sequence, spanning 1.07 Mb, extends 126 kb 5' to the Vα1.1 element and 15 kb 3' to the Cα gene. All of the V elements apart from Vδ3 lie 5' to the Dδ elements. The Vδ3 gene segment lies 3' to the Cδ gene and is in opposite transcriptional orientation to all other TCR elements in the locus. The rearranging δrec element (de Villartay et al., 1998) lie to the 3' end of the V gene segments except for Vδ2, which is found just upstream of the first Dδ element. No other genes are found in the region lying between 850 and 950 kb. In contrast, the human β TCR locus has five trypsinogen genes in the Vβ to Dβ gap (Rowen et al., 1996). The TCR coding regions constitute 3-4% of the locus extending from the Vα1 to the Cα elements. Since all of the V elements have been identified, we propose a new nomenclature numbering the V gene segments with consecutive increasing numbers from 5' to 3' with members of each subfamily given successively higher decimal designations (e.g., member Vα13.1 is 5' to member Vα13.2). A translation from the old to new designations is given in Table 2. The new designations will be used throughout this paper.

Non TCR Elements

Five olfactory receptor genes lie intercalated among the 5' Vα1.1 and Vα1.2 gene segments (Figure 1). These are denoted olf 1-5, in 5' to 3' orientation. Olfactory genes 1, 3, and 4 appear functional, whereas 2 and 5 are pseudogenes. This family is divided into two subfamilies--olf 1, 2, and 3 are most closely related to a rat olfactory receptor, and olf 4 and 5 are most closely related to a chicken olfactory receptor (Figure 2). Members of these subfamilies differ 7-33% and 45% of their amino acid sequence, respectively, and they differ from the members in the other family by approximately 60%. Olfactory receptor genes are encoded by many small multigene families

scattered across the mammalian genome (Chess et al., 1992 and Ben-Arie et al., 1994), although it is interesting to note that several have also been identified in a second immune receptor locus, the class I region of the major histocompatibility locus (Fan et al., 1995). The invasion of the α/δ locus by olfactory receptor genes raises questions as to whether these genes may have any functional or regulatory relationship to the TCR elements.

About 15 kb 3' to the Cα gene lies the DAD gene, an anti-apoptosis gene. The location of this gene is conserved from chickens to mammals, evolutionary lines that diverged over 350 million years ago. The DAD gene is expressed in many different tissues, including the thymus. It is intriguing to speculate that DAD may play a role in avoiding the apoptotic pathway for thymocytes that have successfully rearranged their α TCR genes. The α genes rearrange later in T cell development than the β genes and are, accordingly, the final point of decision in determining whether a T cell has functional TCRs.

Table 3 lists other genes and pseudogenes that are present in this locus, especially at the 5' end. Only one of these, the zinc finger gene, appears to be functional. Several of the others contain at least one intron and, hence, represent defective chromosomal and not processed mRNA genes. It will be interesting to determine how highly conserved in evolutionary time these gene locations are and whether in other species some of these genes may be functional. The presence of multiple intercalated chromosomal genes within the α/δ locus raises the possibility that these genes have been copied and integrated from other chromosomal locations.

The intriguing question underlying the long term association of the DAD, and possibly the olfactory and α/δ genes, is whether the association is

trivial and inadvertent, or whether genes that remain associated over long periods of evolutionary time share functional and/or regulatory constraints.

Features of the V Gene Segments

All of the V gene segments contain a 5' promoter region, exon 1 encoding the major part of the signal peptide, an intron ranging in size from 90 to 459 base pairs, exon 2 encoding the V segment and a DNA rearrangement sequence immediately 3' to exon 2 (Figure 3). The amino acid sequences, including the leader peptide, of the 48 presumably functional V elements are given in Figure 4. Several features are particularly noteworthy. (*i*) There is a conserved 20-mer, possibly a transcription factor binding site, in the promoter region (Gary Stormo, personal communication). This 20 nucleotide sequence is found approximately 200 bp upstream of the initiation codon in most of the functional Vα gene segments and in three of the five Vαδ gene segments, whereas it has not been found in the three Vδ elements (Figure 3). (*ii*) The sizes of the introns correlate nicely with the evolutionary relatedness of the V elements, e.g., the Va26.1 and the Va26.2 elements have the same intron length (Figure 3). (*iii*) Three apparently functional new V gene segments have been identified (Vα7, Vα9, and Vα18). (*iv*) The Vδ2 and Vδ3 elements have heptamer/nanomer DNA rearrangement signals that are different from one another and those of their Vα and Vαδ counterparts (Figure 3). Of course, the same may be said of many of the Vα DNA rearrangement signals. Conversely, Vδ1, some Vαδ and Vα rearrangement signals are quite similar to one another (apart from the spacer regions). Accordingly, it appears unlikely that these subtle differences could contribute to the Cδ-specific associations or the differential patterns of developmental expression of these three Vδ genes. (*v*) The α/δ locus contains 41 Vα or Vαδ

subfamilies and 3 Vδ subfamilies. Seven families are multi-numbered, ranging in size from two to seven members.

cDNA Comparisons to α/δ Chromosomal Coding Regions

About 250 α/δ cDNAs from Genbank have been compared against their germline counterparts. Several observations are striking. (*i*) No pseudogenes are expressed as cDNAs, yet almost all apparently functional V elements are. This implies efficient mechanisms for degrading mRNAs that have premature stop codons, as has been described for yeast and nematode (Leeds et al., 1992, and Pulak and Anderson, 1993). Three V elements appear to be functional, but are not expressed as mRNAs (Vα7, Vα9, and Vα18). In the TCRβ region (Rowen et al., 1996), two V gene segments that appear functional but are not found as cDNAs, were found to incorporate amino acids which would hinder the three-dimensional structure of the TCR (*ii*) Different Vα and Vδ genes appear to be expressed at different levels (Figure 1). The data in Genbank are somewhat skewed because many Vα sequences are derived from particular immune responses employing one or a few Vα elements. Several studies (Robinson, 1992, and Moss et al., 1993) examining the Vα T cell receptor usages in peripheral T cells from individual humans suggest striking differences in expression (e.g., Vα12s, Vα13s, and Vα21 are highly expressed and Vα3 and Vα24 are poorly expressed). There are no obvious Vα expression patterns correlated with chromosomal positions within the α/δ gene family. (*iii*) The 4 Jδ and 61 Jα gene segments also exhibit differing patterns of expression--some high and others low. None of the three pseudo Jα gene segments as determined by in frame stop codons (Koop et al., 1994) was found utilized in the cDNAs, again suggesting effective mRNA degradation for preterminating transcripts. (*iv*) There are two mechanisms for generating junctional diversity (e.g., diversity at the Vα and

Jα or Vδ, Dδ and Jδ boundaries), apart from the combinatorial joining of different gene segments: (1) the addition of N nucleotides to the junctional regions by the enzyme TdT, and (2) the removal of the ends of the cleaved gene segments by an exonuclease. The relative contributions of these mechanisms are given in Figure 5. The Vδ genes have enormous potential for diversification with the joining of up to all three Dδ elements with four distinct N regions (Figure 5).

Three distinct chromosomal domains.

The α/δ locus is divided into three distinct domains: The 5' 150 kb is rich in Alu repeats, and poor in LINE elements. This corresponds to a higher GC level found in this region (Figure 6). Across the V gene segments from 150 to 890 kb, is the sequence rich in LINE elements, and poor in Alu and GC content. At the 3' end, the GC content is back up at 45%, whereas almost no repeats are found. Virtually, no LINE elements are found, and the Alu elements present seem to go back to before the divergence of mouse and man. What is striking about the last two domains is that they also correlate with highly conserved (3') and significantly less conserved (5') domains evident when the α/δ loci of human and mouse are compared (Figure 6). The ~130 kb of the 3' domain compared exhibits an average of 71% similarity, even though only ~4% of this segment represents coding regions.

Several groups have described long-range GC% mosaic structures related to chromosomal bands. The long-range regions constant in GC% are called isochores (Bernardi, 1993). Giemsa-dark G bands are composed mainly of AT-rich sequences and T bands (a subgroup of the Giemsa-pale R bands) mainly of GC-rich sequences. Ordinary R bands are intermediate. Several have suggested that gene diversity, codon usage, CpG island diversity, DNA

replication timing, repeat sequence diversity, chromosomal condensation, and even recombination and mutation rates are related to long-range GC% structures (Ikemura, 1985, Holmquist, 1987, Korenberg and Rykowski, 1988, and Wolfe 1989). We have shown this to be true of several of these features. The GC content, CpG islands, mutation rate (see below), and genome wide repeats appear to correlate either with the domains or their boundaries. The TCR α/δ region has been located to chromosome 14q11.2, which represents an R-band. This is in agreement with the two isochores observed here. Several other of these features can now be examined. We should note that a sequence highly homologous (90% over 600 base pairs) to the pseudoautosomal boundary at the short arms of human sex chromosomes, a PAB: XY-like sequence, was detected at 580.5 kb -- 300 kb from the boundary described above.

Homology units.

A comparison of the entire sequence against itself, revealed large blocks of recently duplicated regions at the 5' end of the V gene segments (Figure 7). A 50 kb region including seven V gene segments is highly similar (60-100%) to an adjacent 50 kb region. Twenty kb of one of these regions has further been duplicated once. The major differences in the homology units are due to insertions of genome wide repeats, especially seen in the duplicated 20 kb sequence. These homology units explain the existence of the majority of the members in different subfamilies, Vα8, Vα12, and Vα13. Some of the gene segments have over time become pseudo-genes or relics. The homology units found here, suggest that the V gene segment repertoire evolves initially through duplication of long stretches of DNA, involving several V gene segments, and that these V gene segments later diversify. Similar long range

duplications have been observed in the murine TCR α/δ locus (Wang et al., 1994). Whereas, one can align the human and mouse V gene segments for the most 3' V gene segments, and a few V gene segments at the very 5' end, the middle region in both mouse and human contains different long duplicated regions, indicating that these duplications have occurred after the divergence of the two species.

Polymorphisms.

The YAC, BAC, and PAC clones were all constructed from differing human DNAs and, accordingly, six different chromosome 14 haplotypes were represented in these libraries. Sequence analysis of two overlapping clones from the same library would have a 50% chance of comparing the same haplotypes, thus giving an indication of the error rate, and a 50% chance of comparing different haplotypes, thus providing an estimate of the rates of variations among these haplotypes. Comparing overlapping clones from different libraries would provide an estimate, once again, of the rate of variations. The overlapping regions and their types of variation are given in Table 3. The overall single base polymorphism rate is 170 variations in 172 kb or about 1 polymorphism per kb. However, the rate varies with position, and it should be noted that the mutation rate at the 3' end is extremely low in the chromosomal domain highly conserved between human and mouse. The implication is that this conserved domain may have a very low rate of polymorphism.

Simple sequence repeats (microsatellites) are scattered more or less evenly across the locus (Figure 1). Ninety microsatellites, mostly di and tetra nucleotides, were found that contained more than eight repeated units. We

have chosen eight of these repeats scattered evenly across the locus for analysis against the CEPH families.

The developmentally controlled expression of Vδ genes.

The Vδ gene segments in both mouse and human are expressed in a highly ordered fashion during fetal development and before any α gene rearrangement takes place (Allison and Havran, 1991 and Krangel et al., 1990). We analyzed the promoter and recombinational signals of all the V gene segments (Figure 3). A conserved 20-mer (CCA/TCAAGRGGGCRRTGTTTC) was found approximately 200 bp upstream of the start codon in the majority of the promoters of the Vα gene segments, whereas it was not present in the promoters of the three Vδ gene segments. The consensus was found in three of the five V gene segments, which had been found to rearrange to generate either α or δ chains. However, seven of the forty Vα gene segments did not contain it either. No known DNA binding proteins have been found which could recognize this or part of this sequence.

The heptamers in the recombinational signals for Vδ2 and Vδ3, were found to differ from the consensus sequence. However, so do some of the heptamers in the recombinational signals of the Vα gene segments. Vδ1 contains a perfect recombinational signal, and thus other mechanisms must be responsible for its preferred rearrangement to the δ region. It should be noted, that Vδ1 as the only V gene segment contains a T nucleotide as the last base before the heptamer (Figure 3). Preferential expression of Vδ2 could be explained by its location. Vδ2 and Vδ3 both fall within the highly conserved 3' domain in a comparison between human and mouse (Figure 6). Vδ2 seems to possess the same specific expression pattern as does Vδ1 in mouse. However, in the dot matrix analysis of the two regions, mVδ1 was not aligned

with hVδ2, in fact mVδ1 was not located to the sequenced region in mouse. The human Vδ2 element aligned to some unknown mouse sequence with high similarity especially in the promoter region. The Vδ3 in human is the orthologue of Vδ5 in mouse, both located in an inverted orientation at the 3' end of the Cδ-region. Vδ1 has similarity to two mouse genes, both of which have been found expressed with either α or δ. Likewise, are some of the human V gene segments with similarity to mouse Vδ gene segments found to be either αδ or α gene segments. The fact that Vδ2 is present within the third domain as defined by GC content (Figure 6) may effectively determine its rearrangement and expression pattern. Different chromosomal domains differ in many different aspects, such as chromatin structure, early versus late replication, transcription levels, genome-wide repeats, etc. It is possible that recombination of Vδ2 is facilitated by its location in this chromosomal domain.

Comparison of the human α/δ and β loci.

The number of V gene segments in the TCR α/δ locus are similar to the number found in the β locus. The β locus contains 65 V gene segments, 46 of which apparently are functional whereas the other 19 are pseudo genes (Rowen et al., 1996). It further contains 22 relics. However, whereas the Vα/δ gene segments constitute 44 subfamilies, the β locus includes only 30 subfamilies. On average each of the β families contains more members than the α/δ subfamilies. This is mainly due to a series of more recent duplication events. The region spanned by Vβ gene segments is also shorter, resulting in one V gene segment every 8 kb, whereas the average is one V gene segment every 13 kb for the TCR α/δ locus. The recombinational signals found in each case have very similar consensus sequences, which are also found around the

D, the J, and the immunoglobulin gene segments. Approximately two thirds of the Vβ promoters contain a conserved 14-mer, with a proposed CREB site. Few α promoters seem to contain a CREB element, but instead a 20-mer conserved region has been identified. Whether this 20 nucleotide long sequence is present in the promoters of the Vβ gene segments is under investigation.

Like the TCR α/δ locus, the β region contains other multi-member families. Two groups of trypsinogen genes were found (Rowen et al., 1996). One was located 5' to the V gene segments similar to the olfactory receptors found in TCR α/δ, whereas the other group was found at the 5' end between the V gene segments and the Dβ gene segments. This region in the TCR α/δ locus appears void of any genes, except for Vδ2 located just 5' of the Dδ elements. Larger genomic regions outside the V gene segments have not been sequenced in the β region yet, so it is not known whether as many non-TCR related pseudogenes as seen in the α/δ region are present.

System Analysis

The challenge of biology as we move into the 21st century is to analyze complex systems and networks and understand their so-called emergent properties. Emergent properties for the immune system are, for example, tolerance and immunity, for both of these features arise from the complex network of lymphocytes, antigen presenting cells, etc. Tolerance or immunity, systems properties, can never be understood by studying individual molecules and/or cells in isolation. Two challenges arise for the analysis of complex systems and networks. (1) How does one divide extremely complex systems (e.g., humans have perhaps 10^{12} lymphocytes) into analyzable subsystems whose emergent properties still reflect those of the

whole system? (2) Are there bottlenecks in the system that play a key role in integrating information and, thus, serve as key sites for deciphering or manipulating biological complexity. The T helper lymphocyte is such a bottleneck point in the immune system, for it plays an early role in generating both immunity and tolerance. The T helper cell may be manipulated by virtue of its particular unique cell surface addresses, the T cell receptors, and they, accordingly, represent an analyzable subsystem. This analysis of the human α/δ T cell receptor family has given us the tools to effectively interrogate the subsystem. For example, unique PCR primers can be placed outside each functional V element to analyze the nature of their polymorphisms in the human population and determine whether any of these correlate with immune-related diseases such as multiple sclerosis or allergies. Second, a unique PCR primer can be placed in the coding region or each V gene segment to be used in conjunction with a single Cδ or Cα primer to interrogate the Vα or Vδ repertoire during T cell development, the induction of immunity or tolerance, or at autoimmune disease sites. In each case, all of the functional V element coding and flanking sequences must be known before unique primers can be designed for each element. The important point is that the entire subsystem (the α/δ V, D, and J elements) can be analyzed in response to a systems property. For example, we have carried out preliminary analyses for the expression of the 61 Jα gene segments in mouse and human (Hood et al., 1993). We have also used this approach to examine 30 Vα elements in 10 different individuals for their polymorphisms (Boysen et al., 1996b). In a similar vein, the distribution of simple sequence repeats across the human α/δ locus allows us to identify polymorphisms at any site in the family that may predispose to immune-related diseases.

SUMMARY

The sequence of the complete human α/δ T cell receptor locus has been determined. There are 115 functional T cell elements (V, D, J, and C). Fascinating genes, the olfactory receptor genes and the DAD genes, are associated with the 5' and 3' regions of the family, respectively. The locus is divided into three chromosomal domains by GC content, LINE1 and other genome-wide repeat content, and sequence conservation across species. Complete knowledge of this locus provides powerful tools (individual V or J element PCR, microsatellites) for interrogating the response of the entire family to the emergent properties of the immune system--development, immunity, and tolerance.

ACKNOWLEDGMENT

We would like to thank Lee Rowen for many interesting discussions and help in the analysis. We would also like to thank Todd Smith, Phil Green, and Gary Stormo for providing help with different computer programs. We highly appreciate the work of the technicians making the sequencing of this locus possible. Finally, we thank Tawny Biddulph for typing this manuscript. This work was supported by a grant from the Department of Energy (DOE).

REFERENCES

Allison, J. P., and Havran, W. L.: The immunobiology of T cells with invariant γδ antigen receptors. Annu Rev Immunol 9: 679-705, 1991.

Arden, B., Clark, S. P., Kabelitz, D., and Mak, T. W.: Human T-cell receptor variable gene segments families. Immunogenetics 42: 455-500, 1995.

Ben-Arie, N., Lancet, D., Taylor, C., Khen, M., Walker, N., Ledbetter, D.H., Carrozzo, R., Patel, K., Sheer, D., Lehrach, H., and North, M.A.: Olfactory receptor gene cluster on human chromosome 17: possible duplication of an ancestral receptor repertoire. Hum. Mol. Genet. 3: 229-235, 1994.

Bernardi, G.: The isochore organization of the human genome and its evolutionary history - a review. Gene 135: 57-66, 1993.

Boysen, C., Simon, M. I., and Hood, L.: The use of bacterial artificial chromosomes (BACs) as mapping and sequencing reagents. Submitted, 1996a.

Boysen, C., Carlson, C., Hood, E., Hood, L., and Nickerson, D.A.: Identifying DNA polymorphisms in human TCRA/D variable genes by direct sequencing of PCR products. Immunogenetics (*in press*), 1996b.

Chess, A., Buck, L., Dowling, M.M., Axel, R., and Ngai, J.: Molecular biology of smell: Expression of the multigene family putative odorant receptors. Cold Spring Harbor Symp. Quant. Biol. LVLL: 505-516, 1992.

Davis, M. M., and Bjorkman, P. J.: T-cell antigen receptor genes and T-cell recognition. Nature 334: 395-402, 1988.

De Villartay, J.P, Hockett, R.D., Copran, D., Korsmeyer, S.J., and Cohen, D.I.: Deletion of the human T-cell receptor δ-gene by a site-specifc recombination. Nature 335: 170-174, 1988.

Fan, W.F., Liu, Y.C., Parimoo, S., Weisman, S.M.: Olfactory receptor-like genes are located in the human major histocompatibility complex. Genomics 27: 119-123, 1995.

Griesser, H., Champagne E., Tkachuk D., Takihara, Y., Lalande, M., Baillie, E., Minden, M., and Mak, T.W.: The human T cell receptor α-δ locus: a physical map of the variable, joining and constant region genes. Eur J Immunol 18: 641-644, 1988.

Haas, W.: Gamma/delta cells. Annu. Rev. Immunol. 11: 637-85, 1993.

Hata, S., Clabby, M., Devlin, P., Spits, H., De Vries, J. E., and Krangel, M. S.: Diversity and organization of human T cell receptor δ variable gene segments. J. Exp. Med. 169: 41-57, 1989.

Hollemann, T., Schuh, R., Pieler, T., and Stick, R.: Xenopus Xsal-1, a vertebrate homolog of the region specific homeotic gene spalt of Drosophila. Mech. Develop. 55: 19-32, 1996.

Holmquist, G.P.: Role of replication time in control of tissue specific gene expression. Am. J. Hum. Genet. 40: 151-173, 1987.

Hood, L., Koop, B. F., Rowen, L., and Wang, K. : Human and mouse T cell receptor loci: The importance of comparative large scale DNA sequence analyses. Cold Spring Harbor Symp. Quant. Biol. LVIII: 339-348, 1993..

Ibberson, M. R., Copier, J. P., and So, A. K.: Genomic organization of the human T-cell receptor variable α (TCRAV) cluster. Genomics 28: 131-139, 1995.

Ikemura, T.: Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2: 13-34, 1985.

Isobe, M., Russo, G., Haluska, F.G., and Croce, C.M.: Cloning of the gene encoding the d subunit of the human T-cell receptor reveals its physical organization within the α-subunit locus and its involvement in chromosome translocations in T-cell malignancy. Proc. Natl. Acad. Sci. USA 85: 3933-3937, 1988.

Kabelitz, D.: Function and specificity of human γ/δ-positive T cells. Crit. Rev. Immunol. 11: 281-303, 1992.

Koop,B. F., Wilson, R.K., Wang, K., Vernooij, B., Zaller, D., Kuo, C.L., Seto, D., Toda, M., and Hood, L.: Organization, structure and function of 95 kb of DNA spanning the murine T-cell receptor Cα/Cδ region. Genomics 13: 1209-1230, 1992.

Koop, B. F., Rowen, L., Wang, K., Kuo, C. L., Seto, D., Lenstra, J. A., Howard, S., Shan, W., Deshpande, P., and Hood, L.: The human T-cell receptor TCRAC/TCRDC (Cα/Cδ) region: Organization, sequence, and evolution of 97.6 kb of DNA. Genomics 19: 478-493, 1994.

Koop, B. F., and Hood, L.: Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. Nature Genetics 7, 48-53, 1994.

Korenberg, J.R., and Rykowski, M.C.: Human genome organization: Alu, Lines, and the molecular structure of metaphase chromosome bands. Cell 53: 391-400, 1988.

Krangel, M.S., Yssel., H., Brocklehurst, C., and Spits, H.: A distinct wave of human T cell receptor γ/δ lymphocytes in the early fetal thymus: Evidence for controlled gene rearrangement and cytokine production. J. Exp. Med. 172: 847-859, 1990.

Kuhnlein, R.P., Frommer, G., Friedrich, M., Gonzalez-Gaitan, M., Weber, A., Wagner-Bernholz, J.F., Gehring W.J., Jackle, H., and Schuh, R.: Spalt encodes an evolutionarily conserved zinc finger protein of novel structure which provides homeotic gene function in the head and tail region of the Drosophila embryo. EMBO 13: 168-179, 1994.

Leeds, P., Wood, J.M., Lee, B.S., and Culbertson, M.R.: Gene-products that promote messenger-RNA turnover in Saccharomyces-cerevisiae. Mol. Cell. Biol. 12: 2165-2177, 1992.

Lieber, M. R.: The mechanism of V(D)J recombination: A balance of diversity, specificity, and stability. Cell 70: 873-876, 1992.

Loh, E.Y., Cwirla, S., Serafini, A.T., Phillips, J.H., and Lamier, L.L.: Human T-cell-receptor δ chain: Genomic organization, diversity, and expression in populations of cells. Proc. Natl. Acad. Sci. USA 85: 9714-9718, 1988.

McVay, L., Carding, S.R., Bottomly, K., and Hayday, A.C.: Regulated expression and structure of T cell receptor γ/δ transcripts in human thymic ontogeny. EMBO 10: 83-91, 1991.

Morita, C.T., Parker, C.M., Brenner, M.B., and Band, H.: TCR usage and functional capabilities of human γδ T cells at birth. J. Immunol. 153: 3979-3988, 1994.

Moss, P. A. H., Rosenberg, W. M. C., Zintzaras, E., and Bell, J. I.: Characterization of the human T cell receptor α-chain repertoire and demonstration of a genetic influence on Vα usage. Eur. J. Immunol 23: 1153-1159, 1993.

Nomura, N., Nagase, T., Miyajima, N., Sazuka, T., Tanaka, A., Sato, S., Seki, N., Kawarabayasi, Y., Ishikawa, K., and Tabata, S.: Prediction of the coding sequences of unidentified human genes. II. The coding sequences of 40 new genes (KIAAoo41-KIAA0080) deduced by analysis of cDNA clones from human cell line KG-1. DNA Res. 1: 223-229, 1994.

Pulak, R., and Anderson, P.: Messenger-RNA surveillance by the Caenorhabditis-Elegans SMG genes. Genes & Developm. 7: 1885-1897, 1993.

Qian, Y.W., Wang, Y.C., Hollingsworth, R.E. Jr., Jones, D., Ling, N., and Lee, E.Y.: A retinoblastoma-binding protein related to a negative regulator of Ras in yeast. Nature 364: 648-652, 1993.

Robinson, M.A.: Usage of human T-cell receptor V-beta, J-beta, C-beta and V-alpha gene segments is not proportional to gene number. Hum. Immunol. 35: 60-67, 1992.

Rowen, L., Koop, B.F., and Hood, L.: The complete 685 kilobase DNA sequence of the human beta T cell receptor locus. Science *(in press)*, 1996.

Satyanarayana, K., Hata, S., Devlin, P., Roncarolo, M. G., De Vries, J. E., Spits, H., Strominger, J. L., and Krangel, M. S.: Genomic organization of the human T-cell antigen-receptor α/δ locus. Proc Natl Acad Sci USA 85: 8166-8170, 1988.

Seto, D., Koop, B.F., Deshpande, P., Howard, S., Seto, J., Wilk, E., Wang, K., and Hood, L.: Organization, sequence, and function of 34.5 kb of genomic DNA encompassing several murine T-cell receptor α/δ variable gene segments. Genomic 20: 258-266, 1994.

Tanaka, T., Shibasaki, F., Ishikawa, M., Hirano, N., Sakai, R., Nishida, H., Takenawa, T., and Hirai, H.: Molecular cloning of bovine actin-like protein, actin2. Biochem. Biophys. Res. Commun. 187: 1022-1028, 1992.

van der Stoep, N., de Krijger, R., Bruining, J., Koning, F., and van der Elsen, P.: Analysis of early fetal T-cell receptor δ chains in humans. Immunogenetics 32:331-336, 1990.

