1995

Evolutionary Dynamics At Two Loci Of The Human Genome As Assessed By Examination Of Nucleotide Sequence Diversity And Organization

Kathleen Allen Hill

Evolutionary Dynamics at Two Loci of the Human Genome as Assessed by Examination of Nucleotide Sequence Diversity and Organization

by

Kathleen Allen Hill

Department of Zoology

Submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy

Faculty of Graduate Studies
The University of Western Ontario
London, Ontario
July, 1995

ISBN 0-612-03460-7

Canada

Name KATHLEEN A HILL

Dissertation Abstracts International is arranged by broad, general subject categories. Please select the one subject which most nearly describes the content of your dissertation. Enter the corresponding four-digit code in the spaces provided.

BIOLOGY GENETICS

SUBJECT TERM

0369

SUBJECT CODE

U·M·I

Subject Categories

THE HUMANITIES AND SOCIAL SCIENCES

COMMUNICATIONS AND THE ARTS

Architecture 0729
Art History 0377
Cinema 0900
Dance 0378
Fine Arts 0357
Information Science 0723
Journalism 0391
Library Science 0399
Mass Communications 0708
Music 0413
Speech Communication 0459
Theater 0465

EDUCATION

General 0515
Administration 0514
Adult and Continuing 0516
Agricultural 0517
Art 0273
Bilingual and Multicultural 0282
Business 0688
Community College 0275
Curriculum and Instruction 0727
Early Childhood 0518
Elementary 0524
Finance 0277
Guidance and Counseling 0519
Health 0680
Higher 0745
History of 0520
Home Economics 0278
Industrial 0521
Language and Literature 0279
Mathematics 0280
Music 0522
Philosophy of 0998
Physical 0523
Psychology 0525
Reading 0535
Religious 0527
Sciences 0714
Secondary 0533
Social Sciences 0534
Sociology of 0340
Special 0529
Teacher Training 0530
Technology 0710
Tests and Measurements 0288
Vocational 0747

LANGUAGE, LITERATURE AND LINGUISTICS

Language
- General 0679
- Ancient 0289
- Linguistics 0290
- Modern 0291

Literature
- General 0401
- Classical 0294
- Comparative 0295
- Medieval 0297
- Modern 0298
- African 0316
- American 0591
- Asian 0305
- Canadian (English) 0352
- Canadian (French) 0355
- English 0593
- Germanic 0311
- Latin American 0312
- Middle Eastern 0315
- Romance 0313
- Slavic and East European 0314

PHILOSOPHY, RELIGION AND THEOLOGY

Philosophy 0422
Religion
- General 0318
- Biblical Studies 0321
- Clergy 0319
- History of 0320
- Philosophy of 0322

Theology 0469

SOCIAL SCIENCES

American Studies 0323
Anthropology
- Archaeology 0324
- Cultural 0326
- Physical 0327

Business Administration
- General 0310
- Accounting 0272
- Banking 0770
- Management 0454
- Marketing 0338

Canadian Studies 0385
Economics
- General 0501
- Agricultural 0503
- Commerce-Business 0505
- Finance 0508
- History 0509
- Labor 0510
- Theory 0511

Folklore 0358
Geography 0366
Gerontology 0351
History
- General 0578
- Ancient 0579
- Medieval 0581
- Modern 0582
- Black 0328
- African 0331
- Asia, Australia and Oceania 0332
- Canadian 0334
- European 0335
- Latin American 0336
- Middle Eastern 0333
- United States 0337

History of Science 0585
Law 0398
Political Science
- General 0615
- International Law and Relations 0616
- Public Administration 0617

Recreation 0814
Social Work 0452
Sociology
- General 0626
- Criminology and Penology 0627
- Demography 0938
- Ethnic and Racial Studies 0631
- Individual and Family Studies 0628
- Industrial and Labor Relations 0629
- Public and Social Welfare 0630
- Social Structure and Development 0700
- Theory and Methods 0344

