# Large-Scale Sequencing and Data Analysis for Eukaryotic Genomes

Masaaki Yamazaki

DOCTOR OF SCIENCE

Department of Genetics
School of Life Science
The Graduate University for Advanced Studies

1998

# Large-Scale Sequencing and Data Analysis for Eukaryotic Genomes

Masaaki Yamazaki

# ACKNOWLEDGEMENTS

The author wishes to express his sincere thanks to Prof. Yoshio Tateno (National Institute of Genetics; NIG) for his kind and excellent guidance and generous participation in his advisory roles for the author' s doctorate.

The author's heartfelt appreciation is addressed to Dr. Eiichi Soeda (The Institute of Physical and Chemical Research) for providing the opportunity to propose this significant and valuable research subject. Without these people the present work would have been impossible.

Special thanks are extended to Prof. Hidetoshi Inoko (School of Medicine, Tokai University) for providing significant information and valuable experimental materials for this work.

The author is deeply grateful to Dr. Yasufumi Murakami and Dr. Toshihiko Eki (The Institute of Physical and Chemical Research) for their liberal support in providing superior experimental ingredients, technical information and friendship.

The author expresses his great appreciation to President Rintaro Fujii (Fujiya Co., Ltd.) and company executives for providing this special opportunity, financial support and their encouragement for this research.

The author extends his sincere thanks to Director Kuniaki Sasaki and Manager Hiroyuki Tashiro (Fujiya Co., Ltd.) for their countless and thoughtful arrangements for smooth accomplishment of this research and numerous suggestions, discussion and advice, and to all the others who kindly supported the author in areas of technical assistance and services during the study.

Last but not least, the author acknowledges with the deepest gratitude, the understanding and encouragement of his parents, Mr. Kouichi Yamazaki and Mrs. Mutsue Yamazaki, which have been given lovingly throughout the many years of the author's studies.

# CONTENTS

# Introduction

The eukaryotic genome is a unique resource for the maintenance and duplication of life, dominating all cellular activity. The ordered bases in the genome contain the complete set of instructions for genetic inheritance. In this sense, the genome can be said to be a " master copy " of life programs. Since the days of Watson & Crick, the issue of what part of the genome is responsible for each activity has been one of the central problems in molecular biology. Therefore, it is quite natural that a great number of scientists have concentrated on this problem focusing on a variety of genomes from viruses, prokaryotes and eukaryotes. If the genome is the " master copy " of life programs, the way to understand it is to decipher and analyze its nucleotide sequence. Thus, sequence analysis of the genome plays an important role in modern biology. First, by revealing the similarities of homologous genes, it provides insights into the possible regulation and function of these genes as well as into their evolutionary history. Second, sequencing human genome of 3 billion base pairs (bp) containing 50- to 100 thousand genes, will provide valuable information which will lead to an understanding of diseases related to genetic variation and have an enormous impact not only on biological research but also on the practice of medicine, as Dulbecco stated in regards to understanding of cancer (Dulbecco, 1986). Thus, a new field of biology, genome biology has been born.

In genome biology, two general forms of DNA have most often been the target of sequencing efforts : (i) genomic DNA ; and (ii) cDNA generated as a copy of messenger RNA. cDNAs typically range in size from 300 to 8000 bp and, accordingly, can generally be cloned and sequenced more easily than genomic DNA, which often requires significant physical mapping and extensive subcloning. Some scientists have concentrated their efforts on characterizing cDNA as an efficient approach to identifying the sequences that directly encode genes. Yamazaki *et al.* (1988) sequenced 308 cDNA clones randomly sampled from human cDNA libraries in an attempt to isolate non-identified genes homologous to known genes using sequences newly obtained as probes. They succeeded in isolating and identifying a full-length cDNA encoding a new human heat-shock protein, the HSP 90α gene (Yamazaki *et al.*, 1989, 1990), demonstrating that public DNA databases managed under international collaboration (Tateno and Gojobori, 1997) provide sufficient information for isolating and identifying new genes. In 1993, Adams *et al.* reported over 3400 human cDNA sequences and classified expression genes. Later, those sequences were designated Expression Sequence Tags (ESTs) and were frequently utilized in gene mapping. However, as the cDNA libraries they employed did not originate from a normalized mRNA source, they isolated many redundant candidates for unidentified genes. In addition, the expression manner of respective genes was not touched although it is most important when understanding genes. In contrast, Okubo *et al.* (1992) focused on expression bias by sequencing 1000 or more partial cDNAs per representative human cell or tissue. Construction of a "body map" represented to some extent, which genes work in the respective cells and tissues. However, many genes that are only expressed at a low level or in a cell- or tissue-restricted manner would have been missed using such approaches, as would all of the important regulatory sequences, which are not expressed as RNA. Furthermore, such studies shed little light on genomic evolutionary history.

On the other hand, genomic sequence analysis would potentially provide an insight into the ultimate question of biology, what factors permit organisms to exist as a species. In fact, virologists in the past have sequenced a lot of viral genomes to learn the substance of virus particles which can duplicate only in particular host cells, and those works have made an enormous contribution to the study of DNA replication, host-vector systems and viral-infectious diseases. In case of lower life forms such as prokaryotes, the complete nucleotide sequence (580,070 bases) of the *Mycoplasma genitalium* genome, the smallest known genome of any free-living organism, was determined quite recently (Frazer *et al.* 1995). And in the same year, the complete nucleotide sequence of the *Haemophilus influenzae* genome (1,830,137 bases) was determined (Fleischmann *et al.*, 1995). It is not too much to say that these works are monumental for molecular biology because scientists have obtained for the first time, the specifications for life from genomic features.

Although a variety of genomic DNAs have been the target of sequencing efforts, those of eukaryotic organisms, especially human, have been relatively difficult to sequence. This is because the genome is divided into a number of huge chromosomes on which some genes are located organized into multi-copy gene families, and short and long interspersed repetitive elements appear frequently. Because of this, a large-scale sequencing and analysis of the genome is indispensable. However, the complexity and size have been barriers to deciphering long continuous regions through multiple loci or an entire chromosome. Recently, physical mapping efforts have enabled us to provide ordered sets of cloned DNA which accurately represent the genome. Olson *et al.* (1991a, 1993) and his collaborators constructed a physical map of *Saccharomyces cerevisiae* chromosomes I, III, V, VI, VIII and IX by aligning sets of ordered clones on those chromosomes using lambda phages and cosmids. Also, Mizuki *et al.* (1994) constructed a complete cosmid contig spanning about 0.6 million bases from YAC (Yeast artificial chromosomes) clone Y109 (Imai and Olson, 1990) that was thought to be located on human chromosome 6 to reveal the precise organization of the human major histocompatibility complex genes by DNA sequencing. In addition, some progress in sequencing technologies such as polymerase chain reaction (PCR) , automated fluorescence sequencing machines, and advanced computer software for rapid base-calling has facilitated mass-sequencing programs. Nevertheless, if we consider the current limit of sequence size readable by Sanger's method (Sanger *et al.*, 1977), successful reconstruction of a large number of small sequences into the entire whole is necessary. Several approaches have been investigated for constructing short sequences into larger one, such as the shotgun method (Anderson, 1981, Deininger, 1983), nested deletion method (Poncz *et al.*, 1987, Hattori *et al.*, 1993), primer-walking method (Bankier *et al.*, 1989), and oligomer-hybridization method (Drmanac *et al.*, 1993). Among these, the shotgun approach has two powerful advantages for practical genomic sequencing when compared with others : (i) it provides massive sequencing with continuous flow and massive production of data with no interruption to the final stage ; and (ii) its applicability to wide-ranged sequencing materials from clones of 10 to over 100 kilo-base pairs (kbp) such as lambda phages, cosmids, P1 plasmids, P1-derived artificial chromosomes (PACs), Bacterial artificial chromosomes (BACs) and YACs. However, success in shotgun sequencing projects with larger targets depends on the efficient preparation of a shotgun library and on sufficient computer assembly for the reconstruction of a large number of sequences.

This thesis is composed of three parts among which the last part (Part 3) is the main component. The theme of the thesis is twofold, replication and duplication of genome sequences.

In Part 1, I will review the procedure for constructing a shotgun library and data assembling algorithm; I played the major role in developing the procedure (Yamazaki *et al.*, 1995). This part is concerned particularly with the need of an efficient preparation of a shotgun library and a sufficient computer assembly for the reconstruction of a large number of sequences mentioned just above. In Part 2, I will revisit the result of the application of the procedure to one of the smallest eukaryotic chromosomes, the *Saccharomyces cerevisiae* chromosome VI (Murakami, Yamazaki *et al.*, 1995; Naitou, Yamazaki *et al.*, 1995, 1996; Eki, Yamazaki *et al.*, 1996). It is noted that the density of the coding regions of the yeast genome is thought to be relatively high among the eukaryotic genomes, and that the chromosome VI is the only chromosome for which the autonomously replicating sequence (ARS) activity has been studied (Shirahige *et al.*, 1993) prior to the completion of sequencing among the sixteen chromosomes of this species. I will focus particularly on the organization of genetic and ARS elements on the chromosome in this part.

In Part 3, I will discuss the organization of multi-copy genes in the human leukocyte antigen (HLA) class I region, the duplication of genome structures of this region in evolution, and the discovery of possible genes in it. I and my colleagues first sequenced the region spanning 385 kb around the centromeric end of the HLA class I region by use of the procedure mentioned above. We then analyzed this region with respect to the three aspects just mentioned. The major part of Part 3 will soon be published (Yamazaki *et al.*, 1999)

# Part 1.

# Improvement of Shotgun Sequencing Approach

## Advantage of the approach

Since the current size of sequences obtainable from a single Sanger's reaction set is limited to several hundred to perhaps 1000 bases, the characterization of long regions of DNA, such as found in most eukaryotic genes, must be accomplished through successive generation of many smaller sequences. The manner in which these smaller sequences are obtained from the larger whole is a fundamental issue in choosing of overall sequencing approaches. Current approaches fall into two general categories, directed and random (table 1-1). The former permit the direct and sequential sequence analysis of a large DNA fragment from one end to the other. Representative methods are the primer-directed (walking) method (Bankier *et al.*, 1989) and nested deletion (ND) method (Poncz *et al.*, 1987, Hattori *et al.*, 1993). Direct sequencing minimizes the redundancy required and permits one to concentrate on the regions of greatest interest. However, the primer method may not be necessarily successful when sequencing highly repetitive regions and, the ND method requires prior knowledge of the relation of each fragment to the original target, highly skilled technicians to deal with the generation of nested deletion clones, and optimizing of sequencing reactions in problem regions. In addition, due to the chemical properties of Sanger's reaction, direct approaches are applicable to 100 kb or shorter sequencing templates for sufficient reading. On the other hand, random or "shotgun" approaches generate a library of subclones through random cleavage or shearing of a large piece of DNA (Anderson *et al.*, 1981, Deininger *et al.*, 1983). The appeal of random strategies lies in the absence of a need for prior information about particular subclones. This allows projects to be undertaken with a great deal of automation. In addition, it has the capability for massive sequencing with continuous flow and massive production of data with no interruption to the final stage. The second advantage is the overdetermination of most of the sequence, which minimizes the number of remaining errors. Also, the process is convenient and data accumulate relatively fast during the entire phase of sequencing. The third advantage is its wide availability for various kinds of targeted clones. No other single approach is applicable to a variety of sequencing materials from clones of 10 to over 100 kbp, such as λ-phages, cosmids, P1-

plasmids, BACs, PACs and YACs. Recently, a novel random approach categorized as sequencing by hybridization was investigated (Drmanac *et al.*, 1993). However, it requires a highly discriminative hybridization procedure to distinguish between perfect match and one base mismatch, and is more error-prone than Sanger's reaction.

Thus, the shotgun approach is the most practical strategy for large-scale genomic sequencing and this is why I have used the method since 1988.

## Improvement of the shotgun approach

### Principle of the approach

In the shotgun sequencing approach, small fragments are generated from a vast number of identical target sequences, so the resulting library from which they are selected for further analysis is redundant. Therefore, individual fragments may overlap in the sense that they commonly possess some bit of the target sequence. The presence of such overlaps allows retrospective determination of which fragments represent adjacent target sequences. When enough overlapping fragments have been analyzed, the original sequence may be deduced (Figure 1-1). In the case of sequencing a cosmid clone, how many random fragments are sufficient to cover the original whole ? The relationship between the target length and the number of fragments required is shown in figure 1-2 and 1-3(A). According to the model, about 500 fragments of 500 bp-long sequences are required to cover 99% of the target region. In general, a target region not covered in any fragment is a gap. Adjacent contigs (islands) are thus separated by gaps. Figure 1-3(B) shows the relationship between the number of islands and the degree of redundancy with which it is assumed to have been sequenced. Cosmid sequencing requires redundancies around sevenfold for closure according to the mathematical model. Agreement with experimental results show that cosmid sequencing requires redundancies around six- to sevenfold of the target to reach closure (Yamazaki *et al.*, 1995, Mizuki *et al.*, 1997b). However, there is an exponentially increasing cost in redundancy used to close the final several gaps in the shotgun approach (Roach, 1995). Therefore, when read with redundancies of around five to six, beginning direct sequencing (primer-walking) for the remaining gaps is fundamental and practical.

## Construction of a single high-quality library

When constructing a shotgun library, quality must be evaluated according to three important criteria regarding generated clones; how long an insert they have, how randomly they are distributed, and how many are obtained. In addition, it is necessary to accomplish this in as few steps as possible to preserve the advantages of this approach. In view of this, shearing by sonication (Deininger *et al.*, 1982, Yamazaki *et al.*, 1995) is the most practical approach. In comparison with other methods for generating random fragments from large target DNA such as enzymatic treatments, shearing by sonication is more convenient, has fewer parameters to optimize and produces overlapping clones more randomly. My experimental results showed that the size of fragments generated by sonication depends only on the shearing time when the target DNA molecules are treated. Furthermore, the introduction of a cup-horn sonication system with high throughput (Yamazaki *et al.*, 1995) enabled us to make the target into the desired size in a few seconds (Figure 1-4). Size fractionation of the sheared fragments is indispensable to obtaining useful clones with a sufficient insert size, longer than at least 1kbp, for final gap closure to be considered possible. Fragments that are too small are also a disadvantage because they can possibly be cloned together in the same recombinant. Spin-colunm chromatography using sephacryl S-400 is a more rapid, efficient and rather convenient method for size selection than gel electrophoresis separation (Table 1-2, Figure 1-5). In optimizing of the end-repairing reaction of the sheared DNA fragments for blunt end ligation for cloning into an appropriate cloning vector, subsequent enzymatic reactions have been considered necessary (Maniatis *et al.*, 1989). Nevertheless, my experimental data produced a surprising results that showed, even a brief reaction using only the Klenow enzyme works well enough to obtain sufficient recombinants to satisfy the requirement estimated mathematically (Table 1-2, Figure 1-5). A plasmid vector is superior to a single-stranded vector such as the M13 phage because of its capability to reduce the number of clones to be picked up since it possible to read from both ends of the insert. Using the cycle sequencing method (Murray *et al.*, 1989), a double-stranded template can currently be read with good resolution and also permits the length to be extended over 1kb with a set of universal primers.

Some researchers have recommended constructing multiple shotgun libraries from target DNA (Church *et al.*, 1988, Chen *et al.*, 1993) based on the false belief that such a strategy leads to rapid and effective reconstruction of the small sequences into larger ones. However, massive sequence throughput can be successfully obtained with improvement in the construction of a single high-quality library as described above.

### Novel algorithm for Data Assembly

In the shotgun approach, little prior information except that of the vector sequence is given before any two fragment sequences are overlapped. This, sometimes raises problems in computer assembly. If two sequences far apart are highly similar to each other, the data assembly fails due to a false connection between the two. This problem often occurred when sequencing a region with many copies of a short unit of repetitive sequences (Figure 1-6). To avoid such a problem, I searched for effective procedures and assembling algorithms which would have lower probabilities of false sequence connections. In cosmid sequencing for example, 500 sequences from a 500 base fragment are enough for sixfold equivalents of the target assuming the total length to be determined is 40 kb. Taking into consideration the unnecessary sequences in vector region, approximately 300 random plasmid clones from a shotgun library must be picked up and read by the forward primer. Then, the clones whose sequences are identical to the cosmid vector region (actually, approximately 15% of total sequences) should be omitted. The remaining unique set of clones is sequenced from the opposite end using a reverse primer and, repetitive sequences of high frequency, for example the Alu sequences of human, are to be pooled separately. Sequences other than the cosmid vector and pooled sequences are subjected assembly together. The islands in this phase might be separated by large numbers of gaps, but the gaps represent cosmid vector or repeat unit regions. In the last stage of data assembly, the pooled sequences are included to bridge the islands. This two-stage assembling algorithm can reduce false connections with non-identical repetitive copies (Figure 1-7). Another to avoid such problem is to use more stringent parameters in computer assembly. However, this produces many more islands and forces the gaps to be filled with less information. Raw data from an individual run will not always suffice (usually, the final 50 base-sequence out of 500 bases will contain

5% or more errors). In case of an absence of short units of repetitive sequences such as in the yeast genome, the assembling device might be unnecessary. Theoretically, even if final gaps remain, some clones in the high-quality shotgun library mentioned above would supply a unique template for gap closure. The scheme for procedures and the algorithm are summarized in figure 1-8. An example of an actual sequence distribution using the improved shotgun approach is shown in figure 1-9. Most of the sequences obtained in this way are distributed in random manner.

**Table 1-1.** **DNA Sequencing Strategies**

| Strategy | Category | Advantage | Drawback | Applicable range |
|---|---|---|---|---|
| Primer-walking | Direct | Casual utility | Costly<br>Need prior information | <100kb |
| Nested deletion | Direct | Low redundancy | Complex procedure<br>Need prior information | <20kb |
| Shotgun | Random | Concise procedure<br>Massive throughput | High redundancy | No restriction |
| Oligomer-hybridization | Random | No need for subcloning | Error-prone in repetitive sequences | No restriction |

**Table 1-1.** DNA sequencing strategies. Four typical sequencing strategies currently available and the characteristics are shown. The multiplex sequencing method (Church and Kieffer-Higgins, 1988) is not listed because it is a derivative of the shotgun method. All method except oligomer-hybridization require a Sanger's reaction.

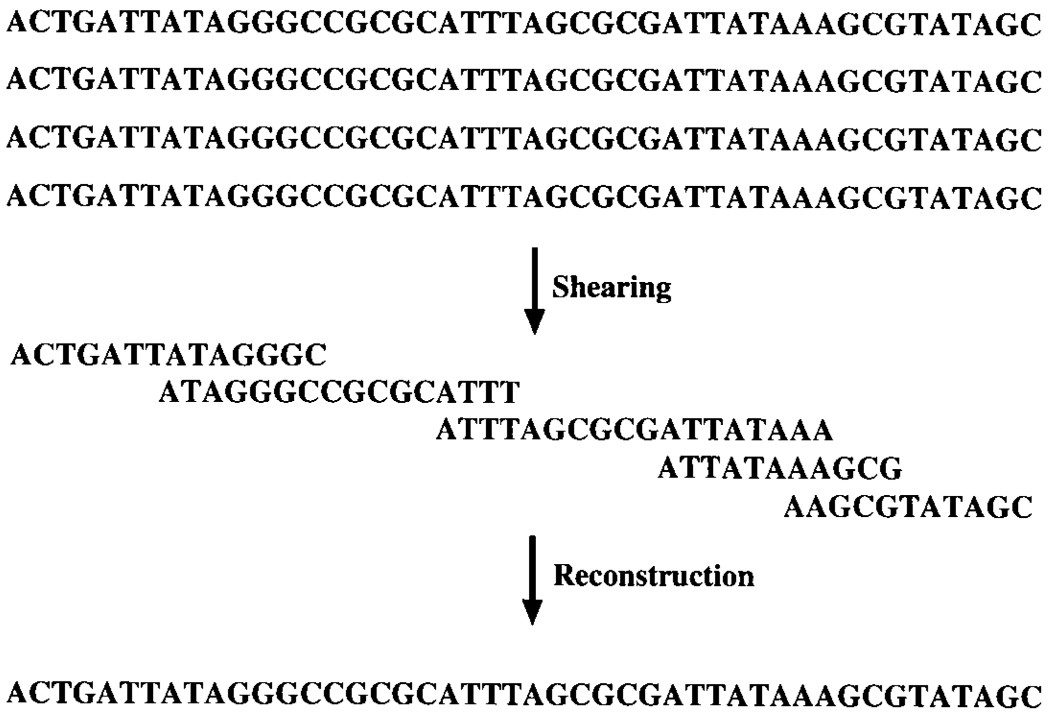
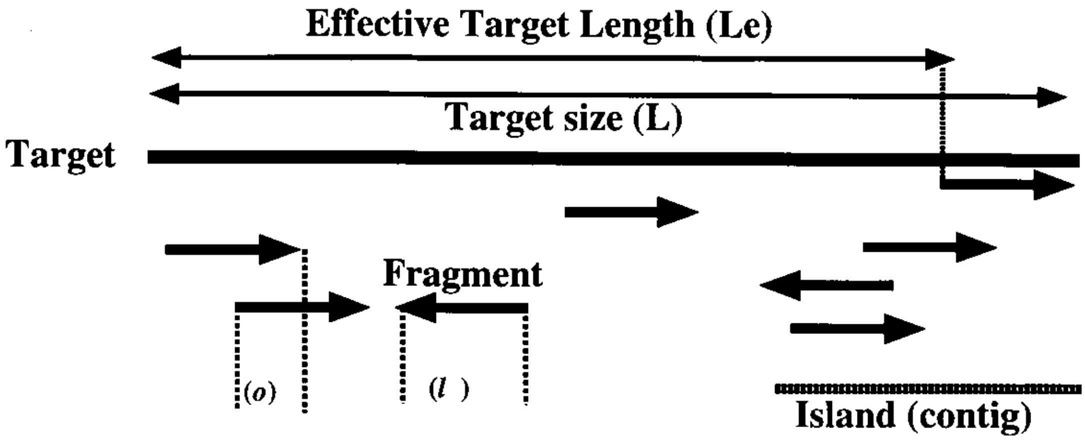

ACTGATTATAGGGCCGCGCATTTAGCGCGATTATAAAGCGTATAGC
ACTGATTATAGGGCCGCGCATTTAGCGCGATTATAAAGCGTATAGC
ACTGATTATAGGGCCGCGCATTTAGCGCGATTATAAAGCGTATAGC
ACTGATTATAGGGCCGCGCATTTAGCGCGATTATAAAGCGTATAGC
Shearing
ACTGATTATAGGGC
ATAGGGCCGCGCATTT
ATTTAGCGCGATTATAAA
ATTATAAAGCG
AAGCGTATAGC
Reconstruction
ACTGATTATAGGGCCGCGCATTTAGCGCGATTATAAAGCGTATAGC

**Figure 1-1.** Scheme of shotgun sequencing approach. A large number of identical but unknown target sequences are randomly fragmented. These fragments are analyzed and aligned based on unique overlapping sequences. When enough fragments have been analyzed, the original target sequences may be deduced. The number of fragments is typically much larger than depicted here.