Wang. K., Klotz, J. L., Kiser, G. Bristol, G., Hays, E., Lai, E., Gese, E., Kronenberg, M., and Hood, L.: Organization of the V gene segments in mouse T-cell antigen receptor α/δ locus. Genomics 20, 419-428, 1994.

Wolfe, K.H., Sharp, P.M., and Li, W.-H.: Mutation rates among regions of the mammalian genome. Nature 337: 283-285, 1989.

Table 1. Characteristics of pseudo V gene segments in the TCR α/δ locus. We define a nucleotide sequence as a pseudo V gene segment, when after splicing of putative exons the resulting amino acid sequence can be aligned with other V gene segments. In some cases where frameshifts have occurred it might be necessary to translate in different reading frames to make the alignment. If more than two frameshifts had to be made for alignment, we call the sequence a relic.

V gene segment	Defects
Vα11.1	No start codon, bad heptamer
Vα8.5	96 % similar to Vα8.3, but contains a MER11A/B insertion in exon2.
Vα15.1	1 frameshift, 1 stop
Vα8.7	Bad heptamer
Vα28.1	1 frameshift
Vα31.1	2 frameshifts
Vα32.1	2 frameshifts, 1 stop, Cys missing
Vα33.1	2 frameshifts, 1 stop, Tyr missing
Vα37.1	1 frameshift

Table 2. Translation from the new nomenclature used in this report to the old nomenclature previously reported (Arden et al., 1995).

New nomenclat.	Old nomenclat.	New nomenclat.	Old nomenclat.
AV1S1	AV7S1	AV19S1	AV12S1
AV1S2	AV7S2	AV20S1	AV30S1
AV2S1	AV11S1	AV21S1	AV23S1
AV3S1	AV16S1	AV22S1	AV13S1
AV4S1	AV20S1	ADV23S1	ADV17S1
AV5S1	AV15S1	AV24S1	AV18S1
AV6S1	AV5S1	AV25S1	AV32S1
AV7S1	New	AV26S1	AV4S2
AV8S1	AV1S1	AV26S2	AV4S1
AV8S2	AV1S5	AV27S1	AV10S1
AV8S3	AV1S4	AV28S1	New,pseudo
AV8S4	AV1S2	ADV29S1	ADV21S1
AV8S5	New,pseudo	AV30S1	AV29S1
AV8S6	AV1S3	AV31S1	New,pseudo
AV8S7	New,pseudo	AV32S1	New,pseudo
AV9S1	New	AV33S1	New,pseudo
AV9S2	AV22S1	AV34S1	AV26S1
AV10S1	AV24S1	AV35S1	AV25S1
AV11S1	New,pseudo	ADV36S1	ADV28S1
AV12S1	AV2S3	AV37S1	New,pseudo
AV12S2	AV2S1	AV38S1	AV14S2
AV12S3	AV2S2	ADV38S2	ADV14S1
AV13S1	AV8S1	AV39S1	AV27S1
AV13S2	AV8S2	AV40S1	AV31S1
ADV14S1	ADV6S1	AV41S1	AV19S1
AV15S1	New,pseudo	DV101S1	DV101S1
AV16S1	AV9S1	DV102S1	DV102S1
AV17S1	AV3S1	DV103S1	DV103S1
AV18S1	New		

Table 3. Non-TCR genes found in or around the TCR α/δ region.

Gene	Location	Properties
Zinc finger protein	30 kb	Single open reading frame encoding 1,003 amino acids. Probably functional. In comparison with *Sal* (Kuhnlein et al., 1994) and*Xgal1* (Hollemann et al., 1996) small up and downstream exon should be found. Contains one aminoterminal CC/HC zinc finger, three double CC/HH zinc fingers connected by H/C links, and one single CC/HH zinc finger in connection with the middle double zinc finger. ESTs match the 3' end of the gene. CpG island upstream.
Retinoblastoma binding protein (pseudo)	69 kb	Almost 100% identical to mRNA RbAp48 (Qian et al., 1993). Contains 1 intron. Two frame shifts leading to stop codons.
Olfactory receptors		Please see Figure 1 and text.
Actin2 (pseudo)	90 kb	Highly similar to bovine mRNA for actin2 (Tanaka et al., 1992). Contains 2 introns. Three frame shifts and one stop codon.
Ubiquitin-conjugating enzyme, E2. Drosophila bendless gene product (pseudo)	118 kb	Almost 100% identical to human epidermoid carcinoma mRNA for Drosphila bendless gene product (Genbank accession no.: D83004). Contains a single C nucleotide deletion leading to a stop codon.
Open reading frame (pseudo)	154 kb	Highly similar to human mRNA for ORF (Nomura et al., 1994). Contains one intron. Two frame shifts.

continued

Table 3, continued

Gene	Location	Properties
Cosmid-like sequence	187 kb	Highly similar to several regions in cosmid L191F1 (Genbank accession no.: Z68756). Contains several CpG islands. No apparent coding region. Possible unidentified repeat.
Ribosomal protein (pseudo)	255 kb	Almost a 100% identical to human mRNA for ribosomal protein (Genbank accession no.: D23660). Two frame shifts and two stop codons.
DNA-binding proteins (pseudo)	392 + 465 kb	Similar to DNA-binding proteins (CROC-1A and 1B) (Genbank accession no.: U39360 and U39361). Frame shifts and stop codons.
Defender against death (DAD)		Please see Figure 1 and text.

Table 4. Sequence variations in regions of overlapping clones.

Overlapping clones	Same (s) or different (d) haplotypes	kb	Single base substitutions	Indels	Microsatellites (n>8)
BAC129-PAC161	d	52	76	8	14
PAC161-BAC956	d	37	37	6	4
BAC956-PAC230	d	32	39	2	1
PAC230-Cosmid1	d	32	6	0	2
BAC480-BAC378	d	2	6	1	1
BAC378-BAC810	s	20	0	0	0
BAC810-Cosmid9	d	16	0	0	0
Cosmid9-CosmidX	d	17	6	0	1

FIGURE LEGENDS

Figure 1. A schematic diagram of the TCR α/δ locus. a) The upper panel indicates the location of each TCR element, black bars (the 61 Jα elements are drawn as a big open box at the 3' end of the locus). Presumed functional V gene segments are indicated by long bars, pseudo V gene segments by two thirds the height, and relics by one third the height. The nomenclature used is explained in the text, and a conversion to earlier nomenclature is given in Table 2. The DAD gene, five olfactory receptor genes (Olf1-Olf5), and a gene encoding a zinc finger protein are included in this map, whereas other presumably non-fuctional genes have been omitted (Table 3). The arrows indicate the transcriptional orientation for each element. b) Number of cDNAs for each V gene segment found in Genbank release 94, April 96. The bars represent from 1-14 cDNAs, except for two cases, ADV14S1 and AV17S1with 35 and 42 cDNAs, respectively. These two are overrepresented in Genbank due to a thyroiditis study where these two gene segments are dominant. c) Distribution of genome wide repeats. d) Microsatellite repeats. Ninety microsatellites with eight or more repeated units are shown. e) Sequenced clones. The 3' cluster of cosmid clones has been sequenced in a previous study (Koop et al., 1994). The other cosmids, BACs and PACs have been sequenced by the high redundancy shotgun approach. They might constitute as many as 6 different haplotypes.

Figure 2. Alignment of the amino acid sequences for five olfactory receptors present in the TCR α/δ locus with their rat (Genbank accession no.: U50949) and chicken (X94744) counterparts. The seven transmembrane domains are

indicated by boxes. Highly conserved cysteine residues, supposedly involved in disulfide bridges in the extra cellular loops are indicated with a dot. Olf2 has an internal stopcodon at position 85, whereas Olf5 contains a frame shift at codon 209.

Figure 3. Alignment of promoter, splice sites and recombinational signals for 48 presumably functional V gene segments. The V gene segments have been divided into three groups. Those expressed in TCR δ chains, those expressed in either α or δ chains, and those only found in α chains. A conserved 20-mer has been identified approximately 200 bp upstream of the start codon in the majority of the Vα gene segments, and in some of the Vαδ gene segments, whereas it has not been found in the Vδ elements.

Figure 4. Alignment of the translated amino acid sequence of the 48 presumably functional V gene segments.

Figure 5. Nucleotide nibling and additions in the junctional regions of the TCR α/δ genes. The genomic sequence of the V gene segments was compared to 246 cDNAs. In the analysis of exonuclease activity and N nucleotide addition, nucleotides found in the genomic sequence was counted as such, rather than N nucleotides. In aligning the δ gene segments with their cDNA counterparts, the Dδ gene segments are included if three or more nucleotides matched. Otherwise, the nucleotides are indicated as belonging to the N-regions.

Figure 6. Distribution of LINE1 and Alu repeats and GC content in the TCR α/δ and two human-mouse comparisons. Based on GC content and the

genome-wide repeats, the locus is divided in three domains: 150 kb 5' with high GC content and high Alu concentration, but very few LINE1 elements; 740 kb (150-890 kb) with low GC content and Alu concentration, whereas LINE1 elements are rich in this region; The last 175 kb has a high GC content, but contains a very low amount of all kinds of genome wide repeats. Comparison to mouse sequences (Seto et al., 1994) located in the middle domain (680-720 kb) indicates no conservation outside the V gene segments, whereas comparison to 130 kb at the 3' end of mouse (Koop et al., 1992 and Lee Rowen, personal communication) indicates a striking degree of similarity (71%), even if the coding regions only occupy about four percent of this region.

Figure 7. Homology units in the TCR α/δ locus. The entire sequence was compared against itself, to identify larger regions of similarity. a) Dotplot of duplicated region of 50 and 20 kb blocks around 333-485 kb. b) The sequences involved in the duplications were aligned to one another, indicating the V gene segments involved in the duplication and the insertion of genome wide repeats.

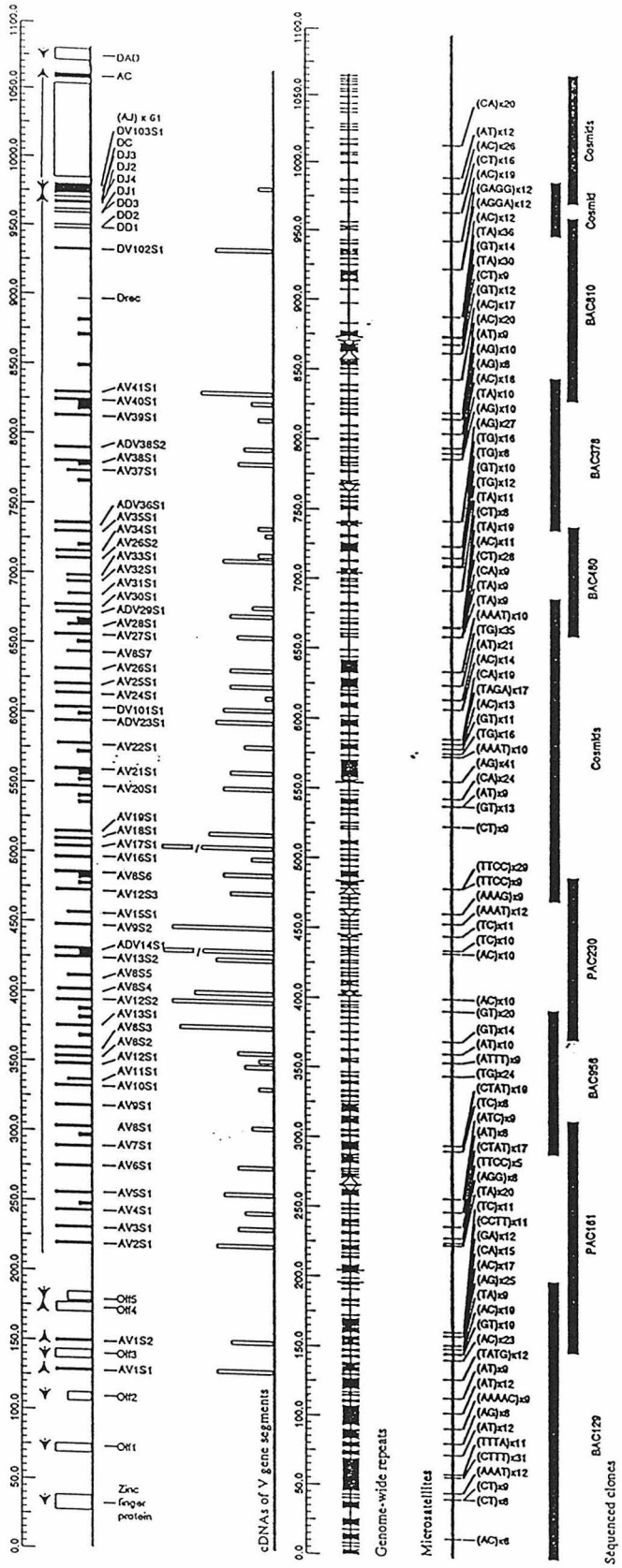

DAD
AC
(AJ) x G1
DV103S1
DC
DJ3
DJ2
DJ4
DJ1
DD3
DD2
DD1
DV102S1
Drec
AV41S1
AV40S1
AV39S1
ADV38S2
AV38S1
AV37S1
ADV36S1
AV35S1
AV34S1
AV26S2
AV33S1
AV32S1
AV31S1
AV30S1
ADV29S1
AV28S1
AV27S1
AV8S7
AV26S1
AV25S1
AV24S1
DV101S1
ADV23S1
AV22S1
AV21S1
AV20S1
AV19S1
AV18S1
AV17S1
AV16S1
AV8S6
AV12S3
AV15S1
AV9S2
ADV14S1
AV13S2
AV8S5
AV8S4
AV12S2
AV13S1
AV8S3
AV8S2
AV12S1
AV11S1
AV10S1
AV9S1
AV8S1
AV7S1
AV6S1
AV5S1
AV4S1
AV3S1
AV2S1
Otf5
Otf4
AV1S2
Otf3
AV1S1
Otf2
Otf1
Zinc finger protein
cDNAs of V gene segments
Genome-wide repeats
Microsatellites
(CA)x20
(AT)x12
(AC)x26
(CT)x16
(AC)x19
(GAGG)x12
(AGGA)x12
(AC)x12
(TA)x36
(GT)x14
(TA)x30
(CT)x9
(GT)x12
(AC)x17
(AC)x20
(AT)x9
(AG)x10
(AG)x8
(AC)x16
(TA)x10
(AG)x10
(AG)x27
(TG)x16
(TG)x8
(GT)x10
(TG)x12
(TA)x11
(CT)x8
(TA)x19
(AC)x11
(CT)x28
(CA)x9
(TA)x9
(TA)x9
(AAAT)x10
(TG)x35
(AT)x21
(AC)x14
(CA)x19
(TAGA)x17
(AC)x13
(GT)x11
(TG)x16
(AAAT)x10
(AG)x41
(CA)x24
(AT)x9
(GT)x13
(CT)x9
(TTCC)x29
(TTCC)x9
(AAAG)x9
(AAAT)x12
(TC)x11
(TC)x10
(AC)x10
(AC)x10
(GT)x20
(GT)x14
(AT)x10
(ATTT)x9
(TG)x24
(CTAT)x10
(TC)x8
(ATC)x9
(AT)x8
(CTAT)x17
(TTCC)x5
(AGG)x8
(TA)x20
(TC)x11
(CCTT)x11
(GA)x12
(CA)x15
(AC)x17
(AG)x25
(TA)x9
(AC)x10
(GT)x10
(AC)x23
(TATG)x12
(AT)x9
(AT)x12
(AAAAC)x9
(AG)x8
(AT)x12
(TTTA)x11
(CTTT)x31
(AAAT)x12
(CT)x9
(CT)x8
(AC)x8
BAC129
PAC161
BAC958
PAC230
Cosmids
BAC480
BAC878
BAC610
Cosmid
Cosmids
Sequenced clones

Figure 1.

Figure 2.

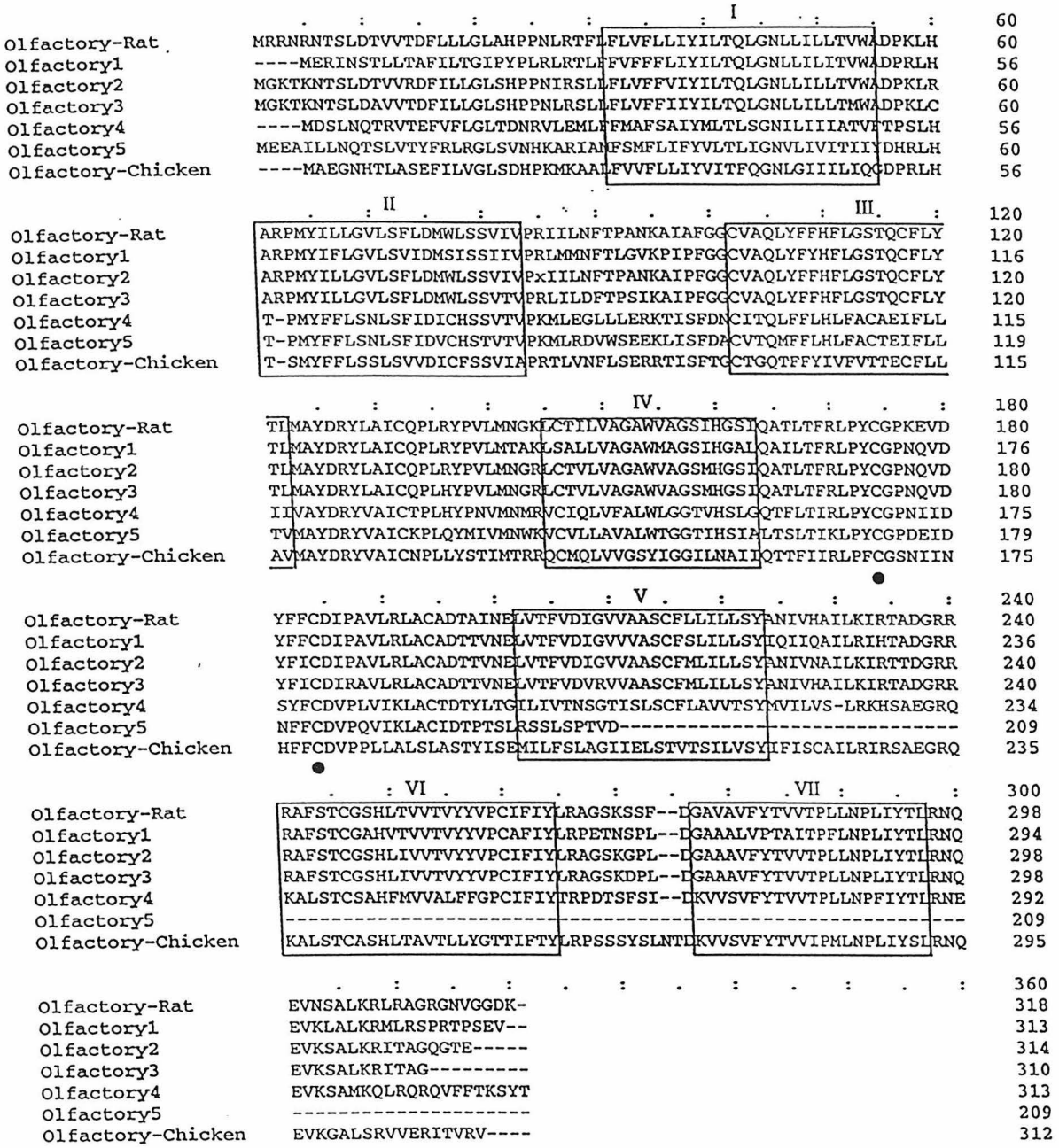

I
II
III
IV
V
VI
VII
Olfactory-Rat
Olfactory1
Olfactory2
Olfactory3
Olfactory4
Olfactory5
Olfactory-Chicken

Figure 3.

Name	Promoter	5' splice	Intron	3' splice	3'-end of V	Heptamer	Spacer	Nonamer
DV101		CTGGTATGGATG 223		CCCACAGGATC	TACTTTTGTGCTCTTGGGGAACT	CACAGTG	TTTGAAGTGATAGTAAAAGCAAA	ACAAAAACC CTAG
DV102		CAGGTAAGGCAG 152		CTCTCAGGAGT	TACTACTGTGCCTGTGACACC	CACCCTG	CTGCAGCTCTACTTCTGAGCAGC	TCAAAAACC ACTG
DV103		TCTGTAAGTAGT 233		TTCCCCAGACAG	TACTACTGTGCCTTTAG	CACTATG	ATGCAGGTGCCCAGGAAGTCATA	ACACAAACT CCTG
ADV14		TAGGTACGGGTG 164		CCTTCAGGACC	TATTTCTGTGCAATGAGAGAGGG	CACAGTG	ACAGAACTGTGGAGGGAGGTGT	ACAAAGCC CTGG
ADV23	ACAAAGTGGGGAAGTGCTTC	GCTGTGAGTTAT 146		TAAGCAGGGGT	TACTTCTGTGCAGCAAGCA	CACAGTG	CTCCCCAGGCACCTGAAGCCTGT	ACCCAAACC TGCA
ADV29	ACTGTAGGGGGCAGTGCTTC	ACTGTGAGTTGT 164		CAAACAGGGGT	TACTTCTGTGCAGCAAGCG	CACAGTG	CTTCTCAGGCACCTGCAGCCCGT	ACTCAAACC TGCT
ADV36	TCACAAGATGGCGGTGTTCT	GCTGTAAGTAGG 153		TACACAGGGGT	TACCTCTGTGCTGTGGAGG	CACAGTG	CTCCCTAGTCACCTGCAGCCTGT	ACTCAAATT CTAC
ADV38S2		TTGGTAAGGAAG 249		CCCACAGAATT	TATTTCTGTGCTTTCATGAAGCA	CACAATG	AGATGAGCAGCAGGGAGAGGCTT	ACAGAAACC TCAG
AV1S1		GAGGTAAGTCTC 338		GTCATAGGCAC	TACTTCTGCGCTGTGAGAGA	CACAGTG	ACTATGAGGCCTCCTTAACTGTG	CCAAAATTC AAAA
AV1S2		GAGGTAAGTCTC 288		GTCTTAGGCAC	TACCTCTGTGCTGTGAGAGA	CACGGTG	ACTATGAGGCCTCTTTAGCTGCA	CCAAAATTC AAAA
AV2	CCGCTAGGTGGCGGAGGTTA	AGGGTGAGTAGC 175		TCTGTAGTTGC	TACTACTGTGCTGTGGAGGA	CACAGAG	GCAGGGAACCCATGAAGAGCTGA	ACAGAAACA GAGA
AV3	GCTCTAGGGGGCGATGCAGT	TGAGTGAGTAAT 90		TTTACAGGTGG	TACTTCTGTGCTGTGAGAGACA	CACACTG	ATAGGGGCTGCAGGGGGAGCAGA	ACACAAACT CTTG
AV4		TGAGTGAGTTAT 436		CCCATAGGTAC	TACTACTGCCTCGTGGGTGACA	CACAGTG	AGACAGATGGGCCTGCACCTGTG	CCGTTTTCC TCTG
AV5	CCACAAGGTGGCAGAGCTTC	ACTGTGAGTCGA 176		TGCACAGGTAT	TACTTCTGTGCAGAGAGTA	CACATTG	CTTCTCAGGCACCTGTATCCTGT	ACCCAAACC TGCA
AV6	ACACAAGAGGGCAGTATTTC	AGGGTGAGTAGC 188		TGCCTAGGGGT	TACTACTGTGCTCTAGACA	CACAGTA	GTGCCCTGGCAGCTGCTTCCTGC	ACCCAAACT CTGC
AV7	CCACAAGGTGACAGGGGTTC	GCTGTAAGTAGG 154		TACACAGGGGT	TACCTCTGTGCTGTGGAGG	CACAGTG	CTCCCTAGTCACCTGCAGCCTGT	ACTCAAATT CTAC
AV8S1	CCTCCAGAGGGTGACGTTGC	TGAGTGAGTAAC 127		TTTTCAGGAGA	TACTTCTGTGCCGTGAATGC	CACAGTG	TCTGGGACTGCCAAAGGGAGCTGA	ACACAAACT TCCT
AV8S2	TGGGTGAGTAAC 119			GTTGCAGGAGG	TACTTCTGTGTTGTGAGTGA	CACAGTG	CTTGAGACTGCAGGAGAGCTGA	ACACAAGCC TCCT
AV8S3	CCTCCAGAGGCGGATGCTGC	TGAGTGAGTAAA 108		TTTCCAGGAAC	TACTTCTGTGCTGTGGGTGC	CACAGTG	CCTGGGTCTGCAGGGGGAGCTGA	ACACAAACT GCCT
AV8S4	CCTTCAGGGGGCGACGTTGC	TGGGTTAGTAAC 129		GTTGCAGGAGG	TACTTCTGTGCTGTGAGTGA	CACAGTG	CTTGAGACTGCAGGAGAGGTGA	ACATAAACC TCCT
AV8S6	CCTCTAGAGGGCGATGTTGC	TGGGTGAGTAAC 101		GTTGCAGGAGG	TACTTCTGTGCTGTGAGTGA	CACAGTG	CCTGAGACTGCAGGAGAGCTGA	ACACAAACC TCCT
AV9S2	CCACAAAGGGGTGCATCTTC	TTGGTAAGTAAA 138		TTTTCAGGGGG	TACTTCTGTGCTCTGAGTGA	CACAGTG	ACAGGGACTGCAGGGGAAGCTGA	GCACAAACT CTGA
AV9S2	CCACCAGAGGGAGTGGCTTC	TTGGTTAGTGAG 142		TCCCTAGGAAG	TACTTCTGTGCTCTGAGTGA	CACAGTG	ACAGGGACTGCAGGGGGAGCTGA	GCACAAACT CTGA
AV10	CCCCAAGAGGGCAGTATTTG	ATAGTAAGTTAG 225		TGAATAGGGGG	TACATCTGTGGTGTGAGCG	CACTGTG	CTCCACAGGCACTTGAAGCCAGT	ATGCAAACC TGCA
AV12S1	CCACAAGATGGCAGTGTGTT	GCTGTGAGTTGG 195		TTCACAGGGGT	TACCTCTGTGTGGTGTGAACA	CACAGTG	CTCCCCAGACACCTGCAGTCTGT	ACCCAAACC TGCT
AV12S2	CCACAAGATGGCAGTGTGTT	GCTGTGAGTTTG 201		TTCACAGGGGT	TACCTCTGTGCCGTGAGACA	CACAGTG	CTCCCCAGACACCTGCAGTCTGT	ACCCAAACC TGCC
AV12S3	CCTCAAGATGGCAGTGTCTT	GCTGTGAGTTTT 213		TTCACAGGCGT	TACCTCTGTGCAATGACCG	CACAGTG	CTCCCCAGACACTTGCAGTCTGT	ACCCAAACC TGCT
AV13S1	CCACAAGATGGCAGTGCTCC	ACTGTGAGTTGA 167		TGCATAGTGGT	TACTTCTGTGCAGCAAGTA	CACATTG	CTTCCCAGGCACCTGCTACCCGT	ACACAAACC TGAG
AV13S2	CCACAAGATGGCAGTGCTTC	ACTGTAAGTTTA 159		TGCACAGGGGT	TACTTTTGTGCAGCAATA	CACATTG	CTTTCCAGGCATCTGTAACCATC	ACCCAAACC TGAG
AV16	CCTCTAGAGAGAGATGTTAC	TCAGTAAGTAAA 114		GTTATAGGAGG	TACTGTGCTCTAAGTGG	CACAGTA	GCTGGTTTTGCAAGGAAGCAGA	ACACAAACC CTTT
AV17	TCACAAGAGGGCAGTGCCTC	CTAGTGAGTTGG 144		TGCCTAGGGGT	TACTTCTGTGCTACGGACG	CACAGTG	TTCCCCAGGAACCTGCAGCCTCT	ACGCAAACC CTGC
AV18	CCTCAAGAGGGCCGCTGCGC	TCAGTGAGTAAT 127		TTTTTAGGAAG	TACTACTGCGCTCTGAGAGA	CAGAGTG	GGAGGGACTGCAGCGAGAGCCCA	GCACAAACC CTGG
AV19	CCTCTTGAGTGCAAGGTGGA	TTGGTAAGCGTG 203		TTCACAGTATC	TACTTCTGTGCTCTGAGTGAGGC	CACAGTG	AGATGGGTGCCTGTGGGAGCCCT	ACAAAAACC TCAA
AV20	CCACGAGGTGGCAGTGGTTC	GCTGTAAGTTGG 154		TGAACAGGGTT	TATCTCTGTGCTGTGCAGG	CACAGCG	TTCCCCAGGCACCTGCAACTTGT	ATCAAAACC CTGC
AV21		AATGTGAGTTAG 160		TGAACAGGGGT	TACCTCTGTGCTGTGAGG	CACAGTG	CACAACAGGCACCTGCAACCAAT	ACCCAAACT CTAT
AV22		GCTGTGGGAGGA 188		TGACCAGGTGT	TATTTCTGTGCTGTGGAGC	CACAGTG	CTCCCCAGGCACCTGCGGCCTGT	ACACAAACC CTCA
AV24	TGAACATAGGGCAGAGCGTT	ACTGTAAGTCCA 161		AAAACAGGCGT	TACCTCTGTGCTACGGACG	CACAGTG	CTGTTCAGGCACCTGCACGCCAT	ACGCAAACC TGTG
AV25	CCACAGGGGGGCAGCATCGG	CACGTGAGTTTG 282		TAAACAGAGGT	TACTTCTGTGCTGAGGG	CACAGTG	CTCCCCAAGCACCTGCTGCCTGT	CTCCAAATC TTGC
AV26S1		TTGGTAAGTGCT 431		CCCTTAGGAAC	TACTATTGCATCGTCAGAGTCG	CACACTG	GGACAGATGGGGCTGCACCTGT	GCAATATCT CCCT
AV26S2	CCTGATGTGGGATGTGCTGT	TGGGTAAGTGTT 459		CCCATAGGTAT	TACTACTGCATCCTGAGAGAC	CACAGTG	GGACAGATGGGGCTGCAGCTGT	GCAATATCT CCCT
AV27	CATGTGAGTTGA 214			TAAACAGGGGT	TACCTCTGTGCAGGAG	CACAGTG	CTCTTGAGGCACCTGCTGCCTGC	ACCCAAACC CTGC
AV30	CCACAGGGGAGAGCTGTTTA	CCTGTGAGTACA 223		TGATCAGGGGT	TACTTCTGCGGCACAGAGA	CACAGTG	ATACCCAGGCCTCCAAGACCTGT	ACTCAAACC TAAA
AV34	CCAAAGAGTGGCAGTGTTAA	CCTGTGAGTATG 259		TGAACAGGGGT	TACCTCTGTGGAGCAGACA	CACAGCG	ATCTTCAGGCCTCTATCAGCTGT	CTCCAAACC TGCA
AV35	CCATATGGGGGCAGTGTCTC	CATGTGAGTCTT 250		TAAACAGGGGT	TACTTCTGTGCTGGGCAG	CACAGTG	CTCCCCAAACACCTGCAGCCTGT	ACTCAAACT TGCA
AV38S1		TTGGTAAGGAAG 243		TCCACAGAATC	TATTTCTGTGCTTATAGGAGCG	CACAGTG	AGACAAGCAACAGGGAGAGGCTT	ACAGAAACC TCAG
AV39	CCTGAAGAGGGCACTGTTTC	ACCGTGAGCTGG 169		TGGACAGGGTT	TACTTCTGTGCCGTGGACA	CACAGTG	CTCCCCTGACGCCACCAGTCTGT	ACCCAAACC TGCA
AV40	CCACTTGGGGGTGCTATTCA	TTGGTAAGTACA 113		CCTACAGGAGG	TACTACTGTCTTCTGGGAGA	CACTGTG	TTAAAAGCACAGTGGGAGCTAT	ACAAAAACC TCAA
AV41	CCACTAGGTGGCAATGTGGG	GCTGTAAGTTGG 168		AGAACAGGTGT	TACATCTGTGCTGTCAGA	CACAGTG	CTCCCCAGGCACCTGGAGCCCGT	ACCTAAACT CTAA
Consensus					TYR CYS			
60%	CC C AG GGGCA TG T	GT AGT		CAGG G	TACTTCTGTGC GTG	CACAGTG T	A	CTGT AC CAAACC
70%	CC C AG GGC G	GT AGT		CAGG	TACTTCTGTGC T	CACAGTG	A	C G AC AAAC
80%	C G GG G	GT AGT		AGG	TAC TCTGTG	CACAGTG		C AC AAAC
90%	C G	GT AG		AG	TA CTG	CAC TG		AAA