Transportation 0709
Urban and Regional Planning 0999
Women's Studies 0453

THE SCIENCES AND ENGINEERING

BIOLOGICAL SCIENCES

Agriculture
- General 0473
- Agronomy 0285
- Animal Culture and Nutrition 0475
- Animal Pathology 0476
- Food Science and Technology 0359
- Forestry and Wildlife 0478
- Plant Culture 0479
- Plant Pathology 0480
- Plant Physiology 0817
- Range Management 0777
- Wood Technology 0746

Biology
- General 0306
- Anatomy 0287
- Biostatistics 0308
- Botany 0309
- Cell 0379
- Ecology 0329
- Entomology 0353
- Genetics 0369
- Limnology 0793
- Microbiology 0410
- Molecular 0307
- Neuroscience 0317
- Oceanography 0416
- Physiology 0433
- Radiation 0821
- Veterinary Science 0778
- Zoology 0472

Biophysics
- General 0786
- Medical 0760

EARTH SCIENCES

Biogeochemistry 0425
Geochemistry 0996
Geodesy 0370
Geology 0372
Geophysics 0373
Hydrology 0388
Mineralogy 0411
Paleobotany 0345
Paleoecology 0426
Paleontology 0418
Paleozoology 0985
Palynology 0427
Physical Geography 0368
Physical Oceanography 0415

HEALTH AND ENVIRONMENTAL SCIENCES

Environmental Sciences 0768
Health Sciences
- General 0566
- Audiology 0300
- Chemotherapy 0992
- Dentistry 0567
- Education 0350
- Hospital Management 0769
- Human Development 0758
- Immunology 0982
- Medicine and Surgery 0564
- Mental Health 0347
- Nursing 0569
- Nutrition 0570
- Obstetrics and Gynecology 0380
- Occupational Health and Therapy 0354
- Ophthalmology 0381
- Pathology 0571
- Pharmacology 0419
- Pharmacy 0572
- Physical Therapy 0382
- Public Health 0573
- Radiology 0574
- Recreation 0575
- Speech Pathology 0460
- Toxicology 0383

Home Economics 0386

PHYSICAL SCIENCES

Pure Sciences
Chemistry
- General 0485
- Agricultural 0749
- Analytical 0486
- Biochemistry 0487
- Inorganic 0488
- Nuclear 0738
- Organic 0490
- Pharmaceutical 0491
- Physical 0494
- Polymer 0495
- Radiation 0754

Mathematics 0405
Physics
- General 0605
- Acoustics 0986
- Astronomy and Astrophysics 0606
- Atmospheric Science 0608
- Atomic 0748
- Electronics and Electricity 0607
- Elementary Particles and High Energy 0798
- Fluid and Plasma 0759
- Molecular 0609
- Nuclear 0610
- Optics 0752
- Radiation 0756
- Solid State 0611

Statistics 0463

Applied Sciences
Applied Mechanics 0346
Computer Science 0984
Engineering
- General 0537
- Aerospace 0538
- Agricultural 0539
- Automotive 0540
- Biomedical 0541
- Chemical 0542
- Civil 0543
- Electronics and Electrical 0544
- Heat and Thermodynamics 0348
- Hydraulic 0545
- Industrial 0546
- Marine 0547
- Materials Science 0794
- Mechanical 0548
- Metallurgy 0743
- Mining 0551
- Nuclear 0552
- Packaging 0549
- Petroleum 0765
- Sanitary and Municipal 0554
- System Science 0790

Geotechnology 0428
Operations Research 0796
Plastics Technology 0795
Textile Technology 0994

PSYCHOLOGY

General 0621
Behavioral 0384
Clinical 0622
Developmental 0620
Experimental 0623
Industrial 0624
Personality 0625
Physiological 0989
Psychobiology 0349
Psychometrics 0632
Social 0451

ABSTRACT

Accurate and comprehensive measurement of the extent and pattern of nucleotide diversity is necessary to refine theories on the dynamics of evolution. It is possible to determine the sequence of any genomic region for numerous individuals using the polymerase chain reaction (PCR) and associated techniques of DNA sequence analysis. The two regions of the human genome examined in this study were the third exon of the highly conserved (two major alleles) alcohol dehydrogenase, *Adh2* locus and the second exon of the highly polymorphic (26 alleles) human leukocyte antigen, *HLA-DQβ1* gene. Sequence information was determined from 25 individuals from Southwestern Ontario and 26 Dogrib individuals from the Northwest Territories of Canada. The Southwestern Ontario population is heterogeneous in ancestry and predominantly European while the Dogrib population is homogeneous and of Asian ancestry. Intra-allelic nucleotide diversity was characterized at two regions of the genome in two different human populations using PCR with direct sequencing and chaos representation of sequence organization.