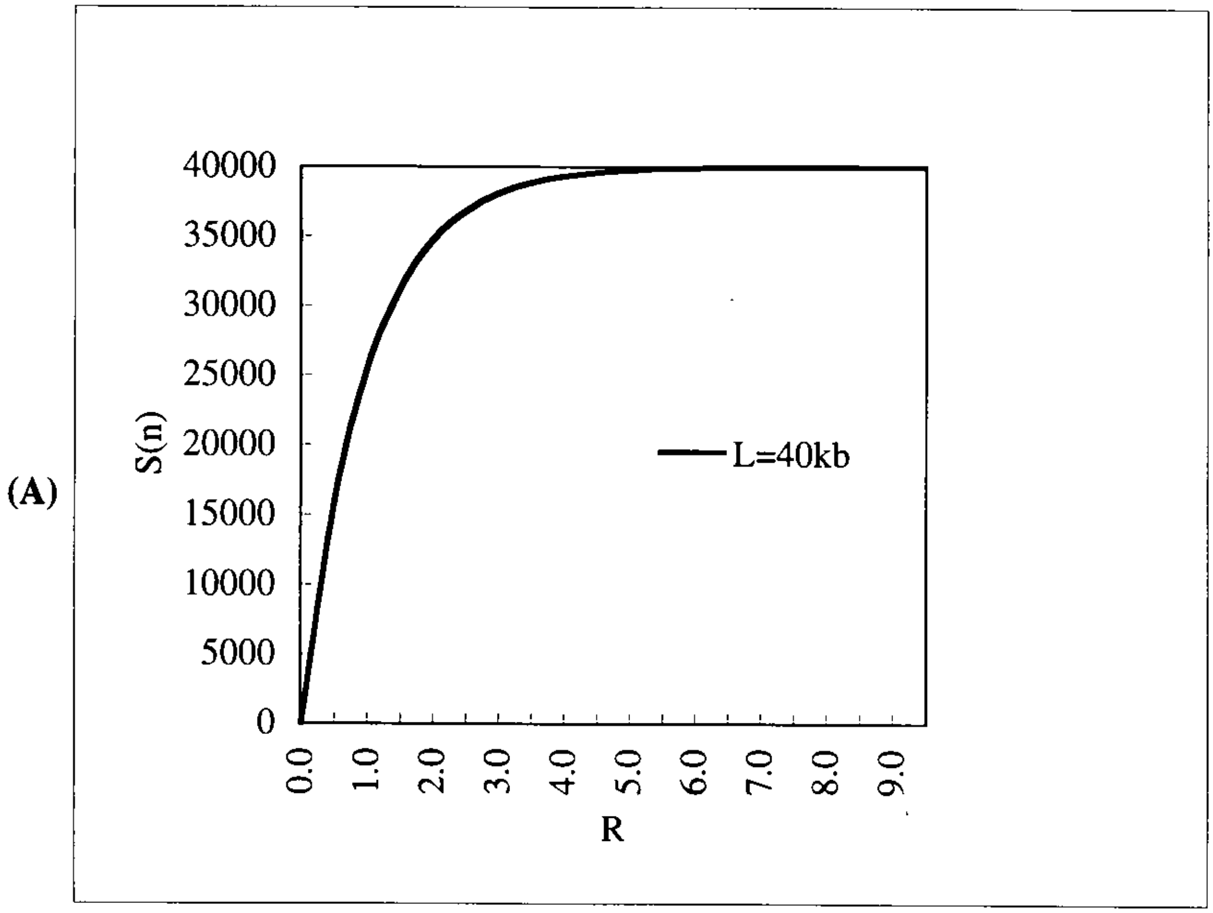


$$S(n) = L[1-\{1-(l-o)/Le\}^{n}] \quad \text{--- (1)}$$

$$Nis = 1+(n-1)\{1-(l-o)/(L-l+1)\}^{n} \quad \text{--- (2)}$$

(When $o = 0$ and $L >> l$, $S(n) = L\{1-(1-l/L)^{n}\}$

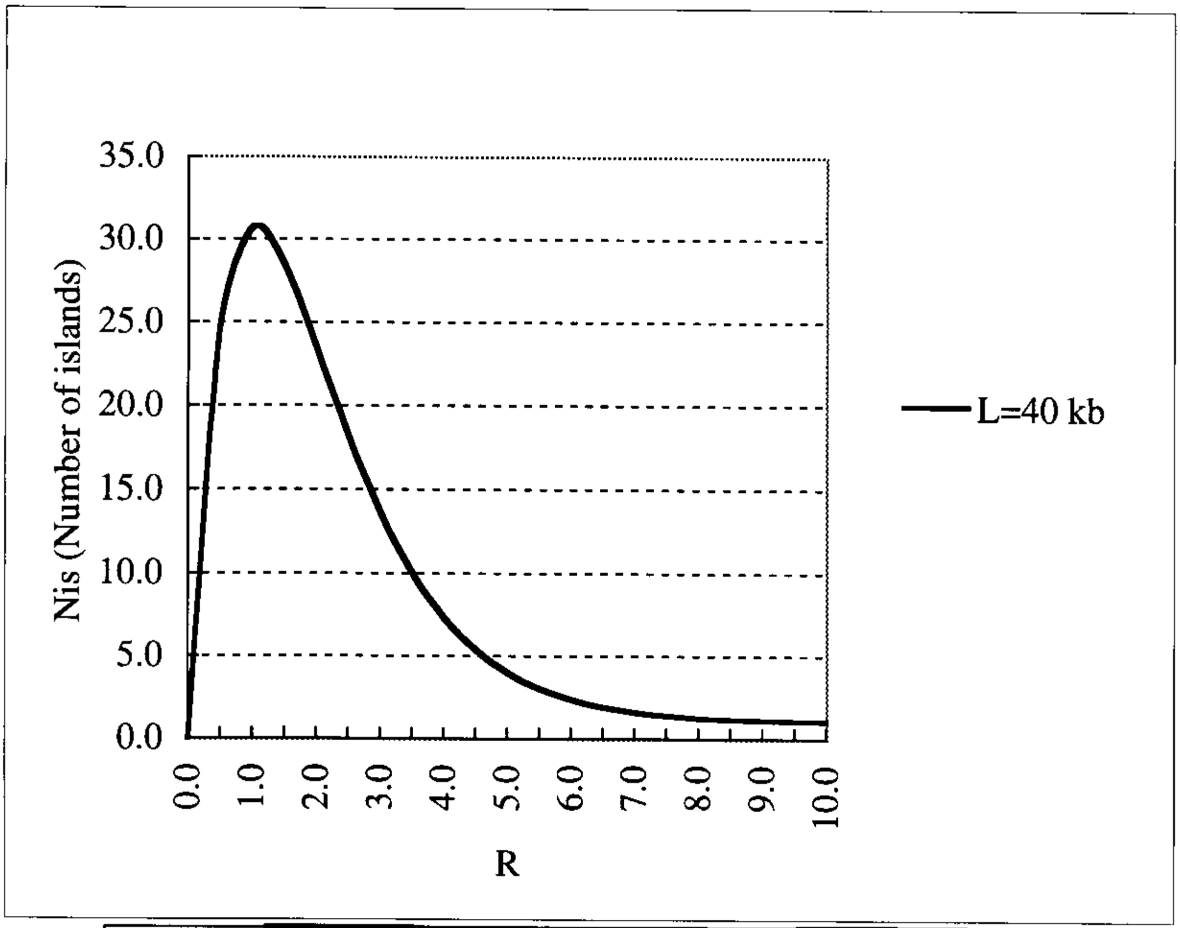
**Figure 1-2.** Mathematical model of sequence acquisition in random sequencing (Roach, 1995). A linear discrete target of length ' L' is assumed. For a given project, 'n' fragments of constant length '$l$' are generated from the target and analyzed in a manner in which overlaps between fragments are detectable. All fragments are generated from distinct identical copies of ' L'. No fragments may start within '$l$-1' bases of the last, right-most base of ' L', as such fragments would not be contained entirely within ' L'. Thus, the effective length ' Le' available for fragment start sites is ' L-$l$+1'. An assumption is made that an overlap of a length of at least '$o$' is necessary and sufficient to detect the adjacency of two fragments. Redundancy, 'R', is defined as 'n$l$/L'. Thus the total accumulated sequence, 'S(n)', is given as, $S(n) = L[1-\{1-(l-o)/Le\}^{n}]$ (equation 1). The expected number of islands 'Nis' is given as, $Nis = 1+(n-1)\{1-(L-o)/(L-l+1)\}^{n}$ (equation 2).



| n | R | S(n) | S(n)/L | Nis |
|---|---|---|---|---|
| 0 | 0.0 | 0 | 0% | 0.00 |
| 80 | 1.0 | 25,377 | 63.443% | 30.71 |
| 160 | 2.0 | 34,654 | 86.636% | 23.48 |
| 240 | 3.0 | 38,046 | 95.115% | 13.71 |
| 320 | 4.0 | 39,286 | 98.214% | 7.38 |
| 400 | 5.0 | 39,739 | 99.347% | 4.00 |
| 480 | 6.0 | 39,905 | 99.761% | 2.35 |
| 560 | 7.0 | 39,965 | 99.913% | 1.59 |
| 640 | 8.0 | 39,987 | 99.968% | 1.26 |
| 720 | 9.0 | 39,995 | 99.988% | 1.11 |
| 800 | 10.0 | 39,998 | 99.996% | 1.05 |

**Figure 1-3.** Theoretical curves associated with sequence acquisition in shotgun cloning (A): Total accumulated sequence S(n) is plotted versus R (redundancy) according to the equation, assuming L=40 (kb) and *l*=0.5 (kb). When *o*=0 and *l* << L, S(n) is approximated to L{1-(1-*l*/L)$^{n}$}. (see also figure 1-2 and the legend). (B): Nis is plotted versus R with the same definition and parameters in figure 1-2 except *o*=20.

(B)



| n | R | S(n) | S(n)/L | Nis |
|---|---|---|---|---|
| 0 | 0.0 | 0 | 0% | 0.00 |
| 80 | 1.0 | 25,377 | 63.443% | 30.71 |
| 160 | 2.0 | 34,654 | 86.636% | 23.48 |
| 240 | 3.0 | 38,046 | 95.115% | 13.71 |
| 320 | 4.0 | 39,286 | 98.214% | 7.38 |
| 400 | 5.0 | 39,739 | 99.347% | 4.00 |
| 480 | 6.0 | 39,905 | 99.761% | 2.35 |
| 560 | 7.0 | 39,965 | 99.913% | 1.59 |
| 640 | 8.0 | 39,987 | 99.968% | 1.26 |
| 720 | 9.0 | 39,995 | 99.988% | 1.11 |
| 800 | 10.0 | 39,998 | 99.996% | 1.05 |

**Figure 1-3.** (continued)

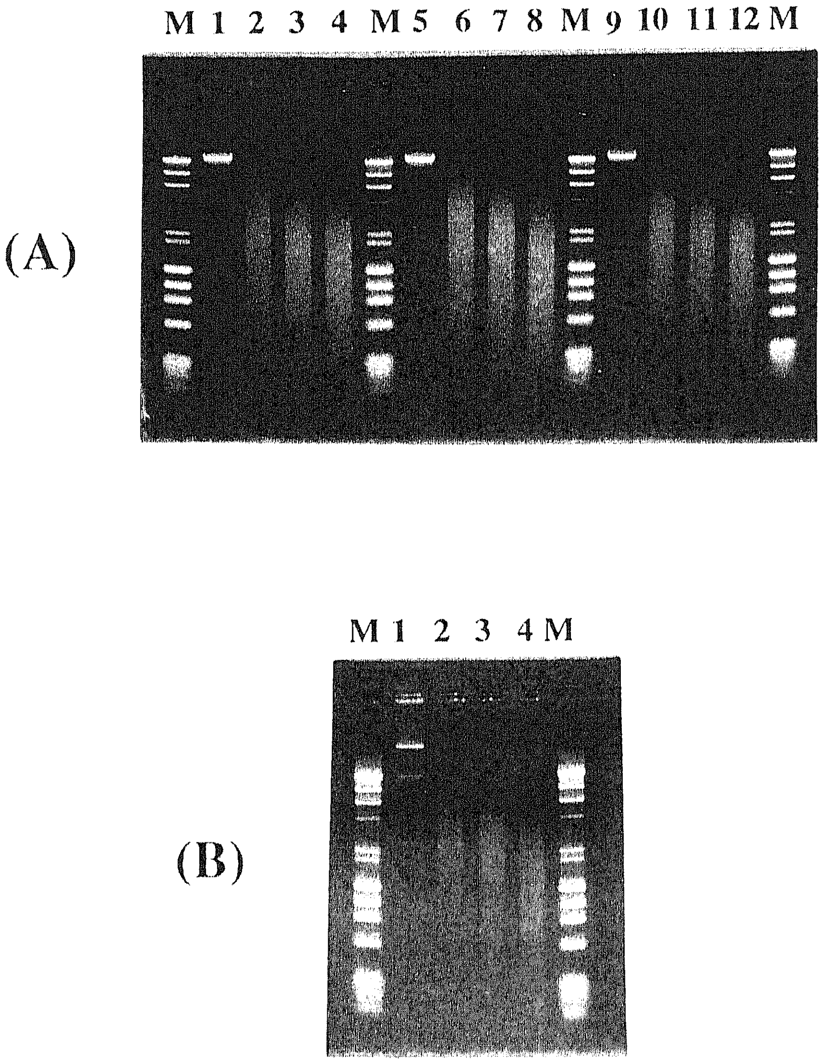


**Figure 1-4.** Size distribution of sonicated DNAs. (A): Linear DNA fragment (48 kb)was sonicated by cup-horn type sonication system (Sonifier-450, Branson) at varied DNA concentrations and sonication times. Sonication was performed at 180w of throughput and 27μm of amplitude in 50μl. M:Size marker λ/*Hind* III + φχ174/*Hae* III. Lanes 1, 5 and 9 : Intact DNA solution at 100ng/μl, 30ng/μl and 10ng/μl. Lanes 2-4 : Sonication products of lane 1 for 3 sec.(lane 2), 6 sec.(lane 3) and 12 sec.(lane 4). Lanes 6-8 and 10-12 are the sonication products of lane 5 and 10 for 3 sec., 6 sec. and 12 sec., respectively. (B): Closed circular DNA (cosmid pM67) was sonicated at 100ng/μl. Lane 1, Intact cosmid DNA. Lanes 2-4, Sonication products of lane 1 for 3 sec., 6 sec. and 12sec., respectively.

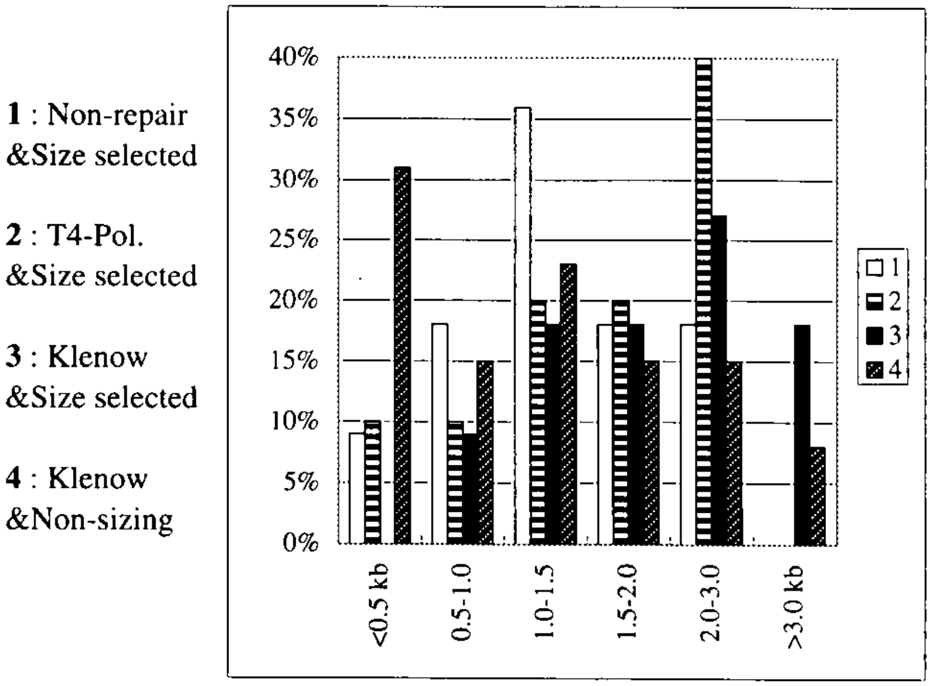
**Table 1-2. End-repair of sonicated fragments and library efficiency**

| | Non-repair | Repaired by T4 Pol. | Repaired by Klonow | Non-sizing & Klenow | No insert Ligation |
|---|---|---|---|---|---|
| Recombinants /Total colonies* | 18/24 | 25/31 | 49/55 | 70/82 | 0/7 |
| Rate of recombinants | 75% | 81% | 89% | 85% | 0% |
| Recombinants /single library | 9.0x$10^3$ | 12.5x$10^3$ | 24.5x$10^3$ | 35.0x$10^3$ | - |
| Recombinants /10ng of insert DNA | 0.60x$10^3$ | 0.83x$10^3$ | 1.63x$10^3$ | 2.33x$10^3$ | - |

***Plating Condition : 10μl of Competent cells (0.9x$10^7$cfu/100μl)**
**1μl/5μl of Ligation Mixture**
**1/10 of Transformation Culture**

**Table 1-2.** End-repair of sonicated fragments and library efficiency. T4 DNA polymerase reaction was performed at 1unit/μg DNA, 37°C for 30 min. Klenow enzyme reaction was done at 1unit/μg DNA, 30°C for 30 min. The reaction was terminated by phenol/chroloform extraction followed by Ethanol precipitation. The ligation reaction was performed in 5 μl of reaction with 66mM Tris-Cl (pH 7.5), 6.6mM $MgCl_2$, 1mM ATP, 150ng of the end-repaired DNA and 50ng of Vector (*Sma* I digested and dephosphorylated pUC 19) DNA. One μl of ligation mixture was transfected into 10 μl of competent DH5α cells and 1/10 volume of the cells was spread onto LB medium containing 100 μg/ml of ampicillin.

| | <0.5 kb | 0.5-1.0 | 1.0-1.5 | 1.5-2.0 | 2.0-3.0 | >3.0 kb | Avarage(kb) |
|---|---|---|---|---|---|---|---|
| 1 | 9% | 18% | 36% | 18% | 18% | 0% | 1.41 |
| 2 | 10% | 10% | 20% | 20% | 40% | 0% | 1.43 |
| 3 | 0% | 9% | 18% | 18% | 27% | 18% | 2.05 |
| 4 | 31% | 15% | 23% | 15% | 15% | 8% | 1.31 |



(A)

**Figure 1-5.** Insert size distribution of shotgun clones obtained by varied end-repairing procedures. Sonicated DNA is repaired to be blunt ended and size-selected passing through Sephacryl S-400 spin-column chromatography (Chromaspin-1000, Clontech). (A), 1: Non-repaired DNA. 2: Repaired by T4 DNA polymerase treatment. 3: Repaired by Klenow enzyme treatment. 4 is negative control of size selection for 3. (B), comparison of the size distribution and average size of clones obtained.

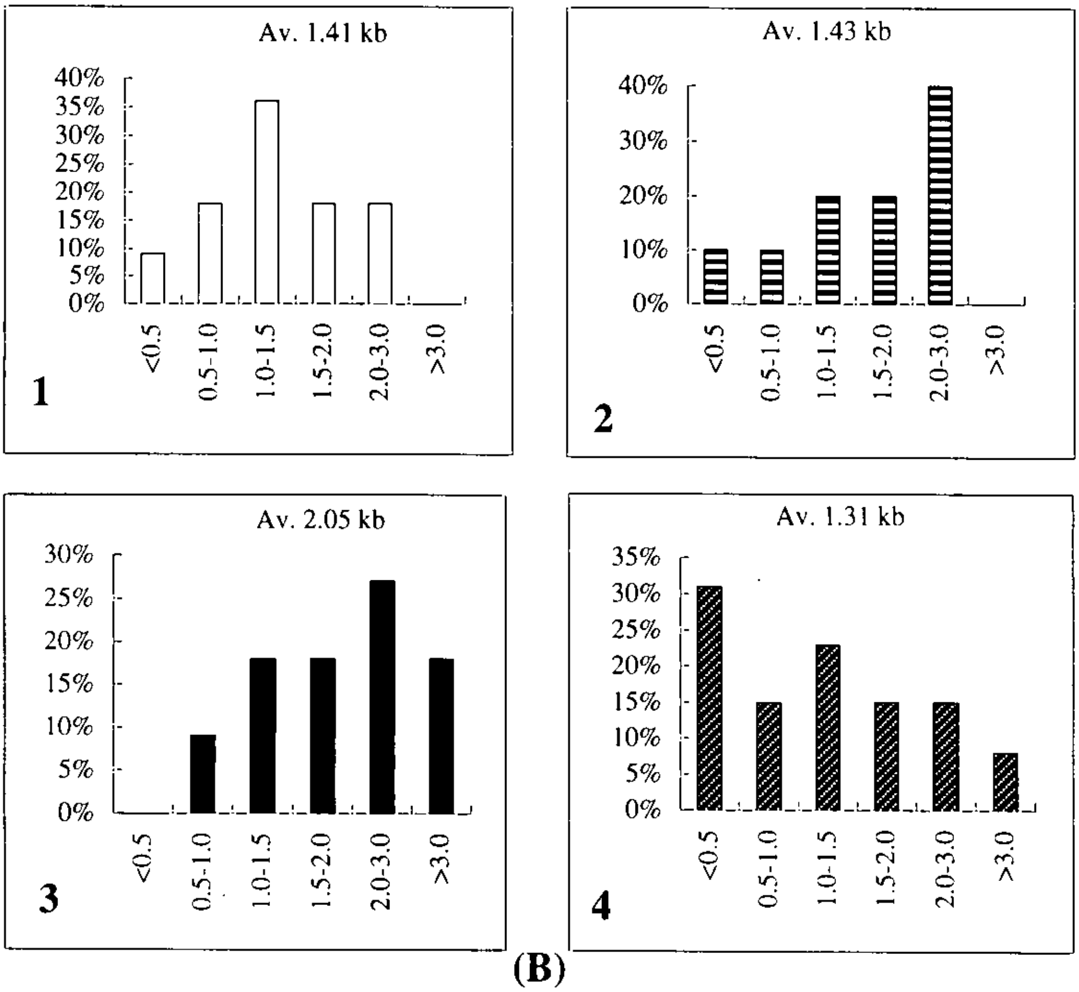


**Figure 1-5.** (continued)

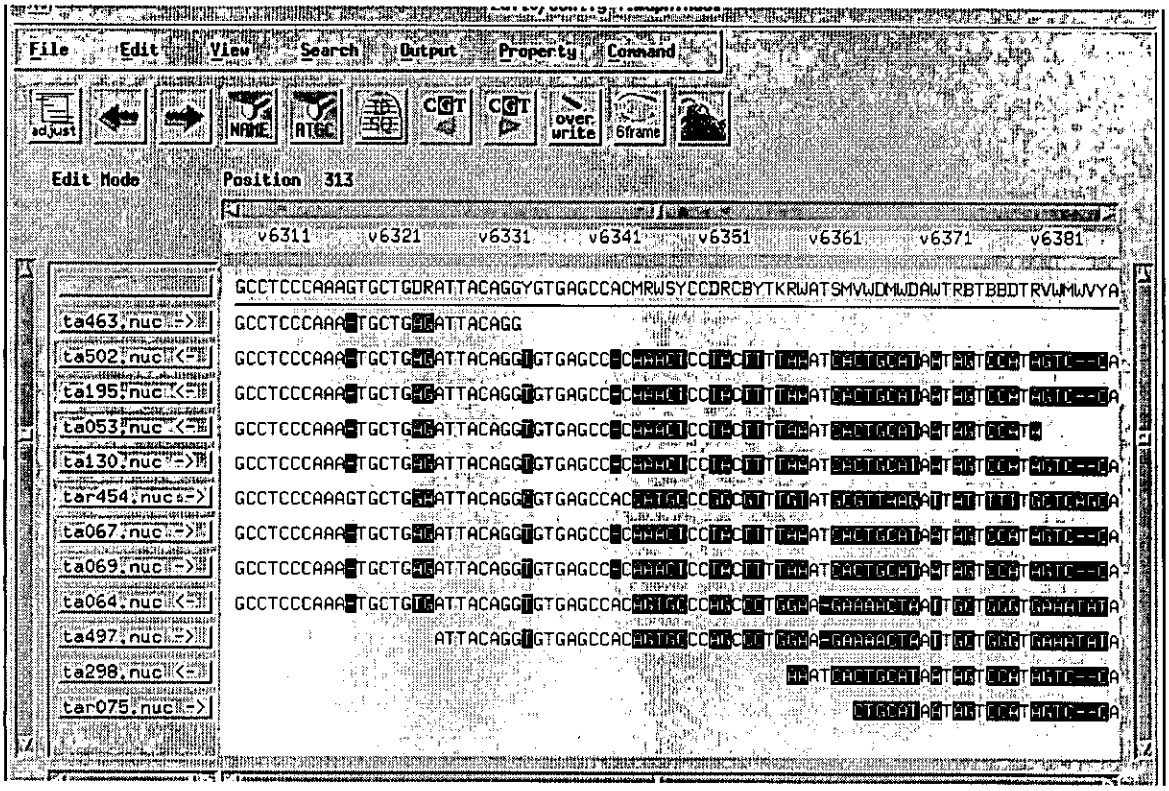
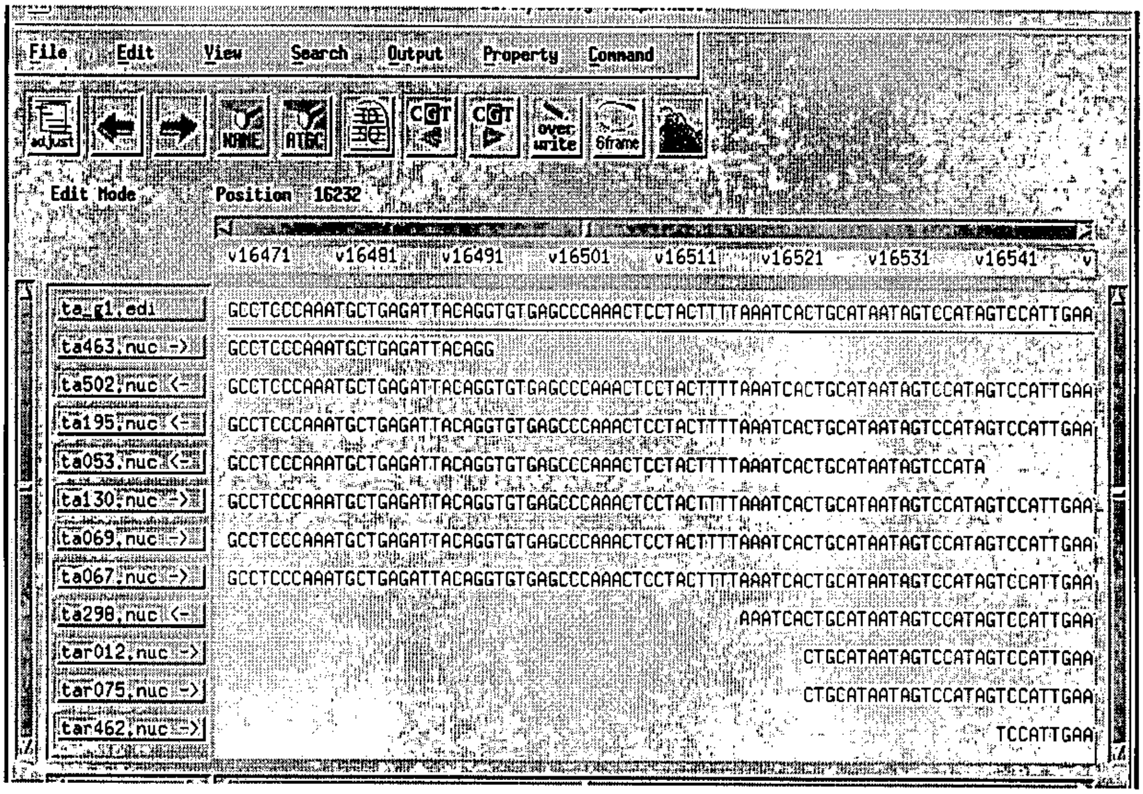


**Figure 1-6.** An example of false alignment in a problem region with an Alu-dense sequence. 648 random sequences obtained by shotgun method from cosmid pM56 were assembled by GENETYX-Σ/SQ (Software Development Corporation, Tokyo). A few sequences containing non-identical Alu repetitive sequence species, tar454, ta064 and ta497 were incorrectly connected on the same region.