Figure 4.

	$E_{V\delta}$	$N_{V\delta}$	E^*	E^*	N_1^*	E	E	N_2	E	E	$N_{J\delta}$	$E_{J\delta}$
Vδ1, avr.	4.3	5.8	1	4	2	0.8	1.8	4.5	3.6	3.2	9.3	3.4
Vδ2, avr.	0.9	4.6				2.3	2.3	2	1.4	4.6	7.5	1.1
Vδ3, avr.	3.0	3.5	0	5	0	4	1	2.5	0.5	6.5	14.5	0
Vαδ's, avr.	2.8	1.8	2.5	2	3	0.5	2	7	1	2.8	6.5	1
Range	0-7	0-10	0-5	0-5	0-3	0-6	0-4	0-9	0-9	0-8	0-15	0-7

(* Only one or two cDNAs available for each subfamily)

	$E_{V\alpha}$	N	$E_{J\alpha}$
Vα, avr.	3.7	5.8	4.1
Range	0-12	0-16	0-21

Figure 5.

Figure 6.

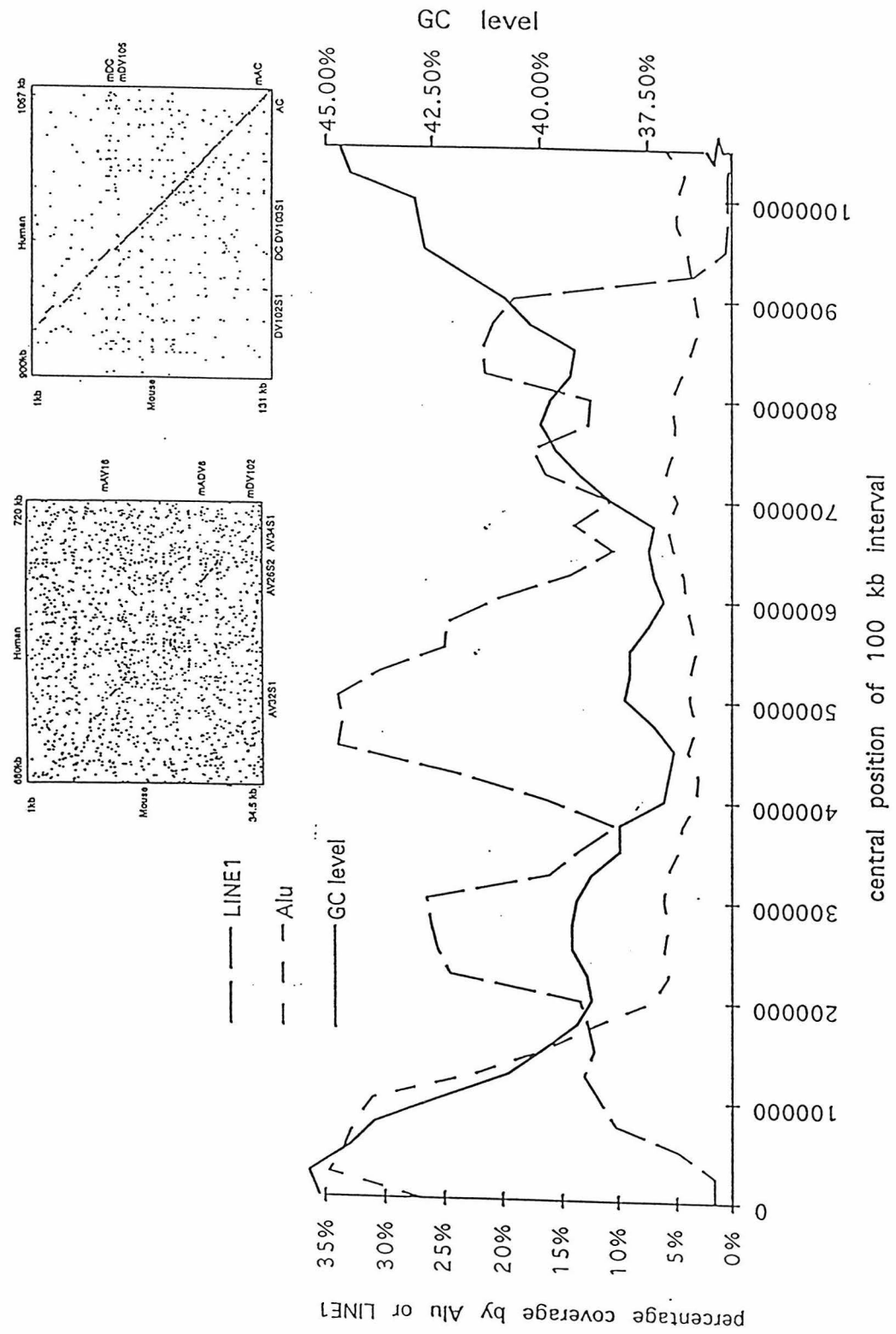
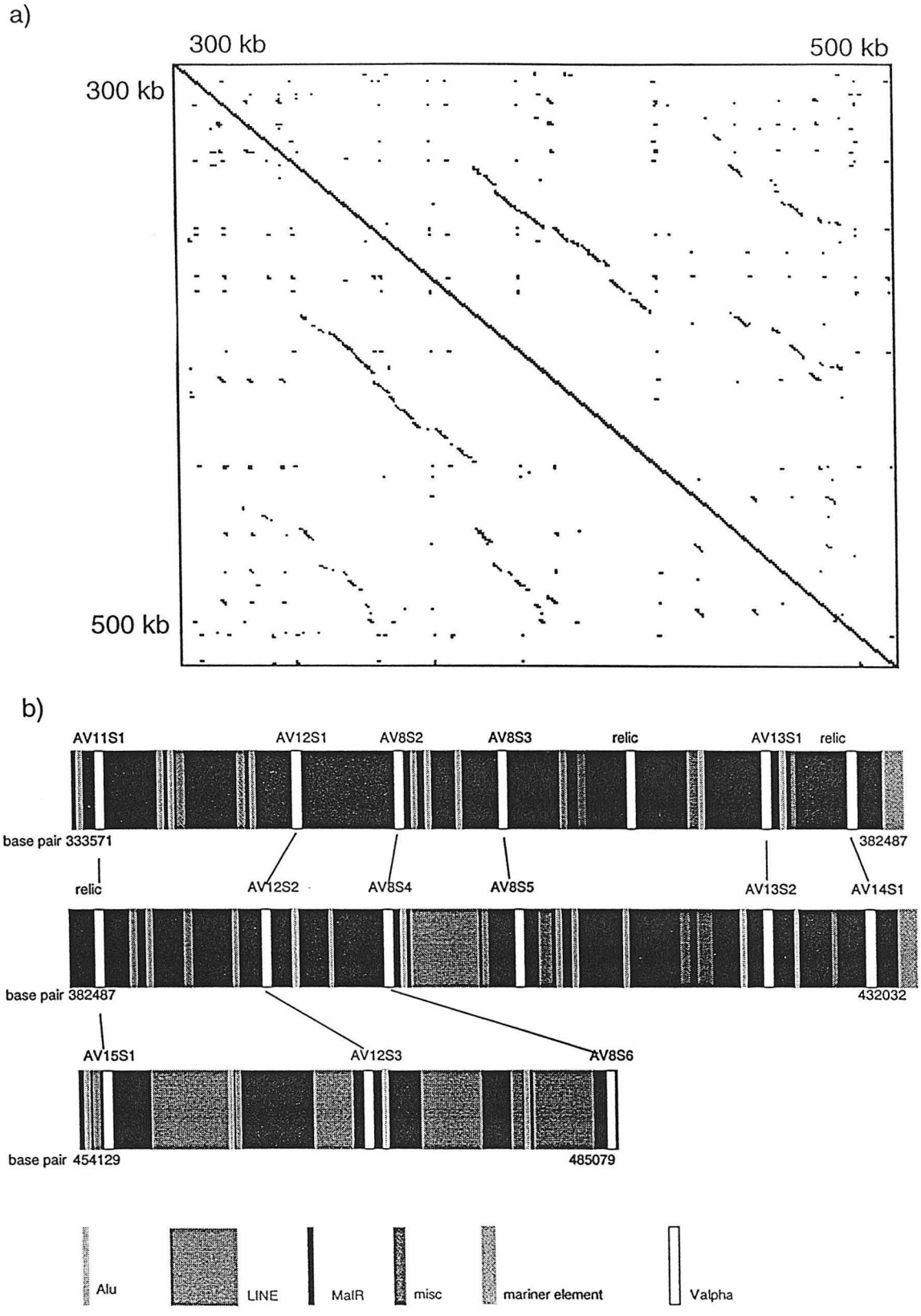

GC level
45.00%
42.50%
40.00%
37.50%
LINE1
Alu
GC level
35%
30%
25%
20%
15%
10%
5%
0%
percentage coverage by Alu or LINE1
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000
central position of 100 kb interval
Human
Mouse
1067 kb
900kb
1kb
131 kb
mDC
mDV102
mAC
AC
DV102S1
DC DV102S1
720 kb
680kb
34.5 kb
mAV16
mADV6
mDV102
AV26S2
AV24S1
AV22S1

Figure 7.


a)
300 kb
500 kb
300 kb
500 kb
b)
AV11S1
AV12S1
AV8S2
AV8S3
relic
AV13S1
relic
base pair 333571
382487
relic
AV12S2
AV8S4
AV8S5
AV13S2
AV14S1
base pair 382487
432032
AV15S1
AV12S3
AV8S6
base pair 454129
485079
Alu
LINE
MalR
misc
mariner element
Valpha

Appendix. Alignment of the translated open reading frame found 5' to the TCR locus with the zinc finger protein XSal1 from Xenopus. A small exon1 and exon 3 were not found, since there is very little homology outside the zinc finger structures.

Identifying DNA polymorphisms in human *TCRA/D* variable genes by direct sequencing of PCR products.

Cecilie Boysen · Christopher Carlson · Eran Hood · Leroy Hood · Deborah A. Nickerson

Cecilie Boysen · Eran Hood
Division of Biology, California Institute of Technology, Pasadena, CA 91125, USA

Christopher Carlson · Leroy Hood · Deborah A. Nickerson ([X])
Box 357730, Department of Molecular Biotechnology, University of Washington, Seattle, WA 98195-7730, USA

The nucleotide sequence data reported in this paper have been submitted to the nucleotide sequence database GenBank and have been assigned the accession numbers U32520-U32549

Abstract

The T cell receptor (TCR) is a highly variable molecule composed of two polypeptide chains that recognize antigenic peptides in the context of major histocompatibility complex (MHC) molecules. In this study, we describe a sequence-based search for germline polymorphisms in the variable (V) gene segments of the human TCR *A/D* locus. Thirty different V gene segments were amplified from six to eight unrelated individuals and sequenced from low melting point agarose. Twenty seven polymorphisms were identified in 15 V gene segments. These polymorphisms are mainly single nucleotide substitutions, but an insertion/deletion polymorphism and a single dinucleotide repeat with variable length were also seen. Of the 15 sequence variations found in the coding regions, six are silent and nine encode amino acid changes. All of the amino acid changes are found at non-conserved residues, frequently in the hypervariable regions, where they may influence MHC and/or peptide recognition. Therefore, it is possible that germline variations in TCR genes could influence an individual's immune response, and may also contribute susceptibility to diseases such as autoimmunity.

Introduction

T cells recognize foreign peptides presented by class I or class II major histocompatibility molecules (MHC) through a heterodimeric receptor composed of alpha and beta chains, or gamma and delta chains. At the protein level, these chains are composed of an antigen recognition or variable (V) domain, and a constant (C) domain, encoded by three discrete gene families: *TCRA/D, TCRB*, and *TCRG* (Marrack and Kappler 1990).

T cell receptor (TCR) diversity is generated by a variety of mechanisms: 1) the multiplicity of discrete germline gene segments encoding the V and joining (J) regions for alpha and gamma chains, and V, diversity (D), and J regions for beta and gamma chains; 2) the combinatorial joining of these gene segments during T cell development; 3) the non-germline (N) addition of nucleotides between the boundaries of the joined gene segments during T cell development (Lieber 1992); 4) the combinatorial association that occurs between the alpha and beta chains, or gamma and delta chains when forming TCRs; and 5) the presence of germline polymorphisms in the gene segments encoding the variable domains.

Structurally, T cell receptors are believed to fold like their antibody counterparts. The V domains each have three hypervariable regions, presumably folding to constitute the walls of the peptide and MHC binding sites. Therefore, germline polymorphisms in these regions could alter potential specificities. Because the structure of the V gene segment is so complex, V region polymorphisms may cause many different types of changes. Each V gene segment contains a promoter region, two exons (a leader and V exon), an intron, and at the 3' end of the last exon a hexamer-spacer-nanomer sequence that mediates DNA rearrangements. Thus, V gene polymorphisms may affect transcription levels (promoter), translation levels (regulatory), compartmentalization (leader), affinity for peptides or MHC molecules, or interactions with the beta chain (variable), RNA stability, the frequency of DNA rearrangements, or the probability of successful RNA splicing (intron). Examples of

some of these effects have been reported for germline polymorphisms in TCRB genes (Gahm et al. 1991, Charmley et al. 1993, Posnett et al. 1994, Vissinga et al. 1994). In constrast, the analysis of germline polymorphisms in *TCRA/D* genes has been less extensive.

The analysis of germline polymorphisms in V gene segments is interesting for two reasons. First, coding polymorphisms may alter certain potential T cell receptor repertoire interactions with MHC and/or peptide antigens. Second, V gene polymorphisms may predispose to autoimmune diseases such as multiple sclerosis or rheumatoid arthritis (Sinha et al. 1990). MHC polymorphisms have clearly been implicated in predisposing to certain autoimmune diseases (Banga et al. 1989, Martin et al. 1992). Studies associating T cell receptor polymorphisms and autoimmune diseases have been far more equivocal (Hillert and Olerup 1992 and Steinman et al. 1992). However, in many cases, these studies have used only one or few markers, taken mainly from the C region (Oksenberg et al. 1988, Hillert et al. 1992, Funkhouser et al. 1992, Hashimoto et al. 1992, and Lynch et al. 1992). Due to the size of the major loci encoding the *TCRA/D* and *TCRB* chains and rate of recombination over time (Robinson and Kindt, 1987), the linkage disequilibrium between markers is generally poor, and it is difficult to draw conclusions based on negative results from using a few random markers (Oksenberg et al. 1988, Hillert et al. 1992, Funkhouser et al. 1992, Hashimoto et al. 1992, and Lynch et al. 1992). Therefore, new polymorphic markers particularly in *TCRA/D* genes, would aid these analyses, and would generate a greater understanding of the type and nature of germline polymorphisms in the TCR.

Material and Methods

PCR. Human genomic DNA from 8 unrelated individuals was used in the identification of DNA polymorphisms in germline TCR sequences. Specific V gene segments (primers given in Table 1) were amplified in 100 ul containing 10 to 100 ng of DNA, 0.3 μM of each primer, 40 μM of each of the four deoxynucleotides (dATP, dCTP,

dGTP, and dTTP), and 25 U/ml Taq DNA polymerase in a buffer composed of 10 mM Tris-HCl pH 8.3, 50 mM KCl, 1.5 mM $MgCl_2$, 0.001% gelatin. The reactions were overlaid with oil and heated to 94°C for 1 min before 35-40 cycles of denaturation at 94°C for 30 sec, annealing for 45 sec, and extension at 72°C for 90 sec followed by a final extension step at 72°C for 5 min. The annealing temperature depended on the primer pair (see Table 1).

DNA sequencing. 100 µl amplification product was ethanol precipitated, resuspended in 10 µl dH_2O and electrophoresed through a 1% low melting point (LMP) agarose, and the band was excised from the gel. The gel plugs were used without further purification by melting the plug immediately prior to adding it as the DNA template for DNA sequencing using the dideoxynucleotide chain termination method as detailed by Kretz et al. (1989). The primers used in the amplification reaction also served as primers for the sequencing reactions. Sequencing reactions were performed in 96-well microtiter plates using either different waterbaths, or more conveniently a 96-well programmable thermocycler (MJ Research). Ten µl of melted agarose plug were added to 5 µl of primer (1.5 µM) and overlaid with oil. The samples were heated to 94°C for 5 min followed by annealing at 37°C for 2-5 min. The temperature was kept at 37°C during the remainder of the reaction. Eight and a half µl sequencing reaction mixture (50 mM Tris-HCl, pH 7.5, 12.5 mM $MgCl_2$, 25 mM dithiothreitol, 0.13 µM of each of the three deoxynucleoside triphosphates, dCTP, dGTP, and dTTP, 10 µCi α-35^S-dATP (>1000 Ci/mmol), and 0.25 U/µl Sequenase Version 2.0 (United States Biochemical)) were added and the reactions incubated for 5 min. During this incubation, 2 µl of each of the four termination mixes were added to separate microtiter wells on the same plate that contained 80 µM of each dATP, dCTP, dGTP, and dTTP, 50 mM NaCl, and 8 µM of their respective dideoxy-nucleotide. Four µl of the reaction mixture were transferred to each of the termination mixes, the samples overlaid with oil, and incubated for 5 min, before the reactions were stopped with 5 µl of stop solution (formamide, 10 mM EDTA, xylene cyanol FF, and

bromophenol blue). The sequencing products were heated to 94°C, and 2 µl of each loaded onto a 6% polyacrylamide gel which was electrophoresed for 2-5 hours at 90 Watts, transferred to Whatman filter paper, dried, and exposed overnight to Kodak X-OMAT AR film.

Results

Identification of DNA polymorphisms in *TCRA/D* V gene segments.

From cDNA analysis, the *TCRA/D* locus is known to contain about 50 V gene segments which are classified into different V gene families based on sequence homology (members of individual V gene families have greater than 75% sequence similarity with one another). To search for germline sequence polymorphisms in *TCRA/D* genes, specific PCR assays were developed for at least one member from most of the gene families (see Table 1). The reported *DV4S1, DV5S1, DV6S1, DV7S1*, and *DV8S1* genes are the same as *AV6S1, AV21S1, AV17S1, AV28S1,* and *AV14S1*, respectively, and have not been listed with the other *TCRD* genes (*DV101S1-DV103S1*). A few V gene segments were excluded from our analysis since they had been examined extensively at the sequence level (Wright et al. 1991, and Charmley et al. 1994a).

In developing PCR assays, the forward primer was chosen in the leader sequence when sequence information was available from this area (exon 1), while the reverse primer was chosen from sequences at the 3' end of exon 2 to maximize the number of base pairs screened for sequence polymorphisms. In some cases, especially within the *AV1* and *AV2* multigene subfamilies, specific primer pairs were difficult to obtain and those V gene segments were not routinely included in this analysis.

Screening for germline DNA polymorphisms by sequencing PCR products obtained from 6 to 8 individuals was simplified by loading the gel with all the A reactions from each individual next to each other, all the C reactions next to each other, and so on. This loading scheme made it easy and fast to scan for sequence polymorphisms in V gene

segments as shown in Figure 1. All PCR products were sequenced from both ends. Using this approach, the majority of sequence between the primers could be scanned for common polymorphisms. It was immediately clear whether an individual was heterozygous or homozygous for each allelic variant (Figure 1).

Of 30 V gene segments scanned by this method, 15 were found to contain one or more polymorphisms. A total of 27 DNA sequence polymorphisms were detected altogether (Table 2). In approximately 6500 nucleotides of coding region, 15 polymorphisms were found (Table 3), approximately one variation every 430 bps, when sequences from several individuals were scanned. The same frequency of DNA polymorphisms was found in the intronic sequences, where 12 polymorphisms were found in the 5200 bps studied. Nucleotide diversity, i.e. the number of differences per nucleotide site for all pair-wise combination of two random chromosomes among the individuals scanned (6 to 8 individuals, Table 2), for *TCRA/D* genes was calculated to be approximately one variation every 1,250 bp (0.08%, Table 3) and again, was similar for intron and exon based sequences.

Nature of the DNA polymorphisms and amino acid changes in the *TCRA/D* V gene segments

The majority of germline polymorphisms were single base substitutions as indicated in Table 2, although a few insertion/deletion variations were identified. Among the latter group was a polymorphic short tandem repeat (STR) in the intron of *AV22S1*. Three different allelic forms of this STR, $(CT)_7$, $(CT)_{11}$, and $(CT)_{12}$, were found among the 16 chromosomes (8 individuals) analyzed. Additionally, we have found that sequences which were previously thought to be two different V gene segments were actually allelic variants of the same gene segment (*AV2S1/S3* and *AV2S4/S5*).

Of the 15 single nucleotide substitution polymorphisms found in the coding region, nine would lead to amino acid changes in the TCR (Table 2). The distribution of these across the V domain is indicated in Figure 2. Conservative substitutions such as those in

DV102S1 (Val-Ile, Ile-Met) and *AV29S1* (Glu-Asp) were detected, as well as more dramatic substitutions like the glutamic acid to glutamine in *AV6S1* (acidic to polar), the polar to non-polar changes in *AV2S1/S3* (Val-Gly, Ser-Phe), *AV4S2* (Thr-Pro), and *AV6S1* (Gln-Pro), and a basic arginine to an uncharged, polar glycine in *AV29S1*. Furthermore, a number of these amino acid substitutions (4 of 9) were found to be located in hypervariable regions (Figure 2).