No intra-allelic variation was observed in the 39,000 nucleotides examined for both exons. There was no evidence for higher substitution rates for highly polymorphic loci. The maintenance of a large number of alleles at the *HLA-DQβ1* locus in populations is attributed to selective forces, in particular heterozygote advantage, while admixture and stochastic forces such as founder effects, and bottlenecks could account for observed population-specific allele frequencies at the two loci. Nucleotide diversity was nonrandom and influenced by nearest-neighbor nucleotide associations. Dinucleotide representation also accounted for the major features of the global organization in DNA sequences. Analysis of 56

large sequences from 10 species, 28 mitchondrial DNAs and 31 viral genomes identified, for the first time, that the global structure of DNA is under selective constraints that are genome-type specific and related to an as yet unknown force(s) or factor(s).

Selection plays a predominant role in determining the gene-specific variability and genome-type specific sequentiality of DNA. Thus, the evolution of the DNA sequence of a gene or genome should be viewed in the dual context of the constraints on its specific function and the genome-type specific global organization.

ACKNOWLEDGMENTS

I would like to thank my supervisor. Dr. S.M. Singh and the members of my advisory committee, Dr. M. Clarke, Dr. M. Coulter-Mackie, Dr. G. Kidder and Dr. D. McMillan for their generous assistance throughout the course of my research program.

I thank Dr. E. Szathmary, Department of Anthropology, McMaster University, Hamilton, Ontario and Dr. George, University Hospital, London Ontario for their generous gifts of DNA samples. I am appreciative of the individuals who participated in the study for their interest in contributing to our knowledge regarding human population genetics. I gratefully acknowledge the computer expertise of A. Devito, G. Stafleu, R. Govindarajan and N. Schisler. I am grateful to Roger Frappier for generously providing his photographic skills. Special thanks to Mary Martin of the Zoology Department for all her help and kindness. I also thank many co-workers who have helped me including David Ribble, Steven Chao, Steven Gallant, Mike Topping, Andor Kiss, Michael Coulthart, Lisa Bogue and Rebecca Ott.

The friendship of Grace Trentin and Indira Pillay has been supremely supportive. I have learned from Grace the joy in tackling life with an aerobics instructor's enthusiasm. Indira's genuine concern for the welfare of others is inspirational.

I am sincerely appreciative of the friendship and enthusiastic encouragement of Dr. Shari Bond. Shari has been a mentor, guiding me with her experiences, reviewing the thesis and 'being me' in London while I was living in Rochester. I enjoy the friendship of Shari's family, Ian, Alexandra and Buttons. Special thanks to Allie for trying to read the 'A's of the sequencing films.

I am truely blessed to have a family who have been a constant source of encouragement. I and my friends wish to thank my parents for the "IUTS Care Packages". These packages provided the essentials for completing graduate work such as chocolates, fudge, cookies and words of encouragement.

Financial support was provided by an operating grant to S.M. Singh from The Natural Sciences and Engineering Research Council of Canada (N.S.E.R.C.). K. A. Hill was supported by scholarships from N.S.E.R.C., The Canadian Federation of University Women and a Graduate Research Fellowship from The University of Western Ontario.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