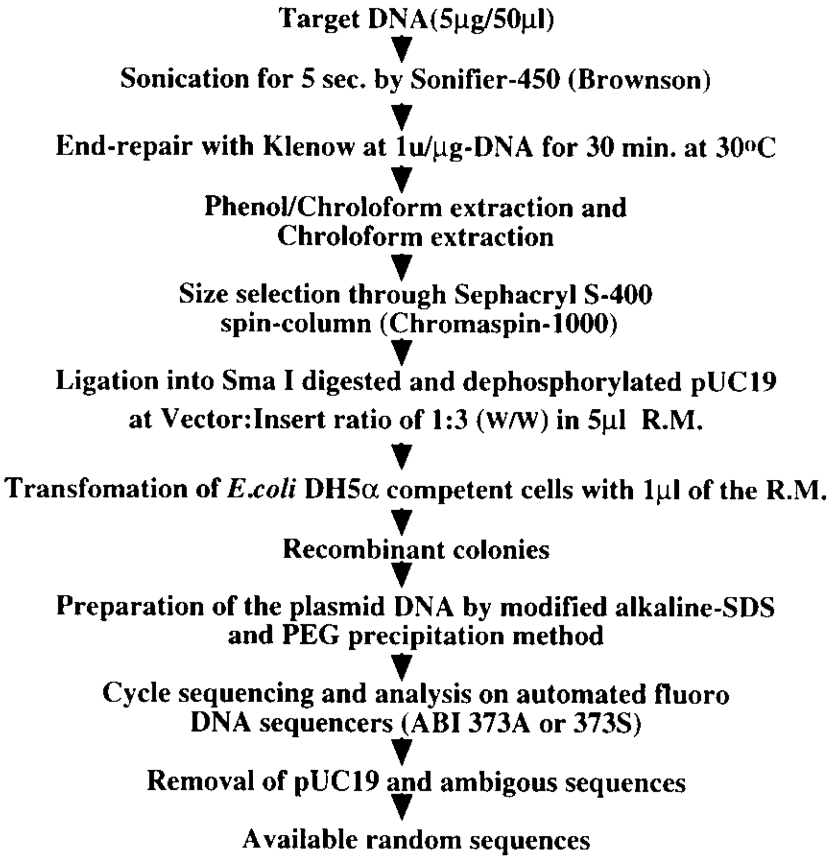


```
Edit Mode      Position 16232

               v16471    v16481    v16491    v16501    v16511    v16521    v16531    v16541    v
ta_g1.edi      GCCTCCCAAATGCTGAGATTACAGGTGTGAGCCCAAACTCCTACTTTTAAATCACTGCATAATAGTCCATAGTCCATTGAA
ta463.nuc ->   GCCTCCCAAATGCTGAGATTACAGG
ta502.nuc <-   GCCTCCCAAATGCTGAGATTACAGGTGTGAGCCCAAACTCCTACTTTTAAATCACTGCATAATAGTCCATAGTCCATTGAA
ta195.nuc <-   GCCTCCCAAATGCTGAGATTACAGGTGTGAGCCCAAACTCCTACTTTTAAATCACTGCATAATAGTCCATAGTCCATTGAA
ta053.nuc <-   GCCTCCCAAATGCTGAGATTACAGGTGTGAGCCCAAACTCCTACTTTTAAATCACTGCATAATAGTCCATA
ta130.nuc ->   GCCTCCCAAATGCTGAGATTACAGGTGTGAGCCCAAACTCCTACTTTTAAATCACTGCATAATAGTCCATAGTCCATTGAA
ta069.nuc ->   GCCTCCCAAATGCTGAGATTACAGGTGTGAGCCCAAACTCCTACTTTTAAATCACTGCATAATAGTCCATAGTCCATTGAA
ta067.nuc ->   GCCTCCCAAATGCTGAGATTACAGGTGTGAGCCCAAACTCCTACTTTTAAATCACTGCATAATAGTCCATAGTCCATTGAA
ta298.nuc <-                                                   AAATCACTGCATAATAGTCCATAGTCCATTGAA
tar012.nuc ->                                                         CTGCATAATAGTCCATAGTCCATTGAA
tar075.nuc ->                                                         CTGCATAATAGTCCATAGTCCATTGAA
tar462.nuc ->                                                                           TCCATTGAA
```

**Figure 1-7.** An example of successful data assembling on the region indicated in figure 1-6. Two stage data assembling algorithm using the same software (GENETYX-Σ/SQ) works well to connect identical Alu sequences on the proper site in pM56 (see also figure 1-6).

## (A) Procedure for high-quality shotgun library construction



Target DNA(5μg/50μl)

Sonication for 5 sec. by Sonifier-450 (Brownson)

End-repair with Klenow at 1u/μg-DNA for 30 min. at 30°C

Phenol/Chroloform extraction and
Chroloform extraction

Size selection through Sephacryl S-400
spin-column (Chromaspin-1000)

Ligation into Sma I digested and dephosphorylated pUC19
at Vector:Insert ratio of 1:3 (W/W) in 5μl R.M.

Transfomation of *E.coli* DH5α competent cells with 1μl of the R.M.

Recombinant colonies

Preparation of the plasmid DNA by modified alkaline-SDS
and PEG precipitation method

Cycle sequencing and analysis on automated fluoro
DNA sequencers (ABI 373A or 373S)

Removal of pUC19 and ambigous sequences

Available random sequences

**Figure 1-8.** Improved shotgun sequencing procedure. (A): Flow-chart of high-quality shotgun library construction. From the single library according to the procedure, more than twenty thousand recombinants (average 2 kb insert) were obtained which are sufficient to determine 40 cosmid clone equivalents (> 1Mb in total length). (B): A novel and practical algorithm to assemble accurately a large number of fragment sequences. Two stage assembling algorithm is shown. Alu-dense or tandem Alu regions can be assembled in high fidelity by the algorithm (see figure 1-6, 1-7)

## (B) Novel algorithm for sequence data assembling

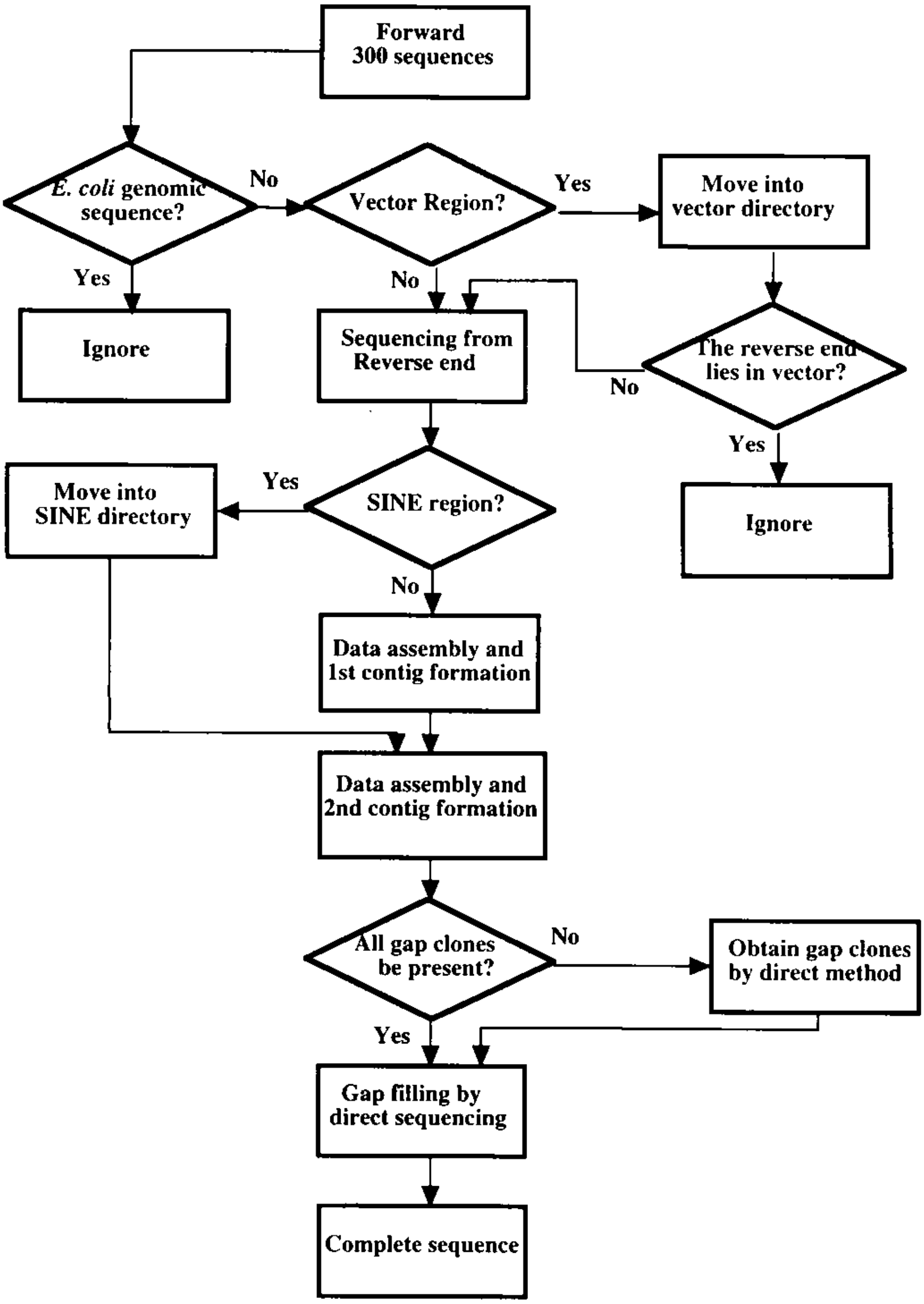

Forward 300 sequences
*E. coli* genomic sequence?
No
Yes
Ignore
Vector Region?
Yes
No
Move into vector directory
The reverse end lies in vector?
No
Yes
Ignore
Sequencing from Reverse end
SINE region?
Yes
No
Move into SINE directory
Data assembly and 1st contig formation
Data assembly and 2nd contig formation
All gap clones be present?
No
Yes
Obtain gap clones by direct method
Gap filling by direct sequencing
Complete sequence

**Figure 1-8.** (continued)

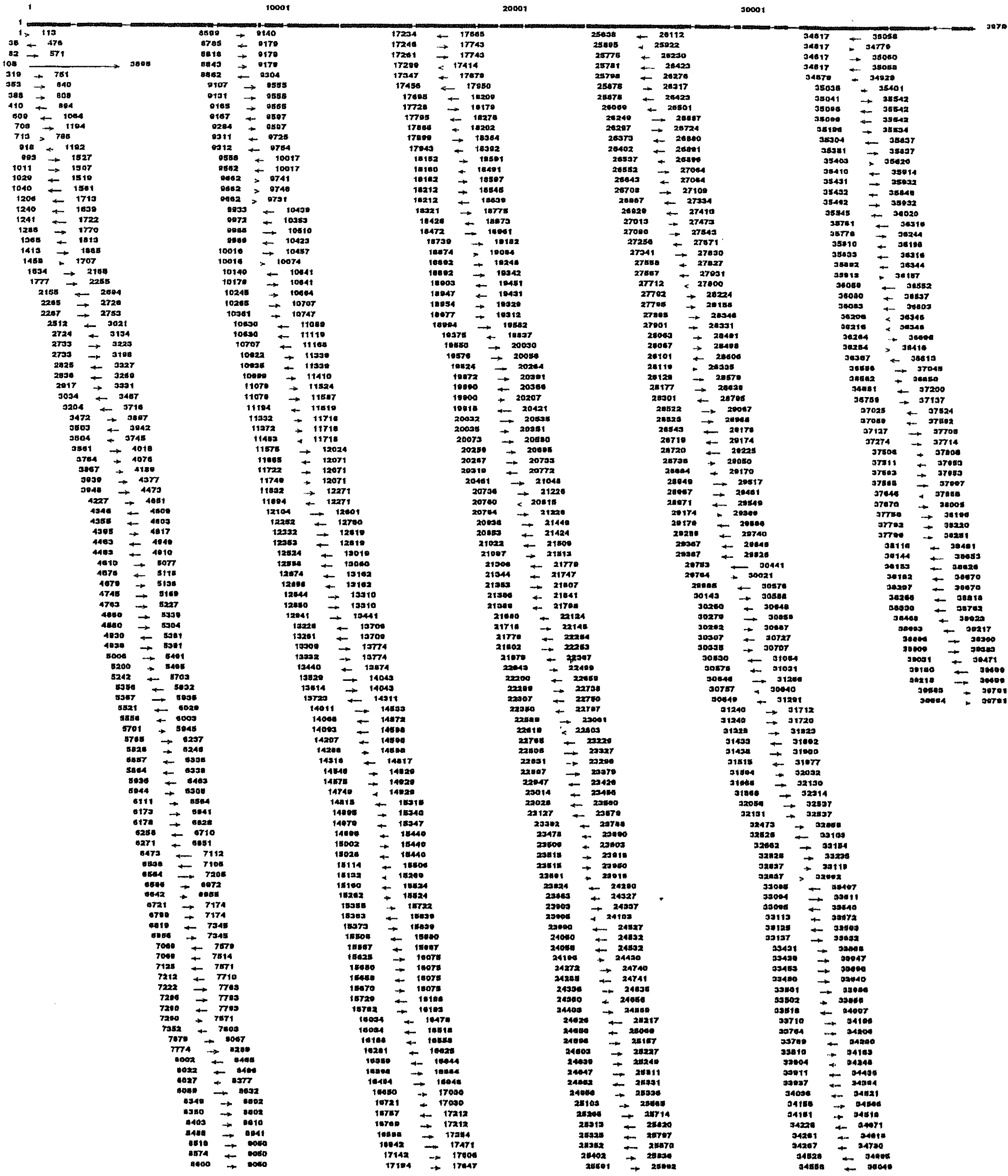


**Figure 1-9.** Result of sequence assembly obtained from the improved shotgun approach (cosmid pM213-5; Insert size 39,791bp). Numbered arrows indicate regions sequenced with their orientations. Each sequence obtained in this approach appeared to exist in a random manner. A 3.8 kb-sequence (1-3896) shows overlapping region with adjacent cosmid clone, pM67.

## Summary

For the large-scale sequencing of ordered sets of genomic clones from eukaryotes, several improvements in the shotgun approach were achieved : (i) Introduction of a cup-horn type sonication system with high-throughput capable of rapid shearing to allow sufficient fragmentation to desired size ; (ii) Optimization involving end-blunting and sizing procedures using the Klenow enzyme in a short time treatment and spin-column chromatography on sephacryl S-400 ; (iii) Construction of a shotgun library based on the estimation from a mathematical theory of random subcloning ; (iv) Double-stranded sequencing which can reduce the number of clones to be sequenced and can provide successful gap filling at the final phase and ; (v) A unique assembling algorithm (two-stage assembly) to separate repetitive sequences which are to be assembled at the final stage so as not to produce a false connections.

These improvements made it possible to produce a high-quality shotgun library of accurately collected random subclones and an accurate reconstruction of small sequences into an entire whole.

# Part 2.

# Nucleotide sequencing and analysis of the entire chromosomeVI from *Saccharomyces cerevisiae*

# Yeast chromosome VI

The budding yeast Saccharomyces cerevisiae is an important model organism for basic biological processes of higher eukaryotes. Although the yeast genome is relatively small (16 chromosomes totaling a 13Mb genome size ; Olson *et al.*, 1991a,b), its molecular mechanisms for control of cellular growth, DNA replication, transcription, signal transduction and DNA repair are thought to be similar to those of higher eukaryotes. Since the density of the coding region is relatively high and an ordered set of cosmid clones has already been aligned on most of the chromosomes, genome sequencing of this organism has been carried out by a unique international collaboration. Six reports have revealed that sequencing yeast chromosomes is a very efficient procedure for finding novel genes because two-thirds of the yeast genes identified through these studies have not been previously identified (Oliver *et al.*, 1992; Dujon *et al.*, 1994; Johnston *et al.*, 1994; Feldmann *et al.*, 1994; Murakami *et a*l., 1995; Bussey *et al.*, 1995). Complete sequencing the yeast genome greatly facilitated our understanding of yeast chromosome organization (Goffeau *et al.*, 1997). Here I revisit DNA sequencing and analysis of yeast chromosome VI (270kb) as the fruit of the improved shotgun sequencing approach mentioned in part 1 and, describe 92 novel genes identified through the analysis. The structural features of chromosome VI will be described including, G+C composition, gene density, and distribution of ARS elements.

Among the 16 yeast chromosomes, this chromosome was the only one which had been subcloned in its entirety into plasmids that were tested for ARS activity. As a result, eight active ARS elements were isolated (ARS 8 was further divided into two elements) and the loci were mapped (Shirahige *et al.*, 1993) prior to sequencing of the entire chromosome. In addition, the local DNA sequence analysis of the elements revealed two common features : the presence of an 11-bp consensus sequence (core sequence or domain A) : 5'-(A/T)TTTA(C/T)(A/G)TTT(A/T)-3' and the presence of a domain which has a higher A+T content than bulk yeast DNA usually found 3' to the core consensus sequence (Shirahige *et al.*, 1993). The construction and analysis of the effects of point mutations, small deletions, and small substitutions within this consensus sequence have demonstrated that it is essential for ARS function. Although ARS elements do not necessarily have to contain a perfect consensus, all the elements should have at least one element which has a 10/11 match with the consensus sequence (Deshpande *et al.*, 1992). In chromosome VI, 373 elements which fulfilled this criterion were found. Among these 373 elements, sixteen were 100% identical to the consensus sequence. To investigate the mechanism by which active ARS elements are selected from such numerous consensus loci, complete sequencing of the chromosome and systematic analysis of the sequence motifs around the consensus sequences were carried out.

# Materials and Methods

## Materials

*Escherichia coli* strain DH5α (supE44 ΔlacU169(f80lacZΔM15) hsdR17 recA1 endA1 gyrA96 thi-1relA1) was used for all subcloning and sequencing steps. Lambda clones and cosmid clones of yeast chromosome VI shown in figure 2-1 were isolated and mapped by Olsons *et al.* (1993). Gap A, Gap B and Gap C clones were isolated by Iwasaki *et al.* (1992) These clones were kindly supplied by Dr. M. Olson (Washington University) and H. Yoshikawa (Nara Institute of Science and Technology). A plasmid clone containing the right telomere of this chromosome, pEL174, was kindly supplied by E. Louis (John Radcliffe Hospital).

## Methods

*Preparation of shotgun libraries*

The insert DNA of the phage clones (clone 3193, 4121, 6781, 3068, 6552, 4682, 4233 and 4231 were isolated and recloned into the Charomid 9-28 vector (Saito *et al.*, 1986) at the *Sma* I site. The cosmid clones (clone 9993, 9765 and 9965) were directory sonicated to construct the shotgun libraries. The charomid DNA, plasmid DNA and cosmid DNA were purified by the alkaline lysis method following the CsCl ultra-centrifigation. One to several hundred micrograms of purified DNA was obtained from an overnight culture (500ml) grown in CircleGrow medium (Funakoshi Co.). The purified DNA was subjected to sonication, size fractionation, and treatment with Klenow enzyme to generate repaired blunt ends according to the improved method as described in part 1. The blunt ended DNA was then ligated into *Sma* I digested, dephosphorylated pUC19 vector DNA at an insert to vector molar ratio of 5:1. The ligation mixture was then transfected into competent DH5α cells prepared as described by Hanahan *et al.* (1983). Several thousand recombinants were regularly obtained from ten micrograms of target DNA. The ratio of clones which had yeast DNA inserts was as high as 90% in a typical preparation.

*Preparation of the sequencing templates*

Plasmid DNA was purified from the overnight culture by the alkaline lysis method followed by the Polyethyleneglycol precipitation method (Maniatis *et al.*, 1989) or using an automated plasmid DNA purifier (Model PI-100, Kurabo, Kurashiki). 80 to 192 random clones were purified into plasmid DNA in a day. On average, approximately 30μg of purified plasmid DNA were obtained per 5ml of overnight culture under the conditions. This was enough for up to ten sequencing reactions.

*DNA sequencing*

Sequencing reactions were performed using 1μg of double-stranded plasmid DNA. The DNA sequences were determined using an Amplitaq polymerase dye primer cycle sequencing method (Craxton *et al.*, 1991) and were analyzed using an ABI 373A automated fluoro-sequencers (Perkin Elumer).

Oligonucleotide primers were synthesized on a DNA synthesizer (ABI 394 DNA synthesizer ) and purified with OPC columns according to the manufacturer's protocol. Dye terminator cycle sequencing reactions were carried out to fill the gaps of the contigs which had been assembled using the software, Shotgun (Mitsui Knowledge Inc., Tokyo) or ATSQ (Software development Co., Tokyo). To determine the order of the contigs, sequence reactions were carried out in each direction with the universal, forward (-21M13) and reverse (M13RP1) primers flanking the cloning site.

*Sequence data assembly*

Raw sequence data were transmitted to Unix workstations (Sun 4/10 or SPARCstation 10) through the TCP/IP protocol and were assembled using the software, Shotgun (Mitsui Knowledge Inc., Tokyo) or ATSQ (Software development Co., Tokyo). Determination of the final sequence was performed by comparing the chart obtained from the ABI 373A sequencer and the results of the assembled sequence data. All bases were covered by more than five fragments. The entire chromosome was fully covered by sequence data, the charts from the sequencer were carefully compared and subclones were reanalyzed or analyzed with synthetic primers from alternative sequencing start positions. To verify the sequence data, comparison of the sequence between overlapping clones were carried out. There were only four bases discrepancies out of 20 kb of data. These conflicts were resolved by re-sequencing. Although the overall accuracy of the sequence determined was estimated at 99.98%, the author's group will continue to revise the sequence for chromosome VI through DDBJ.

*Analysis of the sequence and database submission*

The final sequence was initially analyzed using GENETYX software package (Software development Co., Tokyo) on Macintosh computers. ORFs of longer than 300 bases were analyzed for similarities with known sequences in the GenBank, EMBL, PIR and SwissProt batabanks.

The sequence data reported has been submitted to DDBJ/EMBL/GenBank data libraries under accession numbers, D50617 (270 kb full sequence), D44603 (clone 9993), D44594 (clone 3193), D44598 (clone 4121), D44595 (clone 6781), D44601 (clone GapA), D44596 (clone 3068), D31600 (clone 6552), D44604 (clone GapB), D44600 (clone 4682), D44599 (clone 4233 and 4231), D44606 (clone GapC), D44602 (clone 9765) and D44597 (clone 9965), respectively.

# Results

## Sequencing and physical map

To determine the complete nucleotide sequence of yeast chromosome VI, an ordered set of phage, plasmid and cosmid clones was used (Figure 2-1). Most regions were covered by phage and cosmid clones (Olson *et al.*, 1993) ; the gaps between phage contigs were filled with plasmid clones (Iwasaki *et al.*, 1992). Due to the improved shotgun method (see also part 1), two sets of sequence data were easily obtained from both sides of the plasmid vector. Gaps found after random sequencing were filled by primer walking on clones which linked two contigs. The complete nucleotide sequence was 270,148 bp by both strand sequencing and all bases were read at least five times for determination.

The restriction map obtained from the sequence data was compared with a one previously reported (Riles *et al.*, 1993). The *Eco* RI and *Hind* III restriction sites in both maps matched with the exception of one fragment located approximately 170 kb from the left end (Figure 2-2). Reinvestigation of the local restriction map indicated that the value obtained in this sequencing project is the correct size ; therefore the 3373-bp fragment should be corrected to 4645 bp.

## Gene Identification

The distribution of open reading frames (ORFs) on the chromosome is illustrated in figure 2-2. The sequence contained 129 non-overlapping ORFs greater than 300 bp. The density of ORFs calculated from this number was 0.48 per kb. This value is consistent with those obtained for yeast chromosomes previously reported (Oliver *et al.*, 1992, Dujon *et al.*, 1994, Johnston *et al.*, 1994, Feldmann *et al.*, 1994, Bussey *et al.*, 1995). The average size of the genes identified was 476 codons, and the longest ORF (YFR 019W) spanned 6834 bp (2278 codons).

A similarity search among genes identified in this project and those in the public databanks (GenBank, EMBL, PIR, and SwissProt)are summarized in table 2-1. Genes having FASTA optimum score higher than 200 were regarded as highly homologous. Among the 129 ORFs identified here, 37 (28%) are identical to previously identified genes. Of the remaining 92 novel ORFs, 39 (30%) were highly similar to known genes in yeast or other organisms. One half of the ORFs (53 out of 129) appear to encode proteins that have no similarity to known sequences.

Comparative analysis of the genetic map (Mortimer *et al.*, 1993) and the physical map newly constructed by the sequence revealed two inversions, one between *sup* 11 and *suf* 20 (figure 2-3). Two genes, *pho* 4 and *cdc* 26, were very close on the physical map while they were distant on the genetic map.