Discussion

The identification of DNA sequence variations plays a central role in the analysis of the relationship between genome structure and function. This is particularly true in regard to genes such as those encoding the MHC proteins which exhibit significant diversity in human populations. Recently, similar analyses have been undertaken with other immune genes such as the TCR. However, the diversity of these does not appear as extensive as the MHC locus (Marsh and Bodmer 1993, Zemmour and Parham 1993). In general, human DNA polymorphisms are estimated to occur on average once in every 500 to 1,500 bp (Cooper et al. 1985, Li and Sadler, 1991). In human *TCRA/D* genes, we found one polymorphic site on average every 433 bp in both exons and introns when 12-16 chromosomes are being compared (Table 3). The overall nucleotide diversity in *TCRA/D* genes was found to be 8.0×10^{-4} (0.08 %), or one variation in every 1250 bp. This diversity is similar to that previously reported by Li and Sadler (1991) who compared sequences from 49 different genes representing approximately 75,000 unique bp of human DNA sequence. It is also similar to the levels of germline polymorphism reported previously in *TCRB* genes using direct sequence analysis (Posnett 1990, Cornelis et al. 1993, Charmley et al. 1994b, Wei et al. 1995, Charmley and Concannon 1995).

A number of approaches have been applied to finding DNA polymorphisms in TCR V genes. These include methods that: i) compare the cleavage patterns in DNA sequences following treatment with a restriction enzyme (Robinson and Kindt 1987, Grier et al.

1990, and Zhang and LeFranc 1993, Oksenberg et al. 1988), ii) determine whether there are differences in the melting temperatures of specific TCR sequences (Nickerson et al. 1992, Charmley et al. 1994a), or iii) detect changes in the sequence conformation following denaturation and renaturation under conditions to promote the annealing of single strands (Cornelis et al. 1993, and Ibberson et al. 1995). In many cases, the sequence basis of these RFLP, melting, or conformational variants has been subsequently determined. Ten *TCRA* V gene segments have previously been shown to contain 20 polymorphisms using a combination of these strategies to identify polymorphisms (Wright et al. 1991, Charmley et al. 1994a, Moss et al. 1993, Reyburn et al. 1993, Cornelis et al. 1993, Ibberson et al. 1995). However, when compared side by side with other approaches, we found DNA sequencing to be the most rapid and direct approach for identifying new DNA polymorphisms. Furthermore, the speed, automation, and accuracy of DNA sequencing is rapidly improving particularly in regard to the identification of human DNA variations (Kwok et al. 1994).

The identification of DNA polymorphisms by direct sequencing offers several other advantages. First, it is the most sensitive scanning approaches available and can identify all the variations present in the sequences using a single set of assay conditions (Nickerson et al. 1992, and Kwok et al. 1994). This is difficult to achieve with other approaches and often requires the development of special PCR primers (Sheffield et al. 1989), or numerous gel runs under varying conditions to achieve maximum sensitivity which in the end may not approach 100% (Leren et al. 1993). In this study, 27 polymorphisms were identified by direct sequence analysis of 30 V genes. The majority of these were not previously detected (19 of 27 polymorphisms) when other approaches were applied to these gene segments (Charmley et al. 1994a, Cornelis et al. 1993, Ibberson et al. 1995). In addition to its sensitivity, direct sequence analysis can also provide new sequence information, i.e. when cDNA sequences are used to develop PCR primers, new sequences from intervening

introns can be obtained. In fact, more than 4500 bps of previously unknown intronic sequence was uncovered during this analysis of the *TCRA/D* genes.

Another advantage to finding polymorphisms by DNA sequence analysis is that it provides precise information on the nature and location of the variation. Of the coding region variations found in the *TCRA/D* genes, nine would lead to amino acid changes in the TCR. Although the atomic structure of the human TCR has not yet been determined, it is thought to be very similar to that of the immunoglobulin structure based on sequence comparisons (Chothia et al. 1988, Davis and Bjorkman, 1988). Alignment of V sequences for the alpha chain indicates there are approximately 40 conserved residues of 92 amino acids total. None of the amino acid changes reported here were located in the conserved residues but polymorphisms were found in hypervariable and non-conserved sites in the TCR (Figure 2). Some of these polymorphisms are conservative substitutions, while others are less conservative substitutions like the glutamic acid to glutamine in *AV6S1* (acidic to polar), or the three polar to non-polar changes in *AV2S1/S3* and *AV6S1*. It is worth noting that these latter substitutions occur primarily in regions equivalent to the hypervariable regions in immunoglobulin (Figure 2). Residues in these regions are thought to be involved in binding to the MHC molecules (Davis and Bjorkman, 1988), and therefore, could influence thymic selection of T cells bearing these receptors. However, further studies will be required to determine whether these changes in fact influence the functional TCR repertoire in individuals.

Finally, once the sequence of a DNA variation is known, it can be typed in human populations on a large-scale using high-throughput and semi-automated methods such as the oligonucleotide ligation assay, OLA (Nickerson et al. 1990), genetic bit analysis (Nikiforov et al. 1994), or by allele-specific oligonucleotide hybridization during PCR (Taqman-ASO, Livak et al. 1995). In fact, we have already developed semi-automated typing formats (PCR/OLA) for a number of these polymorphisms. In this regard, genetic diversity in MHC genes has been linked to a number of disease susceptibilities, including

several human autoimmune diseases (Sinha et al. 1988 and Todd et al. 1988). Similar studies examining DNA variations from the *TCRA/D* have been contradictory (Hillert and Olerup 1992 and Steinman et al. 1992). Therefore, the sequence variations reported here may prove useful in further assessing the relationship between germline TCR polymorphisms and genetic susceptibility to disease in human populations.

Acknowledgements

We thank Dr. Masa Toda for his helpful suggestions. This work was supported by a grant from the National Institutes of Health (HG 00464) and DIR 8809710 from the National Science Foundation.

References

Banga, J. P., Barnett, P. S., Mahadevan, D., and McGregor, A. M. Immune recognition of antigen and its relevance to autoimmune disease: recent advances at the molecular level. Eur J Clin Invest 19: 107-116, 1989.

Charmley, P., Wang, K., Hood, L., and Nickerson, D.A. Identification and physical mapping of a polymorphic human T cell receptor V beta gene with a frequent null allele. J Exp Med 177: 135-143, 1993.

Charmley, P., Nickerson, D., and Hood, L. Polymorphism detection and sequence analysis of human T-cell receptor Vα-chain-encoding gene segments. Immunogenetics 39: 138-145, 1994a.

Charmley, P., Nepom, B. S., and Concannon, P. HLA and T cell receptor β-chain DNA polymorphisms identify a distinct subset of patients with pauciarticular-onset juvenile rheumatoid arthritis. Arthritis & Rheumatism 37: 695-701, 1994b.

Charmley, P. and Concannon, P. PCR-based genotyping and haplotype analysis of human TCRBV gene segment polymorphisms. Immunogenetics 42, 254-261, 1995.

Chothia, C., Boswell, D. R., and Lesk, A. M. The outline structure of the T-cell α/β receptor. EMBO 7: 3745-3755, 1988.

Cooper, D.N., Smith-B.A., Cooke, H. J., Niemann, S., and Schmidtke, J. An estimate of unique DNA sequence heterozygosity in the human genome. Hum Genet 69: 201-205, 1985.

Cornelis, F., Pile, K., Loveridge, J., Moss, P., Harding, R., Julier, C., and Bell, J. Systematic study of human αβ T cell receptor V segments shows allelic variations resulting in a large number of distinct T cell receptor haplotypes. Eur J Immunol 23: 1277-1283, 1993.

Davis, M. M., and Bjorkman, P. J. T-cell antigen receptor genes and T-cell recognition. Nature 334: 395-402, 1988.

Funkhouser, S. W., Concannon, P., Charmley, P., Vredevoe, D. L., and Hood, L. Differences in T cell receptor restriction fragment length polymorphisms in patients with rheumatoid arthritis. Arthritis & Rheumatism 35: 465-471, 1992.

Gahm, S.-J., Fowlkes, B. J., Jameson, S. C., Gascoigne, N. R. J., Cotterman, M. M., Kanagawa, O., Schwartz, R. H., and Matis, L. A. Profound alteration in an αβ T-cell antigen receptor repertoire due to polymorphism in the first complementarity-determining region of the β chain. Proc Natl Acad Sci USA 88: 10267-10271, 1991.

Grier, A. H., Mitchell, M. P., and Robinson, M. A. Polymorphism in human T cell receptor alpha chain variable region genes. Exp Clin Immunogenetics 7: 34-42, 1990.

Hashimoto, L. L., Mak, T. W., and Ebers, G. C. T cell receptor α chain polymorphisms in multiple sclerosis. J Neuroimmunol 40: 41-48, 1992.

Hillert, J. and Olerup, O. Germ-line polymorphism of TCR genes and disease susceptibility- fact or hypothesis? Immunol Today 13: 47-49, 1992.

Hillert, J., Leng, C., and Olerup, O. T-cell receptor α chain germline gene polymorphisms in multipe sclerosis. Neurology 42, 80-84, 1992.

Ibberson, M. R., Copier, J. P., and So, A. K. Genomic organization of the human T-cell receptor variable α (TCRAV) gene cluster. Genomics 28, 131-139, 1995.

Kretz, K. A., Geoffrey, S. C. and O'Brien, S. O. Direct sequencing from low-melt agarose with Sequenase. Nucleic Acid Res 17: 5864, 1989.

Kwok, P.Y., Carlson, C., Yager, T.D., Ankener, W., Nickerson, D.A. Comparative analysis of human DNA variations by fluorescence-based sequencing of PCR products. Genomics 23:138-144, 1994.

Leren, T.P., Solberg, K., Rodningen, O. K., Ose, L., Tonstad, S., and Berg, K. Evaluation of running conditions for SSCP analysis: application of SSCP for detection of point mutations in the LDL receptor gene. PCR Methods Appl 3: 159-162, 1993.

Li, W.-H. and Sadler, L. A. Low nucleotide diversity in man. Genetics 129: 513-523, 1991.

Lieber, M. R. The mechanism of V(D)J recombination.: A balance of diversity, specificity, and stability. Cell 70: 873-876, 1992.

Livak, K.J., Marmaro, J., and Todd, J.A. Towards fully automated genome-wide polymorphism screening. Nat Genet 9: 341-342, 1995.

Lynch, S. G., Rose, J. W., Petajan, J. H., and Leppert, M. Discordance of the T-cell receptor alpha-chain gene in familial multiple sclerosis. Neurology 42,839-844, 1992.

Marrack, P. and Kappler, J. W. The T cell receptors. Chem. Immunol. 49: 69-81, 1990.

Marsh, S. G., Bodmer, J.G. HLA class II nucleotide sequences. Immunogenetics 37: 79-94, 1993.

Martin, R., McFarland, H. F., and McFarlin D. E. Immunological aspects of demyelinating diseases. Annu Rev Immunol 10: 153-187, 1992.

Moss, P. A. H., Rosenberg, W. M. C., Zintzaras, E., and Bell, J. I. Characterization of the human T cell receptor α-chain repertoire and demonstration of a genetic influence on Vα usage. Eur J Immunol 23: 1153-1159, 1993.

Nickerson, D. A., Whitehurst, C., Boysen, C., Charmley, P., Kaiser, R., and Hood, L. Identification of clusters of biallelic polymorphic sequence-tagged sites (pSTSs) that generate highly informative and automatable markers for genetic linkage mapping. Genomics 12: 377-387, 1992.

Nickerson, D. A., Kaiser, R., Lappin, S., Stewart, J., Hood, L., and Landegren, U. Automated DNA diagnostics using an ELISA-based oligonucleotide ligation assay. Proc Natl Acad Sci USA 87: 8923-8927, 1990.

Nikiforov, T. T., Rendle, R. B., Goelet, P., Rogers, Y. H., Kotewicz, M. L., Anderson, S., Trainor, G. L., and Knapp, M. R. Genetic bit analysis: a solid phase method for typing single nucleotide polymorphisms. Nucleic Acids Res 22: 4167-4175, 1994.

Oksenberg, J. R., Gaiser, C. N., Cavalli-Sforza, L. L., and Steinman, L. Polymorphic markers of human T-cell receptor alpha and beta genes. Family studies and comparison of frequencies in healthy individuals and patients with multiple sclerosis and myasthenia gravis. Human Immunol 22: 111-121, 1988.

Posnett, D. N. Allelic variations of human TCR V gene products. Immunol Today 11: 368-373,1990.

Posnett, D.N., Vissinga, C.S., Pambuccian, C., Wei, S., Robinson, M.A., Kostyu, D., Concannon, P. Level of human TCRBV3S1 (V beta 3) expression correlates with allelic polymorphism in the spacer region of the recombination signal sequence. J Exp Med 179: 1707-1711, 1994.

Reyburn, H., Cornélis, F., Russell, V., Harding, R., Moss, P., and Bell, J. Allelic polymorphism of human T-cell receptor V alpha gene segments. Immunogenetics 38: 287-291, 1993.

Robinson, M. A. and Kindt, T. J. Genetic recombination within the human T-cell receptor α-chain gene complex. Proc Natl Acad Sci USA 84: 9089-9093, 1987.

Sheffield, V.C., Cox, D.R., Lerman, L.S., Myers, R.M. Attachment of a 40-base-pair G + C-rich sequence (GC-clamp) to genomic DNA fragments by the polymerase chain reaction results in improved detection of single-base changes. Proc Natl Acad Sci USA 86: 232-236, 1989.

Sinha, A. A., Brautbar, C., Szafer, F., Friedmann, A., Tzfoni, E., Todd, J. A., Steinmann, L., and McDevitt, H. O. A newly characterized HLA DQβ allele associated with pemphigus vulgaris. Science 239: 1026-1029, 1988.

Sinha, A. A., Lopez, M. T., and McDevitt, H. O. Autoimmune diseases: The failure of self tolerance. Science 248: 1380-1388, 1990.

Steinman, L., Oksenberg, J. R., and Bernard, C. C. A. Association of susceptibility to multiple sclerosis with TCR genes. Immunol Today 13: 49-51, 1992.

Todd, J.A., Acha-Orbea, H., Bell, J. I., Chao, N., Fronek, Z., Jacob, C. O., McDermott, M., Sinha, A. A., Timmerman, L., Steinmann, L., McDevitt, H. O. A molecular basis for MHC class II-associated autoimmunity. Science 240: 1003-1009, 1988.

Vissinga, C.S., Charmley, P., Concannon, P. Influence of coding region polymorphism on the peripheral expression of a human TCR V beta gene. J Immunol 152: 1222-1227, 1994.

Wei, S., Charmley, P., Birchfield, R. I., and Concannon, P. Human T-cell receptor Vβ gene polymorphism and multiple sclerosis. Am J Hum Genet 56, 963-969, 1995.

Wright, J. A., Hood, L., and Concannon, P. Human T-cell receptor Vα gene polymorphism. Hum Immunol 32, 277-283, 1991.

Zemmour, J., and Parham, P. HLA Class I nucleotide sequences. Immunogenetics 37: 239-250, 1993.

Zhang, X. M. and LeFranc, M.-P. PCR-based detection of one BamHI polymorphic site in the human T cell receptor delta gene TCRDV2. Hum Genet 92: 100, 1993.

Table 1. Amplification and sequencing primers.

Gene	5' Primer	3' Primer	Annealing Temperature	Genbank Accession
DV101S1	AAGGTTACTCAAGCCCAGTC	CTGTAAGGCTGAAATGGTTAAG	60°C	U32547
DV102S1	TTGGTGCCTGGACACCAAAC	GATGGTGCAAGTATCTTAAGTA	60°C	U32548
DV103S1	CAGAGTTCCCCGGACCAGAC	CCTTACTGGAGAGATCACCA	60°C	U32549
AV1S1	TCATACCAGTGCTGGGGA	GGCACAGAAGTACTCAGCT	58°C	U32520
AV2S1/S3	TGAAATCCTTGAGAGTTTTACTA	GAGAAACATACTGGCTGG	60°C	U32539
AV2S4/S5	GGCATCTCTGTAGAAACATA	GTCTCTGATGAACAAGGAGAT	60°C	U32538
AV3S1	CTGGGAGTGTCTTTGGTGATT	AAGGAACTGCTTTTCTTGGAAG	60°C	U32540
AV4S2	GGCTGGTGGCAAGAGTAACTG	GCGTAGCGTGGGGCAGGA	60°C	U32541
AV5S1	GGAGACGAATGGAGTCATCC	GGCTGTGATATGAAACAAACTC	58°C	U32542
AV6S1	GTCACTTTCTAGCCTGCTGA	CAGTTGTGAAGCGGAGATGACA	60°C	U32543
AV7S2	GTGGGGAGTTTTCCTTCTT	TTCATCTGGAGCTCCTTCAA	58°C	U32544
AV8S2	ATTCGAGCTTTATTTATGTACTTGT	AATTTGCAGAGAGAGATGTTTCA	60°C	U32545
AV9S1	ATGAAGCCCACCCTCATCT	TGAGTCTTCCTCTTGAGCAA	58°C	U32546
AV10S1	GTCCTGAAATTCTCCGTGTCCA	AGTGATGTGGAGAGAACTGTC	60°C	U32521
AV11S1	ATGGCTTTGCAGAGCACTCTGG	CTGGAGGATGAGCAGCGATG	60°C	U32522
AV13S1	GGACCTCTGCTGGGGCTC	TGTGGTCTGGGAAGAGGAA	60°C	U32523
AV14S1	CCTGGCTTCCTGTGGGCA	GTGAGTCTGAGATCTTGAGAC	60°C	U32524
AV16S1	GCCTCTGCACCCATCTCGA	CAAAGCGGAGTCGCTCACAA	58°C	U32525
AV17S1	TAGTTCTGTGGCTTCAACTATG	AGTCTCCAGGCTGGGAATCCA	60°C	U32526
AV18S1	GAATCCTTTGGCAGCCCCATTA	AAATAGCTGTAACCCTCCTTGG	60°C	U32527
AV19S1	AGATCCGGCAATTTTTGTTGGCT	GGGAGCTGTGCTTTTCCTGTA	60°C	U32528
AV20S1	GGCAAGTGGCGAGAGTGATC	AGTGTCGCTCAGGGAAACCCG	60°C	U32529
AV22S1	AGGCTTAGTATCTCTGATACTC	ACACCGCTGAGTCTGACACT	59°C	U32530
AV23S1	GTCTAAGTGACAGAAGGAATG	AATGTATAAAGTACTACGTCCTGA	60°C	U32531
AV24S1	GCATCTGACGACCTTCTTGGT	ATGTAGGAGGCTGAATCGCTGAG	60°C	U32532
AV25S1	ATGCTCCTTGAACATTTATTA	GTAGATGCCTACATCACTAGG	60°C	U32533
AV26S1	GAGACTGTTCTGCAAGTACTCCTA	AGGTAGATGCCTGCATGGCTGG	60°C	U32534
AV27S1	ATGAAGAAGCTACTAGCAATG	GTAGGTGGCAGAGAGGTCATG	60°C	U32535
AV28S1	ATGATGAAGTGTCCACAGGCT	GGTAGACGGCCGAGTCTCCGG	60°C	U32536
AV29S1	ATGGAGACTCTCCTGAAAGTGC	AAGTAGGTTCCTGAGTAACTG	60°C	U32537

Genbank accession numbers for sequences, introns, and alternative alleles.

Table 2. Polymorphisms identified by sequencing V genes.

V gene	Allele 1	Polymorphism	Allele 2	Frequency of Allele 2[a]	Reference
DV102S1	GGGGTCCCT Gly Val Pro	G-A Val-Ile 16	GGGATCCCT Gly Ile Pro	3/12	Zhang & LeFranc 1993
DV102S1	ACAATCACT Thr Ile Thr	C-G Ile-Met 45	ACAATGACT Thr Met Thr	4/12	
AV2S1/S3	TAATGTACA	G-T Intron	TAATTTACA	8/16	Charmley et al. 1994a
AV2S1/S3	CGAGTTTCC Arg Val Ser	T-G Val-Gly 28	CGAGGTTCC Arg Gly Ser	11/16	
AV2S1/S3	ATGTCCATA Met Ser Ile	C-T Ser-Phe 48	ATGTTCATA Met Phe Ile	4/16	Charmley et al. 1994a
AV2S4/S5	GAAACATGA	C-T Intron	GAAATATGA	2/16	
AV3S1	CAAACTTTA	C-T Intron	CAAATTTTA	3/12	
AV4S2	ATAACTAAC	C-T Intron	ATAATTAAC	2/16	
AV4S2	CCCACCTCC Pro Thr Ser	A-C Thr-Pro 8	CCCCCCTCC Pro Pro Ser	15/16	
AV5S1	GAGGCCCTG Glu Ala Leu	C-T Silent	GAGGCTCTG Glu Ala Leu	3/16	
AV6S1	GGTACGGGT	C-T Intron	GGTATGGGT	11/16	
AV6S1	TAAATCTTC	T-C Intron	TAAACCTTC	10/16	
AV6S1	GATCAAAGT Asp Gln Ser	A-C Gln-Pro 29	GATCCAAGT Asp Pro Ser	10/16	Reyburn et al. 1993
AV6S1	GGTCTATTC Gly Leu Phe	A-C Silent	GGTCTCTTC Gly Leu Phe	1/16	Reyburn et al. 1993
AV6S1	GACGAGCAA Asp Glu Gln	G-C Glu-Gln 55	GACCAGCAA Asp Gln Gln	10/16	Reyburn et al. 1993
AV7S2	GCACCCACA Ala Pro Thr	C-T Silent	GCACCTACA Ala Pro Thr	4/16	
AV10S1	CAGCTGCTG Gln Leu Leu	G-A Silent	CAACTGCTG Gln Leu Leu	1/14	
AV14S1	TTCTCTCTC	C-G Intron	TTCTGTCTC	12/16	
AV18S1	TTTTTTAAAAAAA	Del/Ins Intron	TTTTTTAAAAAAAA	2/12	
AV22S1	AAACAGA$(TC)_n$	CT-repeat Intron	n=7, 11, or 12	7/16, 2/16, 7/16	
AV22S1	CGTAAAGAA Arg Lys Glu	A-G Silent	CGTAAGGAA Arg Lys Glu	2/16	
AV23S1	TGCTCTTTT	C-G Intron	TGCTGTTTT	8/16	
AV23S1	GTGACACAG Val Thr Gln	A-G Silent	GTGACGCAG Val Thr Gln	8/16	
AV27S1	ATGCCTCCT	C-T Intron	ATGCTTCCT	3/16	
AV27S1	AAATGTTCT	G-A Intron	AAATATTCT	2/16	
AV29S1	AAGCGTCAT Met Arg Arg	C-G Arg-Gly 57	AAGGGTCAT Met Gly Arg	1/12	Moss et al. 1993
AV29S1	CATGAAAAA Arg Glu Lys	A-C Glu-Asp 59	CATGACAAA Arg Asp Lys	4/12	Moss et al. 1993

[a] Frequency of allele 2 among the analyzed chromosomes.

Table 3. Types and frequency of DNA polymorphisms in *TCRA/D* genes.

		Polymorphism Type				
Location	# of bp scanned	Substitutions	InDel	# of Polymorphisms	Frequency[a]	Nucleotide diversity[b]
Intron	5200	10	2	12	1/433	0.087 %
Exon	6500	15	0	15	1/433	0.075 %
Total	11700	25	2	27	1/433	0.080 %

[a]The frequency of polymorphism per base-pair of scanned sequence using 6 to 8 individuals (12 to 16 chromosomes), approximately 1 polymorphism detected every 433 bp scanned.

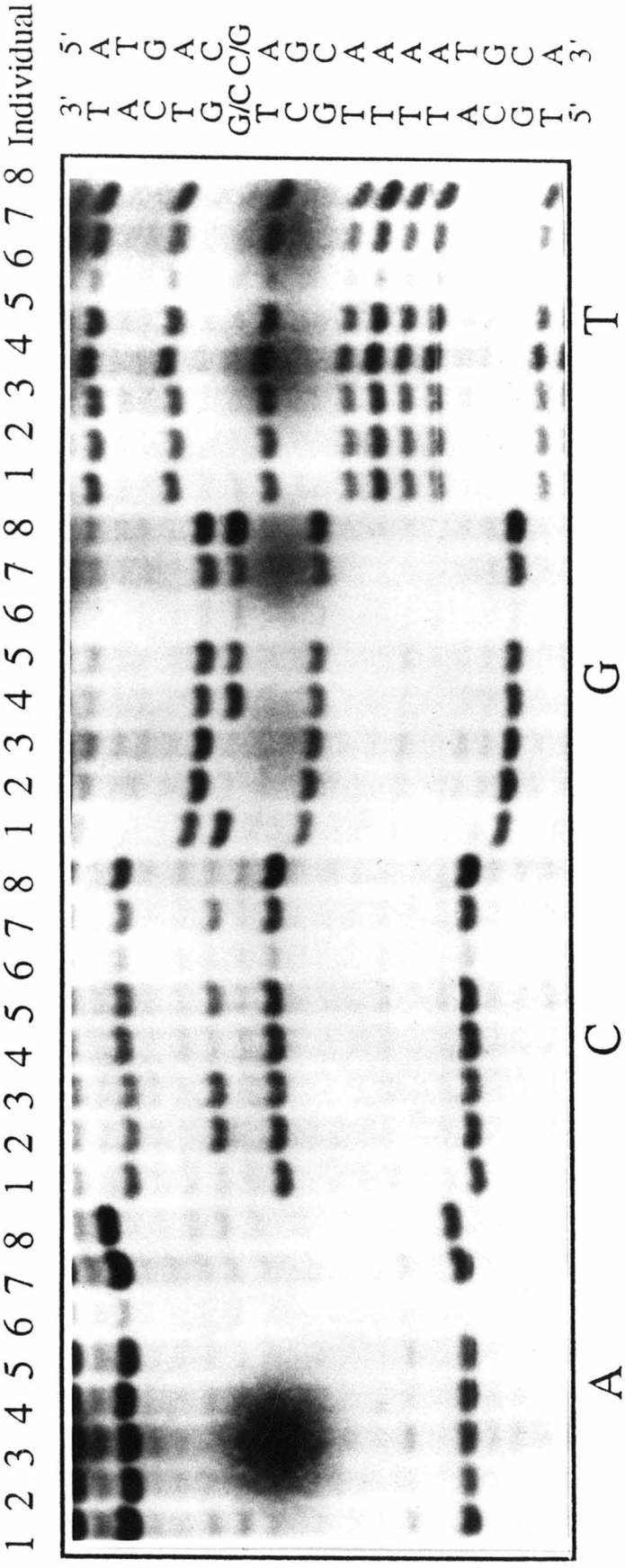
[b]The average number of differences per nucleotide site taken from all pair-wise combinations of two random chromosomes among the individuals scanned.

Figure Legends

Figure 1. Human *AV6S1* PCR product sequenced with the 3' PCR primer for eight individuals. A transversion (G to C) in exon 2 leading to a Glu -> Gln substitution in amino acid 55 of the mature peptide is shown.

Figure 2. A schematic diagram of the locations of the V gene segment polymorphisms in the mature alpha or delta peptide.

Figure 1.


Individual
1 2 3 4 5 6 7 8 1 2 3 4 5 6 7 8 1 2 3 4 5 6 7 8 1 2 3 4 5 6 7 8
3' 5'
T A
A T
C G
T A
G C
G/C C/G
T A
C G
G C
T A
T A
T A
T A
A T
C G
G C
T A
5' 3'
A
C
G
T

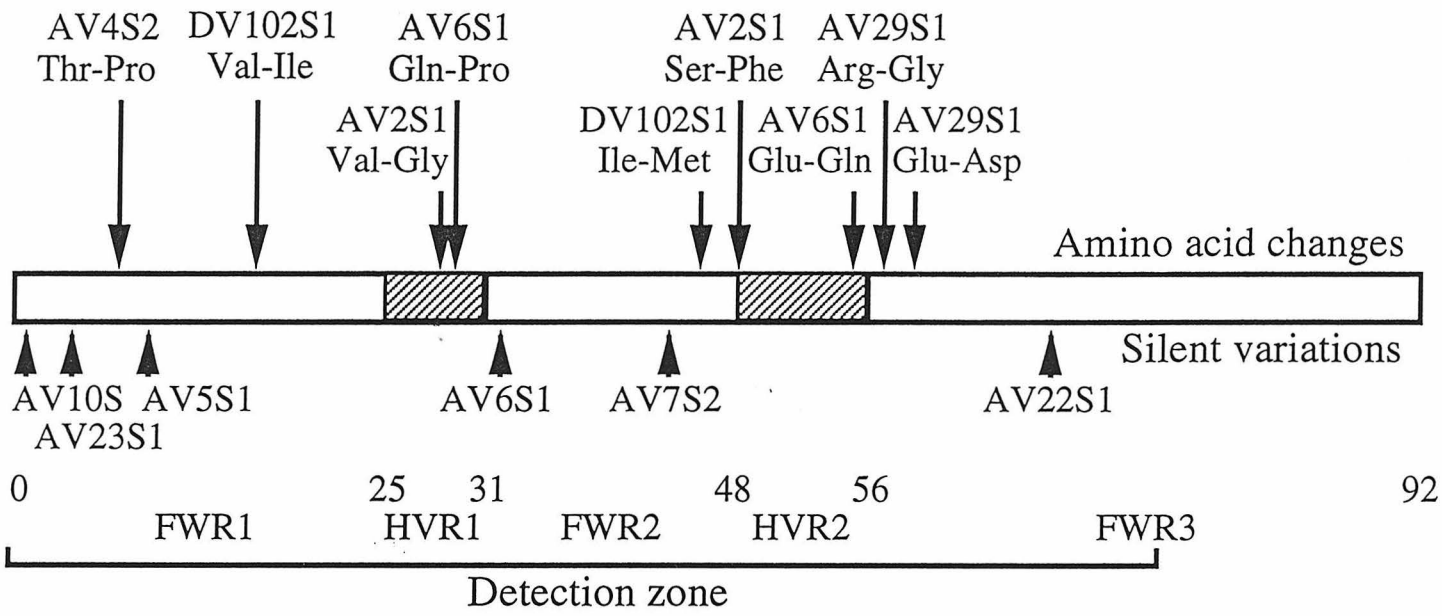


Figure 2. Position of coding region TCR α/δ V gene segment polymorphisms in relation to the Vα domain. FWR: Frame work region, HVR: Hyper variable region, as suggested by Chothia et al.,1988. Numbers indicate amino acid position in the Vα domain.

The use of Bacterial Artificial Chromosomes (BACs) as Mapping and Sequencing Reagents

Cecilie Boysen[1], Melvin I. Simon[1], Leroy Hood[2]

1 Division of Biology 147-75, California Institute of Technology, Pasadena, CA 91125

2 Department of Molecular Biotechnology 357730, University of Washington, Seattle, WA 98195

Correspondence to Leroy Hood

Abstract

Bacterial artificial chromosome (BAC) clones were used to map and sequence the human T cell receptor α/δ locus. Seventeen BAC clones were analyzed in detail. They were found to be excellent mapping and sequencing reagents, exhibiting the following properties: (*i*) The 17 BAC clones covered the 1.1 megabase (Mb) region with the exception of one small gap expected from a 3.7 fold library. (*ii*) The ends of the BAC inserts were randomly distributed. (*iii*) The BAC clones faithfully represented the genomic DNA with the exception of a single clone. (*iv*) The sequence from the ends of the BAC inserts could be obtained directly from the BAC DNA by the chain terminator method. (*v*) The complete BAC inserts could be sequenced directly by the shotgun approach. These properties have led to a new approach to sequence the human genome.

Introduction

As the genome project moves into its sequencing phase, effective integration of large-scale physical mapping and large-scale DNA sequencing becomes increasingly important. Currently, this process is typically carried out in three discrete steps. (*i*) A low resolution physical map is developed from genome-wide or chromosome-specific yeast artificial chromosomes (YACs) ranging in size from ~100 kilobases (kb) to 1 (Mb) (Chumakov et al., 1995, and Doggett et al., 1995). The insert DNA is prepared by partial restriction enzyme digestion, thus generating somewhat random overlaps. Typically, 2- to 10-fold coverage has been achieved. (*ii*) Sequence-ready maps are prepared by subcloning YACs, after partial restriction digestion, into cosmid vectors with insert sizes typically ranging from 30-40 kb. Generally, 5- to 10-fold coverage is sought. (*iii*) A minimum tiling (overlap) path of cosmid clones is selected for sequence analysis. Individual cosmid clones are then randomly sheared into ~1 kb fragments for subcloning into phage M13 vectors and DNA from the M13 clones sequenced using M13 vector primers. Typically 600-900 forward sequence reads are assembled into contigs; contig closure and editing generally requires additional sequences from reverse reads, primer directed sequencing, or selected PCR amplification. This approach or an alternative where genome-wide or chromosome-specific cosmid libraries are the starting material has been employed by most large-scale DNA sequencing laboratories.

This approach has several severe limitations. (*i*) YAC, and even cosmid inserts, are often chimeric and/or suffer from deletions (Green et al., 1991). The determination of clone fidelity is time consuming and small deletions can be difficult to identify. (*ii*) YAC clones (or their subclones) must

be purified from yeast DNA. (*iii*) Cosmid inserts may be shorter than tandemly repeated DNA arrays, thus rendering physical mapping and sequence closure difficult. For example, the human β T cell receptor locus has five tandemly arrayed 10 kb homology units 90-92% similar to one another (Rowen et al., 1996). (*iv*) The need to subclone DNA three times and create maps at two different levels adds complexity and expense to the potential automation of large-scale DNA sequencing.

Bacterial artificial chromosome (BAC) inserts appear to offer an interesting alternative to YAC inserts for physical mapping and sequencing in that preliminary results indicated that they are relatively stable (Shizuya et al., 1992) (infrequent deletions), rarely are chimeric (Julie Korenberg, personal communication), and can be readily separated from bacterial DNA. To test the advantages of BACs both as mapping and sequencing reagents, we generated a physical BAC map of the human α/δ T cell receptor locus from a library with 3.7-fold coverage of the entire genome.

The human α/δ T cell receptor locus is ideal for testing BACs as mapping and sequencing reagents. First, it is approximately 1 Mb in length and, thus, offers a significant mapping challenge. Second, the locus encodes approximately 50 variable (V) gene segments, 61 Jα gene segments, 3 Dδ gene segments, 4 Jδ gene segments, and the Cδ and Cα genes. The 3' ~100 kb encompassing all of the coding elements from Cδ to Cα including all the Jα gene segments has been sequenced (Koop et al., 1994), thus offering a superb control for many of the experiments described below. Third, while the region encompassing the V elements has not been sequenced, the relative order of most of the V gene segments has been determined as a consequence of the deletional rearrangement process that joins the Vα and Jα or Vδ, Dδ and Jδ elements (Ibberson et al., 1995). Fourth, the V elements fall into distinct

subfamilies that exhibit 75% or more sequence homology. Members of these subfamilies, ranging in size from 2 to 5, are located across the locus and, hence, probes from a few subfamilies can be used to identify BAC clones across the entire V element region. Finally, this region has locus-specific repeats (homology units) that pose challenges typical for much human DNA for both mapping and sequencing.

The BAC inserts appeared to be faithful replicas of genomic DNA. Five BACs, ranging in size from 86 to 208 kb, were successfully sequenced directly by the shotgun approach. Not only do BACs appear to be excellent mapping and sequencing reagents, they also suggest a new approach to sequencing the human genome.

Materials and methods

DNA source. BAC clones were obtained from a human BAC library at California Institute of Technology. This library was developed from a normal human male fibroblast cell line, (ATCC: CRL 1905: CCD-978Sk) (Shizuya et al., 1992). This cell line was also used in PFGE analysis of human genomic DNA

BAC library screening. To obtain specific BACs we used PCR amplified T cell receptor variable gene segments as probes. These were labeled with P-32 using a random labeling approach (T7 QuickPrime, Pharmacia, or Multiprime DNA Labeling System, Amersham) and hybridized overnight at 65°C to the BAC library membranes in SET (0.6 M NaCl, 0.02 M EDTA, 0.2 M Tris-HCl [pH 8.0], 2% SDS, and 0.1% pyrophosphate). The membranes were washed 10 minutes in 1 x SSC + 0.1% SDS, followed by 2-3 washes in 0.1 x SSC + 0.1%

SDS at 65°C for 10-20 minutes each. Positive clones were identified after exposure at -70°C to Kodak X-AR film with intensifying screen overnight or longer. Specific PCR-probes were also made for the ends of different clones, and the PCR product labeled and used as above. Whole cosmids and BACs were also used as probes in hybridization to the BAC library. Cosmid and BAC DNAs were digested with NotI to separate the vector from the inserts, and run on a PFGE (see below). The inserts were cut out of the gel and the DNA extracted from the agarose using beads (Sephaglas BandPrep, Pharmacia or Qiaex, Qiagen). When using P-32 labeled cosmids or BACs as probes, cold vector DNA, human Cot-1 or total placental DNA, and total E. *coli* DNA were used to suppress hybridization of repeat sequences and contaminating vector and E. *coli* DNA.

DNA Preparation. Total human genomic DNA from the same cell line used to make the BAC library was prepared in low melting point (LMP) agarose. Cells were washed twice in phosphate buffered saline (PBS) and resuspended to 10^8 cells/ml in PBS. The cells were then warmed to 37°C before they were mixed with an equal volume of melted 1% LMP-agarose and poured into molds. The solidified plugs were incubated overnight at 50°C in a solution of 0.5 M EDTA [pH 9.0], 1% Sarcosyl, and Proteinase K (0.5 mg/ml). This step was repeated once for one more overnight incubation. The plugs were then rinsed and stored in 0.5 M EDTA.

BAC DNA was prepared using standard alkaline lysis procedures (Sambrook et al., 1989). Minipreparations were made either by hand without organic extractions or by an automated minipreparation machine, Autogen 740, Integrated Separation Systems.

Southern Blots and Hybridizations. BAC DNA was digested with various restriction enzymes according to the suggestions of the manufacturer. The DNA was run in a 0.8% agarose gel, and transferred to nylon membranes by capillary action using 0.4 N NaOH. The membranes were rinsed twice in 2 X SSC before use. PFGE gels were irradiated at 254 nm ultraviolet light for 45 seconds in the presence of ethidium bromide before blotting. The blots were prehybridized in hybridization solution (50% formamide, 5 X SSC, 0.02 M sodium phosphate [pH 6.7], 100 μg/ml denatured salmon sperm DNA, 1% SDS, 0.5% nonfat dry milk, and 10% dextran sulfate) at least half an hour prior to hybridization. As above, probes were labeled and hybridized at 65°C overnight followed by washing, although the washing conditions would vary in their concentration of SSC from 0.1 X SSC to 2 X SSC, dependent on the desired stringency.

Pulsed Field Gel Electrophoresis. Large DNA molecules were separated in 1% agarose in 0.5 X TBE at 14°C using different PFGE apparatuses, either homemade or from Biorad. Voltage applied was 6 V/cm, switch times and total time depended on the sizes separated (Birren and Lai, 1993).

Sequencing. BAC DNA was completely sequenced by the random shotgun method (C. Boysen, in preparation). In short, BAC DNA was sonicated to generate fragments of 1-3 kb in length. The sonicated DNA was repaired with Mung Bean Nuclease (vendor), and run on an agarose gel. Fragments of 1-3 kb were cut out and the DNA purified with beads as above. The purified DNA was subcloned into HincII or SmaI cut, dephosphorylated M13 vector (Sigma). Single stranded DNA was prepared from white plaques, and

sequenced using different versions of Perkin Elmer's dye primer cycle sequencing kits.

End sequencing or primerwalking directly on BACs was performed as described by Boysen et al., 1996 using the ABI PRISM Dye Terminator Cycle Sequencing Ready Reaction Kit with AmpliTaq DNA polymerase, FS, Perkin-Elmer.

Results and Discussion

High quality BAC DNA can readily be purified. High quality DNA was purified from smaller cultures (1.5 ml-5 ml) by hand or by automated procedures, Autogen 740, Integrated Separation Systems, using standard alkaline lysis procedures (Sambrook et al., 1989). DNA from larger volumes (100-250 ml) was also prepared by alkaline lysis and a further purification step performed either by CsCl banding or by passing the DNA over a Sepharose column. The most critical step is the alkaline lysis where care should be taken not to shear E.coli host DNA. This DNA was readily cleaved by restriction enzymes, readily sequenced at either end by primer directed sequencing, and was suitable for random shearing and subcloning into M13 bacteriophage vectors.

The 1 Mb α/δ human T cell receptor locus was readily covered by 17 BACs from a 3.7x BAC library apart from a single gap. Seventeen BAC clones were identified by hybridization to probes for certain V and C gene segments, probes from cosmids previously mapped in this region, or probes obtained from end sequence information from other BACs. Pulsed field gel

electrophoresis suggested that these clones ranged from 85 kb to 240 kb with an average insert size of 137 kb. This provides an average 2.3x coverage. The BACs were obtained from a 3.7x human BAC library with an average insert size of 139 kb (Kim et al., 1996). These BACs were analyzed against V gene segments either by hybridization to restricted BAC DNA or via V gene segment specific STSs (Figure 1). This resulted in three BAC contigs, two of which overlapped with only 2 kb and therefore the overlap was not detected until end sequence information from the BACs at the ends of the contigs was used to make STSs and probes to use against BACs from the other contigs. This leaves the BAC map with a single gap--consistent with what might be expected from a 3.7x library. The middle region of the α/δ family was not exhaustively screened for BACs because we had already obtained a detailed cosmid map after subcloning from YAC clones. We have now covered the last gap with a PAC clone.

The BAC clones exhibit striking genomic fidelity. The fidelity (chimeras, deletions, rearrangements) of the BAC inserts has been checked by six different methods. (*i*) All 17 BAC clones were fingerprinted with three different restriction enzymes (Hind III, EcoRI, and Bam HI or PstI) (Figure 2). The fingerprints of overlapping clones were compared against one another and against an array of cosmid clones spanning 600 kb in two contigs of the α/δ locus. Almost all fragments in overlapping regions could be matched, except for end fragments when cut with EcoRI, Bam HI, and PstI, a possible polymorphic HindIII site in BAC628, and BAC196, which later by limited sequence information was determined to have undergone an internal deletion of 68.2 kb. The single 3' BAC 705 clone has been checked against the 3' sequence for fidelity. BAC clones 129 and 116 show similar patterns at the

5' end. Data indicates that none of the 17 BACs are chimeric and that only BAC196 has a major rearrangement greater than 1 kb. (*ii*) V gene segments were ordered on the BACs via PCR or hybridization to Southern blots made with the restriction enzymes mentioned above (Figure 1a). Their locations, with a few exceptions, corresponded with that previously determined by deletional mapping and pulsed field gel electrophoresis (PFGE) studies using genomic DNA (Ibberson et al., 1995). Furthermore, in all cases where genomic DNA were included on the Southern blot, the hybridization bands in the BAC clones matched their genomic counterparts except in cases were the BAC insert ended close to the probe used. For example, BAC363 ends in a HindIII site found in the middle of a V gene segment, and thus whereas that V gene segment probe gave two bands for HindIII in BAC378, BAC274, and genomic DNA, BAC363 only showed one band corresponding to one of the two. Likewise, this probe differed in its EcoRI pattern from BAC363 since one of the EcoRI sites was missing. (*iii*) Sequences have been determined for both ends of each BAC. This information was used to generate STSs and probes for 15 of the ends, and in all cases gave they positive results when tested on overlapping BACs. (*iv*) Appropriately placed rare cutting restriction sites have been identified across the locus in both genomic and BAC DNAs (Figure 1). These data suggest the BACs faithfully reflect genomic DNA sites. (*v*) The entire α/δ locus is now sequenced using 5 BACs, 2 PACs, and 9 cosmids (Figure 1b). When comparing the overlap regions between sequenced BACs (22 kb), BACs and PACs (120 kb) or between BACs and cosmids (78 kb), we have found no discrepancies that can not be accounted for by polymorphisms. (*vi*) Thirty-two of the 34 end sequences from the 17 BACs matched against the complete sequence of the α/δ locus (Figure 1b), while the last two ends extend outside the sequenced region. The sizes of the BAC inserts

determined this way matched the sizes determined by PFGE except for BAC196, which size was determined by PFGE to be 100 kb, and the insert size determined by aligning the BAC ends with the final sequence suggest the original clone was 163.7 kb. In fact, we went back to the original library, and streaked the BAC196 clone again. The new BAC196 clone were analyzed by Not1 digestion and PFGE analysis, and gave a band at 170 bp, suggesting that this was the original clone. We further digested it with EcoRI and HindIII to compare it with the old BAC196 clone, and clearly, the old clone was a deleted version of the original. This suggests that there are no chimeras, major rearrangements, or deletions greater than 1 kb apart from the deletion in BAC196.

BAC inserts can directly be sequenced at their 5' and 3' ends. We have successfully sequenced all 34 insert ends directly from DNA of the 17 BAC clones. Indeed, in total we have sequenced 34 BAC clones and 22 PAC clones using the T7 and SP6 primers. 110/112 sequences were successful on the first attempt (98% success). By comparing the end sequences with their final high redundancy sequenced counterparts, the average high quality read length was 445 base pairs with an error rate of 0.36%.

Twenty-three of 70 end sequences in the α/δ T cell locus (33%) contained genome-wide repeats (Table 1). Fifteen of these sequences contained 100 bp or more of unique sequence. Only 11% of the end sequences contained no unique sequences.

BAC inserts can be effectively sequenced by the shotgun approach. We successfully sequenced five BACs ranging in size from 86 kb to 208 kb by the shotgun approach. The BACs were randomly sheared, cloned into M13

bacteriophage, and the M13 fragments were sequenced to an average coverage of 6- to 8-fold (Table 2). Closure was generally achieved either by reverse sequence reads from appropriate located M13 clones or by synthesizing DNA primers at the gap edge and using these for walking on appropriate M13 clones or directly on the BAC DNA itself by chain terminator DNA sequencing. No difficulties were experienced in this sequence analysis. The order of the V elements in these BAC inserts is totally consistent with that obtained from the physical map analyses mentioned above.

BAC inserts exhibit several features that facilitate physical mapping. BAC clones are single copy vectors and, accordingly, exhibit relative stability in clonal growth. BAC clones have several features that are attractive for physical mapping. (*i*) They faithfully represent chromosomal sequences. By V gene segment analysis, restriction enzyme analyses, and end sequence analysis, all 17 BAC clones lying across the human α/δ T cell receptor locus, apart from BAC 196, faithfully reflect genomic sequence to the varying levels of discrimination analyzed. (*ii*) BAC inserts appear to be rarely chimeric--indeed, none of the BACs we analyzed were chimeric. In contrast, 6 of 9 YAC clones obtained across this region were clearly chimeric. Moreover, Julie Korenberg has mapped by chromosomal *in situ* hybridization more than 2,000 BAC clones; less than 4% exhibit more than one site of hybridization. Most of these probably represent double clones rather than chimeras. (*iii*) The BAC clones appear to delete or rearrange rarely (only 1/17 clones exhibited a deletion). In contrast, 38% of the 234 cosmid clones analyzed across the human β T cell receptor locus contained defects (deletions, chimeras, rearrangements, failure in end sequencing, etc.) (Lee Rowen, personal communication). This appears high, but the important point is that

most cosmid clones have not been carefully checked before sequencing because of the time consuming nature of careful checks. (*iv*) The average BAC clone is of sufficient length to span most locus-specific clusters of tandem repeats. In the human β T cell receptor locus, we identified one tandem cluster of five 21 kb repeats (105 kb). Having mapping (and sequencing) reagents that span these similar clusters significantly facilitates the mapping process. (*v*) BAC clones seem to be randomly distributed across the 1 Mb TCRα/δ region. One gap of 3 kb was not covered, but this is expected from a library with a 3.7 fold coverage of the human genome. A critical question is whether the even distribution will extend across the entire genome. We would point out that loci with lots of homology units probably present one of the worst case scenarios for genomic cloning. (*vi*) For shotgun sequencing, BACs can be used directly to prepare sequence-ready maps. Thus, one subcloning step (YACs to cosmids) and one mapping step (cosmid physical map) is eliminated. This increased efficiency will greatly facilitate the automation necessary for large-scale sequencing projects. (*vii*) BAC ends can readily be sequenced, thus suggesting a strategy that completely eliminates physical mapping in the large-scale DNA sequencing procedures. (*viii*) An arrayed BAC library allows the easy placement of other landmark features on the BAC clones (e.g. STSs, ESTs, polymorphic satellites, etc.). This will permit the transfer of all of the previously identified landmarks to BACs.

BACs are good sequencing reagents. We have been successful in the shotgun sequence analysis of 5 BACs ranging in size from 86-208 kb. Furthermore, the 208 kb BAC (BAC 129) has significant genome-wide and locus-specific repeats--yet we were able to sequence this large insert without difficulty. The shotgun clone coverage is quite evenly distributed, suggesting that BACs can be

randomly sheared. The assembly of sequence contigs as large as these BACs from randomly sequenced M13 inserts has been made possible by new methods developed for base calling, quality assessment, and assembly by Phil Green (personal communication). Apart from the fact that BACs can be sequenced by shotgun analysis, BAC clones do have several advantages for sequencing, in part similar to those mentioned above for mapping. (*i*) BAC clones appear to faithfully represent the genome. (*ii*) BAC clones rarely delete, rearrange, or are chimeric. (*iii*) The average BAC clone can readily traverse the largest locus-specific repeats identified to date. (*iv*) The BAC vector is only 7.5 kb in length, thus representing a significantly lower percentage of the insert than found in the clones currently used most frequently for shotgun sequencing--cosmids (4-8/30-40 kb).

In summary, BACs are attractive sequencing reagents because of their genomic fidelity, size, stability, and potential for eliminating one cloning and one mapping step in traditional shotgun sequencing. BAC inserts also offer the possibility of sequencing the human genome without the need for any physical mapping.

The human genome may be sequenced by the sequence tagged connector (STC) approach. This approach is outlined schematically in Figure 3. (*i*) A 15-fold BAC library of randomly cloned human DNA will be prepared and arrayed in 384 well microtiter plates. This would require 300,000 BAC clones with 150 kb average insert sizes. (*ii*) BAC DNAs will be prepared for end sequence and single restriction enzyme analysis of each BAC insert. The end sequences, averaging 500 base pairs, will be randomly spaced every 5 kb and will represent ~10% of the genome sequence. The fingerprints will be useful

for determining BAC insert fidelity with respect to the genome. (*iii*) STS, polymorphic microsatellite, or EST landmarks can readily be placed on the BAC clones to position them with regard to already defined markers. (*iv*) DNA sequencing can begin with one seed BAC clone. After a 150 kb BAC insert has been sequenced, it will be connected to other 30 BAC clones through the sequence tagged connectors (end sequences) of BAC clones overlapping this region. At this point, the fingerprints of the overlapping clones can be compared to detect artifacts (chimeras, deletions, rearrangements, etc.). Clones with minimal overlaps with the 5' and 3' ends of the seed BAC can be selected to efficiently extend the sequence analyses in either direction. After each successive minimally overlapping BAC clone is sequenced, the next minimally overlapping clone can be selected for sequencing. Obviously, seed clones can be simultaneously sequenced in larger sequencing centers, permitting efficient sequence analyses, for example, from landmarks scattered across particular chromosomes.

This proposal has several striking advantages. (*i*) BACs are excellent mapping and sequencing reagents. This approach eliminates the need for physical mapping and the need to construct the intermediate cosmid library. Accordingly, only two procedures, DNA purification and sequence reactions, need to be automated for large-scale DNA sequencing. (*ii*) The existing EST and STS landmarks can easily be identified on the arrayed BAC clones. Thus, this positional information can readily be transferred to the STC approach. (*iii*) The random distribution of BAC ends throughout the genome can be facilitated by creating a library using two (or three) different restriction enzymes. (*iv*) The STC approach is ideal for sequencing interesting multigene families and for the identification of genes obtained by positional

cloning. The sequenced seed BAC readily gives access to other 5' and 3' BACs with minimal clone overlaps through the STCs. Thus, DNA can rapidly be obtained (and sequenced) across interesting regions. (*v*) Big and small laboratories alike can readily benefit from the STC strategy. Big laboratories can start simultaneously with many seed BACs at different locations. Small laboratories do not require a large mapping infrastructure to sequence interesting regions. (*vi*) Since the STCs will constitute 10% of the genome analyzed by single pass sequencing, many interesting features can be identified just by the analysis of these sequences. The STCs will also match many previously identified chromosomal landmarks. (*vii*) The STC approach can support an international effort to sequence the human genome from arrayed and accessible BAC clones.

The STC approach has raised several concerns. (*i*) A good high resolution STS framework map already exists. This map can be used with a deep BAC library to create a deep BAC physical map. Thus, the STC approach is unnecessary. In fact, the STS map is not at a sufficiently high resolution to give a sequence-ready BAC map. To get to this level of resolution will require considerable additional work. Hence, we would argue that the STC approach in the long run with the advantages cited above will be more cost effective. (*ii*) One should not depend on a single clone library. New, better libraries could come along. It may not represent a random distribution of fragments. As noted above, the library could be prepared using two (or more) enzymes. Moreover, the end sequence analysis could easily be done within two years (by 20 377 ABI sequencers). These data would be immediately useful, even at 3-5-fold coverage. Hence, the idea that one would have to wait two years to use the information is wrong. It could be useful within the first 3-4 months.

New libraries could be integrated into the old arrays. (*iii*) How could one be assured that the seed BAC does not have an artificial insert (chimera, deletion, or insertion). After a 1-2x sequence coverage of the BAC, STC could identify many of the overlapping BAC. The fingerprints of these could be compared to look for artifacts. On balance, the STC approach appears to have considerable promise.