## Base composition of chromosome VI

As observed in previously sequenced chromosomes, base composition was clearly not uniform along chromosome VI (Figure 2-4b,c). The C+G composition of the central domain (108-173 kb) was significantly lower than that at the ends (figure 2-4a). Alternation of window sizes between 10 kb and 50 kb did not significantly change the pattern of C+G composition (data not shown). In chromosome II (Feldmann *et al.*, 1994) and XI (Dujon *et al.*, 1994), high gene density was observed predominantly in regions where the C+G composition was higher than average; however, no correlation between high C+G content and high gene density was observed in this chromosome (Figure 2-4d). On the contrary, the central A+T-rich region exhibited a relatively high gene density. In addition, no such correlation between high G+C content and gene density was observed in the recently sequenced chromosome VIII (Johnston *et al.*, 1994), thus the phenomenon observed in chromosome II and XI is not universally true of all yeast chromosomes. Of further note, the top strand is preferentially utilized for encoding genes in the A+T-rich central domain of the chromosome (Figure 2-4b,c). The gene density of the right end of the chromosome was significantly lower than that in other regions. The left end of the chromosome contained many small ORFs.(figure 2-4d). A similar observation has been reported in the recently sequenced chromosome I (Bussey *et al.*, 1995). ; the gene density of both ends of the chromosome was significantly lower than that of the central region.

## Distribution of ARS elements

Chromosome VI had been previously subcloned, and all the subclones had been assayed for their ability to replicate autonomously (assayed for active ARS elements, which are candidates of the chromosomal replication origin). As the result of the complete sequencing in the present study, those elements were revealed to be distributed in 30 kb intervals on average. Using the complete sequence of chromosome VI, a systematic comparison was made between the features of active ARS element sequences and inactive ones that had the complete core sequence. A mutational analysis of the ARS 307 consensus sequences and quantitive analysis of the consensus of HO (HO gene:mating-type interconversion endonuclease gene)(Kipling *et al.*, 1990) had indicated that the ARS core sequence should be as follows : 5'-(A/T)TTTA(T/C)(A/G)TTT(A/T)(T/C/G)-3'. All active ARS elements mapped to regions with lower than average G+C content with the exception of ARSS1, located near the right telomare (Figure 2-5). When one base mismatch with the core sequence was allowed, ARS consensus sequences were found all over the chromosome at a density of one per 800 bp (Murakami, Yamazaki *et al.*, 1995). To determine the *cis*-factor governing ARS activity, the 3' flanking regions of active and inactive ARS elements were examined. Eleven inactive ARS elements were found which contained perfect matches with the above core sequence (table 2-2). All but one active element (ARSS6) had an additional ARS-like consensus sequence in the 3' flanking region, whereas nine of eleven inactive loci had no additional ARS-like consensus sequences. Further analysis of the distribution of nuclear scaffold binding site consensus sequences in the region adjacent to the ARS core sequences (Amali *et al.*, 1988, Hoffman *et al.*, 1989) did not locate any prominent differences in the distribution pattern of nuclear scaffold binding domains between active and inactive loci. Some additional *cis*-elements including transcription factor ABF1 binding sites (Shore *et al.*, 1987, Diffley *et al.*, 1988), and topoisomerase (topo) II cleavage sites (Spitzner *et al.*, 1988) were also found in this region. Interestingly, no active ARS element has an ABF1 binding site, and there are more Topo II cleavage sites in the 3' flanking region of inactive ARS elements.

The relationship between the position of ARS core sequences and the distribution of ORFs were also investigated. All but one active ARS element (ARSS5) were mapped in non-coding regions while five of eleven inactive loci were mapped in coding regions. This suggests that transcriptional regulation may play some role in the activation of ARS core sequences.

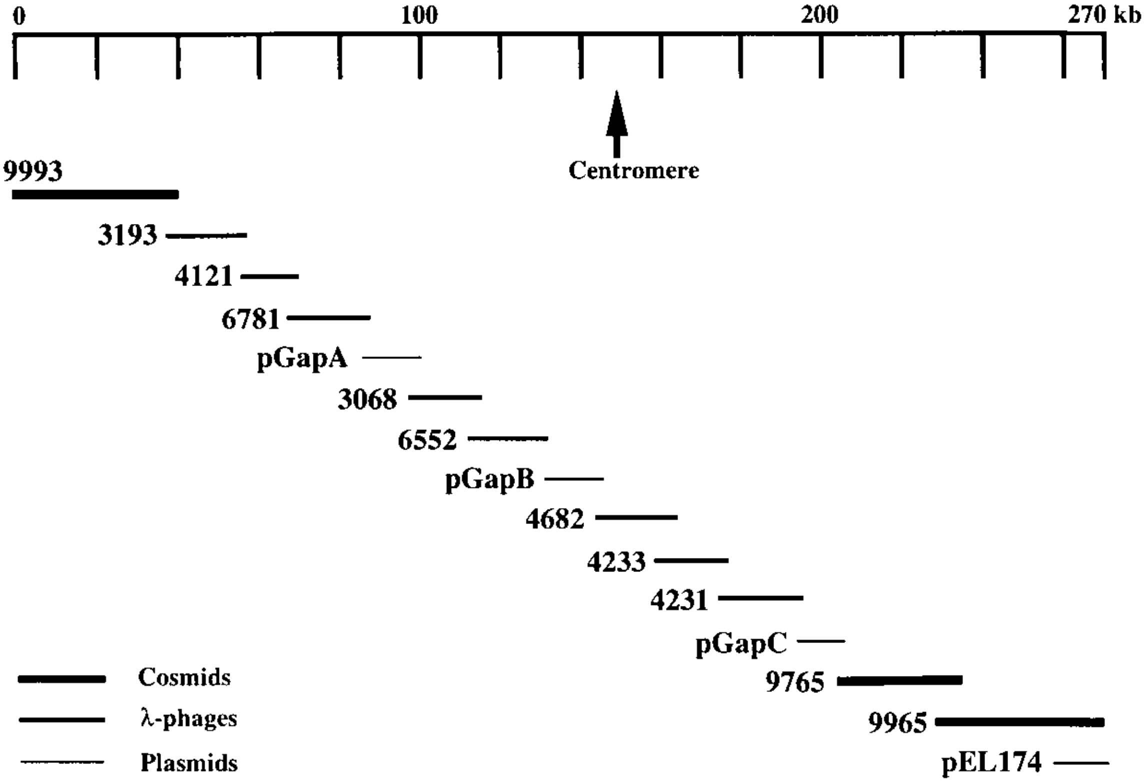


**Figure 2-1.** Schematic representation of the lambda phage, plasmid and cosmid clones sequenced in this study. The nucleotide sequence data in this project will appear in the GSDB, EMBL, NCBI and DDBJ sequence databases with the accession number D50617.

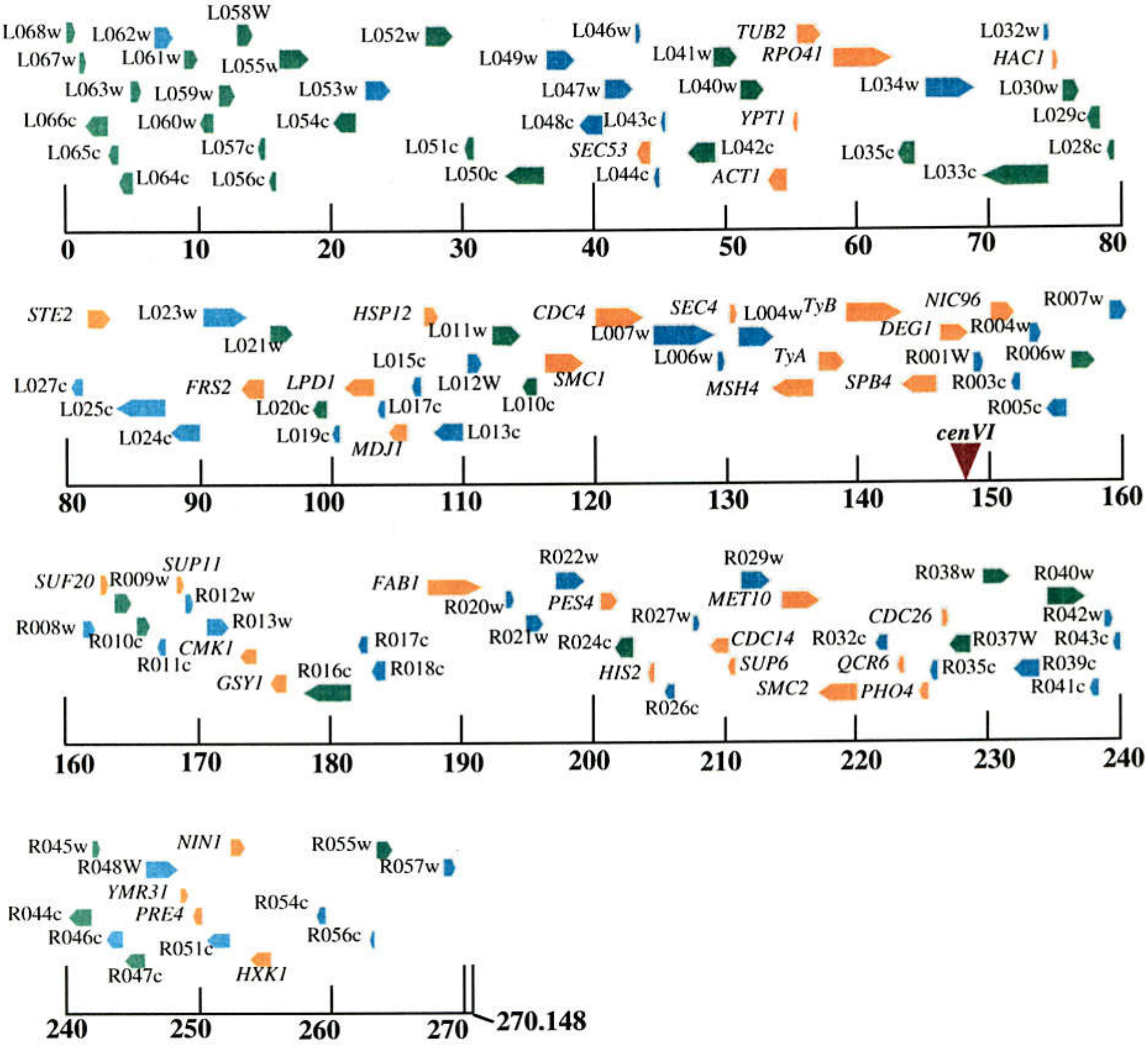


**Figure 2-2.** Distribution of ORFs on Yeast chromosome VI. All ORFs larger than 300 bases are indicated. Arrows represent ORFs or ARS elements. Orange arrows indicate previously identified genes, green arrows indicate ORFs highly homologous to database entries in SwissProt, PIR, GenBank or EMBL. ORFs of which FASTA optimum scores were larger than 200 were regarded as having a high degree of homology. Blue arrows indicate ORFs with no detectable similarity to sequences in the databases.

Table 2-1. List of genes and features of chromosome VI

| Position | ORF ID | Locus | Function or Homology | FASTA score | Acc. no. | Data base |
|---|---|---|---|---|---|---|
| 1 | Y' element | | Y' subtelomeric repeat | | | S |
| 4685 | telomere | | Telomeric repeat (C(1-3)A) | | | S |
| 4823 | X element | | X subtelomeriv repeat | | | S |
| 836 | YFL067w | | Period clock protein(fragment) | 202 | P08399 | S |
| 2615 | YFL066c | | Hypothetical 137.7 kd protein in Y' repeat region | 2371 | P24089 | S |
| 3338 | YFL065c | | General amino acid permerase | 587 | P24088 | S |
| 3846 | YFL064c | | Hypothetical 31.5 kd protein in CBP2 5' region | 751 | P24088 | S |
| 5066 | YFL063w | | Hypothetical 13.3 kd protein in URA1 5' region | 383 | P36030 | S |
| 6424 | YFL062w | | Hypothetical 45.2 kd protein in NAL3S 3' region | 179 | P25354 | S |
| 9545 | YFL061w | | Cyanamide hydratase (EC 4.2.1.69)(urea hydrolyase) | 303 | P22143 | S |
| 10969 | YFL060c | | Hypothetical 21.4 kd protein in DACA-SERS intergenic region | 233 | P37528 | S |
| 11363 | YFL059w | | Hypothetical 31.6 kd protein in DACA-SERS intergenic region | 803 | P37527 | S |
| 12929 | YFL058w | | No message in thiamine protein1 | 1191 | P36597 | S |
| 14763 | YFL057c | | Hypothetical 40.9 kd protein in HMR 3' region | 809 | P25612 | S |
| 15431 | YFL056c | | Hypothetical 40.9 kd protein in HMR 3' region | 687 | P25612 | S |
| 17004 | YFL055w | | General amino acid permerase | 895 | P19145 | S |
| 22787 | YFL054c | | Glycerol uptake facilitator protein | 426 | P11244 | S |
| 28232 | YFL052w | | Maltose fermentation regulatory protein MAL6R | 1904 | P10508 | S |
| 30540 | YFL051c | | Hypothetical 122.2 kd protein in SIR1 3' region precursor | 372 | P36170 | S |
| 35848 | YFL050c | | Hypothetical 109.7 kd protein in NUP100-MSN4 intergenic region | 299 | P35724 | S |
| 36803 | YFL049w | | NPL6 protein | 195 | P32832 | S |
| 42815 | YFL046w | | Myosin heavy chain A | 114 | P12844 | S |
| 44392 | YFL045c | *SEC53* | Phosphmannomutase(EC 5.4.2.8)(PMM) | 1291 | P07283 | S |
| 45560 | YFL044c | | Dystrophin | 116 | P11532 | S |
| 47745 | YFL042c | | Hypothetical 149.7 kd protein in IRE1-KSP1 intergenic region | 609 | P38800 | S |
| 49140 | YFL041w | | Iron transport mulicopper oxidase | 1303 | P38993 | S |
| 51351 | YFL040w | | Glucose transport protein | 389 | P15729 | S |
| 54696 | YFL039cs | *ACT1* | Actin | 1781 | P02579 | S |
| 55986 | YFL038c | *YPT1* | GTP-binding protein(protein YP2) | 998 | P01123 | S |
| 56336 | YFL037w | *TUB2* | Tublin beta chain | 2154 | P02557 | S |
| 58782 | YFL036w | *PRO41* | Mitochondrial DNA-directed RNA polymerase(EC 2.7.7.6) | 6498 | P13433 | S |
| 63795 | YFL035cs | | Hypothetical 27.4 kd protein in PFK26-SGA1 intergenic region | 422 | P40484 | S |
| 74426 | YFL033c | | Protein kinase CEK1(EC 2.7.1.-) | 598 | P38938 | S |
| 75178 | YFL031w | *HAC1* | HAC1 gene | 2325 | D26506 | G |
| 76829 | YFL030w | | Soluble hydrogenase, small subunit(EC 1.12.-.-) | 286 | P14776 | S |
| 79159 | YFL029c | | Protein kinase | 208 | P36615 | S |
| 80211 | YFL028c | | Protein kinase CSK1(EC 2.7.1.-) | 222 | Q00564 | S |
| 82578 | YFL026w | *STE2* | Pheromone alpha factor receptor | 2022 | P06842 | S |
| 87232 | YFL025c | | NADH-ubiquinone oxidoreductase chain 4(EC 1.6.5.3) | 136 | P33511 | S |
| 90343 | YFL024c | | Hypothetical 195.1 kd protein in DNA43-UBL1 intergenic region | 101 | P40457 | S |
| 90984 | YFL023w | | Glutamic acid-rich protein | 197 | P13816 | S |
| 95008 | YFL022c | *FRS2* | Cytoplasmic phenylalanyi-tRNA synthetase beta chain(EC 6.1.1.20) | 2454 | P15625 | S |
| 95964 | YFL021w | | Nitrogen regulatory protein GLN3 | 303 | P18494 | S |
| 99593 | YFL020c | | Hypothetical 13.0 kd protein in URA1 5' region | 502 | P35994 | S |
| 103121 | YFL018c | *LPD1* | Dihydrolipoamide dehydrogenase precursor(EC 1.8.1.4) | 2298 | P09624 | S |
| 104456 | YFL017c | | Protease synthetase and sporulation nagative regulatory protein PAL1 | 109 | P21340 | S |
| 106230 | YFL016c | *NDJ1* | MDJ1 protein precursor | 2452 | P35191 | S |
| 107250 | YFL014w | *HSP12(GLP1)* | Heat shoch protein 12(glucose and lipid-regulated protein) | 475 | P22943 | S |
| 109924 | YFL013c | | Nucleolin(Protein C23) | 105 | P13383 | S |
| 112339 | YFL011w | | High-affinity glucose transporter HXT2 | 2369 | P23585 | S |
| 115737 | YFL010c | | Hypothetical 98.3 kd protein R10 in chromosome III | 116 | P34552 | S |
| 116139 | YFL009w | *CDC4* | Cell division control protein 4 | 3646 | P07834 | S |
| 119424 | YFL008w | *SMC1* | Chromosome segregation protein | 5660 | P32908 | S |
| 123474 | YFL007w | | RNA polymerase (EC 2.7.7.48)(L protein) | 101 | P33453 | S |
| 130328 | YFL005w | *SEC4* | Ras-related protein | 995 | P07560 | S |
| 131804 | YFL004w | | Hypothetical 14.4 kd protein in RNR1-ILV1 intergenic region | 176 | P40046 | S |
| 137151 | YFL003c | *MSH4* | MUTS protein homologue 4 | 4100 | P40965 | S |
| 138198 | Ty element | *TyA* | Transposon Ty1-17 49.8 kd hypothetical protein | 2008 | P25383 | S |
| 139471 | Ty element | *TyB* | Transposon Ty1-17 154.0 kd hypothetical protein | 6390 | P25384 | S |
| 146928 | YFL002c | *SPB4* | Putative RNA helicase | 2967 | P25808 | S |
| 147125 | YFL001w | *DEG1* | Depressed growth-rate protein | 2195 | P31115 | S |
| 148503 | cenVI | *(CDE I)* | | | | |
| 148512 | cenVI | *(CDE II)* | | | | |
| 148597 | cenVI | *(CDE III)* | | | | |
| 149104 | YFR001w | | Myosin heavy chain, clone 203(fragment) | 133 | P39922 | S |
| 150010 | YFR002w | *NIC96* | 96 kd nucleoporin-interacting component | 1997 | P34077 | S |
| 156138 | YFR006w | | X-pro dipeptidase(EC 3.4.13.9)(proline dipeptidase)(prolidase) | 607 | P12955 | S |
| 160528 | YFR008w | | Centromeric protein E(Cenp-E protein) | 109 | Q02224 | S |
| 162222 | tRNA(G) | *SUF20* | Yeast SUF20(+) frameshift supressor gene for tRNA-Gly | 2321 | X05270 | G |
| 162481 | YFR009w | | Probable ATP-dependent transporter YER036c | 918 | P40024 | S |
| 165059 | YFR010w | | Queuine trna-ribosyltransrefase(EC 2.4.2.29) | 325 | P40826 | S |
| 167429 | tRNA(Y)s | *SUP11* | Yeast Tyr-tRNA gene(Sup11) | 557 | J01380 | G |
| 173868 | YFR014c | *CMK1* | Calcium/calmodulin-dependent protein kinase type I (EC 2.7.1.123) | 2071 | P27466 | S |

Table 2-1. (continued)

| Position | ORF ID | Locus | Function or Homology | FASTA score | Acc. no. | Data base |
|---|---|---|---|---|---|---|
| 176382 | YFR015c | *GSY1* | Glycogen synthetase, isoform 1 (EC 2.4.1.11) | 3516 | P23337 | S |
| 180734 | YFR016c | | Neurofilament triplet M protein (160 kd neurofilament protein) | 251 | P12839 | S |
| 184489 | YFR019w | *FAB1* | FAB1 protein | 10567 | P34756 | S |
| 199861 | YFR023w | *PES4* | PES4 protein (DNA polymerase epsilon supressor4) | 2606 | P39684 | S |
| 203068 | YFR024cs | | Hypothetical 41.8 kd protein in ARG4 3' region | 969 | P32793 | S |
| 204737 | YFR025c | *HIS2* | Histidinyl-phosphatase(EC 3.1.3.15) | 1701 | P38635 | S |
| 210055 | YFR028c | *CDC14* | Probable protein-tyrosine phosphatase(EC 3.1.3.48) | 1659 | Q00684 | S |
| 210694 | tRNA(Y)s | *SUP6* | tRNA-Tyr(SUP6-o) | 361 | X07534 | G |
| 213299 | YFR030w | *MET10* | Sulfate reductase (NADPH) flavoprotein component(EC1.8.1.2) | 4893 | P39692 | S |
| 220093 | YFR031c | *SMC2* | Chromosome segregation protein (DA-box protein SMC2) | 5388 | P38989 | S |
| 222946 | YFR032c | | Polyadenylate-binding protein PABP | 145 | P31209 | S |
| 224756 | YFR033c | *QCR6* | Ubiquinal-cytochrome C reductase 17 kd protein(EC 1.10.2.2) | 695 | P00127 | S |
| 225945 | YFR034c | *PHO4* | Phosphate system positive regulatory protein | 1338 | P07270 | S |
| 226949 | YFR036w | *CDC26* | Cell division control protein SCD26(mutance of CDC26) | 512 | P14724 | S |
| 229172 | YFR037c | | Transcription regulatory protein SWI3 | 493 | P32591 | S |
| 229366 | YFR038w | | Hypothetical 128.5 kd protein CCR4-TPD3 intergenic region | 822 | P31380 | S |
| 223531 | YFR039c | | Hypothetical 38.1 kd protein in BCR 5' region | 101 | P33915 | S |
| 234521 | YFR040w | | Hypothetical 121.4 kd protein in BCK1 5' region | 886 | P40856 | S |
| 238243 | YFR0041c | | DnaJ protein | 123 | P17631 | S |
| 241425 | YFR044c | | Hypothetical Trp-Asp repeats containing protein in DPB3-MRPL27 | 478 | P38149 | S |
| 242450 | YFR045w | | Putative mitochondrial carrier YBR31 precursor | 224 | P38152 | S |
| 245153 | YFR047c | | Nicotinate-nucleotide pyrophosphorylase(EC 2.4.2.19) | 302 | P30012 | S |
| 248510 | YFR049w | *YMR31* | Mitochondrial ribosomal protein YMR31 precursor | 602 | P19955 | S |
| 249853 | YFR050c | *PRE4* | Proteosome component Pre4(EC 3.4.99.46)(Macropain subunit PRE4) | 1246 | P30657 | S |
| 252493 | YFR052w | *NIN1* | Nuclear integrityprotein 1 | 1323 | P32496 | S |
| 255037 | YFR053c | *HXK1* | Hexokinase A(EC 2.7.1.1)(Hexokinase PI) | 2323 | P04806 | S |
| 264192 | YFR055w | | Cystathionine beta-lyase(EC 4.4.1.8)(beta-cystathionase) | 706 | P06721 | S |
| 270012 | telomere | | telomere (TG1-3) | | | |

**Table 2-1.** List of genes and features of chromosome VI. Genes which had no homology (FASTA score less than 100) were omitted from the table. Column 1, Nucleotide position of the start of each designated element (ATG for ORFs, the first nucleotide of all other elements). For the LTRs of the *Ty* elements, the beginning of the left LTR and the right LTR are listed. Column 2, Genes are named according to established conventions: Y, yeast; F, chromosome VI; L and R, left or right chromosomal arm, respectively; w and c, genes encoded on the top or bottom strand, respectively; and superscript 's' genes predicted to be spliced. Genes are numbered from the centromere (CEN) towards the telomere (TEL). Transfer RNA designations also follow convention: t indicates tRNA; the next letter is the one-letter code for the amino acid inserted by the tRNA. Column 3, Genetic names of genes identified previously. Column 4, A description of the function of the genes. Description of proteins most similar to the other genes are also listed. Column 5, The FASTA (Pearson, 1990) score for the alignment of the encoded protein to its closest homologue. Column 6, Database accession number of the closest homologue. Column 7, Name of the database from which the entry shown in column 6 is derived. S, Swissprot; G, Genbank. Similarity search of this table was carried out using Wisconsin GCG Sequence Analysis software package or GENETYX-Mac software package.