Acknowledgment

We would like to thank Hiroaki Shizuya for useful hints in the handling of BAC clones and Tawny Biddulph for typing this manuscript. This work was supported by a grant from the Department of Energy (DOE).

References

Birren, B. and Lai, E. 1993. Pulsed field gel electrophoresis: A practical guide. Academic Press, Inc.

Boysen, C., Simon, M. I., and Hood, L. 1996. Fluorescent sequencing directly from bacterial or P1-derived artificial chromosomes. (Submitted).

Chumakov, I. M., Rigault, P., Le-Gall, I., Bellann'e-Chantelot, C. Billault, A., Guillou, S., Soularue, P., Guasconi, G., Poullier, E., Gros-I., et al. 1995. A YAC contig map of the human genome. Nature 377 (6547 Suppl), 175-297.

Doggett, N. A., Goodwin, L. A., Tesmer, J. G., Meincke, L. J., Bruce, D. C., Clark, L. M., Altherr, M. R., Ford, A. A., Chi, H. C., Marrone, B. L., et al. 1995. An integrated physical map of human chromosome 16. Nature 377 (6547 Suppl), 335-65.

Green, E. D., Riethman, H. C., Dutchik, J. E., Olson, M. V. 1991. Detection and characterization of chimeric yeast artificial-chromosome clones. Genomics 11, 658-69.

Ibberson, M. R., Copier, J. P., and So, A. K. 1995. Genomic organization of the human T-cell receptor variable α (TCRAV) cluster. Genomics 28, 131-139.

Kim, U.-J., Birren, B., Sheng, Y.-L., Slepak, T., Mancino, V., Boysen, C., Kang, H.-L., Simon, M. I., and Shizuya, H. 1996. Construction and characterization of a human Bacterial Artificial Chromosome library. Genomics (in press).

Koop, B. F., Rowen, L., Wang, K., Kuo, C. L., Seto, D., Lenstra, J. A., Howard, S., Shan, W., Deshpande, P., and Hood, L. 1994. The human T-cell receptor TCRAC/TCRDC (Cα/Cδ) region: Organization, sequence, and evolution of 97.6 kb of DNA. Genomics 19, 478-493.

Rowen, L., Koop, B. F., and Hood, L. 1996. The complete 685 kb sequence of the human beta T cell receptor locus. Science (*in press*).

Sambrook, J., Fristsch, E. F., and Maniatis, T. 1989. Molecular cloning: A laboratory manual. (Second Edition). Editors: N. Ford, C. Nolan, and M. Ferguson. Cold Spring Harbor Laboratory Press. Chapter 1, 1.25-1.28.

Shizuya, H., Birren, B., Kim, U.-J., Mancino, V., Slepak, T., Tachiiri, Y, and Simon, M. 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proc Natl Acad Sci USA 89, 8794-8797.

Table 1. Genome-Wide Repeats in the 23 (33%) End Sequences from 70 BAC Ends

Type	#	Similarity (%)	Some Unique	Repeat Only
Alu	10	86-93	10/10 (>100 bp)	0
LINE	9	77-92	2/9 (>300 bp)	7
LTR	4	84-93	3/4 (>250 bp)	1

Table 2. Shotgun sequencing of BACs

BAC-clone	129	956	480	378	810
Size	208 kb	113 kb	86 kb	119 kb	143 kb
Redundancy	7.2	8.4	6.0	10.8	7.0
No. of contigs before walking	17	8	10	2	5

Figure legends

Figure 1. Physical map of BACs across the human TCR α/δ region. a) Based on fingerprinting patterns of BACs and on localization of TCR gene segments, as determined either by hybridizations to Southern blots of restricted BACs (●), or by PCR assays (■). b) Sequences from the ends of BAC inserts compared to the final assembled high redundancy shotgun sequence obtained across the locus from BACs, PACs and cosmids as indicated.

Figure 2. Restriction digests of BAC DNA. BAC DNA were cut with HindIII and run on a 0.8% agarose gel. HindIII cuts out the vector resulting in a common vector band, which for the BACs numbered below 450 is 6.8 kb and for BACs with numbers over 450 it is 7.5 kb.

Figure 3. Sequence Tagged Connector strategy. A 15x genomic equivalent BAC library is constructed, roughly 300.000 clones with an average insert size of 150 kb. A fingerprint is obtained and both insert ends are sequenced for all of these clones. This information is stored for future use. Several seed BACs are sequenced completely. This sequence is then compared to the end sequences in the database, and all overlapping BACs are then compared to each other via their fingerprints to determine eventual defects in the clones. BACs with minimal overlaps in each direction are then picked for complete sequencing to extend the sequence from the seed BAC, etc.

Figure 1.

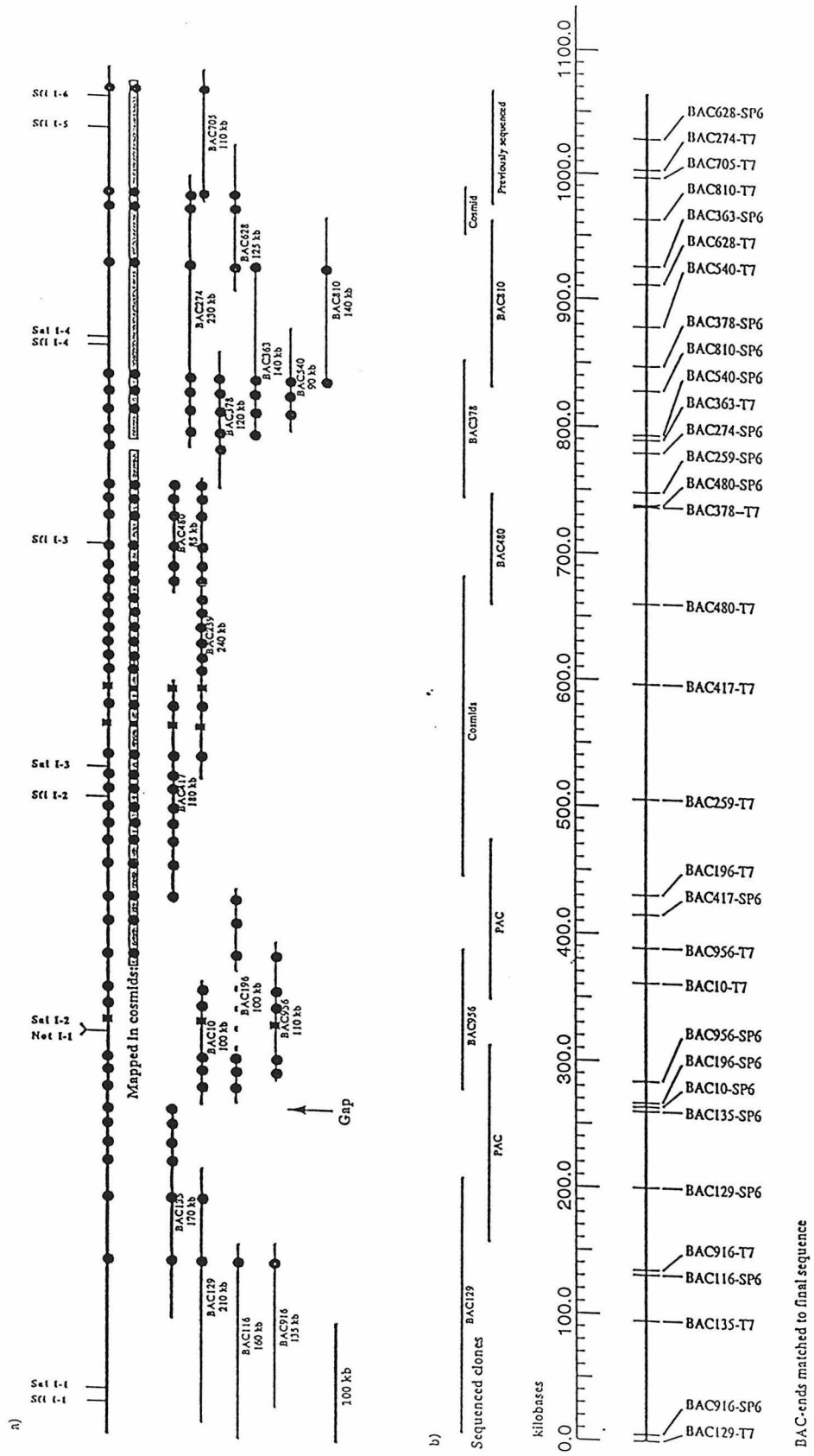

a)
Sal I-6
Sal I-5
Sal I-4
Sfi I-4
Sfi I-3
Sal I-3
Sfi I-2
Sal I-2
Not I-1
Sal I-1
Sfi I-1
Mapped in cosmids:
BAC705 110 kb
BAC628 125 kb
BAC274 230 kb
BAC810 140 kb
BAC540 90 kb
BAC163 140 kb
BAC378 120 kb
BAC480 85 kb
BAC259 240 kb
BAC417 180 kb
BAC10 100 kb
BAC196 100 kb
BAC956 110 kb
Gap
BAC135 170 kb
BAC129 210 kb
BAC116 160 kb
BAC916 135 kb
100 kb
b)
Sequenced clones
BAC129
BAC956
PAC
PAC
Cosmids
BAC480
BAC378
BAC810
Cosmid
Previously sequenced
Kilobases
0.0
100.0
200.0
300.0
400.0
500.0
600.0
700.0
800.0
900.0
1000.0
1100.0
BAC628-SP6
BAC274-T7
BAC705-T7
BAC810-T7
BAC363-SP6
BAC628-T7
BAC540-T7
BAC378-SP6
BAC810-SP6
BAC540-SP6
BAC363-T7
BAC274-SP6
BAC259-SP6
BAC480-SP6
BAC378-T7
BAC480-T7
BAC417-T7
BAC259-T7
BAC196-T7
BAC417-SP6
BAC956-T7
BAC10-T7
BAC956-SP6
BAC196-SP6
BAC10-SP6
BAC135-SP6
BAC129-SP6
BAC916-T7
BAC116-SP6
BAC135-T7
BAC916-SP6
BAC129-T7
BAC-ends matched to final sequence

Figure 2.

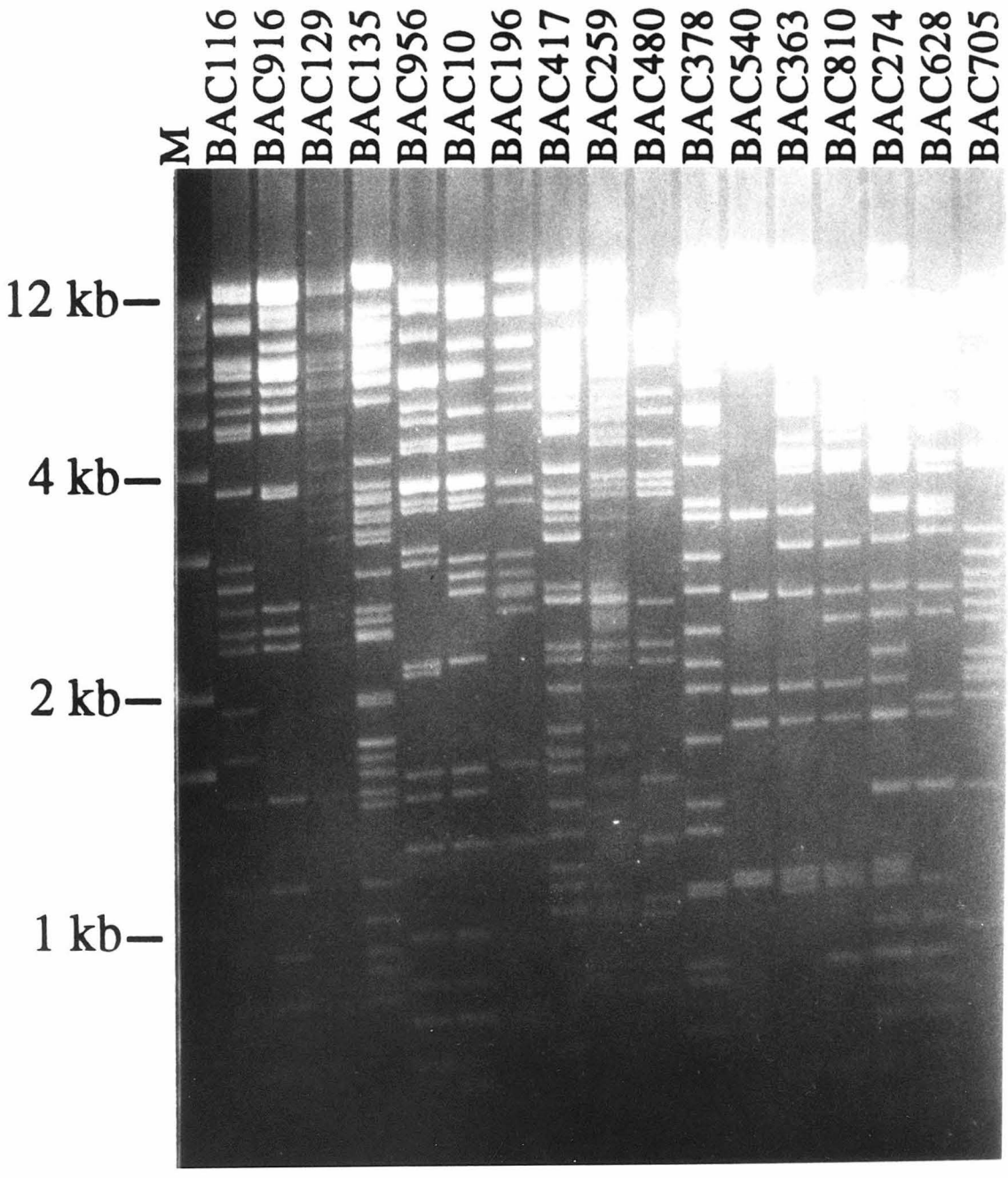
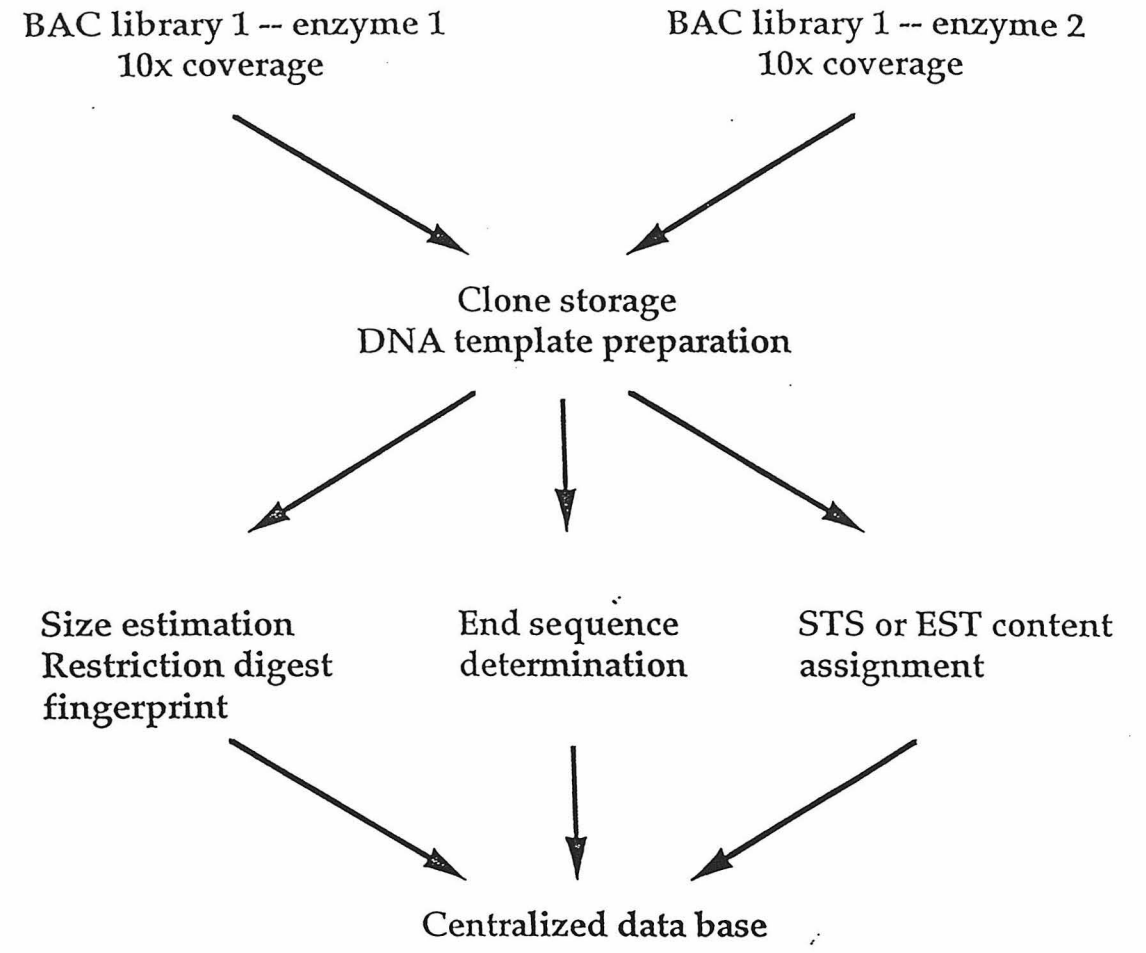
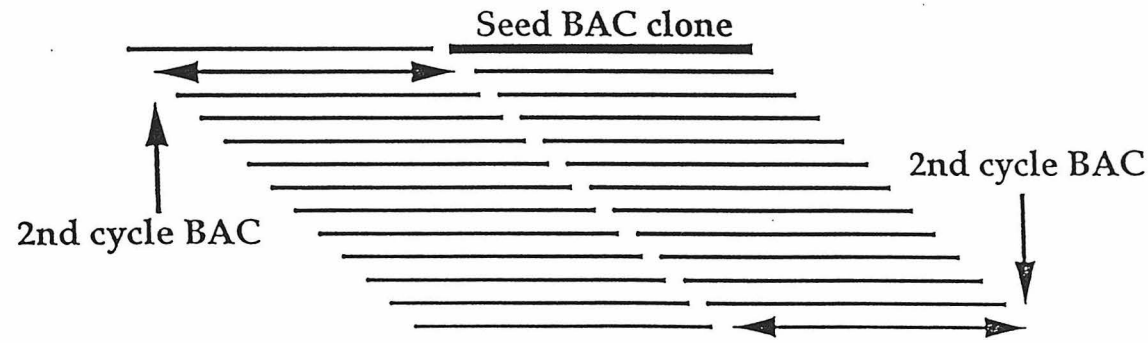

M
BAC116
BAC916
BAC129
BAC135
BAC956
BAC10
BAC196
BAC417
BAC259
BAC480
BAC378
BAC540
BAC363
BAC810
BAC274
BAC628
BAC705
12 kb
4 kb
2 kb
1 kb

Figure 3.


BAC library 1 -- enzyme 1
10x coverage
BAC library 1 -- enzyme 2
10x coverage
Clone storage
DNA template preparation
Size estimation
Restriction digest
fingerprint
End sequence
determination
STS or EST content
assignment
Centralized data base


Seed BAC clone
2nd cycle BAC
2nd cycle BAC

Fluorescent sequencing directly from bacterial and P1-derived artificial chromosomes

Cecilie Boysen†, Melvin I. Simon†, Leroy Hood‡.

† Division of Biology 147-75, California Institute of Technology, Pasadena, CA 91125

‡ Department of Molecular Biotechnology 357730, University of Washington, Seattle, WA 98195

Correspondence to Leroy Hood

ABSTRACT

Bacterial and P1-derived artificial chromosomes, BACs and PACs, are being used increasingly in the human genome project as mapping and sequencing tools. Using fluorescent terminator cycle sequencing, we have developed a method to obtain the sequences directly from the ends of these clones. Of 112 end sequences analyzed to date, more than 98 % have been successfully sequenced. The average read length employing the standard T7 and SP6 primers is 495 bp with an error rate less than 0.36 %. This technique can also be used with custom-made primers. This approach is useful in the initial characterization of BAC and PAC clones, and in closing gaps or obtaining more sequence in interesting, partially sequenced regions.

INTRODUCTION

The human genome project to date has extensively relied on yeast artificial chromosome (YAC) and cosmid clones for mapping and sequencing (Chumakov et al., 1995; Doggett et al., 1995; Levy, 1994; and Rowen et al., 1996). These systems have limitations. Many YAC clones are chimeric (Green et al., 1991) and both YAC and cosmid clones have a tendency to delete or rearrange (Lee Rowen, personal communication). BAC and PAC inserts, ranging in size from 50-300 kb, appear relatively stable and infrequently chimeric (Shizuya et al., 1992; Ioannou et al., 1994) due, in part, to the fact that they are single copy plasmid vectors. BAC and PAC inserts are being increasingly used in physical mapping projects (Ashworth et al., 1995; Boysen et al., 1996). BAC and PAC inserts can be sequenced directly by the random shotgun method (C. Boysen, in preparation). This offers a significant time savings over cosmid sequencing approaches, where YAC clones are subcloned into cosmids, or used to bin cosmids from genome-wide or chromosome-specific cosmid libraries.

In large sequencing projects a minimal overlap between clones is critical to avoid extra redundancy in sequencing. Most mapping projects determine overlaps between clones by STS content mapping or by restriction enzyme digests. These approaches often require significant overlaps to be statistically significant. Obtaining sequence information from the ends of the large insert clones, and in turn using this information to detect smaller overlaps between clones would identify minimally overlapping clones and thus make large scale sequencing more efficient. End-sequences have been obtained from YACs by either the vectorette system (Riley et al., 1990) or the Alu-vector technique (Nelson et al., 1991). Additional techniques have been developed to obtain the ends from P-1 clones (Liu and Whittier, 1995). However, all of these techniques require several steps before the actual sequence is obtained or have a relatively high failure rate. To facilitate the process of obtaining end-sequence analysis from BACs and PACs, we have modified the dye terminator cycle sequencing protocol from Perkin-Elmer. This process only involves two steps: DNA preparation of the clone and sequencing.

MATERIALS AND METHODS

DNA Sources. BAC clones were obtained from a Caltech human genomic BAC library (Shizuya et al., 1992) made using DNA from a normal male fibroblast cell line, ATCC: CRL 1905: CCD-978Sk. PAC clones were from the human genomic PAC library (normal male fibroblast cell line, HSF7) at Genome Systems, Inc. (Ioannou et al., 1994).

DNA Preparation. Each BAC/PAC clone was streaked to obtain single colonies. One single colony was used to inoculate 20 ml Luria Broth containing the appropriate antibiotic and grown at 37°C for 18-22 hours in 50 ml plastic tubes. In some cases, larger cultures of 200 ml were grown. The DNA was prepared by alkaline lysis. In most cases we used an automated miniprep machine, the Autogen 740, Integrated Separation Systems,

following the manufacturers BAC protocol. Otherwise, the DNA was prepared by hand with no organic extractions (Sambrook et al., 1989). In some cases this procedure was followed by a PEG precipitation. DNA was resuspended in double distilled (dd) H2O (approximately 150 µl for a 20 ml prep).

Sequencing. Twenty-two µl DNA (equivalent to DNA from a 3 ml culture, 1-2 µg) were used for each fluorescent terminator sequencing reaction which further contained 50 pmol primer and 16 µl terminator ready reaction mix (ABI PRISM Dye Terminator Cycle Sequencing Ready Reaction Kit with AmpliTaq DNA polymerase, FS, Perkin-Elmer). Oligo's were synthesized, deprotected, dried, and resuspended in ddH2O to either 25 µM or 50 µM. For the end sequencing standard T7 and SP6 primers were used (T7: TAATACGACTCACTATAGGG and SP6: ATTTAGGTGACACTATAG). Cycling was performed in the thermal cycler, GeneAmp 9600, Perkin-Elmer, following the instructions in Perkin-Elmer's protocol P/N 402078, with the addition of an initial denaturation step. The thermal cycler was heated to 96°C before inserting the tubes. These were kept at 96°C for four minutes, followed by 25 cycles of ten seconds at 96°C, five seconds at 50°C, and four minutes at 60°C. After cycling, the reaction was either precipitated directly with ethanol and salt or purified by passing over a CentriSep spin column (Princeton Separations), dried, and run on a 373 DNA Sequencer, Stretch (Applied Biosystems) using either 36 or 48 cm well to read plates (4.75 % or 4 % acrylamide gel, respectively).

RESULTS

Sequencing protocols

In order to find the optimum conditions for sequencing directly from BAC and PAC DNA, we tested different DNA purification methods as well as different sequencing conditions. DNA purified by the minipreparation robot, Autogen 740, sequenced more

consistently than DNA prepared manually. Our success rate for the DNA prepared by the robot was 98 % (see later); our manual preparations succeeded 80 % of the time. The DNA prepared by the robot did in general also provide somewhat longer and cleaner reads. In order to improve the manually purified DNA preparations, we added a PEG precipitation step. This did not noticeably change the quality of the sequence, but needs to be explored further.

To optimize the sequencing procedure we used DNA prepared by the Autogen 740. We varied the concentration of DNA, the volume of the reaction, the amount of primer, and the length of the denaturation step. Furthermore, we tested different primers, four at each vector end, but found that the standard T7 and SP6 primers (sequences are given under Materials and Methods) gave the best results. Therefore, these were used in the experiments described below. The first 14 T7 primer reactions listed in Table 1 provide a summary of some sequences obtained during the initial experiments. The average accurate read length (332 bp) of these is lower, than when the final optimized conditions were used. The optimal conditions are described in Materials and Methods. In general, we double the volume of a standard cycle sequencing terminator reaction from 20 to 40 µl, containing 16 µl terminator sequencing premix, 50 pmoles primer, and DNA prepared from three mls of culture. After cycling, reactions are purified by passing over a spin column. This step could be omitted and an ethanol precipitation performed, but the leftover terminators obscure the first 40-60 bp of the sequence read, and an artifact of 12-14 T-nucleotides is seen around 250-300 bp, although generally the sequence is identifiable. These reaction conditions give a weak but clean signal (Figure 1). Using a smaller reaction volume, whether the same or half the amount of DNA or primers was used, decreased the success rate to about 75 %. One of the reasons that the same amount of DNA did not work in a smaller reaction volume is probably because of impurities in the DNA. In the larger volume, the impurities are diluted out. For half the amount of DNA, the signal was weaker and the signal to noise ratios compromised. Preliminary experiments suggest that half the

amount of DNA, primer and premix, gives good reads on the 377 DNA sequencer, whereas a quarter reaction provides noisy data on the 377 DNA sequencer. We tested different concentrations of primers (3.2, 10, 25, 50, and 100 pmoles total in either 20 or 40 µl reactions), and found 50 pmoles to be optimal for the 40 µl reaction. Finally, we found that instead of starting the cycling directly as called for in the sequencing procedure, better results are obtained, when the reaction tubes are inserted into the already hot, 96°C, thermal cycler, and denatured for four minutes at this temperature, before the normal cycling begins.