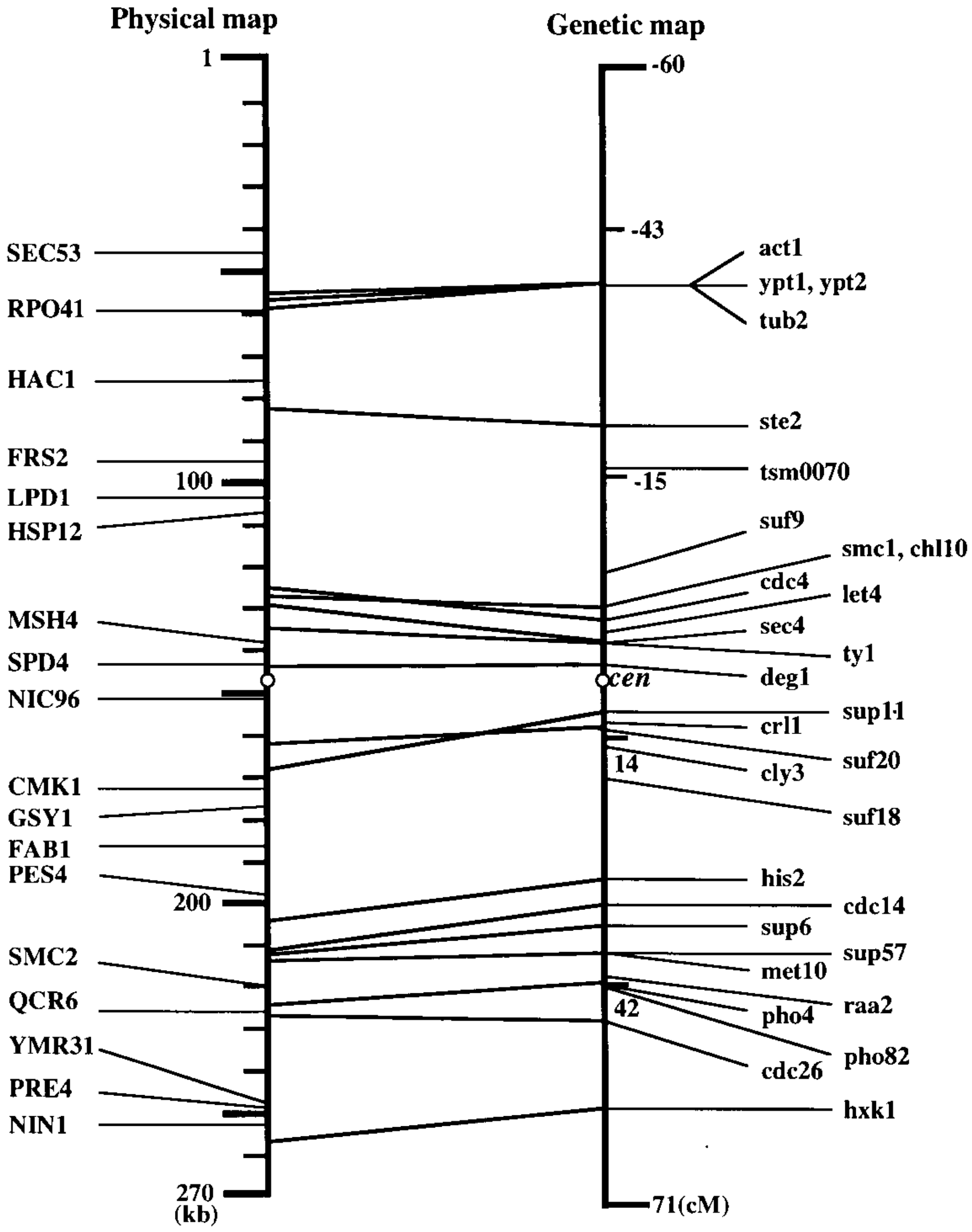


**Figure 2-3.** Genetic and physical maps of Yeast chromosome VI. The true location of the genes mapped previously on the genetic map are indicated by lines connecting them to the scale (in base pairs). Note the two minor discrepancies (two inversions between *cdc4* and *smc1* as well as between *sup11* and *suf20*) in the genetic map. This chromosome is divided into two arms; the region above the centromere is defined as the left arm and the region below the centromere is defined as the right arm (Mortimer *et al.*, 1993).

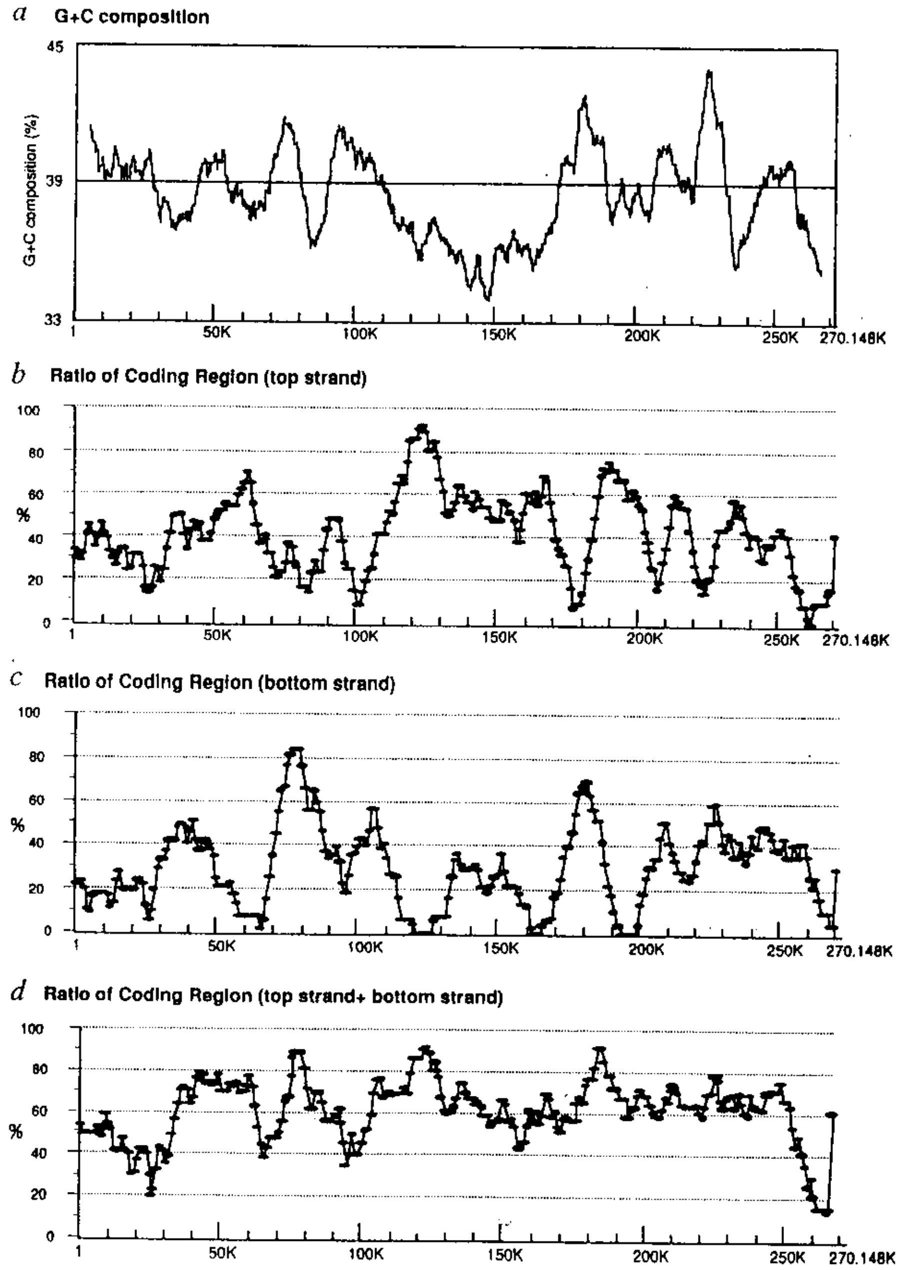


**Figure 2-4.** Plot of coding density and G+C composition over the length of chromosome VI. *a*, Overall G+C composition was calculated over 10 kb windows spaced every 100 bp. The horizontal line marks the average G+C composition (38.5%). Utilization of sequence for protein coding: *b*, top strand (a strand which has 5' terminus at the left telomere); *c*, bottom strand (a strand which has 5' terminus at the right telomere); *d*, both strands. The ratios of coding to non-coding sequence, calculated with 10 kb windows spaced every 100 bp, are plotted.

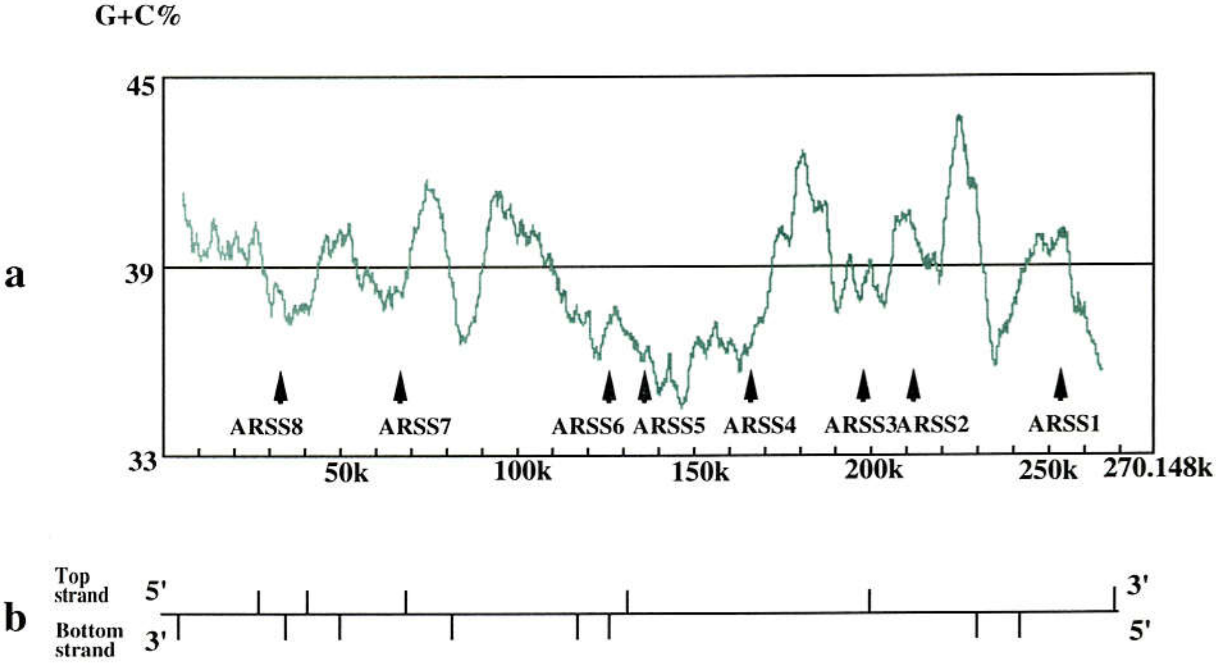


**Figure 2-5.** Analysis of distribution of ARS core sequence and core-like sequence. a, G+C composition and distribution of previously reported active ARS elements. Each arrow indicates the position of an active ARS element. b, Distribution of ARS 12-bp consensus sequences.

Table 2-2.

|  | ARS core sequencs |  | Number of sequence motif |  |  |  | Relative position of ARS to ORF$^c$ |
|---|---|---|---|---|---|---|---|
| ARS elements | Position (bp) | Orientation$^a$ | ARS like (10/11) | SAR$^b$ | Topo II | ABF I |  |
| YSCARSS 1 | 256373 | c | 4 | 0 | 0 | 0 | ◄—▼◄— |
| YSCARSS 2 | 216458 | w (11/12) | 2 | 2 | 1 | 0 | —►▼◄— |
| YSCARSS 3 | 199403 | w | 1 | 1 | 2 | 0 | —►▼—► |
| YSCARSS 4 | 167731 | c (11/12) | 1 | 2 | 1 | 0 | ▼—► |
| YSCARSS 5 | 135567 | c (11/12) | 1 | 3 | 0 | 0 | ◄▼— |
| YSCARSS 6 | 127866 | c | 0 | 1 | 0 | 0 | —▼—► |
| YSCARSS 7 | 68857 | w (11/12) | 1 | 2 | 0 | 0 | —►▼◄— |
| YSCARSS 8 | 32708, 32971 | c+w (11/12) | 2 | 2 | 0 | 0 | ▼▼◄— |
| Inactive perfect match |  |  |  |  |  |  |  |
| 1 | 5492 | c | 2 | 1 | 1 | 1 | —▼—► |
| 2 | 27963 | w | 0 | 1 | 0 | 0 | ▼—► |
| 3 | 43487 | w | 0 | 5 | 1 | 0 | —►▼—► |
| 4 | 51029 | c | 0 | 2 | 1 | 0 | —►▼—► |
| 5 | 80489 | w | 0 | 2 | 2 | 0 | ◄▼— |
| 6 | 118748 | c | 1 | 3 | 0 | 0 | —►▼—► |
| 7 | 195135 | w | 0 | 3 | 1 | 1 | —▼—► |
| 8 | 229906 | c | 0 | 0 | 1 | 0 | —▼—► |
| 9 | 242428 | c | 0 | 2 | 0 | 1 | ▼—► |
| 10 | 258900 | c | 0 | 0 | 1 | 0 | ◄▼— |
| 11 | 269757 | w | 0 | 1 | 4 | 1 | —►▼ |

**Table 2-2.** Analysis of DNA sequence motifs involved in ARS activity. $^a$ARS core sequences located on top (w) or bottom (c) strand. (11/12) : one base mismatch to ARS 12-bp consensus sequence. $^b$Nuclear scaffold attachment region. $^c$Horizontal arrow indicates ORF direction from 5' to 3'. Vertical arrow indicates the position of ARS. The domains were analyzed, each spanning one kb region upstream and downstream of the ARS core sequence.

## Discussion

Analysis of the entire sequence of yeast chromosome VI revealed a unique distribution pattern of ORFs ; there was preferential utilization of one strand in the central region of the chromosome. Further analysis of the possible function of these aligned genes and analysis of sequences upstream of these unidirectional genes may reveal the occurrence of polycistronic control which has been reported recently in a nematode genome study (Zorio *et al.*, 1994). Such an analysis should be conducted in the near future.

In chromosome VI, no relationship was observed between gene density and C+G composition similar to that previously seen in chromosomes II and XI (Feldmann *et al.*, 1994, Dujon *et al.*, 1994). Also, the apparent organization of chromosomes II and XI into regularly spaced intervals of G+C-rich and -poor segments was not observed in chromosome VI. The results from analysis of both chromosomes VIII and VI suggest that the generality of those phenomena (Johnston *et al.*, 1994) are unlikely. A recent report on chromosome I (Bussey *et al.*, 1995) showed that both ends of the chromosome were gene-poor and contained many non-functional gene fragments. In chromosome I, the region 10 kb to 25 kb from the right end is duplicated. Interestingly, the right end of chromosome VI contains few ORFs and the left end has a large number of short ORFs. The region 5-8 kb from the left end of the chromosome is almost identical to a part of chromosome II and the region between 14-15 kb from the left end of chromosome VI had a common sequence with chromosome III. These results suggest that the end regions of such small chromosomes exhibit a high frequency of recombination events. A previous report (Bussey , 1995), in conjunction with this observation, suggested that the role of the end region of such small yeast chromosomes is to increase the length of the chromosomes to ensure mitosis and stability of the chromosome. Functional analysis of the short ORFs on the left end of chromosome VI may confirm this possibility.

To investigate the mechanism through which a small number of ARS elements are selected for activation from a large number of candidate loci, a systematic analysis of the sequence motifs of the flanking regions around ARS core sequences was carried out. The results indicated that all but one ARS element have additional ARS-like consensus sequences in the 3' flanking region, whereas nine of eleven inactive loci do not have any additional sequences that have 10-of-11-base matches to the consensus sequence around there. It is interesting to note that ARSS6, the active consensus sequence that does not have an ARS-like consensus sequence, was reported previously to have markedly low ARS activity (Shirahige *et al.*, 1993). This suggests that the presence of such additional ARS-like sequences may be involved in the variation of ARS activities. Unexpectedly, no marked positional association of ARS consensus sequences with nuclear scaffold binding sites were detected, no active ARS element had an ABF1 binding sites, and there were more Topo II cleavage sites in the 3' flanking region of inactive ARS elements. Therefore, more detailed analysis regarding the distribution of the motif elements around ARS elements and three-dimensional interaction with the activation factors is required before reaching a definite conclusion.

The composition of G+C content with various window sizes showed no marked differences in G+C content around active ARS elements and inactive loci. This argues against the importance of an A+T-rich domain for ARS activity.

The relationship between the position of active ARS elements and the distribution of ORFs was also analyzed. Interestingly, all but one active ARS element (ARSS5) mapped to non-coding regions while five of eleven inactive loci mapped to areas with coding sequences. These results suggest that transcriptional regulation may also play a role in the activation of ARS core sequences.

The most important impact of this study is not only the identification of numerous novel genes as found in previous analyses of other chromosomes but also the analysis of the ARS element, a candidate for replication origin in eukaryotes, in relation to the entire chromosome. Since the gene density in the yeast genome is relatively high, sequencing chromosomal DNA leads to the immediate identification of structural or regulatory genes and elements. International collaboration has led to the sequencing of the whole yeast genome, and more and more interesting features of the smallest eukaryotic genome have been revealed. Upon completion of sequencing efforts, all the yeast genes and replicons will be identified, which will have far-reaching effects on the studies of biological processes in higher eukaryotes.

## Summary

The complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome VI (270 kb) was determined by the improved shotgun strategy (see part 1) and revealed that it contains 129 predicted or known genes, thirty-seven of which have been identified previously. Among the ninety-two novel genes, thirty-nine are highly homologous to previously identified genes.

On this chromosome, eight active ARS elements previously reported were revealed to be distributed in 30kb intervals on average. Local sequence motifs were compared to active ARS regions and inactive loci with perfect ARS core sequences to examine the relationship between these motifs and ARS activity. Additional ARS sequences were predominantly observed in the 3' flanking sequences of active ARS loci.

# Part 3.

# Nucleotide sequencing and analysis around the centromeric end of the HLA class I region

## The HLA class I gene region on human chromosome 6

The human major histocompatibility complex (MHC) genes encode highly polymorphic human leukocyte antigens (HLA) responsible for antigen presentation to T cells. The HLA gene complex is located on the short arm of chromosome 6, 6p21.3, and covers up to a 4 Mb (Mega base pairs) segment that seems to have been generated through repeated gene duplication and conversion during evolution. The HLA region is conventionally divided into three areas, class II (1 Mb), class III (1 Mb) and class I (2 Mb) from centromere to telomere. The HLA class I and class II antigens involved in genetic control of immune response are encoded by the corresponding regions. There are presently 19 HLA or HLA-like expressed class I and class II genes, more than 160 non-HLA expressed genes and 25 pseudogenes and gene fragments localized within the HLA region (Campbell and Trowsdale, 1997).

There is one gene detected in every 30-40 kb in the HLA region on average. However, it is hard to ascertain whether or not gene density observed in the region is remarkably high, since other regions of the human genome have not yet been characterized in sufficient detail. The number of genes so far identified in the class I region is smaller than those in the class II and III regions which especially include a gene-dense region of approximately 700 kb between the INT3 and BAT1 genes where more than 40 expressed genes have been identified (Campbell and Trowsdale, 1993). This is mainly because the class I region has not been studied so extensively, but it is predicted that the number of genes within the region will continue to increase as more sophisticated means of detecting coding sequences become available. According to recent HLA gene mapping, additional multi-gene structures, such as PERB family (Leelayuwat *et al.*, 1992), P-5 family (Vernet *et al.*, 1993) and MIC family (Bahram *et al.*, 1994), have been reported in the class I region. Although many genes responsible or not for immune response have been located in this region, precise organization of these genes remains to be elucidated. There have been some reports on the degree of the genetic polymorphism between the HLA-B and -C (e.g., Bodmer *et al.*, 1997). The surrounding regions and boundaries of the HLA loci, however, have not yet fully studied (Figure 3-1). Therefore, genomic analysis based on large-scale sequencing will greatly facilitate our understanding of human multi-gene organization as the HLA loci arrangement. Accordingly, in order to clarify the genomic structure of the non-characterized class I region, my group has conducted large-scale sequencing and reported the contiguous genomic organization around the centromeric end of the HLA class I region (Accession numbers D83543, D83769, D83770-83771, D83956-D83957, D84394, AB000876-AB000880 and AB000882 ; Mizuki *et al.*, 1997b; Shiina *et al.*, 1998; Yamazaki *et al.*, 1999).

In this part, I will discuss the genomic organization around the centromeric end of the HLA class I region, including the boundaries between the class III and I regions. The total size subjected to the present study is contiguous 385,633 bp. I will then give a new line of evidence for genomic duplication and will also present some results of gene finding analysis for this region.

## Materials and method

### Materials

Two overlapping YAC clones were used to cover around the centromeric end of the HLA class I region, including the boundary between the class III and class I regions. One is 745D12, originated from the CEPH YAC library constructed from the HLA-homozygous cell line, BOLETH (HLA-A2, -B62, -Cw10, -DR4), and the other is Y109 (Imai and Olson, 1990), from CGM1 (HLA-A3, -B8, Cw-, DR3, -DQ2, -DR52 and HLA-A29, -B14, -Cw-, -DR7, -DQ2, -DR53).

Twelve contiguous cosmid clones spanning the 385 kb region, TY1F9, TY1G10, TY2A9, TY3A9, TY2F10, pM67, pM213-5, pM117, pMN125, pMN201, pM30, and pM56, (Figure 3-2) which were subcloned from the two YACs, were selected for nucleotide sequence determination by the shotgun method (Mizuki *et al.*, 1997b; Shiina *et al.*, 1998). For random subcloning by construction of shotgun library from the representative cosmid clones, *Escherichia coli* strain DH5α (supE44 ΔlacU169(f80lacZΔM15) hsdR17 recA1 endA1 gyrA96 thi-1relA1) was used.

### Methods

*Preparation of shotgun libraries*

The cosmid DNA were purified by the alkaline lysis method following CsCl ultra-centrifigation (Maniatis *et al.*, 1989). The purified DNA was subjected to sonication, size fractionation, and treatment with Klenow enzyme to generate repaired blunt ends. The blunt ended DNA was then ligated into *Sma* I digested, dephosphorylated pUC19 vector DNA as an insert to a vector molar ratio of 5:1. The ligation mixture was then transfected into competent DH5α cells prepared as described by Hanahan *et al.* (1983). Several thousand recombinants were regularly obtained from five micrograms of target DNA. The ratio of clones which had human DNA inserts was higher than 95% in a typical preparation (See also Construction of a single high-quality library, described in part 1).

*Preparation of sequencing templates*

Approximately 400 recombinant pUC19 plasmid clones per cosmid clone were randomly selected. The plasmid DNA was purified by the alkaline lysis method followed by Polyethyleneglycol precipitation method (Maniatis *et al.*, 1989). 96 to 192 random clones were purified into plasmid DNA per day. Typically, three to ten µg of high-purity plasmid DNA were obtained per 1 ml of overnight culture in terrific broth (Maniatis *et al.*, 1989), which was enough at least for five sequencing reactions.

*DNA Sequencing and data assembly*

Sequencing reactions were performed using 0.6-1µg of double-stranded plasmid DNA. The DNA sequences were determined using an Amplitaq polymerase dye primer cycle sequencing method and were analyzed on ABI 373A/S automated fluoro-sequencers (Perkin Elumer) following the manufacturer's instructions. To determine the order of the sequence contigs, sequence reactions were carried out in each direction with the universal, forward (-21M13) and reverse (M13RP1) primers flanking the cloning site. Individual sequences were minimally edited to remove vector sequences, transferred to Unix workstations (Sun SPARCstation 10 or SPARCstation 20, Sun Microsystems Inc.) on the TCP/IP protocol and assembled into contigs using GENETYX-Σ/SQ software (Software Development Co., Tokyo). The raw sequence data transferred were then searched against Alu repetitive sequences before assembling as described in part 1. Contigs without Alu sequences were merged by being bridged by the sequences. All bases were covered by more than two fragments. The overlap between adjacent cosmid clones was ascertained at the sequence level. Remaining ambiguities were eliminated according to the electropherogram of 373A/S DNA sequencers. Finally, a contiguous nucleotide sequence of 385,633 bp was determined with the overall redundancy 6.5 per base.

*Analysis of the sequence and gene prediction*

The final sequence of the 385 kb region thus obtained was subjected to homology search against DDBJ/EMBL/Genbank databases and analyzed using the GENETYX-Mac software package (Software development Co., Tokyo) on Macintosh computers. Such repetitive elements as LINEs, SINEs, and LTRs in the entire 385 kb region were extracted and analyzed by RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html). For finding exons in the genomic sequence, I used GenScan (Burge and Karlin, 1997) and GRAIL(Uberbacker *et al.*, 1991).

# Results

## General features in entirety

This region (385 kb) was confirmed to contain fourteen previously known genes and gene-like structures, IkBL, BAT1, MICB, 3.8-1 gene-like fragment, P5-1 gene-like fragment, HLA-X pseudogene-like fragment, MICA, P5-8 gene-like fragment, HLA-17 pseudogene, PERB1 gene-like fragment, DHFR pseudogene, HLA-B, RPL3 pseudogene, and HLA-C (Figure 3-3). Among them, six genes (IkBL, BAT1, MICB, MICA, HLA-B, HLA-C) are apparently expressed (Albertella and Campbell, 1994; Bahram *et al.*, 1994, 1996; Peeman *et al.*, 1995). In addition, five potential genes were supposed to exist in this region as will be mentioned later in the section of Non-identified Gene Hunting). Thus, the average gene density of expressed genes is estimated as one gene per 35.1 kb. It remains to be established whether this gene density is relatively high or not, until large-scale genome sequencing is conducted on other human chromosome regions.

The average G+C content is 44.8% which corresponds to the isochore H1, though the HLA class I region in gross belongs to the GC-richest isocore, H3 (53% on average ; Fukagawa *et al.*, 1995). As seen in figure 1, the G+C content was fairly uniform throughout this area except the segments around the above mentioned six expressed genes (IkBL, BAT1, MICB, MICA, HLA-B and HLA-C) which were remarkably GC rich. Between the MICB and MICA genes (starting at nucleotide position 134,281), a long imperfect repetitive element (GAATA(T/C)ATATATA)$_{195}$ of 2,545 bp, was identified. This repetitive element is unusual in length as well as in nucleotide composition. Generation of such a long repetitious sequence is considered likely to occur at the geneome-wide level in the course of the genome evolution. It is also interesting that there is no linkage disequilibrium observed between the MICB and MICA genes (Ando *et al.*, 1997), suggesting that this long tandem repeat may be a hot spot for genetic recombination.

The HLA class I region has been thought to contain a number of homologous elements, regardless of expressed genes or not (Leelayuwat *et al.*, 1992, 1995; Vernet *et al.*, 1993; Venditii *et al.*, 1994; Pichon *et al.*, 1996). Also, numerous microsatellite sequences (Tamiya *et al.*, 1998) and repetitive elements have been found within the region. My analysis revealed that the 385 kb region contained 185 Alu elements, 24 MERs, and 14 MIRs. Also, at least 148 microsatellites were recognized by Tamiya *et al.*(1998) The content of repetitive elements is listed in Table 3-1. Alu sequences are distributed nearly uniformly throughout the region (Figure 3-3), suggesting that the elements were introduced after the prototype of the HLA class I region had taken shape following repeated duplications in the consequent multi-gene family loci. There are two Alu-dense clusters on the both ends of the 385 kb region; one is located at the telomeric end of the class III region, and the other is found around the 60 kb telomeric region from the HLA-C gene. In the Alu clusters, the average Alu density is higher than 1.0 per kb, which is more than twice of that for the entire region (0.48 per kb). Around these two clusters, a slight increase of the G+C content was observed (Figure 3-3). This observation is consistent with the previous report that Alu sequences exist preferentially in GC rich regions (Mazzarella and Schlessinger, 1997). It is, however, necessary to determine nucleotide sequences of the segments flanking the region on the both sides of the two Alu-dense clusters to draw a definite conclusion.

There are 133 LINEs (114 L1s and 19 L2s) within the 385kb region (Table 3-1) and the density is 0.34 per kb. Among them, 114 L1s occupy as much as 16% of the region, which is remarkably higher than the previously predicted ratio, 4-5% (Fanning and Singer; 1987, Schmid, 1996). The size of L1s ranges from 50 to 5777 bp and the average size is 552 bp. Namely, few LINES have kept the original size, and most have drastically reduced their size in evolution. Note also that the LINEs, SINEs and LTRs in total occupy 47 % of the entire 385 kb region (Table 3-1), which again is much larger than the previously predicted values (19.8-32.3% of the human genome ; Schmid, 1996).

## Evolutionary relationships of segmented sequences

The HLA class I gene composition is thought to have been brought on by repeated duplication in evolution. The repeated duplications are considered responsible for high complexity due to multi-homologous sequences in the HLA gene region (Leelayuwat *et al.*, 1992, 1995; Vernet *et al.*, 1993; Venditti *et al.*, 1994; Pichon *et al.*, 1996). Nevertheless, most data obtained by the previous studies on gene mapping and identification in the class I region are not so informative. It is thus important to carry out a large scale sequencing and analysis to elucidate the evolutionary organization of the region. There have been several reports showing that there are two pairs of duplicated segments in the 385 kb region (Mizuki *et al.*, 1997b; Kulski *et al.*, 1997; Gaudieri *et al.*, 1997a,b; Shiina *et al.*, 1998). One pair is an approximately 52 kb genomic fragment including the MICB gene and a 35 kb fragment including MICA (segment A and A' ; Figures 3-3 and 3-4) and the other is an approximately 43 kb genomic fragment including the HLA-B gene and a 40 kb fragment including HLA-C (segment B and B' ; Figures 3-3 and 3-4). Segment A contains the MICB gene, the P5-1 gene-like fragment and the HLA-X pseudogene-like fragment from telomere to centromere. Similarly, Segment A' contains the MICA gene, the P5-8 gene fragment and the HLA-17 pseudogene in the same order (Figure 3-3). I first confirmed the results of the previous reports, giving a new line of evidence that the segments A and A' have extremely similar LINE compositions as well as arrangements, and so do segments B and B', as shown in Figure 3-3. I then extended my study to elucidating the evolutionary relationship between the two pairs. Though there are some reports on the relationship (Kulski *et al.*, 1997; Gaudieri *et al.*, 1997b), it has still been reminded ambiguous.

I carried out my analysis, particularly paying attention to the inner structure of each segment, because the inner structure is known to have information about evolutionary history of the segment. Kulski *et al.* (1997) estimated the period in which the duplication between segments B and B' occurred on the basis of the evolutionary relationships of the Alu sequences in the two segments. However, since the most of the Alu family members were originated after the duplication of A and A' and that of B and B' (Kulski *et al.*, 1997; Gaudieri *et al.*, 1997a), the family does not give much information about the evolutionary relationship of the two pairs. Thus I focused on the LINE family which has considered to have been originated earlier than the Alu family. The Alu family is restricted within mammalian species, whereas the LINE family is found not only in mammalian species but also in other vertebrate species.

As mentioned above, segments A and A' share almost identical LINE sequences (from telomere to centromere, two L2s - one L1MC4 - one L1MB - two L1MA3s - two L1MBs), and so do segment B and B' (from centromre to telomere, one L1M4 - one L2 - one L1P -

one L1MA - two L1MEs - one L1MB8). This indicates that the divergence of the LINE sequences occurred prior to the duplication of each pair, making it possible to study the evolutionary relationship between the two pairs. Sequence homologies between the corresponding LINE subfamilies observed in the segments A and A' were compared to those in the segments B and B'. I thus computed homology between the corresponding LINE subfamilies found in segments A and A' and between those in segments B and B'. The average homology for segments A and A' was 93.7 %, and that for segments B and B' was 87.0 %. Welch's t-test of the two values concluded that the former was significantly higher than the latter (p = 0.00601). This result indicates that the duplication of B and B' occurred earlier than that of A and A', and thus that divergence of the HLA-B and -C occurred prior to that of the MICA and MICB. The HLA-B and -C genes are considered to belong to the classical HLA class I family (class Ia) along with the HLA-A gene, and an ancient HLA-A gene is believed to be the ancestral form of the HLA-B and -C genes (Huges and Nei, 1989). My phylogenetic analysis of sequences around these class I genes (unpublished) supports the proposal that the HLA-B and -C genes have diverged from the ancestral HLA-A gene.

I also classified the Alu subfamilies in the class I, class II and class III regions as shown in Table 3-2. AluF and AluJ are thought to be old subfamilies, whereas AluY is a younger one (Kapitonov and Jurka, 1996; Schmid, 1996; Batzer *et al.*, 1996; Toda *et al.*, 1997). Accordingly, their dispersed patterns may facilitate our understandings of divergence of genomic segmental duplication (Kulski *et al.*, 1997). As seen in Table 3-2, the older subfamilies were much less frequent in the centromeric class I region than in the class II and III regions. In the class II and III regions, the frequency of the older ones (AluF and AluJ) is approximately two-fold higher than that in the centromeric class I region. Moreover, the ratio, (AluF+AluJ)/AluY in the centromeric class I region is significantly lower than that in the other two regions (table 3-2). This suggests that the class I region has been established more recently than the class II and III regions and has continued to be duplicated during Alu dispersion. The presence of a few copies of an ancestral member of the SINE family (MIR) in the class I region (0.5 % occupancy) also supports the proposal that the class I region is youngest among the three. (generally, 1-2% in the genome; Smit and Riggs, 1995).

### Non-identified gene hunting

Since the class I region has not been studied less extensively than the class II and III, regions, the number of identified genes in the former region is less than those in the latter regions. It is true, however, that some diseases such as Behçhet's, ankylosing spondylitis and psoriasis vulgaries have been established to be strongly associated with particular HLA-B or -C alleles such as B51, B27 and Cw6, respectively (Tiwari and Terasaki, 1985). My group has recently indicated that the GCT triplet repeat polymorphism in the fifth exon of the MICA gene (Figure 3-5) encoding the transmembrane domain is more strongly associated with Behçhet's disease than HLA-B51 (Mizuki *et al.*, 1997a). Also, some additional transcripts designated as new organization associated with HLA-B (NOBs) were recently reported around this region (Mizuki *et al.*, 1997b). Thus, the clarification of uncharacterized genes in this region is an urgent prerequisite for the identification of causative genes for the HLA class I associated diseases.

I first used GRAIL (Uberbacker *et al.*, 1991) and GenScan (Burge and Karlin, 1997) for finding exons in the 385 kb region, as the first step toward identifying the causative genes. Table 3 shows the result of exon prediction by both tools. There are in total 175 possible exons found. Among them, 43 exons were previously identified within the six expressed genes mentioned earlier, that is, 4 in IkBL, 11 in BAT1, 6 in MICB, 6 in MICA, 8 in HLA-B and 8 in HLA-C. Of these, GenScan predicted 34 true exons (79%) and only one false exon, whereas GRAIL 32 true (74%) and 8 false ones indicating in accordance with the comparative study by Burge and Karlin (1997) that GenScan is slightly better than GRAIL. It seems better to utilize both of the two softwares for gene prediction.

I then carried out homology search of the 175 exons predicted by GenScan/GRAIL analyses against GenBank's dbEST, and the result is shown in Table 3-3. In consideration of inaccuracy and incompleteness of EST data, we selected those exons from Table 3-3 which showed more than 99% homology to the extant ESTs. I aligned ESTs with accession numbers AA324358, AA401679 and AA399356 (Table 3-3) and found that the corresponding exons 13, 14 and 15 together could form a single gene. From the information on the location of the exons, I could locate this newly found gene between IkBL and BAT1. By the same token, I found four other possible genes as listed in Table

3-4. AA677652 may split one potential new gene 70 kb telomeric of the HLA-C gene into three exons and one of them lies on the position which only GenScan predicted (Gene 5). Unexpectedly, only this gene was positioned on one of the potential CpG islands defined by Gardiner-Garden (Gardiner-Garden and Frommer, 1987).

Gene 2 shows a partial homology with the gene fragment designated as 3.8-1-hom in the previous report (Shiina *et al.*, 1998). Gene 3 was found to be localized near P5-8 showing 67 % amino acid homology to the human fibroblast growth factor recepter-3 gene (FGFR3; Keegan *et al.*, 1991). Fibroblast growth factors (FGF) are members of a family of multi functional growth factors which are thought to stimulate proliferation of mesenchymal, epithelial, and neuroectodermal cells. FGF also plays a role in other processes such as angiogenesis, promotion for differentiation of adipocytes and neurons, and healing (Keegan *et al.*, 1991), indicating that abnormal events in the gene encoding its receptor may accordingly cause genetic diseases. It is thus important to elucidate the relationship between Gene 3 and FGFR. Gene 3 also showed some homology with PERB1 (Leelayuwat *et al.*, 1996). I have so far found no noticeable homology for the other possible genes against the public nucleotide/protein databases. It is noteworthy, however, that the continuation of homology search against the existing databases would be rewarding, because new data are registered at the databases at every moment.

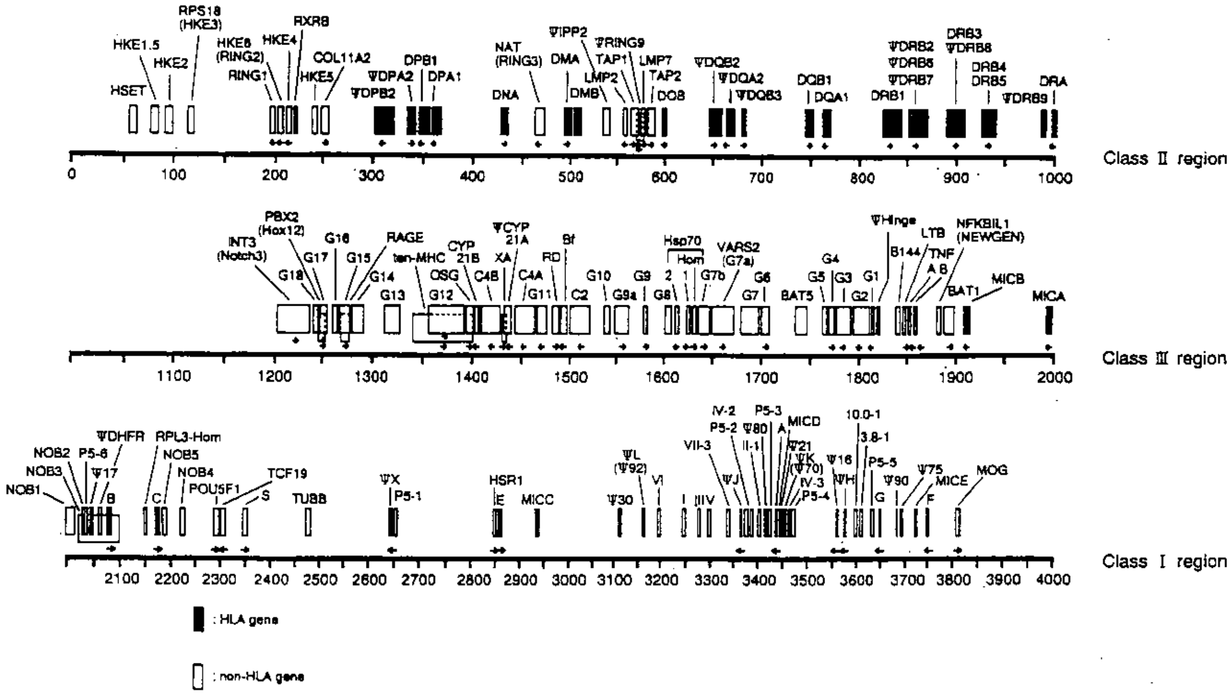


**Figure 3-1.** Genetic map of the human major histocompatibility complex gene region. Boxes indicate genes previously identified in the region. Arrows indicate their transcription orientations. Horizontal lines under the genes show approximate scale in kb.

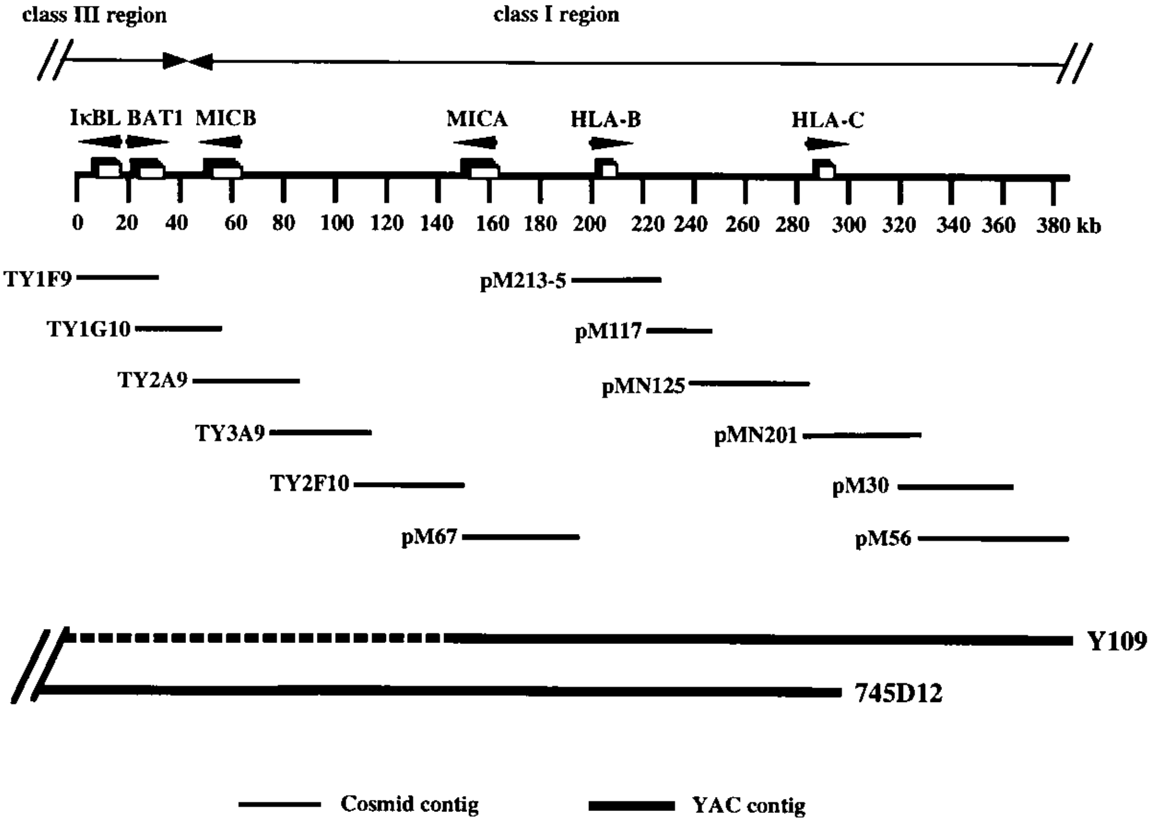


**Figure 3-2.** Physical map of the centromeric end of the HLA class I region. Overlapping cosmid and YAC clones are shown as horizontal lines with their names. Shadowed boxes show the locations of previously known expressed genes, such as IkBL, BAT1, MICB, MICA, HLA-B, and HLA-C genes. Arrowheads indicate their transcription orientations.

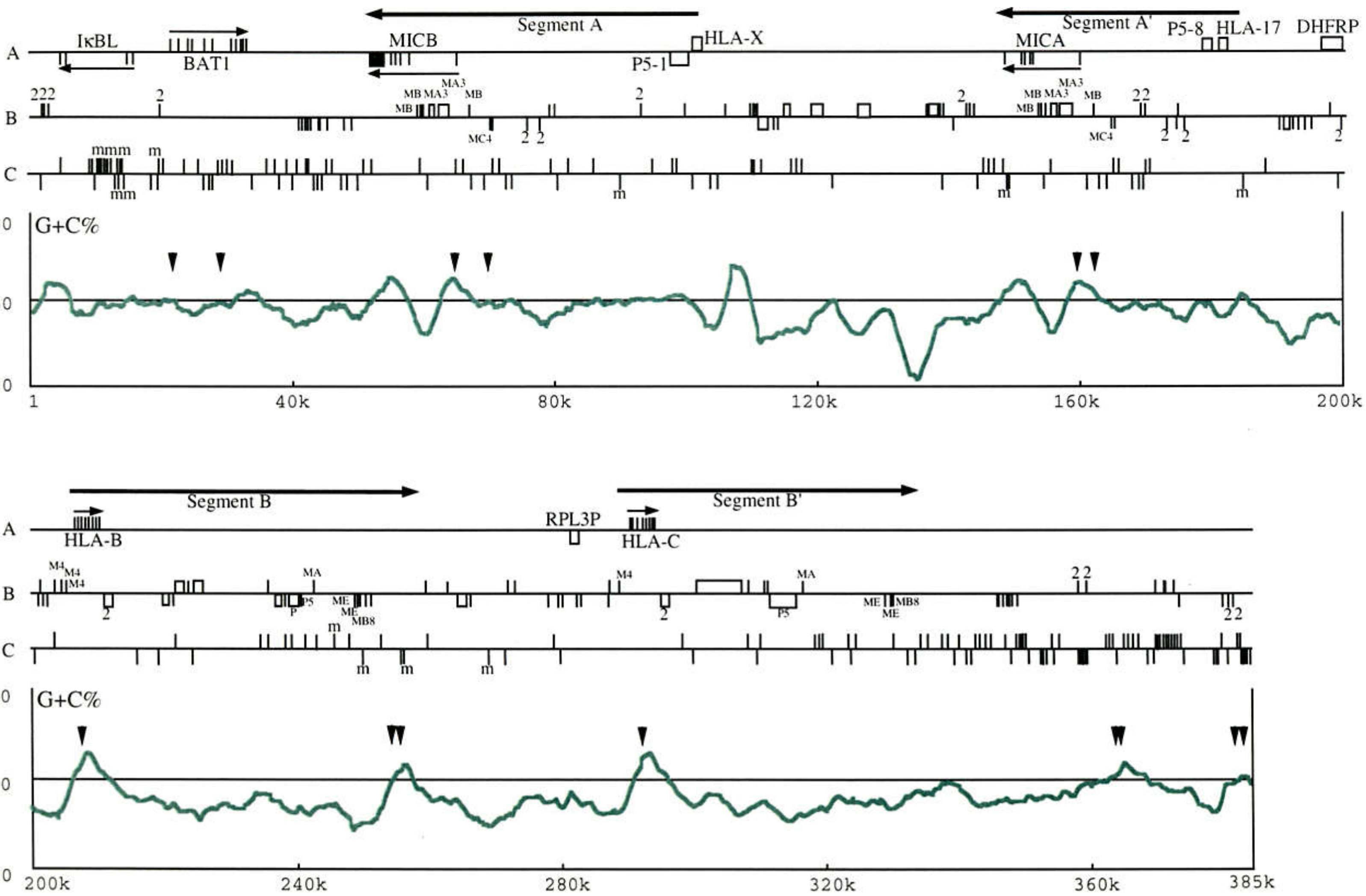

Segment A
Segment A'
A
IκBL
BAT1
MICB
P5-1
HLA-X
MICA
P5-8
HLA-17
DHFRP
B
222
2
MB MA3
MA3
MB
MB
MC4
2 2
2
2
MB MA3
MA3
MB
MB
MC4
22
2 2
2
C
mmm
m
mm
m
m
m
G+C%
80
50
20
1
40k
80k
120k
160k
200k
Segment B
Segment B'
HLA-B
RPL3P
HLA-C
M4
M4
M4
2
MA
P5
P
ME
ME
MB8
m
M4
2
MA
P5
ME
MB8
ME
2 2
22
m
m
m
200k
240k
280k
320k
360k
385k

**Figure 3-3**. Structural features around the centromeric end of the HLA class I region, including the boundary of the class III and the class I regions. A: Gene organization of this region (385,633 bases). The top boldface arrows indicate two pairs of the duplicated segments. Expressed genes are shown with their exons (vertical bars) and transcriptional orientations (horizontal arrows). Open boxes show gene fragments or pseudogenes. B: Distribution of the LINE elements. Vertical boxes and bars show the L1 or L2 sequences (Bars numbered as '2' are L2s). Similar L1 subfamilies observed in the segments A and A' (L1MC4 - L1MB - L1MA3 - L1MB), and in the segments B and B' (L1M4 - L1P - L1MA - L1ME - L1MB8), are shown with their name. C: Distribution of the SINE elements. Vertical bars show the Alu or MIR sequences ('m' indicates MIR). Upper bars or boxes in A, B and C indicate their orientation from centromere to telomere, and lower ones, from telomere to centromere. The bottom plot shows the local G+C content with scale. Vertical arrows indicate potential CpG islands defined according to the Gardiner-Garden's formula.

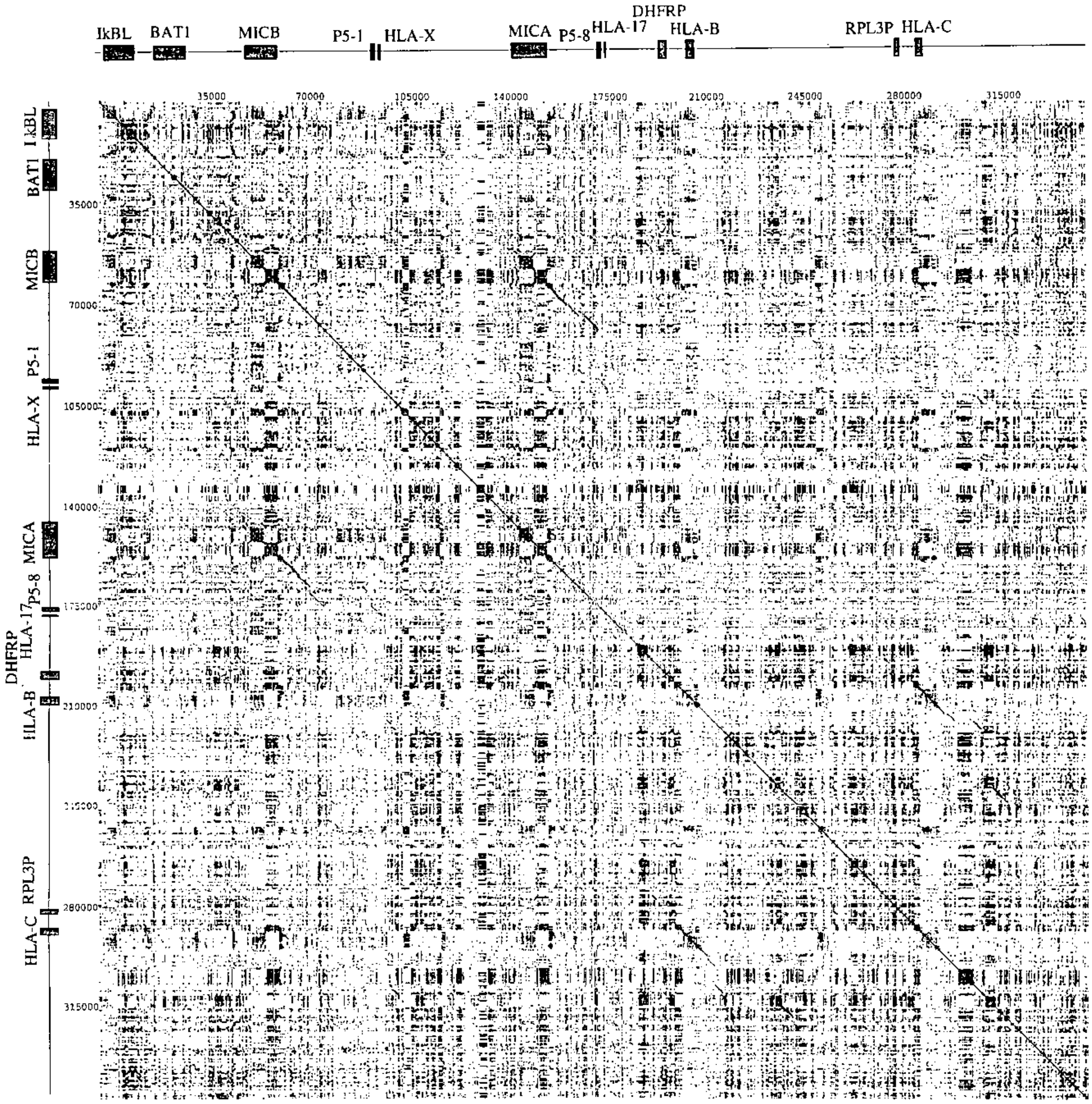


**Figure 3-4**. Dot-matrix analysis around the centromeric end of the HLA class I region versus itself. The 350 kb sequence was plotted against itself with the GENETYX-Mac software package (unit size to compare in huge plot is five). Numbers indicate the distance from the centromeric end of a cosmid clone, TY1F9. Representative genes, gene fragments and pseudogenes are shown on their approximate positions. Dots in the background indicate the common genome-wide repetitive elements, such as LINEs, SINEs and LTRs.