End Sequencing

Using these optimized conditions, we sequenced 34 BACs and 22 PACs from both ends. These sequences were obtained in several different batches and no differences were seen in quality between batches when the same protocol was followed. BAC and PAC clones gave similar results. The 48 cm well to read plates on the 373 DNA sequencer gave significantly better reads out beyond 500 bp than the 36 cm well to read plates (Table 1). Due to plate availability, most of the sequences described here have been run using the 36 cm well to read plates. Of the 112 end sequences, 110 (98 %) were successfully sequenced, that is, they gave more than 250 bp of reliable sequence as judged by evaluating the chromatograms. To estimate the length and error rate, 45 end sequences falling within previously fully sequenced regions, were compared to their high redundancy shotgun analyzed counterparts (Table 1). The first fourteen of these sequenced with the T7 primer were from the initial optimization process and therefore have a lower read length of 332 bp (0.47 % error). After optimization ten and twenty-one more reactions (T7 and SP6 primer, respectively) were compared to the equivalent sequence obtained by the highly redundant shotgun method, and an average read length of 470 bp high quality sequence were obtained for the T7 primer (0.20 % error), whereas 514 bp (0.42 % error) were the average for the SP6 primer. The worst case, gave 262 bp of reliable sequence (2 errors), whereas many

sequences extended out to more than 700 bp, although the error rate does go up towards the end of the read (Figure 1). Most of these errors were due to broader peaks at the end of the read, and better results were obtained when the reactions were run using the long 48 cm well to read plates (Table 1). The majority of the errors internal to the read were due to drop-outs of G-peaks following A-residues. This is a common observation for the terminator mix used here, and is not specific to the BAC or PAC sequencing reactions.

Primer walking sequence

Applying these same conditions for sequencing directly of BACs or PACs, we tested primer walking with custom made primers. These primers were picked from ends of sequence contigs obtained from random shotgun sequencing projects in an effort to close the gaps. Forty-three of these were used on three different BACs and one PAC. Thirty-six (83 %) gave good long reads as above, whereas the other seven either gave noisy data or no results at all. In two cases the primers were in Alu repeats, but there was no obvious reason why the other five failed. The average high quality read length was 525 bp with an error rate of 0.60 % (Table 1). Sixteen of the reactions were run using the 48 cm well to read plates, and the read length for these were 568 bp, whereas the other 20 using the 36 cm well to read plates had an average read length of 492 bp (Table 1).

DISCUSSION

We have developed a method to sequence insert ends directly from BAC or PAC DNA without the intermediate PCR step, which is used in most other end sequencing protocols (Riley et al., 1990, Nelson et al., 1991, and, Liu and Whittier, 1995). Our direct approach only involves two steps, DNA preparation and sequencing, and thus is automatable and generally faster than the PCR based methods. Furthermore, we do not have to rely on specific sequences being present in the DNA close to the ends, such as a

specific restriction site or an Alu sequence. An additional disadvantage to an intermediate PCR step is that if the sequence contains simple repeat tracks, the thermostable enzyme tends to skip bases during the PCR amplification and, therefore, confuses subsequent sequence analyses. Direct sequences from the clones generally permits accurate sequence reads of the repeats. One disadvantage of the direct sequencing method is the requirement of several fold more DNA than is necessary for PCR. Preliminary experiments using the more sensitive 377 DNA sequencer appear to require half as much, or less, DNA.

Useful sequences were obtained from 110 of 112 clone ends sequenced with either the SP6 or T7 primers. After optimization of the reaction conditions, an average read length of about 500 bp with 0.36 % error was obtained. This read length allows one to obtain unique sequence on ends that have Alu repeats (~300 base pairs). Accordingly, unique STSs can generally be generated from the end sequences.

Obtaining sequence at the ends of clone inserts is crucial in many mapping projects to determine minimal overlap between clones that would go undetected by most restriction fragment analyses. As an example, we mapped a 1.1 megabase region using BAC clones by STS content mapping and restriction fragment analysis (Boysen et al., 1996). The initial analysis generated three BAC contigs. In order to determine whether these contigs overlapped, we sequenced the ends of the inserts from the BAC clones at the end of each contig. An STS primer pair was made from each sequenced end, and used in PCR assays of the potential overlapping clones. Alternatively, the PCR products were used in Southern blot analyses of the restricted BACs. Using this approach, two of the contigs overlapped. Later, complete sequencing of the two overlapping BAC clones showed the overlap region to be 2.2 kb in length.

Obtaining sequence from the clone ends at the contig boundaries is also useful in extending maps. In the BAC mapping project described above, the last gap had no known sequences, and thus we had to use clones around the gap to close it. We produced the end

sequences from the two BACs extending into the gap and used them to construct PCR assays and probes to screen a PAC library. Several PAC clones closed the gap.

End sequence analysis can also facilitate the choice of minimal sequence tiling paths. For example, once an initial BAC has been sequenced, the end sequences of all of the overlapping BACs will permit a minimum sequence overlap to be chosen. In the project mentioned above, we obtained the complete sequence of two BAC clones on either side of the gap. By end sequencing all the overlapping PAC clones and comparing the end sequences with the final complete BAC sequences, we could easily choose the smallest PAC clone that would close the gap without resorting to any mapping.

BAC or PAC clones are also excellent substrates for primer directed walking. We have sequenced five BACs and two PACs directly by the shotgun method (C. Boysen, manuscript in preparation). In most cases, after the initial round of M13 insert sequencing, two to ten gaps remained. Most of these gaps were closed by choosing primers about 100 base pairs from the end of the sequence on either side of the gap, and using the primers to walk directly on the large insert clone. One must avoid choosing primers in genome-wide repeats. This is easily avoided by computational analysis of the contig sequences against a complete data file of human repeat sequences (A. Smit, personal communication).

Thus, BAC and PAC clones are excellent substrates for direct end sequencing and primer walking procedures. These simple procedures should greatly facilitate ongoing mapping and sequencing efforts.

ACKNOWLEDGMENT

We would like to thank Harold Garner and David Burbee for useful suggestions, and the sequencing facilities at the California Institute of Technology and the Department of Molecular Biotechnology at University of Washington. This work was supported by grants from the National Science Foundation (NSF) and the Department of Energy (DOE).

REFERENCES

Ashworth, L. K., Alegria-Hartman. M., Burgin, M., Devlin, L., Carrano, A. V., Batzer, M. A. 1995. Assembly of high-resolution bacterial artificial chromosome, P1-derived artificial chromosome, and cosmid contigs. Anal-Biochem 224, 564-71.

Boysen, C., Simon, M. I., and Hood, L. 1996. The use of bacterial artificial chromosomes (BACs) as mapping and sequencing reagents. Submitted.

Chumakov, I. M., Rigault, P., Le-Gall, I., Bellann'e-Chantelot, C. Billault, A., Guillou, S., Soularue, P., Guasconi, G., Poullier, E., Gros-I., et al. 1995. A YAC contig map of the human genome. Nature 377 (6547 Suppl), 175-297.

Doggett, N. A., Goodwin, L. A., Tesmer, J. G., Meincke, L. J., Bruce, D. C., Clark, L. M., Altherr, M. R., Ford, A. A., Chi, H. C., Marrone, B. L., et al. 1995. An integrated physical map of human chromosome 16. Nature 377 (6547 Suppl), 335-65.

Green, E. D., Riethman, H. C., Dutchik, J. E., Olson, M. V. 1991. Detection and characterization of chimeric yeast artificial-chromosome clones. Genomics 11, 658-69.

Ioannou, P. A., Amemiya, C. T., Garnes, J., Kroisel, P. M., Shizuya, H., Chen, C., Batzer, M. A., and de Jong, P. J. 1994. A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. Nature Genetics 6, 84-89.

Levy-J. 1994. Sequencing the yeast genome: an international achievement. Yeast 10, 1689-706.

Liu, Y. G. and Whittier, R. F. 1995. Thermal asymmetric interlaced PCR: automatable amplification and sequencing of insert end fragments from P1 and YAC clones for chromosome walking. Genomics 25, 674-81.

Nelson D. L., Ballabio, A., Victoria, M. F., Pieretti, M., Bies, R. D., Gibbs, R. A., Maley, J. A., Chinault, A. C., Webster, T. D., and Caskey, C. T. 1991. Alu-primed polymerase chain reaction for regional assignment of 110 yeast artificial chromosome clones from the human X chromosome: Identification of clones associated with a disease locus. Proc Natl Acad Sci USA 88, 6157-6161.

Riley, J., Butler, R., Ogilvie, D., Finniear, R., Jenner, D., Powell, S., Anand, R., Smith, J. C., and Markham, A. F. 1990. A novel, rapid method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones. Nucleic Acids Res 18, 2887-2890.

Rowen, L., Koop, B. F., and Hood, L. 1996. The complete 685 kb sequence of the human beta T cell receptor locus. Science (*in press*).

Sambrook, J., Fristsch, E. F., and Maniatis, T. 1989. Molecular cloning: A laboratory manual. (Second Edition). Editors: N. Ford, C. Nolan, and M. Ferguson. Cold Spring Harbor Laboratory Press. Chapter 1, 1.25-1.28.

Shizuya, H., Birren, B., Kim, U.-J., Mancino, V., Slepak, T., Tachiiri, Y, and Simon, M. 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proc Natl Acad Sci USA 89, 8794-8797.

Table 1. Average high quality read length and error rate. Of the 112 BAC and PAC insert ends sequenced directly with SP6 or T7 primers, forty-five of the 110 working sequences were compared to their counterparts in sequence determined by the highly redundant shotgun approach. Thirty-six of 43 custom primers successful yielded sequence. These were likewise compared to the final sequence.

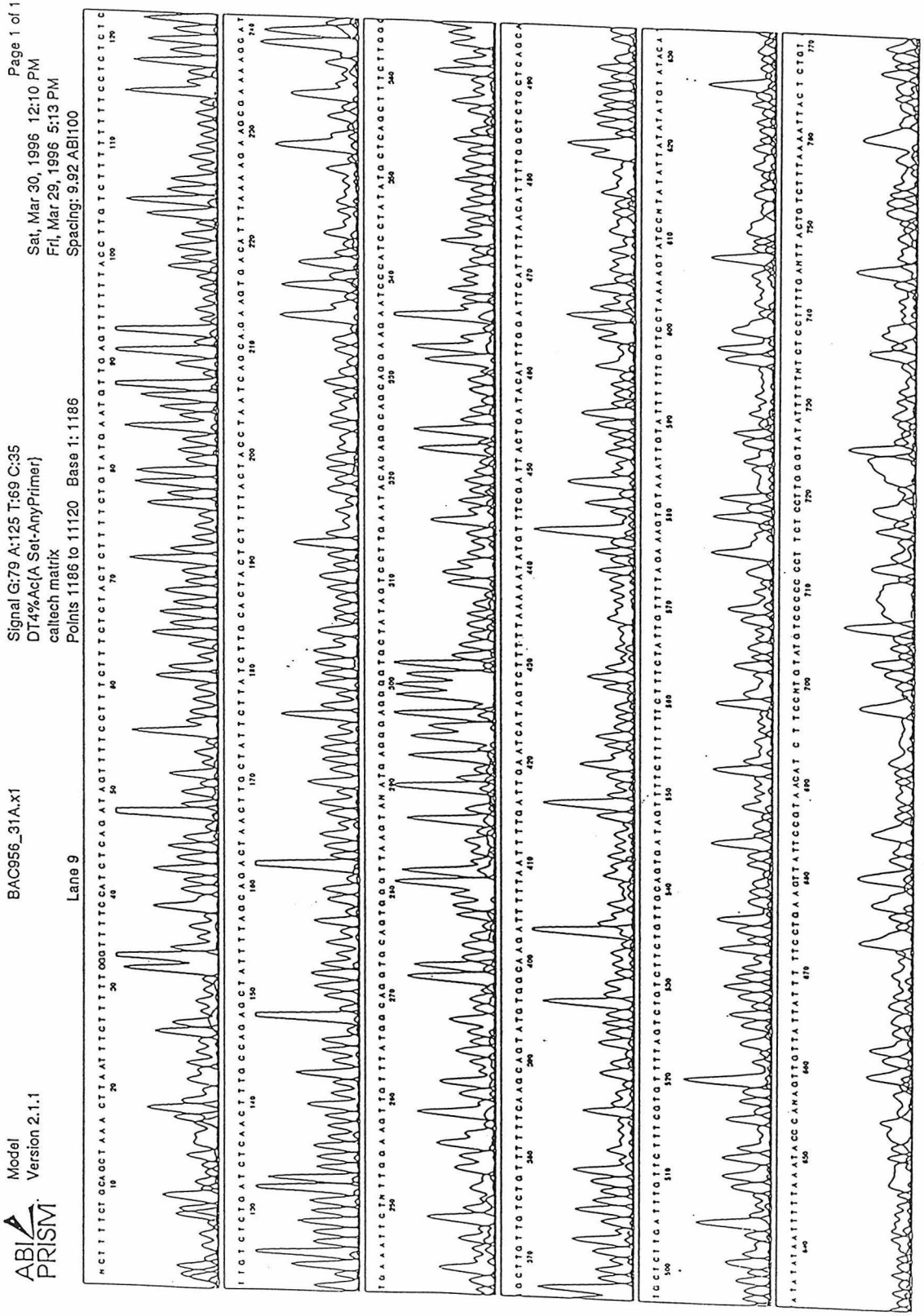
	# of reads	**Average # of high quality bases**	**Error rate**
T7 primer[a]	14	332 bp	0.47 %
T7 primer	10	470 bp	0.20 %
SP6 primer	21	514 bp	0.42 %
Various primers	36	525 bp	0.60 %
Various primers:			
48 cm well to read	16	568 bp	0.90 %
36 cm well to read	20	492 bp	0.34 %

a) Preliminary experiments

FIGURE LEGEND

Figure 1. DNA sequence trace obtained by primer walking with a custom synthesized oligonucleotide directly on total DNA from BAC956. The dye-terminator sequencing reaction was run on a 373 DNA sequencer for 18 hours using the long 48 cm well to read plate with 4 % acrylamide in the gel.

Figure 1.


ABI PRISM
Model
Version 2.1.1
BAC956_31A.x1
Lane 9
Signal G:79 A:125 T:69 C:35
DT4%Ac(A Set-AnyPrimer)
caltech matrix
Points 1186 to 11120 Base 1: 1186
Sat, Mar 30, 1996 12:10 PM
Fri, Mar 29, 1996 5:13 PM
Spacing: 9.92 ABI100
Page 1 of 1

Summary

The work described in this thesis has focused on two goals. First, the development of efficient strategies to use in large scale DNA mapping and sequencing. Second, using these tools to determine the entire DNA sequence of the human TCR α/δ locus and analyze this 1.07 Mb region, including the organization, structure, evolution and polymorphisms of the TCR elements, as well as an analysis of non TCR genes, genome wide repeats, and other chromosomal features.

New approaches to mapping and sequencing

When I started this project five years ago, the major sources of cloned genomic DNA were YAC, cosmid, and phage libraries. YAC and to some extent cosmid clones were used in most mapping projects, despite the fact, that these clone libraries contain a high percentage of chimeras, and that the clones easily rearrange or delete DNA (Green et al., 1991). I originally began my mapping efforts using YAC clones, but quickly came to the conclusion that more stable, non-chimeric clones were needed to generate representative physical maps. For this reason I began working with the newly developed BAC clone system. Preliminary results suggested that the BAC clones were quite stable (Shizuya et al., 1992). I wanted to test the usefulness of BAC clones as a mapping resource for the human TCR α/δ locus. I had already mapped half of the locus in detail using YAC and cosmid clones, and the 3' one hundred kb of the region had been sequenced by Koop et al., 1994. These two results were important for evaluating the fidelity of the BAC clones. I obtained a total of 17 BAC clones across the α/δ TCR region. These were characterized by restriction digest patterns, ability to hybridize to densely located probes across the region, and their STS content (Chapter 2, 5). Later they were further characterized by obtaining sequence information from the ends of the clone inserts (Chapter 6) , and comparing these end sequence data with the final sequence spanning the entire region. By doing this one could

determine the insert size, and compare it to the sizes obtained by PFGE. Using all of these characteristics plus comparison to genomic DNA, YAC and cosmid clones, sixteen of the seventeen BAC clones were found to faithfully represent genomic DNA. The last BAC clone had undergone a deletion of 68 kb in its center. This deletion occurred in one of the first growth cycles after being picked from the library, since when I went back to the original library and picked the same clone number, I got the intact BAC clone with the correct size and restriction digest pattern. The fidelity of the majority of BAC clones is further supported by *in situ* hybridization studies showing less than 4 % chimera in more than 2000 BACs (Julie Korenberg, personal communication). The majority of the 4% of clones resulting in hybridization to two or more chromosomal locations probably arises from one of two reasons. First, it happens that during the library construction two colonies are transferred to the same well by mistake. Since the BAC clones tested by *in situ* hybridization were not restreaked to obtain single colonies, these "double-clones" will give rise to two signals. Second, some BACs contain genes highly homologous to other genes at different chromosomal locations and thus can result in more than one signal. This is in fact used for example to locate new olfactory receptor gene clusters in the genome. Hence, even if 4% of the BAC clones result in more than one signal using *in situ* hybridization, the number of chimeras is probably much smaller.

Another question relating to the usefulness of BAC clones as mapping substrates, is whether they are randomly distributed along the chromosome. The map produced for the 1.1 Mb TCR region suggested an even distribution of the BACs found in this region. One 3 kb gap remained, which is expected from a library with 3.7 fold coverage. Again from the *in situ* hybridizations mentioned above, the BACs were also found to be evenly distributed along the chromosome.

BAC clones have one more advantage over YAC clones. Since BAC clones essentially are big plasmids, their DNA can readily be separated from the *E.*coli host DNA. YAC clones usually is copurified with the yeast chromosomes.

Having the map in hand, the question was now, how to go about obtaining the sequence. Most sequencing efforts had focused on random or shotgun sequencing of cosmids. One of the major bottlenecks in the genome project is going from the YAC map to a high resolution cosmid map (a sequence ready map). First, cosmid clones have to be obtained either directly by subcloning from YACs or by obtaining the cosmid clones from genomic or chromosome-specific libraries for a larger region and then binning these against the YAC map. Second, the cosmids thus obtained will have to be ordered to pick the minimal tiling path across the region for sequencing. To circumvent this major bottleneck, in this case obtaining cosmid clones from the BACs, I attempted to sequence the BAC clones directly by the random shotgun method. One major concern was whether the different computer assembly programs could handle the extra sequences required to assemble a BAC versus a cosmid, since on average the BAC inserts are three to four times as large as the cosmid inserts. However, in parallel with the sequencing efforts, new powerful basecalling, quality assignment, and assembly programs were developed that could handle large numbers of shotgun sequences (Phil Green, in preparation). A total of five BACs and later two PACs were successfully sequenced by the shotgun approach. BACs have an advantages over both cosmids and PACs in shotgun sequencing, in that their cloning vector constitutes a smaller percentage of the total insert DNA.

In the final stages of a mapping project to detect overlaps between contigs or to make certain one picks the minimal overlapping clones for a sequencing tiling path, sequence information from the insert ends of DNA clones is essential. End sequences can be obtained for BACs, PACs, and YACs by different PCR based approaches, the vectorette technique (Riley et al., 1990), the Alu-vector method (Nelson et al., 1991), or a newly developed technique using degenerate primers with vector specific primers (Liu and Whittier, 1995). However, these all involve multiple steps, and require specific sequences to be present near the insert ends. For these reasons, I developed a technique to sequence the ends of BAC inserts directly (Chapter 6). The ability to easily obtain end sequences

from BAC inserts and the complete sequence by shotgun approaches in combination with BAC clones' fidelity and quite even distribution over the genome, have led to a new strategy for sequencing the human genome (Venter et al., 1996). The strategy, called the sequence tagged connector (STC) approach, avoids both the low and high resolution physical mapping procedures now employed. In short, a 15X BAC library with an average insert size of 150 kb is constructed. Sequence information is obtained from both ends of all BAC inserts and stored in a database. One or more "seed" BACs are sequenced in their entirety, and the final sequence compared to the end sequences in the database. On average, 30 overlapping BAC clones will be identified for each sequenced BAC. BAC clones with minimal amount of overlap to either end of the "seed" BAC are picked for the next cycle of complete sequencing. Thus sequencing proceeds in each direction outward from the "seed" BAC clone.

In conclusion, BAC clones are excellent mapping and sequencing reagents, avoiding the technically difficult and time consuming YAC clone to cosmid clone conversion. Alternatively, they could be used in the STC-approach, a procedure that involves almost no mapping at all.

Analysis of 1.07 Mb DNA sequence: The TCR α/δ locus

I analyzed the complete sequence of the human TCR α/δ region, 1.07 Mb, using several different computer programs, designed to find sequence similarities either within the sequenced region itself or against DNA, protein, EST, and DNA repeat databases. This resulted in identification of 57 V gene segments, forty-eight of which seem to be functional. I analyzed many of the V gene segments for polymorphisms. The sequence also revealed several regions with similarity to other non-TCR genes. Five olfactory receptor genes, a gene encoding a zinc finger protein, and the DAD gene were identified. Analysis of genome wide repeats and GC-nucleotide content across the locus revealed an

interesting division into three separate domains. One of these domains correlated with a region of highly conserved sequence between man and mouse. The existence of these domains might have relevance for rearrangement within the α/δ locus.

TCR elements:

Several hundred sequences, mostly from cDNA studies, have been deposited in Genbank for TCR α/δ gene segments in different species. The complete TCR α/δ sequence was screened against Genbank. I also performed a more sensitive search against a library made from all TCR α/δ elements. This analysis revealed all 45 previously known Vα/δ gene segments (Arden et al., 1995) as well as 12 other sequences with similarities to known V gene segments. Three of these sequences appeared to encode functional TCR chains, whereas the other nine were classified as pseudo-V gene segments, since they appear to be non-functional in that they are missing one or more of the following features: start codon, splice sites, or recombinational signals, or they have frameshifts and/or stopcodons (Table 1, Chapter 3). Pseudo gene segments have been included in the numbering system and map in Chapter 3, since they might have functional alleles in other humans. One particular interesting case, is the Vα8.5 gene segment which is 96% similar to Vα8.3 in the coding regions, indicating that one arose from the other by a recent duplication. However, the Vα8.5 gene segment contains a 1.2 kb insertion of a repeat element, MER11, in its second exon thus disrupting the reading frame. This V gene segment could potentially be functional in alleles that never had the MER11 insertion. Another 25 sequences were identified with similarity to V gene segments. These were termed relics and will presumably never gain function again. Each V gene segment is composed of a promoter, a small exon1 (40-55 bp) encoding most of the leader peptide, an intron ranging in size from 90 to 459 bp, exon2 (275-300 bp), and finally a recombinational signal at the 3' end of exon2.

The V gene segments are traditionally grouped into subfamilies based on two or more members sharing 75% or more nucleotide identity. By this criteria the 57 Vα/δ gene segments were grouped into 44 subfamilies. The majority of these are single membered. The seven multimembered subfamilies contain up to seven members. The V gene segments (excluding Vδ2 and Vδ3, see later) are spread evenly over 700 kb, resulting in an average density of one V gene segment per thirteen kb. This is somewhat less densely populated than the TCR β locus, where one V gene segment is found every eight kb (Rowen et al., 1996). The two loci have approximately the same number of V gene segments (the TCR β locus contains 65 V gene segments, 19 of which are pseudo genes). However, the majority of the Vβ gene segments have arisen by more recent duplications, and thus the Vβ gene segments are divided into fewer subfamilies with more members compared to their Vα/δ counterparts.

To get an idea of the number of T cells expressing different V gene segments, the V gene segments were compared against all of the Vα/δ cDNAs present in Genbank. This comparison indicated different levels in usage of V gene segments (Figure 1, Chapter 3). Many of the more 3' V gene segments only had few cDNAs present in Genbank. This however, might be due to their recent discovery and therefore earlier cDNA studies did not include probes for these V gene segments. Three of the potentially functional V elements (Vα7, Vα9, and Vα18) were not found amongst the cDNAs. The reason for this is unclear, but could be due to low levels of expression, or as above that the majority of cDNA studies performed relied on already known sequences. It should be mentioned that the same phenomenon has been seen in the human TCR β locus (Rowen et al., 1996). Two apparently functional V gene segments had no cDNA counterparts. These Vβ gene segments were shown to encode amino acids, that would impair the three-dimensional structure of the TCR. A few V gene segments (Vαδ14 and Vα17) had many cDNA complements in Genbank, but this was due to one study analyzing two highly expressed V gene segments in thyroiditis. Thus comparisons to cDNAs in Genbank does not give an

accurate picture of the normal expression of the different V gene segments. Two studies have reported differential levels of V gene segment usage (e.g., Vα12, Vα13 and Vα21 (Robinson, 1992, and Moss et al., 1993). The studies were performed using RNA obtained from peripheral blood cells from one or a few individuals. Therefore the differences may have arisen from similar thymic selection or clonal expansions in individuals. Increased usage of the Vα12 and Vα13 elements can be explained, since the two studies did not distinguish between the different subfamily members, and thus combined the expression frequencies of several V gene segments into one group. However, the Vα21 element represents a single member family. Thus it is interesting to speculate why this gene segment is found expressed in many peripheral T cells. I could not find any explanation for this, based on location of this V gene segment or on its promoter or recombinational signal sequences, and it is possible it is due to the selection procedure and not increased DNA rearrangement frequencies.

As mentioned in the introduction, one of the more interesting question in immunology concerns the preferential rearrangement and expression of specific V gene segments. How does the premature T cell decide to express an αβ or a γδ TCR? Why is the Vδ2 element expressed in the very first wave of T cells in the fetal thymus followed a few days later by Vδ1 expression? Both occur before Vα gene segment rearrangement. Why does Vδ1, which is found in the middle of the Vα gene segments rearrange almost exclusively to form TCRδ chains? Several studies have shown enhancer and silencer regions located close to the two constant regions, Cδ or Cα (Winoto and Baltimore, 1989, Lauzurica and Krangel, 1994a,b). Each of these regions has several transcription factor binding sites, many of them in common. The majority of transcription factors specific for the binding sites in these regions are not T cell specific, let alone T cell lineage specific. Special combinations of transcription factors may be responsible for some of the differential V gene segment rearrangement. Careful examination of V gene segment promoters or recombinational signals has been difficult, because few have been identified.