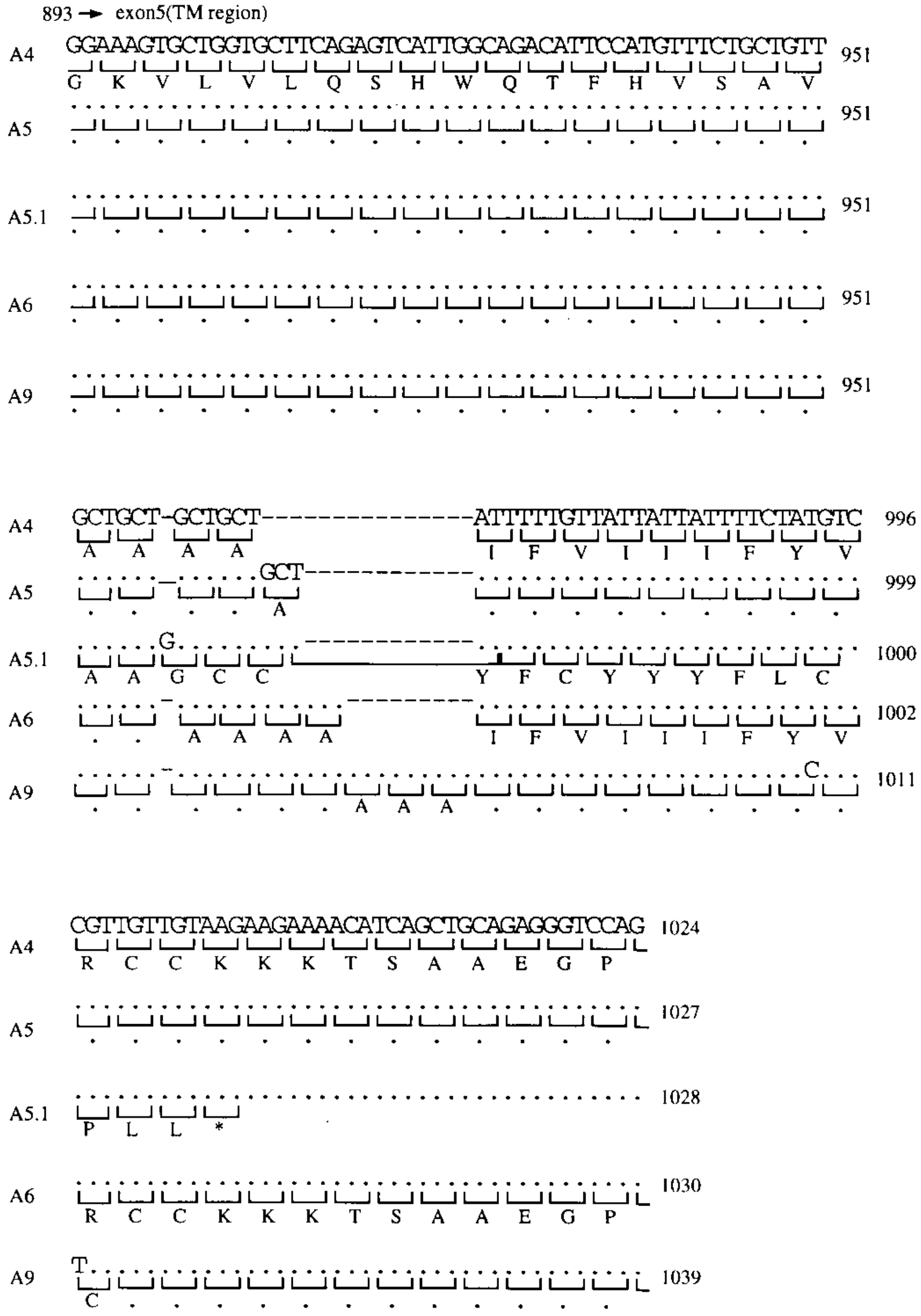

893 → exon5(TM region)
A4
GGAAAGTGCTGGTGCTTCAGAGTCATTGGCAGACATTCCATGTTTCTGCTGTT
951
G K V L V L Q S H W Q T F H V S A V
A5
951
A5.1
951
A6
951
A9
951
A4
GCTGCT-GCTGCT--------------ATTTTTGTTATTATTATTTTCTATGTC
996
A A A A
I F V I I I F Y V
A5
GCT
999
A
A5.1
G
1000
A A G C C
Y F C Y Y Y F L C
A6
1002
A A A A
I F V I I I F Y V
A9
C
1011
A A A
A4
CGTTGTTGTAAGAAGAAAACATCAGCTGCAGAGGGTCCAG
1024
R C C K K K T S A A E G P
A5
1027
A5.1
1028
P L L *
A6
1030
R C C K K K T S A A E G P
A9
T
1039
C

**Figure 3-5.** Microsatellite polymorphism in TM region (exon 5) of the MICA gene. Alphabet characters under the bracket represent amino acid abbreviations. Dots show the same nucleotide or amino acid as the upper one and a perforated line shows a nucleotide deletion site. A6 allele possessing a six GCT repetition (encodes six continues alanines) is strongly associated with Behçhet's disease. (Mizuki *et al.*, 1997a)

Table 3-1. Repetitive sequence composition of the 385 kb region

| Element | Copy numbers | Total bases | Occupancy | |
|---|---|---|---|---|
| | | | observed | (expected) |
| L1 | 114 | 62981 | 16.3% | (~4%) |
| L2 | 19 | 6837 | 1.8% | |
| Alu | 185 | 48521 | 12.6% | (6-13%) |
| MIR | 14 | 1862 | 0.5% | (1-2%) |
| LTR | 5 | 1552 | 0.4% | |
| MER | 51 | 10230 | 0.6% | |
| MaLR | 29 | 6885 | 1.8% | |
| RTV | 64 | 42099 | 10.9% | |
| Others | 36 | 7366 | 1.9% | |
| Total | 517 | 188333 | 46.8% | |

**Table 3-1**. Representative repetitive elements on the entire 385 kb region were extracted and analyzed by RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html). Expected numbers in parenthesis were derived from Schmid (1996).

Table 3-2. Frequency of Alu subfamilies

| Subfamilies | 6p21.3 | | | | | | 21q22.2 | |
|---|---|---|---|---|---|---|---|---|
| | class II | | class III | | Centromeric class I | | CBR region | |
| | Ratio | Density | Ratio | Density | Ratio | Density | Ratio | Density |
| AluF | 9.2% | 0.043 | 8.9% | 0.069 | 5.1% | 0.023 | 3.7% | 0.035 |
| J | 0.9% | 0.004 | 1.0% | 0.008 | 0 | 0 | 0 | 0 |
| Jb | 15.6% | 0.074 | 6.9% | 0.053 | 6.4% | 0.029 | 11.7% | 0.112 |
| Jo | 15.6% | 0.074 | 16.8% | 0.130 | 6.4% | 0.029 | 13.0% | 0.124 |
| AluJ | 32.1% | 0.152 | 24.8% | 0.252 | 12.7% | 0.057 | 24.7% | 0.235 |
| Sc | 3.7% | 0.017 | 4.0% | 0.031 | 5.1% | 0.023 | 1.9% | 0.018 |
| Sg | 8.3% | 0.039 | 11.9% | 0.092 | 15.3% | 0.069 | 11.1% | 0.106 |
| Sp | 5.5% | 0.026 | 10.9% | 0.084 | 9.6% | 0.043 | 4.9% | 0.047 |
| Sq | 6.4% | 0.030 | 5.9% | 0.046 | 10.2% | 0.046 | 10.5% | 0.100 |
| Sx | 22.0% | 0.104 | 20.8% | 0.160 | 17.8% | 0.080 | 36.4% | 0.347 |
| Sg/x | 2.8% | 0.013 | 0 | 0 | 7.0% | 0.031 | 0 | 0 |
| Sq/x | 0 | 0 | 3.0% | 0.023 | 0 | 0 | 0 | 0 |
| AluS | 48.6% | 0.230 | 56.4% | 0.435 | 65.0% | 0.291 | 64.8% | 0.618 |
| AluY | 10.1% | 0.048 | 9.9% | 0.076 | 17.2% | 0.077 | 6.8% | 0.065 |
| Total | 100% | 0.474 | 100% | 0.832 | 100% | 0.449 | 100% | 0.953 |
| AluF+AluJ /AluY | 4.09 | | 3.40 | | 1.04 | | 4.18 | |

**Table 3-2**. The data for the centromeric class I region were based on the entire 385 kb-sequence presented here (44.5 % GC). Sequences for the class II region were taken from DDBJ entries, D84401 and X87344, amounting to 230 kb in total (42.8 % GC), and for class III, D28769, L26261, M16441, X71937, Z15025, Z15026 and the 35 kb centromeric end sequence of the 385 kb region, amounting to 131 kb in total (52.7 % GC). The CBR (carbonyl reductase gene) region sequence (45.8 % GC) serving as control data on human genome was derived from Watanabe *et al.* (1998). Extraction and classification of the Alu sequences were carried out by RepeatMasker (http://ftp.genome.washington.edu/RM/Repeat Masker.html). Ratio: number of each subfamily / number of total Alu sequences. Density: number of each subfamily per kb.

Table 3-3. List of predicted exons and ESTs highly homologous to the genomic sequences of the 385 kb class I region

| Exon ID | Known exon | Actual positions | GenScan Prediction | GRAIL Prediction | ESTs with high identity (>85%) |
|---|---|---|---|---|---|
| 1 | | | | 1216 <- 1107 g | |
| 2 | | | | 2151 <- 1901 g | |
| 3 | | | | 2884 <- 2736 e | |
| 4 | | | | 4503 <- 4373 g | |
| 5 | *IKBL _4* | 5983 <- 5176 | 5983 <- 5394 | 5983 <- 5445 g | |
| 6 | *IKBL _3* | 6376 <- 6155 | 6376 <- 6155 | 6376 <- 6155 e | |
| 7 | | | | 7306 <- 7220 e | |
| 8 | | | | 13440 <- 13350 e | |
| 9 | | | | 15551 -> 15757 g | |
| 10 | *IKBL _2* | 15833 <- 15557 | 15821 <- 15557 | | |
| 11 | *IKBL _1* | 16370 <- 16245 | | | |
| 12 | | | | 16656 -> 16728 g | |
| 13 | | | 17436 -> 17517 | 17436 -> 17517 g | AA324358(100%) |
| 14 | | | 17782 -> 17882 | 17782 -> 17882 e | AA324358(100%),AA401679(100%) |
| 15 | | | 18410 -> 18553 | 18410 -> 18553 e | AA401679(100%),AA399356(100%) |
| 16 | | | 19233 -> 19293 | | |
| 17 | | | 21548 -> 21705 | 21548 -> 21705 g | |
| 18 | *BAT1 _1* | 22012 -> 22041 | | | |
| 19 | *BAT1 _2* | 23326 -> 23668 | 23458 -> 23668 | 23456 -> 23713 e | |
| 20 | | | | 24231 -> 24286 g | |
| 21 | *BAT1 _3* | 24714 -> 24841 | 24714 -> 24841 | 24714 -> 24841 e | |
| 22 | *BAT1 _4* | 25133 -> 25225 | 25133 -> 25225 | 25049 -> 25225 e | |
| 23 | *BAT1 _5* | 27305 -> 27488 | 27305 -> 27488 | 27305 -> 27488 e | |
| 24 | *BAT1 _6* | 28503 -> 28621 | 28503 -> 28621 | 28503 -> 28621 e | |
| 25 | | | | 29134 -> 29356 g | |
| 26 | *BAT1 _7* | 31077 -> 31208 | 31077 -> 31208 | 31077 -> 31212 e | |
| 27 | *BAT1 _8* | 32583 -> 32692 | 32583 -> 32692 | 32583 -> 32692 e | |
| 28 | *BAT1 _9* | 32791 -> 32935 | 32791 -> 32935 | 32807 -> 32935 e | |
| 29 | *BAT1 _10* | 33062 -> 33209 | 33062 -> 33220 | 33070 -> 33209 e | |
| 30 | *BAT1 _11* | 33537 -> 33744 | | | |
| 31 | | | 34240 <- 34052 | 34240 <- 34060 e | |
| 32 | | | 34972 <- 34802 | | |
| 33 | | | | 35895 <- 35711 e | |
| 34 | | | | 37826 <- 37743 e | |
| 35 | | | | 40261 <- 40081 e | |
| 36 | | | | 43919 <- 43701 e | AA311346(91%),D54395(91%) |
| 37 | | | 43687 -> 43966 | | |
| 38 | | | 44290 -> 44498 | | AA628259(90%),AA400081(90%) |
| 39 | | | | 44946 <- 44807 g | |
| 40 | | | 47922 <- 47756 | 47922 <- 47760 e | |
| 41 | | | | 50155 <- 50106 g | |
| 42 | | | | 50868 <- 50746 e | |
| 43 | | | | 52529 <- 52388 g | |
| 44 | | | | 52723 <- 52632 g | |
| 45 | | | | 53830 <- 53729 g | |
| 46 | *MICB _6* | 54199 <- 52862 | | | |
| 47 | | | | 56372 <- 56214 e | |
| 48 | *MICB _5* | 56581 <- 56450 | 56581 <- 56450 | | |
| 49 | *MICB _4* | 56959 <- 56681 | 56959 <- 56681 | 56959 <- 56681 e | |
| 50 | *MICB _3* | 57838 <- 57551 | 57838 <- 57551 | 57677 <- 57551 g | |
| 51 | *MICB _2* | 58364 <- 58110 | 58364 <- 58110 | 58364 <- 58110 e | |
| 52 | *MICB _1* | 65791 <- 65717 | 65786 <- 65717 | 65786 <- 65717 e | |
| 53 | | | | 73480 -> 73610 g | |
| 54 | | | | 74696 <- 74607 e | |
| 55 | | | | 83372 -> 83525 g | |
| 56 | | | | 84566 -> 84678 g | |
| 57 | | | | 87649 -> 87771 e | |
| 58 | | | | 91786 -> 91877 g | AA534999(100%),R93914(100%) |
| 59 | (3.8-1) | 92734 <- 91555 | | 91935 -> 92028 e | AA534999(100%),R93914(100%) |
| 60 | | | | 99512 -> 99682 e | |
| 61 | (P5-1_2) | 100590 <- 98152 | 99638 <- 99442 | | |
| 62 | (P5-1_1) | 100770 <- 100682 | 100745 <- 100682 | 100690 -> 100842 g | |
| 63 | (HLA-X) | 101471 -> 102116 | | | |
| 64 | | | | 112429 <- 112319 g | |
| 65 | | | | 123721 <- 123645 g | |
| 66 | | | | 124137 <- 124077 e | |
| 67 | | | | 124750 <- 124602 g | |
| 68 | | | 125626 <- 125051 | | AA486105(89%),AA613195(86%) |
| 69 | | | | 129261 <- 129180 e | |
| 70 | | | 130035 <- 129643 | | |
| 71 | | | 130787 <- 130699 | 130772 <- 130698 g | R32135(88%),AA691121(85%) |
| 72 | | | 131573 <- 131453 | | |
| 73 | | | | 132271 <- 132237 g | |
| 74 | | | | 133838 <- 133768 g | AA055018(92%),AA001786(91%) |
| 75 | | | | 137855 -> 138238 g | |
| 76 | | | 147752 <- 147507 | 147751 <- 147521 e | |
| 77 | *MICA _6* | 149113 <- 148812 | | 149112 <- 148985 g | |
| 78 | | | 151587 <- 151442 | 151582 <- 151441 g | |
| 79 | *MICA _5* | 151800 <- 151665 | | | |
| 80 | *MICA _4* | 152178 <- 151900 | 152178 <- 151900 | 152177 <- 151899 e | |
| 81 | *MICA _3* | 153053 <- 152766 | 153053 <- 152766 | 153035 <- 152781 g | |
| 82 | *MICA _2* | 153582 <- 153328 | 153582 <- 153328 | | |
| 83 | | | | 157794 -> 157877 g | |
| 84 | *MICA _1* | 160531 <- 160423 | 160492 <- 160423 | | |
| 85 | | | | 167019 <- 166617 g | |
| 86 | | | 169129 -> 169245 | | |
| 87 | | | | 169531 <- 169432 e | AA446140(97%),AA740679(96%) |
| 88 | | | | 170029 <- 169905 g | |
| 89 | | | | 171701 <- 171563 g | |

Table 3-3. *continued*

| Exon ID | Known exon | Actual positions | GenScan Prediction | GRAIL Prediction | ESTs with high identity (>85%) |
|---|---|---|---|---|---|
| 90 | | | | 176189 <- 176041 | |
| 91 | | | | 178373 -> 178554 | |
| 92 | | | 179297 -> 179494 | | N85228(100%) |
| 93 | (P5-8) | 180350 <- 177905 | | | |
| 94 | | | | 181018 -> 181221 g | N59757(100%) |
| 95 | (HLA-17) | 181658 -> 182590 | | 181857 -> 181967 g | |
| 96 | | | | 183646 -> 183708 g | |
| 97 | (PERB1) | 186882 <- 186344 | 186402 <- 186179 | | |
| 98 | | | | 188667 -> 188829 e | |
| 99 | | | 191394 <- 191265 | 191394 <- 191265 g | |
| 100 | | | | 193898 <- 193721 g | |
| 101 | | | 197117 <- 196953 | g | |
| 100 | (DHFRP) | 197156 -> 197890 | 197263 <- 197231 | 197208 -> 197372 g | |
| 101 | | | | 206313 -> 206520 g | |
| 102 | *HLA-B* _1 | 206976 -> 207048 | 206985 -> 207048 | 206976 -> 207048 e | |
| 103 | *HLA-B* _2 | 207177 -> 207446 | 207177 -> 207446 | 207177 -> 207446 e | |
| 104 | *HLA-B* _3 | 207693 -> 207968 | 207693 -> 207968 | 207710 -> 207980 e | |
| 105 | *HLA-B* _4 | 208541 -> 208816 | 208541 -> 208816 | 208541 -> 208816 e | |
| 106 | *HLA-B* _5 | 208910 -> 209026 | 208910 -> 209026 | 208910 -> 209044 e | |
| 107 | *HLA-B* _6 | 209468 -> 209500 | 209468 -> 209500 | 209468 -> 209500 e | |
| 108 | *HLA-B* _7 | 209607 -> 209654 | | 209607 -> 209650 g | |
| 109 | *HLA-B* _8 | 209837 -> 210259 | | | |
| 110 | | | | 213690 -> 213760 e | |
| 111 | | | | 216402 -> 216485 g | |
| 112 | | | | 220098 <- 220060 g | |
| 113 | | | 227345 -> 227403 | 227292 -> 227382 g | |
| 114 | | | 229202 -> 229308 | | |
| 115 | | | 231128 -> 231373 | 231127 -> 231444 g | |
| 116 | | | 236888 -> 237051 | | |
| 117 | | | 245967 <- 245730 | 245967 <- 245730 e | |
| 118 | | | 248557 <- 248472 | 248557 <- 248472 e | |
| 119 | | | 253643 -> 253658 | | |
| 120 | | | 254419 -> 254627 | | |
| 121 | | | 255016 -> 255247 | 254993 -> 255247 g | |
| 122 | | | | 255514 -> 255620 e | |
| 123 | | | | 258761 <- 258630 g | |
| 124 | | | | 261744 -> 261839 e | |
| 125 | | | 261981 -> 262126 | 262007 -> 262126 g | |
| 126 | | | 262704 -> 262856 | | |
| 127 | | | | 268795 <- 268695 g | |
| 128 | | | | 270137 <- 270060 g | |
| 129 | | | 274850 -> 275011 | 274850 -> 275003 g | AA445914(82%),AA774184(82%) |
| 130 | | | | 276792 -> 276941 g | |
| 131 | | | | 279491 <- 279375 g | |
| 132 | (RPL3P) | 283283 <- 282017 | 282419 <- 282075 | | |
| 133 | | | 283277 <- 282489 | 283277 <- 282420 e | AA314925(95%) |
| 134 | | | 285024 <- 284861 | | AA337333(94%),AA319348(92%) |
| 135 | | | 285689 <- 285410 | 285689 <- 285314 g | AA314295(96%),AA651694(96%) |
| 136 | | | 287971 <- 286648 | 287262 <- 286648 g | AA421379(92%) |
| 137 | | | 288318 <- 288314 | | |
| 138 | | | | 290787 -> 290960 e | |
| 139 | *HLA-C* _1 | 291506 -> 291578 | 291515 -> 291578 | 291498 -> 291578 e | |
| 140 | *HLA-C* _2 | 291709 -> 291978 | 291709 -> 291978 | 291709 -> 291978 e | |
| 141 | *HLA-C* _3 | 292229 -> 292504 | 292229 -> 292504 | 292246 -> 292504 e | |
| 142 | *HLA-C* _4 | 293092 -> 293367 | 293092 -> 293367 | 293092 -> 293367 e | |
| 143 | *HLA-C* _5 | 293492 -> 293611 | 293492 -> 293611 | 293492 -> 293611 e | |
| 144 | *HLA-C* _6 | 294052 -> 394084 | 294052 -> 294084 | 294052 -> 294084 g | |
| 145 | *HLA-C* _7 | 294192 -> 294239 | 294192 -> 294239 | 294192 -> 294239 e | |
| 146 | *HLA-C* _8 | 294404 -> 294827 | | | |
| 147 | | | | 297393 -> 297476 e | |
| 148 | | | | 299752 <- 299630 g | |
| 149 | | | | 319916 -> 320000 g | |
| 150 | | | | 321024 -> 321178 g | AA757099(99%) |
| 151 | | | | 323019 -> 323229 g | |
| 152 | | | 326642 -> 326971 | 326642 -> 326971 g | AA601493(99%) |
| 153 | | | 327455 -> 327665 | | AA765804(91%) |
| 154 | | | | 328499 -> 328601 g | |
| 155 | | | 329162 -> 329287 | 329162 -> 329442 g | |
| 156 | | | 329716 -> 329984 | 329716 -> 329984 e | AA676833(85%) |
| 157 | | | | 333251 -> 333387 e | |
| 158 | | | 336998 -> 337107 | 336998 -> 337107 e | |
| 159 | | | 337273 -> 337390 | | |
| 160 | | | | 338500 -> 338603 g | |
| 161 | | | 339030 -> 339157 | 339030 -> 339157 e | |
| 162 | | | | 342653 <- 342570 e | |
| 163 | | | 347366 -> 347491 | 347366 -> 347491 e | |
| 164 | | | | 348800 -> 348865 g | |
| 165 | | | | 351151 <- 351045 g | |
| 166 | | | | 353909 <- 353821 g | |
| 167 | | | | 356241 -> 356380 e | |
| 168 | | | | 356988 -> 357023 g | |
| 169 | | | 359598 -> 359649 | | |
| 170 | | | 360705 -> 360780 | 360705 -> 360866 g | |
| 171 | | | | 361842 -> 361941 e | |
| 172 | | | 366126 -> 366202 | | |
| 173 | | | 366222 -> 366311 | | AA677652(99%) |
| 174 | | | | 375541 -> 375602 g | |
| 175 | | | 377971 -> 378123 | 377971 -> 378077 g | AA293273(96%),AA282915(92%) |

Known exon is defined as gene name, underscore, and exon numbers. Gene fragments, pseudogenes, and not sufficiently characterized gene-like fragments are listed in parentheses. In GRAIL prediction e or g means excellent or good, respectively. Red characters indicate potential genes newly identified in this region.

**Table 3-4. Five potential genes newly mapped on the genomic 385 kb region**

| | Exon ID | Position | Corresponding ESTs |
|---|---|---|---|
| Gene 1 | 13, 14, 15 | 17436 -> 18553 | AA324358, AA401679 and AA399356 |
| Gene 2 | 58, 59 | 91786 <- 92028 | AA534999 and R93914 |
| Gene 3 | 92, 94 | 179297 -> 181221 | N85228 and N59757 |
| Gene 4 | 150, 152 | 321024 -> 326971 | AA757099 and AA601493 |
| Gene 5 | 173 | 366222 -> 366311 | AA677652 |

**Table 3-4.** Five potential genes newly mapped on the genomic 385 kb region. Exon ID is defined as GenScan/GRAIL prediction in Table 3-3. Position shows the potential locations of the newly mapped genes with their transcriptional orientations. Corresponding ESTs are listed in the same table which showing > 99 % identity to GenScan/GRAIL-predicted exons (see also Table 3-3).

# Discussion

Large-scale segmental duplication has contributed to the shaping of the present-day chromosome structure, as observed in the four human syntenic regions including the MHC (6p21.3), 19p13.1-p13.3, 1q21-q25 and 9q33-q34 (Kasahara *et al.*, 1997). On the other hand, for smaller segmental duplications, it was reported that genomic rearrangements have occurred more dynamically in prokaryotes (Watanabe *et al.*, 1997) and in men (Endo *et al.*, 1997). For example, the ancestor of the HLA-B and -C genes is thought to emerge after the duplication of the ancient HLA-A gene (Hughes and Nei, 1989). It has been speculated that this local duplication was caused by introducing a sort of repetitive elements which ever functioned (Schmid, 1996, Britten, 1997). This repeat mediated recombination can be a major driving force in the shaping of mammalian genomes (Charleswoth *et al.*, 1994). The number of repetitive elements, however, can not increase limitlessly by this mechanism, because of negative selection pressure on them (Ohta, 1980). In fact, I observed that the number of pseudogenes is 7, whereas that of expressed gene is at most 11 in this 385 kb region, indicating that the former is extraordinary large. These many pseudogenes are thought to be produced with the interference of this mechanism. Fanning and Singer (1987) estimated the amount of L1 copies to be no larger than 5 % of the total genome. Nevertheless, My finding of L1 copies is too large to be in agreed with their estimation (16 % of the HLA class I region studied). I think large scale genome segmental duplication played major role in increasing the number of L1 up to this many. Since the divergence of the LINE family occurred more extensively than the Alu family prior to the genomic segment duplication, we can expect that the duplication had more impact on increasing the number of the LINE family than those of the Alu family. This is exactly what I observed for the 385 kb region.

I have shown that the duplication of segments B and B' took place earlier than that of segments A and A'. This can explain that the number of LINEs in segment B (21) is lager than that in segment A or A' (14 each). The number in segment B', however, is 11. This could be due to deletion occurred in segment B'. The similar composition of LINEs around the MICB/A and the HLA-B/-C genes can be regarded as fossils of ancient evolutionary events which give us valuable information about the origin and evolution of the genome sequence in which those repetitive sequences reside. As to Alu sequences, the subfamily composition in the centromeric class I region (350 kb) is quite different from that in the HLA class II and class III regions. This suggests that the class I region has continued to be duplicated until recently even after the AluY dispersion, however, still more analyses are necessary to determine whether this trait is extrapolated to the remaining HLA class I region or not. In this regard, my group is now extending our sequencing and analysis to the remaining HLA class I region including the downstream region of the HLA-A gene, toward the completion of the entire 2 Mb HLA class I region to explore their evolutionary features .

The HLA-B and -C genes display the different degrees of genetic polymorphism (to date, at least 186 HLA-B, 48 HLA-C alleles have been officially recognized; Bodmer *et al.*, 1997), probably because HLA-B has undergone more positive evolutionary pressure (Hughes and Nei, 1988,1989). It is also important to investigate the other HLA class I-related gene polymorphisms for elucidating an association between genetic variations and HLA class I linked diseases.

By exon hunting, five possible genes were newly identified which may be responsible for HLA class I associated diseases or others. This finding is encouraging in that my group has already revealed that the GCT triplet repeat polymorphism of the fifth exon of the MICA gene is strongly associated with Behçhet's disease (Mizuki *et al.*, 1997a). Thus, large scale genomic sequencing is an effective and conclusive way also to elucidate such an association of genetic variations with genetic diseases. For establishment the relationship between the possible five new genes and HLA class I related diseases or else, however, the full length cDNA should be obtained first. And second, it should be investigated as to whether there is polymorphism in the locus corresponding to the cDNA in question. If there is, I would proceed to reverse genetics to identify the locus of a patient's genome by using the polymorphic cDNA as a probe.

# Summary

A contiguous nucleotide sequence of 385,633 bp around the centromeric end of the HLA class I region, including the boundaries between the class III and the class I regions, was completely determined by improved shotgun sequencing approach. Analysis of the large sequence data confirmed, giving a new line of evidence, that the following two pairs of the genomic segments were duplicated in evolution ; (i) a 43 kb segment including the HLA-B showing the highest genetic polymorphism among the classical HLA class I loci (class Ia) and a 40 kb-segment including the HLA-C locus showing the lowest among those, and (ii) a 52 kb segment including the MIC (MHC class I chain related) B gene and a 35 kb segment including MICA.

Repetitive elements such as SINEs, LINEs and LTRs occupy as highly as 47 % of nucleotides in this 385 kb region. This unusually high contents of the elements suggests that the repeat-mediated rearrangements have frequently occurred in the evolutionary history of the HLA class Ia region. Analysis of similar LINE compositions within the two pairs of duplicated segments revealed ; (i) LINEs in these regions had been dispersed prior both to the duplication of the HLA-B and -C loci and to that of the MICB and MICA loci, and (ii) the divergence of the HLA-B and -C loci has occurred prior to the duplication of the MICA and MICB loci.

For finding novel genes that may be involved in HLA class I associated diseases, computer analysis applying GenScan and GRAIL to GenBank's dbEST was performed. As a result, at least five yet non-characterized genes were newly mapped on the HLA class I centromeric region studied. These possible genes should further be clarified as to whether they are really responsible for the diseases.

# Conclusion

Large-scale genomic sequencing and analysis provides insight into the dynamic characters of eukaryotic genomes which play a vital role in genome maintenance and duplication .

In part 1, for the large-scale sequencing of ordered sets of genomic clones from eukaryotes, I discussed the achievement of several improvements in the shotgun approach as summarized below ;

(1) Introduction of a cup-horn type sonication system with high-throughput capable of rapid sharing to allow sufficient fragmentation into desired size.

(2) Optimization involving end-blunting and sizing procedures using Klenow enzyme in a short time treatment and spin-column chromatography on a sephacryl S-400.

(3) Construction of a shotgun library based on estimations from a mathematical theory on random subcloning.

(4) Double-stranded sequencing which can reduce the number of clones to be sequenced and can provide successful gap filling at the final phase.

(5) A unique assembling algorithm to separate repetitive sequences which are to be assembled at the final stage so as not to lead to false connections.

Due to the above improvements in shotgun sequencing approach, accurate and successful reconstruction of small sequence fragments into the entire whole, was realized.

In part 2, I discussed the first application of above strategy to one of the smallest eukaryotic chromosome, *Saccharomyces cerevisiae* chromosome VI, which is the only one that had been tested for ARS activity previous to complete sequencing. Using the improved shotgun strategy for representative eight phages, four plasmids and three cosmids, the entire nucleotide sequence of yeast chromosome VI (270 kb) was determined. This revealed that the chromosome contains 129 predicted or known genes, and thirty-seven of which have been identified previously. Among the ninety-two novel genes, thirty-nine are highly homologous to previously identified genes.

On this chromosome, eight active ARS elements previously reported were revealed to be distributed in 30kb intervals on average. Local sequence motifs were compared to active ARS regions and inactive loci with perfect ARS core sequences to examine the relationship between these motifs and ARS activity. Additional ARS sequences were predominantly observed in the 3' flanking sequences of active ARS loci.

In part 3, I described the sequencing and analysis of a 385 kb genomic segment around the centromeric end of HLA class I region including the boundaries between the class I and III region on human chromosome 6. Twelve contiguous cosmids were chosen for the sequencing by the improved shotgun approach to give a single contig of 385,633 bp from the IkBL gene to 91 kb telomeric of HLA-C gene. This region was confirmed as containing six expressed genes, IkBL, BAT1, MICB, MICA, HLA-B, and HLA-C. Additional five novel genes were identified through homology searches and exon prediction analyses.

According to the sequence analysis, the following two pairs of the genomic segments were duplicated in evolution ; (i) a 43 kb segment including the HLA-B showing the highest genetic polymorphism among the classical HLA class I loci (class Ia) and a 40 kb-segment including the HLA-C locus showing the lowest among those, and (ii) a 52 kb segment including the MIC (MHC class I chain related) B gene and a 35 kb segment including MICA. Repetitive elements such as SINEs, LINEs and LTRs occupy as highly as 47 % of nucleotides in this 385 kb region. This unusually high contents of the elements suggests that the repeat-mediated rearrangements have frequently occurred in the evolutionary history of the HLA class Ia region. Analysis of similar LINE compositions within the two pairs of duplicated segments revealed ; (i) LINEs in these regions had been dispersed prior both to the duplication of the HLA-B and -C loci and to that of the MICB and MICA loci, and (ii) the divergence of the HLA-B and -C loci has occurred prior to the duplication of the MICA and MICB loci.

Those which mentioned above showed that large-scale DNA sequencing and analysis throughout multiple loci play an important role also in elucidating the evolutionary history of eukaryotic genomes.

In conclusion, the present work dealt with the strategy of a high efficiency shotgun sequencing approach which provides a framework for large-scale DNA sequencing of higher eukaryotic genomes. In total, more than 655 kb regions of interest from eukaryotic genomes from two species, *Saccharomyces cerevisiae* and *Homo sapiens* have been completely determined. These special works provided unique and valuable information for genome biology by allowing a view of numerous genetic elements on the genome-wide level. Also some sequences, which were assumed to be important in the features specified as chromosomal replication or genome duplication, were confirmed in the long regions. These findings and further study will give us insight into the dynamic role in maintenance and duplication of life in which the genome plays. Especially in the HLA region, relationship between novel genes identified and disease states in immune systems also urgently needs to be examined. Already, the large-scale sequencing strategy in defined regions has been established and is applicable to any other huge and complex chromosomal region, suggesting a direction for the ultimate goal of genome projects which are making significant contributions to genome biology.

# References

Adams, M.D., Dubnick, M., Kerlavage, A.R., Fields, C. and Venter, J.C.(1993) *Nature Genet.* **4**: 256-267.

Albertella, M,R. and Campbell, R,D. (1994) *Hum. Mol. Genet.* **3**:793-799.

Altscul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.A. (1990) *J. Mol. Biol.* **215**: 403-410.

Amati, B.D. and Gasser, S.M. (1988) *Cell* **57**: 967-978.

Anderson, S. (1981) *Nucl. Acids. Res.* **9**: 3015-3027.

Ando, H., Mizuki, N., Ota, M., Yamazaki, M., Ohno, S., Goto, K., Miyata, Y., Wakisaka, K., Bahram,S. and Inoko,,H. (1997) *Immunogenetics* **46**: 499-508.

Bankier, A.T. and Barrell, B.G. (1989) in *Nucleic Acid Sequencing: A Practical Approach, IRL Press, Oxford, pp37-73..*

Bahram,S., Bresnahm,N., Geraghty,D.E. and Spies,T. (1994) *Proc.Natl.Acad.Sci. U.S.A..* **91**: 6259-6263.

Barham, S. and Spies, T. (1996) *Immunogenetics* **43**: 230-233.

Batzer,M., Deininger,P.L., Heilman-Blungberg,U., Jurka,J., Labuda,D., Rubun,C.M., Schmid,C.W., Zietkiewics,E. and Zuckerkandi, W. (1996) *J.Mol.Evol.***42**: 3-6.

Beck, S., Abdullas, S., Alderton, R.P., Glynne, R.J., Gut, I.G., Hosking, L.K., Jacson, A., Kelly, A., Newell, W.R., Sanseau, P., Radley, E., Thorpe, K.L. and Troesdale, J. (1996) *J. Mol. Biol.* **255**: 1-13.

Britten,J.B., (1997) *Gene* **205**: 177-182.

Bodmer,J.G., Marsh,S.G., Albert,E.D., Bodmer,W.F., Bontro,R.E., Wrlich,H.A., Fauchet,R., Maach,B., Mayt,W.R., Pahram,P., Sasaziki,T., Achreuder,G.M., Strominger,G.L., Svegaard,A. and Terasaki,P.I. (1997) *Tissue Antigen* **49**: 297-321.

Browning,M. and Krausa,P. (1996) *Immunol. Today* **17**: 165-170.

Burge,C. and Karlin,S. (1997) *J.Mol.Biol.* **268**: 78-94.

Bussey, H. *et al.* (1995) *Proc. Natl. Acad. Sci. U.S.A.***92**: 3809-3813.

Campbell, R.D. and Trowsdale, J. (1993) *Map of the human MHC, Immunology Today* **14**: 349-352.

Campbell,C. and Trowsdale,J. (1997) *A map of the human major histocompatibility* complex. *Immunol. Today* **18**: Suppl.

Charlesworth,B., Sniegowski,P. and Stephan,W. (1994) *Nature* **371**: 215-220.

Chen, E., Schulessinger, D. and Kere, J. (1993) *Genomics* **17**: 651-656.

Church, G.M. and Kieffer-Higgins, S., (1988) *Science* **240**: 185-189.

Deininger, P.L. (1983) *Anal. Biochem.* **129**: 216-223.

Deshpande, A.M. and Newlon C.S. (1992) *Molec. cell. Biol.* **12**: 4305-4313.

Diffley, J.F. and Stillman, B. (1988) *Cancer Cells* **6**: 235-243.

Drmanac, R., Strezoska, Z., Paunasko, T., Labat, I., Zeremsky, M., Snobby, J., Funkhouser, W.K., Koop, B., Hood, L. and Crkvenjakov, R. (1993) *Science* **260**: 1649-1652.

Dujon, B. et al. (1994) *Nature* **369**: 371-378.

Dulbecco, R. (1986) *Science* **231**: 1055-1056.

Eki, T., Naitou, M., Hagiwara, H., Abe, M., Ozawa, M., Sasanuma, S., Tsuchiya, Y., Shibata, T., Watanabe, K., Ono, A., Yamazaki, M., Tashiro, H., Hanaoka, F. and Murakami, Y. (1996) *YEAST* **12**: 177-190

Endo,T., Imanishi,T., Gojobori,T. and Inoko,H. (1997) *Gene* **205**: 19-27.

Fanning,T,G. and Singer,M.F. (1987) *Biochem.Biophys.acta.* **910**, 203-212.

Feldman, H. et al. (1995) *EMBO J.* **13**: 5795-5809.

Fraser, C.M., Gocaine, J.D., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J., Fritchman, J.L., Weidman, J.F., Small, K.V., Sandusky, M., Utterback, T.R., Saudek, D.M., Phillips, C.A., Merrick, J.M., Tomb, J.F., Dougherty, B.A., Bott, K.F., Hu, P.C., Lucier, T.S., Peterson, S.N., Smith, H.O., Huchison III, C.A. and Vener, J.C. (1995) *Science* **270**: 397-402.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C. J., Tomb, J.F.,Dougherty, B.A., Merrick, J.M., McKenny, K., Sutton, G., FitzHugh, W., Fraser, C.M., Smith, H.O., Vener, J.C. *et al.* (1995) *Science* **269**: 496-512.

Fukagawa, T., Sugaya, K., Matsumoto, K., Okumura, K., Ando, A., Inoko, H. and Ikemura, T. (1995) *Genomics* **25**: 184-191.

Gardiner-Garden,M. and Frommer,,M. (1987) *J.Mol.Biol.* **196**: 261-282.

Gaudieri, S., Leelayuwat, C., Townend, D.C., Kulski, D.C. and Dawkins, R.L., (1997a)

*J.Mol.Evol.* **44** (suppl 1): S147-S154.

Gaudieri, S., Giles K.M., Kulski, D.C. and Dawkins, R.L., (1997b) *Hereditas.* **127**: 37-46.

Goffeau, A. *et al.* (1997) Nature **387**: *The Yeast Genome Directory.*

Hanahan, D. (1983) *J. Mol. Biol.* **166**: 557-580.

Hattori, M. (1993) in *the Human Genome Mapping Workshop 93*, October, Kobe.

Hoffman, J.F.-X., Laroche, T., Brand, A.H. and Gasser, S.M. (1989) *Cell* **57**: 725-737.

Houten J.V. and Newlon, C.S. (1990) *Molec. cell. Biol.* **8**: 3917-3925.

Hughes,A.L. and Nei,M. (1988) *Nature* **335**: 167-170.

Hughes,A.L. and Nei,M. (1989) *Genetics* **122**: 681-686.

Hunkapiller, T., Kaiser, R.J., Koop, B.F. and Hood, L. (1991) *Science* **254**: 59-67.

Imai, T. and Olson, M.V. (1990) *Genomics* **8**: 297-303

Iwasaki, T., Shirahige, K., Yoshikawa, H. and Ogasawara, N. (1992) *Gene* **11**: 81-87.

Johnston, M. et al. (1994) *Science* **265**: 2077-2082.

Kapitonov, A. and Jurka, J. (1996) *J.Mol.Evol.*

Kasahara,M., Nakaya,J., Satta,Y. and Takahata,N. (1997) *Trends in Genetics* **13**: 90-93.

Keegan,K., Johnson,D.E., Williams,L.T. and Hayman, M.J. (1991) *Proc.Natl.Acad.Sci.USA* **88**: 1095-1099.

Kipling, D. and Kearsey, S. (1990) *Molec. cell. Biol.* **10**: 265-272.

Koop, B.F., Rowen, L., Chen, W-Q., Deshpande, M., Lee, H. and Hood, L. (1993) *BioTechniques* **14**: 442-447.

Koop, B.F., Rowen, L., Wang, K., Kuo, C.L., Seto, D., Lenstra, J.A., Hoeard, S. *et al.* (1994) *Genomics* **19**: 478-493.

Kulski, J.K., Gaudueri,S., Bellgard, L., Giles, K., Inoko, H. and Dawkins, R.L. (1997)*J.M.Evol.* **45**: 599-609

Kuwano, Y. and Wool, I.G. (1992) *Biochem. Biophys. Res. Commun.* **187**: 58-64.

Leelayuwat,C., Abraham,L.J., Tabarias,H., Christiansen,F,T. and Dawkins,R,L. (1992) *Immunogenetics* **36**: 208-212.

Leelayuwat,C., Pinelli,M. and Dawkins R.L. (1995) *J.Immunol.* **155**: 692-698.

Leelayuwat,C., Abraham,L.J., Pinelli,M. Townend,D.C., Wilks,A.F. and Dawkins,R.J. (1996) *Tissue Antigens* **48**, 59-64.

Maniatis, T., Fritsch, E.F. and Sambrook, J. (1989) *"Molecular cloning: A Laboratory*

*Manual", 2nd ed.*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Mazzarella,R. and Schlessinger,D. (1997) *Gene* **205**, 29-38.

Mizuki, N., Ota, M., Kimura, M., Ohno, S., Katsuyama, Y., Ando, H., Yamazaki, M., Watanabe, K., Goto, K., Nakamura, S., Barham, S. and Inoko, H. (1997a) *Proc. Natl. Acad. Sci. U.S.A.* **97**: 1298-1303.

Mizuki, N., Ando, H., Kimura, M. Ohno, S., Miyata, S., Goto, K., Ishihara, M., Yamazaki, M., Watanabe, K., Ono, A., Taguchi, S., Sugawara, C., Fukuzumi,Y., Okumura, K., Goto, K., Ishihara, M., Kikuti, Y.Y., Chen, L., Ando, A., Ikemura, T. and Inimo, H. (1997b) *Genomics* **42**: 55-66.

Mortimer, R.K. et al. (1993) Fungi: S. cerevisiae (Nuclear genes). in *Genetic maps, locus maps of complex genome, Sixth Edition Book 3, Lower eukaryotes (Ed. O'Brien, S.J.) pp.36-56*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor Laboratory, NY.

Murakami, Y., Naitou, M., Hagiwara, H., Shibata, T., Ozawa, M., Sasanuma, S., Tsuchiya, Y., Soeda, E., Yokoyama, K., Yamazaki, M., Tashiro, H. and Eki, T. (1995) *Nature Genet.* **10**: 261-268.

Murray, V. (1989) *Nucl. Acids Res.* **17**: 8889.

Naitou, M., Ozawa, M., Sasanuma, S., Kobayashi, M., Hagiwara, H., Shibata, T., Hanaoka, F. Watanabe, K., Ono, A., Yamazaki, M., Tashiro, H., Eki, T. and Murakami, Y. (1995) *YEAST* **11**: 1525-1532.

Naitou, M., Ozawa, M., Sasanuma, S., Kobayashi, M., Hagiwara, H., Shibata, T., Hanaoka, F. Watanabe, K., Ono, A., Yamazaki, M., Tashiro, H., Eki, T. and Murakami, Y. (1996) *YEAST* **12**: 77-84.

Ohta,T. (1980) *Evolution and Variation of Multigene Families.* Spinger-Verlag, New York.

Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y. and Matsubara, K. (1992) *Nature Genet.* **2**: 173-179.

Olson, M.V. and Link, A.J. (1991a) *Genetics* **127**: 681-698.

Olson, M.V. (1991b) in *the Molecular and Cellular biology of the yeast Saccharomyces: genome dynamics, protein synthesis and energetics* (eds. Broach, J.R., Jones, E.W. and Pringle, J.R.) pp. 1-39, Cold Spring Harbor Laboratory Press, Cold Spring

Harbor, NY.

Olson, M.V., Riles, L. *et al.* (1993) *Genetics* **134**: 81-150.

Oliver, S.G., Aart, Q.J.M., Agostoni-Carbone, M.L., Aigle, M., Alberghina, L., Alexandraki, D., Antoine, G., Anear, R., Ballesta, J.P.G., Benit, P. *et al.* (1992) *Nature* **357**: 38-46.

Pohla, H., Kuon, W., Tabaczewsski, P., Oderner, C.and Weiss, E.H. (1989) *Immunogenetics* **29**: 297-307.

Parham, P. and Ota, T. (1996) *Science* **272**: 67-74.

Pearson, W. (1990) *Rapid and sensitive sequence comparison with FASTP and FASTA. in Methods in Enzymology* **183** (ed. Dolittle, R.F.), 63-98.

Peelman,L., Chardon, P., Nunes, M., Renard, C., Geffrotin, C., Vaiman, M., Strominger, J. and Spies. T. (1995) *Genomics* **26**: 210-218.

Pichon,L., Hampe,A., Giffon,T., Carn,G., Legall,J.Y. and David,V. (1996) *Immunogenetics* **44**: 259-267.

Poncz, M. (1982) *Proc. Natl. Acad. Sci. U.S.A.* **79**: 4298-4205.

Riles, L. *et al.* (1993) *Genetics* **134**: 81-150.

Roach, J.C. (1995) *Genome Res.* **5**: 464-473.

Saito, I. and Stark, G.R. (1986) Proc. *Natl. Acad. Sci. U.S.A.* **83**: 8664-8668.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* **74**: 5363-5367.

Schmid,C.W. (1996) *Progress in Nucleic Acid Research and Molecular Biology* **53**: 283-319.

Shiina,T., Tamiya,G., Oka,A., Yamagata,T., Yamagata,N., Kikkawa E., Goto,K., Mizuki,N., Watanabe,K., Fukuzumi,Y., Taguch,S., Sugawara,C., Ono,A., Chen,L., Yamazaki,M., Tashiro,H., Ando,A., Ikemura,T., Kimura,M. and Inoko,H. (1998) *Genomics* **47**: 372-382.

Shirahige, K., Iwasaki, T., Rashid, M.B., Ogasawara, N. and Yoshikawa, H. (1993) *Molec. Cell. Biol.* **13**: 5043-5056.

Shore, D., Stillman, B.J., Brand, A.H. and Nasmyth, K.A. (1987) *EMBO J.* **6**: 461-467.

Sidney,J., Gray,H.M., Kudo,T. and Sette,A. (1996) *Immunology Today* **17**: 261-266.

Smit,A.F.A. and Riggs,A.D. (1995) *Nucl.Acids.Res.***23**: 98-102.

Spitzner, J.R. and Muller, M.T. (1988) *Nucl. Acids Res.* **16**: 5533-5555.

Takahashi, E., Hori, T., O'Connel, P., Leppert, M. and White, R. (1990) *Hum. Genet.* **86**: 14-16.

Takahashi, E., Yamauchi, M., Tsuji, H., Hitomi, H., Meuth, M. and Hori, T. (1991)*Hum. Genet.* **88**: 119-121.

Tamiya,G., Ota,M., Katsuyama,Y., Shiina,T., Oka,A., Makino,S., Kimura,M. and Inoko,H. (1997) *Tissue Antigens in press.*

Tateno, Y. and Gojobori, T. (1997) *Nucleic. Acids Res.* **25**: 14-17.

Tiwari, J.L. and Terasaki, P.I. (1985) *HLA and Disease Associations*, Springer-Verlag, New York.

Toda, Y. and Tomita, M. (1997) *Gene* **205**: 173-176.

Uberbacker, E.C. and Mural, R.J. (1991) *Proc. Natl. Acad. Sci. U.S.A.* **88**: 11262-11264.

Venditti, C.P., Harris, J.M. and Geraghty, M.J. (1994) *Genomics* **22**: 257-266.

Vernet, C., Ribouchon, M-T., Chimini, G., Jouanolle, A-M., Sidibe, I. and Pontarotti P. (1993) *Immunigenetics* **38**: 47-53.

Watanabe,H., Mori,H. and Gojobori,T. (1997) *J.Mol.Evol.* **44** (Suppl. 1): S57-S64.

Watanabe,K., Sugawara,C., Ono, A., Fukuzumi,Y., Yamazaki,M., Tashiro,H. and Nomura,T. (1998) *Genomics* **52**: 95-100.

Wilson, R.K., Koop, B.F., Chen, C., Halloran, N., Sciammis, R. and Hood, L. (1992) *Genomics* **3**: 1198-1208.

Yamazaki, M., Yokoyama, K., Yasunaga, T., Akaogi, K. and Soeda, E. (1988) in *the eighteenthth annual meeting of the Molecular Biology Society of Japan*; *Genetic Analysis of Human cDNA clones* , December, Tokyo.

Yamazaki, M., Akaogi, K., Miwa, T., Imai, T., Soeda, E. and Yokoyama, K. (1989) *Nucleic Acids Res.* **17**: 7108.

Yamazaki, M., Tashiro, H., Yokoyama, K. and Soeda, E. (1990) *Agric. Biol. Chem.* **54**: 3163-3170.

Yamazaki, M., Ono, A., Watanabe, K., Sasaki, K., Tashiro, H. and Nomura, T. (1995) *DNA Res.* **2**: 187-189.

Yamazaki, M., Tateno, Y. and Inoko, H. (1999) *J. Mol. Evol.* **48**:*in press.*

Zorio, D.A.R., Cheng, N.N., Blumenthal, T. and Spieth, J. (1994) *Nature* **372**: 270-272.