I aligned all the DNA recombinational signals grouping them into three categories: those associated with Vδ, Vα, or Vαδ gene segments, respectively (Figure 3 in Chapter 3). No obvious differences were noted. Vδ2 and Vδ3 do have heptamers that differ from the consensus, but only in positions less conserved in the recombinational elements for Vα gene segments. Similar differences have been found in the recombinational signals for some of the Vα gene segments. Vδ1 has a perfect heptamer-23 spacer-nonamer, and so does the majority of the Vαδ gene segments. The Vδ1 element is the only V gene segment containing a 3' T nucleotide as the very last base before the heptamer. Whether this could make a difference in the rearrangement pattern from Jα to Dδ is unknown. In this regard, I studied the recombinational signals on either side of the three Dδ gene segments, compared to those of the Jα gene segments. Only Dδ3 has two perfect recombinational signals, whereas Dδ1 and Dδ2 have recombinational signals somewhat different from the consensus. This may explain why these are used less frequently in the expressed δ genes, especially in the fetal thymus, where Vδ2 almost exclusively rearranges to Dδ3.

I also classified the promoters of Vα, Vδ, Vαδ elements in their respective groups. Five different programs (Prestridge and Stormo, 1993, Chen et al., 1995, Prestidge, 1995, Stormo and Hartzell, 1989, and Lawrence et al., 1993) were employed to screen the sequences extending 1000 bp upstream from the start codon (Gary Stormo, personal communication). Small stretches of similarity could indicate transcription factor binding sites of importance in regulating rearrangement or transcription. No apparent TATA-boxes could be found. This has also been found to be true for the TCRγ genes (Hettmann et al., 1992). A 20 bp sequence was found more or less conserved around 200 bp upstream of the start codon (Figure 3 in Chapter 3) in the majority of the functional Vα gene segments (33 out of 40), in three out of the five Vαδ gene segments, and in none of the Vδ gene segments. It was found conserved in about half of the pseudogenes. This sequence has no or little similarity to other known sites for DNA binding proteins. If the function of this sequence is to bind specific proteins, it is interesting, that the 20 bp is considerable longer

than most eukaryotic protein binding sites. Thus it might contain more than one binding site, to the same or different proteins. Whereas it might be of significance that the 20 bp sequence is not found in the Vδ genes, it does not play an essential role in DNA rearrangement or transcription, because it is not found in all of the expressed of the Vα gene segments. An earlier study has shown that a 600 bp fragment upstream of a Vα gene segment contains T cell specific promoter activity (Luria et al., 1987). What protein(s) if any this sequence may bind is unknown. No transcription factors are currently known to recognize this sequence. In this regard it is interesting to note, that I found a presumably functional gene encoding a novel zinc finger protein upstream of the TCR region (see below). Zinc finger proteins are often transcription factors, and some of them like GATA-3 have been shown to bind to enhancers in the TCR loci (reviewed in Leiden, 1993, Marine and Winoto, 1991, and Ho et al., 1991). Whether these 20 bp are found conserved in other TCR V gene segments is currently under investigation. A conserved CREB element has been observed in more than half of the Vβ gene segments (Rowen et al., 1996). A few of the promoters for the Vα/δ gene segments contain CREB elements, but the majority do not (Gary Stormo, personal communication).

Polymorphisms in the Vα/δ gene segments:

I screened most of the functional Vα/δ gene segments for DNA variations comparing 6-8 individuals (12-16 haplotypes). This project was part of a larger effort to find polymorphic markers evenly distributed across the TCR α/δ locus to use in disease association studies. Thirty V gene segments were amplified by PCR from 8 different individuals, and the products sequenced (Chapter 4). Of these 30 V gene segments, half were found to contain DNA variations for a total of 27 variations. These were later shown to be polymorphisms by typing a hundred individuals (Deborah Nickerson, personal communication). The majority of the polymorphisms were single base nucleotide substitutions, whereas a few microsatellites were found in introns. Twelve of the sequence

variations were found in the introns, and fifteen in exons. The frequency of polymorphisms was the same in introns and exons, one in 433 bp, over the 12-16 chromosomes studied. Based on the frequencies of each allele found in the 6-8 individuals, the chances for identifying two different alleles when comparing two random sequences were calculated for each polymorphism. These values ranged from 0.12 to 0.49, and were utilized to calculate the overall nucleotide diversity, which is defined as the number of differences per nucleotide site, when comparing two random sequences. This was found to be 8.0×10^{-4} (0.08 %) or a little less than one nucleotide in a thousand. This value is comparable to the nucleotide diversity obtained by Li and Sadler (1991), who compared approximately 75,000 bp from 49 genes.

Of the 15 substitutions in the coding regions, six are silent mutations, whereas the other nine lead to amino acid changes. All of the silent substitutions are found at codon position three, whereas the functional substitutions are found at codon position one (4), two (3), or three (2). This is expected, since if there's no codon bias, 95% of substitutions in codon position 1 will lead to amino acid change, all nucleotide changes in position 2 are functional, whereas only 28% of nucleotide changes in position 3 leads to amino acid change (Nei, 1988). Thus 74% of random mutations would lead to amino acid changes. In theory, if no selection is involved, the percentage of synonymous variations per synonymous site should be the same as the percentage of non-synonymous changes per non-synonymous site. Not considering possible codon bias, here the number of functional nucleotide changes per non-synonymous site is smaller than the equivalent number for the synonymous sites. This indicates negative selection for amino acid substitution. However, the distribution of the two kinds of nucleotide changes is not random. All of the silent variations were found in the framework sequences, whereas many of the amino acid changing variations were found in or around the first and second hypervariable domains (Figure 2, Chapter 4). If one only considers the hypervariable regions, only functional sequence variations have been found, which could indicate positive selection for amino acid

changes in these regions. However, the data here are too few to make any conclusions in this regard. Furthermore, the most dramatic amino acid changes seem to fall in these hypervariable regions. In the framework regions, there seems to be negative selection, a higher proportion of silent mutations has been found here. It is particularly noteworthy, that none of the functional substitutions has been found at any of the 40 conserved residues (Chothia et al., 1988). Polymorphic sites have been reported in both human and mouse TCR β V gene segments, leading to non functional alleles (Charmley et al., 1993). Even deletion of larger genomic regions of DNA spanning several V gene segments has been found (Seboun et al., 1989). In all individuals tested, all 30 V gene segments were amplified, suggesting that there were no major deletions including these elements, or genes mutated so strongly that they were no longer recognized in the PCR assay.

The functional mutations described here supposedly lead to functional changes in the TCR. This has been shown for different alleles of V gene segments in the TCR β region (Posnett, 1990). These changes can alter α/β pairing stability, influence positive or negative selection in the thymus, or modify the recognition of MHC/antigen.

The polymorphisms will be useful in studying the possible correlations between specific TCR alleles and susceptibility to autoimmune disease. A careful analysis should include markers in linkage disequilibrium with one another across the entire locus. Preliminary results of the polymorphisms in this study, indicate that even markers close to each other might not be in complete linkage disequilibrium (Deborah Nickerson, personal communication). Additional genetic markers should be developed. One of the tools now available from the entire sequence, is the ability to select microsatellites across the locus. Microsatellites are often highly polymorphic, and so by testing them by PCR against several individuals, the polymorphic candidates can be identified.

Finally, polymorphisms can arise from sequence variations in noncoding sequence. The sequences from the overlapping BAC, PAC, and cosmid clones were compared. If the sequence was obtained from two different haplotypes (Table 4, Chapter 3), the variations

found could be due either to sequencing errors or DNA variations. All differences were checked against the original raw sequence data to determine whether they were errors or DNA variations. DNA variations were found quite frequently (every 750 bp) in most of the overlapping sequences, except for the overlaps at the 3' end of the locus. This might reflect the need for the 3' end to be highly conserved. It could also be a result of different mutation rates between the distinct chromosomal domains in which the overlaps are located (see below).

Evolution of the TCR α/δ region: Homology units, repeats, and chromosome structure.

The Vα/δ gene segments belonging to the same subfamily have in many cases arisen by recent large scale duplications of the genomic DNA. Several large duplication events have occurred including 50 and 20 kb homology units as indicated in Figure 7 in Chapter 3. These regions contain the Vα8, Vα11, Vα12, Vα13, Vα14, and Vα15 subfamilies, three of which (Vα8, Vα12, and Vα13) are multimembered because of these duplications. The other Vα elements involved in the duplication have diverged from each other so they no longer are classified in the same subfamily, some of them have become pseudogenes or even relics over time. Smaller local duplication have also occurred including the Vα1, Vα8, and Vα38 gene segments. In these cases, two V members of the same family are found next to each other (e.g., Vα38.1 and Vα38.2 are 95% identical). In the case of Vα8, the small local duplication has happened before the large scale duplications (Figure 7, Chapter 3). For example is Vα8.2 much more similar to Vα8.4 in the other homology unit, than it is to Vα8.3 right next to it in the same homology unit. These long range duplications give an idea of how the TCR V gene segment repertoire has diversified over time through small and large duplications.

Evidence of other local duplications has been found with the five identified olfactory receptor genes (see below). They can be divided into two subfamilies with two and three members, respectively, based on their nucleotide sequence.

Most of the larger homology units mentioned above have arisen relatively recently. A comparison to the mouse TCR α/δ map (Wang et al., 1994) indicates that the human homology units are not found in mouse. Other large scale duplications involving different V gene segments have been found in mouse. Although the entire sequence for mouse is not yet available, smaller regions have been sequenced. Koop et al., 1992, sequenced almost 100 kb in mouse encompassing the Cδ, Vδ3, Jα, and Cα elements. When this sequence was compared to the equivalent region in human a striking similarity of about 70% was found across the entire 100 kb, even though the coding regions constitute only about five percent. With more sequence now available for both human and mouse, further similarity analysis could be carried out. I compared thirty kb upstream to the previously sequenced 100 kb region with the corresponding region in mouse (Lee Rowen, personal communication). This sequence includes the Vδ2, Dδ, and Jδ elements, and it was of interest how far 5' the highly conserved region extended. The sequence similarity extended across this 30 kb region as well (Figure 6, Chapter 3). However, it was observed that whereas the Vδ3 gene segment in human is the homologue of the Vδ5 element in mouse (both are found 3' to the Cδ gene), the human Vδ2 element was not orthologous to the mouse Vδ1 element. This is surprising, since these two V gene segments both are expressed in the first wave of T cells in thymic development and both map to the approximately same location relatively to the other TCR elements in the human and mouse regions, respectively. It is possible that, if in fact they shared a common ancestor, one of the gene segments was duplicated to another location and there was no longer a need to preserve the original orthologue. I next compared a 35 kb region including the Vα26.2 and Vα34 gene segments to a cosmid sequence containing the mouse Vδ2, Vαδ6, and Vα16 elements (Seto et al., 1994). This region in human is found almost 200 kb upstream from

the conserved 130 kb region. At this point only the V genes themselves exhibited similarity (Figure 6, Chapter 3).

The reason for the high sequence conservation across the 3' end is unknown, but the conserved domain corresponds to striking changes in percent GC content and density of genome wide repeats. Studies have indicated that the human genome can be divided into isochores based on their GC content (Bernardi, 1993). Four isochores have been defined in human, with GC contents of about 40 %, 45 %, 48 %, and 53 %, respectively. Isochores are estimated to be larger than 300 kb in length, although they are usually not as long as chromosomal bands. Chromosomal bands are characterized by their staining intensity using Giemsa and are generally several megabases. G-bands (Giemsa dark) are primarily composed of isochores of the GC-poor type, mostly the 40% isochore, whereas R bands (Giemsa pale) are composed of all four types (the majority being the 40% and the 45% isochores). The T bands are composed of the three most GC-rich isochores. The character of genome wide repeats present varies with GC content, e.g., Alu elements, which are GC rich, are mostly found in GC rich sequence, and LINE elements (AT-rich) are found in the GC-poor regions. Figure 6 in Chapter 3 shows the GC content and distribution of LINE and Alu elements across the entire TCR α/δ region. Three different regions were observed. The first 100 kb has about 45% GC, which drops gradually over the next 100 kb to about 38%. GC content is low (38%) in the 750 kb spanning the V elements. It then increases to 45% again over the last 180 kb. The localization of the 40% and 45% isochores in the TCR α/δ locus corresponds to its cytogenetically determined location on chromosome 14q11.2 (Barbara Trask, personal communication) being an R-band. The LINE elements are present in low levels at the 5' and 3' ends, whereas they make up about 22 % of the DNA sequence in the middle domain. This correlates with the observation that LINEs are more frequent in GC-poor regions. The Alu sequences occupy 36% of the first 100 kb, as expected in more GC-rich regions, and gradually become rarer as the GC content diminishes. Alu elements occupy about 7% of the middle domain,

which is GC-poor. However, the Alu content never rises again even though the GC content does over the last 200 kb. Alu repeats are rare in this last domain. It seems that in this very conserved 3' end, there is even selection against genome wide repeats being inserted. This is further supported by an analysis of the few repeats present here. They seem to go back to before the divergence between mouse and man (Adrianus Smit, personal communication).

Another interesting point about these GC isochores is their relationship to other chromosomal behaviors. Of interest here is the open chromatin structure and early replication of GC rich regions, as well as the occurrence of increased transcription and recombination. These factors have all been implied in the preferential rearrangement of specific V gene segments. GC-rich regions further seem to be associated with a decrease in mutation rate (Wolfe et al., 1988). This could account for some of the difference observed in the frequency of variations in the middle domain (one variation every 750 kb) versus the 3' domain (one variation every 4.5 kb).

The sequence contains genes not related to the TCR:

The TCR α/δ elements extend over 900 kb from the most 5' Vα1.1 gene segment to the Cα region. On both sides of this region are found other genes not related to the TCR elements.

Although not fully sequenced yet, a gene encoding the defender against apoptotic cell death, DAD, has been found 15 kb 3' to the Cα region (Kai Wang, personal communication). This protein is highly conserved in different species (human and hamster 100%) as found by cDNA studies (Nakashima et al., 1993) and has been mapped to chromosome 14q11-q12 in human and chromosome 14 in mouse (Apte et al., 1995) in correspondence to it's location here next to the TCRα/δ locus.

At the 5' end, a family of five olfactory receptor genes were identified. These were found interspersed with the Vα1.1 and Vα1.2 elements, and divided into two groups

based on their sequence similarity. Three of them, one of which is a pseudogene due to a single stop codon, are highly homologous to a rat olfactory gene, whereas the other two, one of which contains a frameshift and therefore are non-functional, are more similar to a chicken olfactory receptor (Figure 2, Chapter 3). The olfactory receptors, each encoded by a single exon, are between 310 and 314 amino acids long. Both of the two clusters have diverged from other human olfactory receptor genes, and thus can be classified as new subfamilies (Ben-Arie et al., 1993). At the protein level, the similarity between olfactory receptor one, two, and three is between 67 and 93%, whereas they are similar to the second subfamily by 37-44%. The three member subfamily is similar to a rat olfactory receptor at 73-87%. Olfactory receptors contain seven transmembrane helices (Figure 2 in Chapter 3), and belong to the superfamily of G-protein coupled receptors. The amino acid differences in the five proteins found here, are primarily found at the N- and C-terminals, and are otherwise spread evenly across the transmembrane domains and the intra- and extra-cellular loops. In a sense, the organization of olfactory receptor genes is like one of the TCR families. It is likely that an original olfactory receptor gene has duplicated and translocated to different chromosomes so as to generate several gene clusters each containing many olfactory receptor genes. Like the V gene segments, many of the olfactory receptor genes have been shown to be pseudo-genes (Crowe et al., 1996), like two of the five found here. It has been suggested, that the regulation of expression is determined by both allelic inactivation and cis-controlling elements to ensure that each neuron expresses only one or a few olfactory receptors (Chess et al., 1994).

Upstream of the olfactory receptor genes is a gene encoding a zinc finger protein. An open reading frame of 1003 amino acids with similarity to a special class of homeotic genes was identified. This class includes the Drosophila *sal* gene (Kuhnlein et al., 1994) and its Xenopus homologue, *Xsal-1* (Hollemann et al., 1996). Like these to genes, the zinc finger gene found here encodes three double zinc fingers of the CC/HH type (Appendix, Chapter 3) as well as a single zinc finger connected to one of the double zinc

fingers. Each of the double zinc fingers is bridged by an H/C-link (Schuh et al., 1986). It furthermore contains a zinc finger of the CC/HC class which has also been found in Xenopus *Xsal-1*, but not in Drosophila *sal*. Except for the zinc finger domains little homology is observed between the three proteins. There are one glutamine-rich stretch found conserved in all three species, and another smaller region has been conserved between human and Xenopus. The genes in Drosophila and Xenopus include a small first and third exon, which I have not been able to locate in the zinc finger gene found here. However, several cDNAs have been identified in the EST databases with perfect matches to the 3' end of this region, indicating this portion to be the long untranslated region also found in the Xenopus and Drosophila genes. The cDNAs have been found in many different both fetal and adult tissues. Upstream of the gene, where a potential first exon may be located is a CpG-island. CpG islands are usually found with household genes. *Sal* and *Xsal-1* have predominantly been found expressed in the central nervous system early in development, and have been suggested to function as transcriptional activators or suppressors. The function and expression pattern of the zinc finger gene found here is unknown.

Besides these genes, several other similarities to known genes were found (Table 3, Chapter 3). The majority of these were found 5' to the TCR V gene segments, and are probably pseudogenes. Many of them contain introns, suggesting they have arisen by gene duplication. They are rendered non-functional by one or a few frameshifts and/or stop codons, but otherwise have an amino acid sequence almost identical to their functional counterparts.

Several other regions identical or with similarity to ESTs or longer unknown sequences in Genbank, may constitute new genes or DNA repeats. Likewise, many open reading frames have been located with unknown, if any function.

In conclusion, the sequence of the human TCRα/δ locus has not only revealed the overall organization and structure of the TCR elements, and given us the tools to further exploit many aspects of the immune system, but it has also revealed interesting chromosomal features, and given insight into the number of genes and pseudogenes in the DNA surrounding this locus.

References:

Apte, S.S., Mattei, M.-G., Seldin, M.F., and Olsen, B.R.: The highly conserved defender against the death 1 (DAD1) gene maps to human chromosome 14q11-14q12 and mouse chromosome 14 and has plant and nematode homologs. FEBS Lett. 363: 304-306, 1995.

Arden, B., Clark, S.P., Kabelitz, D., and Mak, T.W.: Human T-cell receptor variable gene segment families. Immunogenetics 42: 455-500, 1995.

Ben-Arie, N., Lancet, D., Taylor, C., Khen, M., Walker, N., Ledbetter, D.H., Carrozzo, R., Patel, K., Sheer, D., Lehrach, H., and North, M.A.: Olfactory receptor gene cluster on human chromosome 17: possible duplication of an ancestral receptor repertoire. Hum. Mol. Genet. 3: 229-235, 1994.

Bernardi, G.: The isochore organization of the human genome and its evolutionary history - a review. Gene 135: 57-66, 1993.

Charmley, P., Wang, K., Hood, L., Nickerson, D. A.: Identification and physical mapping of a polymorphic human T cell receptor Vβ gene with a frequent null allele. J. Exp. Med 177: 135-143, 1993.

Chen, O.K., Hertz, G.Z., and Stormo, G.D.: Matrix search 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. Comput. Appl. Biosci. 11: 563-566, 1995.

Chess, A., Buck, L., Dowling, M.M., Axel, R., and Ngai, J.: Molecular biology of smell: Expression of the multigene family putative odorant receptors. Cold Spring Harbor Symp. Quant. Biol. LVLL: 505-516, 1992.

Chothia, C., Boswell, D. R., and Lesk, A. M.: The outline structure of the T-cell αβ receptor. EMBO 7: 3745-3755, 1988.

Crowe, M.L., Perry, B.N., and Connerton, I.F.: Olfactory receptor-encoding genes and pseudogenes are expressed in humans. Gene 169: 247-249, 1996.

Green, E. D., Riethman, H. C., Dutchik, J. E., Olson, M. V.: Detection and characterization of chimeric yeast artificial-chromosome clones. Genomics 11: 658-69, 1991.

Hettmann, T., Doherty, P.J., and Cohen, A.: The human T cell receptor gamma genes are transcribed from TATA-less promoters containing a conserved heptamer sequence. Mol. Immunol. 24: 1073-1080, 1992.

Ho, I-C., Vorhees, P., Marin, N., Oakley, B.K., Tsai, S.-F., Orkin, S.H., and Leiden, J.M.: Human GATA-3: a lineage-restricted transcription factor that regulates the expression of the T cell receptor a gene. EMBO 10: 1187-1192, 1991.

Hollemann, T., Schuh, R., Pieler, T., and Stick, R.: Xenopus Xsal-1, a vertebrate homolog of the region specific homeotic gene spalt of Drosophila. Mech. Develop. 55: 19-32, 1996.

Koop,B. F., Wilson, R.K., Wang, K., Vernooij, B., Zaller, D., Kuo, C.L., Seto, D., Toda, M., and Hood, L.: Organization, structure and function of 95 kb of DNA spanning the murine T-cell receptor Cα/Cδ region. Genomics 13: 1209-1230, 1992.

Koop, B. F., Rowen, L., Wang, K., Kuo, C. L., Seto, D., Lenstra, J. A., Howard, S., Shan, W., Deshpande, P., and Hood, L.: The human T-cell receptor TCRAC/TCRDC (Cα/Cδ) region: Organization, sequence, and evolution of 97.6 kb of DNA. Genomics 19: 478-493, 1994.

Kuhnlein, R.P., Frommer, G., Friedrich, M., Gonzalez-Gaitan, M., Weber, A., Wagner-Bernholz, J.F., Gehring W.J., Jackle, H., and Schuh, R.: Spalt encodes an evolutionarily conserved zinc finger protein of novel structure which provides homeotic gene function in the head and tail region of the Drosophila embryo. EMBO 13: 168-179, 1994.

Lauzurica, P and Krangel, M.S.: Enhancer-dependent and in-dependent steps in the rearrangement of a human T cell receptor δ transgene. J. Exp. Med. 179: 43-55, 1994a.

Lauzurica, P and Krangel, M.S.: Temporal and lineage-specifc control of the T cell receptor α/δ gene rearrangement by T cell receptor a and d enhancers. J. Exp. Med. 1913-1921, 1994b.

Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C.: Detecting subtle sequence signals: a Gibbs smapling strategy for mulitple alignment. Science 262: 208-214, 1993.

Leiden, J.M.: Transcriptional regulation of T cell receptor genes. Annu. Rev. Immunol. 11: 539-570, 1993.

Li, W.-H. and Sadler, L. A.: Low nucleotide diversity in man. Genetics 129: 513-523, 1991.

Liu, Y. G. and Whittier, R. F.: Thermal asymmetric interlaced PCR: automatable amplification and sequencing of insert end fragments from P1 and YAC clones for chromosome walking. Genomics 25: 674-81, 1995.

Luria, S., Gross, G., Horowitz, M., and Givol, D.: Promoter and enhancer elements in the rearranged α chain gene of the human T cell receptor. EMBO 6: 3307-3312, 1987.

Marine, J. and Winoto, A.: The human enhancer-binding protein Gata3 binds to several T-cell receptor regulatory elements. Proc Natl Acad Sci USA 88: 7284-7288, 1991.

Moss, P. A. H., Rosenberg, W. M. C., Zintzaras, E., and Bell, J. I.: Characterization of the human T cell receptor α-chain repertoire and demonstration of a genetic influence on Vα usage. Eur. J. Immunol 23: 1153-1159, 1993.

Nakashima, T., Sekiguchi, T., Kuraoka, A., Fukushima, K., Shibata, Y., Komiyama, S., and Nishimoto, T.: Molecular cloning of a human cDNA encoding a novel protein, DAD1, whose defect causes apoptotic cell death in hamster BHK21 cells. Mol. Cell. Biol. 13: 6367-6374, 1993.

Nelson D. L., Ballabio, A., Victoria, M. F., Pieretti, M., Bies, R. D., Gibbs, R. A., Maley, J. A., Chinault, A. C., Webster, T. D., and Caskey, C. T.: Alu-primed polymerase chain reaction for regional assignment of 110 yeast artificial chromosome clones from the human X chromosome: Identification of clones associated with a disease locus. Proc Natl Acad Sci USA 88, 6157-6161, 1991.

Posnett, D. N.: Allelic variations of human TCR V gene products. Immunol. Today 11: 368-373, 1990.

Prestridge, D.S. and Stormo, G.: Signal Scan 3.0: new database and program features. Comput. Appl. Biosci. 9, 113-115, 1993.

Prestridge, D.S.: Predicting Pol II promoter sequences using transcription factor binding sites. J. Mol. Biol. 249: 923-932, 1995.

Riley, J., Butler, R., Ogilvie, D., Finniear, R., Jenner, D., Powell, S., Anand, R., Smith, J. C., and Markham, A. F.: A novel, rapid method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones. Nucleic Acids Res 18: 2887-2890, 1990.

Robinson, M.A.: Usage of human T-cell receptor V-beta, J-beta, C-beta, and V-alpha gene segments is not proportional to gene number. Hum. Immunol. 35: 60-67, 1992.

Rowen, L., Koop, B.F., and Hood, L.: The complete 685 kilobase DNA sequence of the human beta T cell receptor locus. Science *(in press)*, 1996.

Schuh, R., Aicher, W., Gaul, U., Cote, S., Preiss, A., Maier, D., Seifert, E., Nauber, U., Schroder, C., Kemler, R., and Jackle, H: A conserved family of nuclear proteins containing structural elements of the finger protein encoded by Kruppel, a Drosophila segmentation gene. Cell 48: 1025-1032, 1986.

Seboun, E., Robinson, M.A., Kindt, T.J., and Hauser, S.I.: Insertion/deletion-related polymorphisms in the human T cell receptor β chain complex. J. Exp. Med. 170: 1263-1270, 1989.

Seto, D., Koop, B.F., Deshpande, P., Howard, S., Seto, J., Wilk, E., Wang, K., and Hood, L.: Organization, sequence, and function of 34.5 kb of genomic DNA encompassing several murine T-cell receptor α/δ variable gene segments. Genomic 20: 258-266, 1994.

Shizuya, H., Birren, B., Kim, U.-J., Mancino, V., Slepak, T., Tachiiri, Y, and Simon, M.: Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proc Natl Acad Sci USA 89, 8794-8797, 1992.

Stormo, G. D. and Hartzell, G.W.III.: Identifying protein-binding sites from unaligned DNA fragments. Proc. Natl. Acad. Sci. USA 86: 1183-1187, 1989.

Venter, J.C., Smith, H.O., and Hood, L.: A new cooperative strategy for sequencing the human and other genomes. Submitted, 1996.

Wang. K., Klotz, J. L., Kiser, G. Bristol, G., Hays, E., Lai, E., Gese, E., Kronenberg, M., and Hood, L.: Organization of the V gene segments in mouse T-cell antigen receptor α/δ locus. Genomics 20, 419-428, 1994.

Winoto, A. and Baltimore, D.: Developmental regulation of the TCRαδ locus. Cold Spring Harbor Symp. Quant. Biol. LIV: 87-92, 1989.

Wolfe, K.H., Sharp, P.M., and Li, W.-H.: Mutation rates among regions of the mammalian genome. Nature 337: 283-285, 1989.