GLOSSARY

5'	upstream of the nucleotide sequence of interest
3'	downstream of the nucleotide sequence of interest
ADH	alcohol dehydrogenase enzyme
Adh	alcohol deydrogenase gene
Adh1	gene coding for the α ADH subunit in humans
Adh2	gene coding for the ß ADH subunit in humans
Adh2 ß1	typical allele at the alcohol dehydrogenase locus 2
Adh2 ß2	atypical allele at the alcohol dehydrogenase locus 2
bp	base pair
°C	degree Celsius
cDNA	complementary DNA
Ci	Curie
cpm	counts per minute
ddATP	2', 3'-dideoxyadenosine 5'-triphosphate
ddCTP	2', 3'-dideoxycytidine 5'-triphosphate
ddGTP	2', 3'-dideoxyguanosine 5'-triphosphate
ddNTP(s)	2', 3'-dideoxynucleoside 5'-triphosphate(s)
ddTTP	2', 3'-dideoxythymidine 5'-triphosphate
dATP	2'-deoxyadenosine 5'-triphosphate
dCTP	2'-deoxycytidine 5'-triphosphate
dGTP	2'-deoxyguanosine 5'-triphosphate
DNA	2'-deoxyribonucleic acid
dNTP(s)	2'-deoxynucleoside 5'-triphosphate(s)
ds	double-stranded deoxyribonucleic acid
DTT	dithiothreitol
dTTP	2'-deoxythymidine 5'-triphosphate
EDTA	ethylenediaminetetraacetic acid
HLA	human leukocyte-associated antigen
HLA-DQß1	human leukocyte antigen locus encoding the DQß1 subunit
HLA-DQß2	human leukocyte antigen locus, the pseudogene DQß2
kb	kilobase
mtDNA	mitochondrial deoxyribonucleic acid

nt(s)	nucleotide(s)
PCR	polymerase chain reaction
ss	single-strand deoxyribonucleic acid
SSCP	single-strand conformation polymorphism
Tris	tris(hydroxymethyl)aminomethane
Tris·Cl	tris(hydroxymethyl)aminomethane hydrochloride

Chapter 1

INTRODUCTION

1.1 The Measurement of Genetic Variation

Measurement of genetic variation is essential to obtain insight into many issues of population genetics and evolutionary change. In the past, genetic variation was measured by examining inherited phenotypic differences, specifically examined were alterations in morphology (Ford, 1940), cytology (Dobzhansky, 1941), electrophoretic isozymes (Lewontin and Hubby, 1966), immunological components (Sarich and Wilson, 1966) and amino acid sequences (Thatcher, 1980). Many populations, including humans, contain considerable genetic variability (Harris, 1966; Lewontin, 1974). The average heterozygosity for *Homo sapiens* estimated from analysis of electrophoretic isozymes ranges from 7.4 to 14% (based upon 87 and 121 loci; Harris et al., 1977 and Nei and Graur, 1984, respectively). Sequence variation in noncoding regions, such as introns and synonymous codon positions cannot be measured by examining phenotypes, is presumably not subject to natural selection acting through the phenotype of the encoded protein, and is essential for estimating the frequency of variation due to neutral nucleotide substitutions. The problem remains that the total extent of genetic variation cannot be extrapolated from analysis of phenotypic differences.

Ideally, estimates of genetic variation should be based on a direct examination of nucleotide sequences where its extent, including neutral substitutions, can be examined. Generally, measurements of nucleotide sequence variation have been incomplete and are based primarily upon the

presence and absence of recognition sites for restriction endonucleases (Ferris et al., 1981). Such analyses merely identify the alteration of a nucleotide site or sites and do not elucidate the nature of the nucleotide differences. Determination of the degree of nucleotide sequence diversity and the type and pattern of sequence differences are also important for understanding the role of mutation, natural selection and genetic drift in the origin and maintenance of genetic variation.

No extensive study of diversity in the human nuclear genome has been conducted in any population except for a survey by Li and Sadler (1991) of published (Bilofsky and Burks, 1988) cDNA and genomic sequences for 49 human loci. Nucleotide differences are not distributed randomly among the genes nor along the length of the nucleotide sequences examined. The maximum nucleotide diversity is 0.11% and is one order of magnitude lower than that observed in *Drosophila* populations (2.2%, Aquadro, 1991; Li and Sadler, 1991). This level of nucleotide diversity translates into an average expected heterozygosity of 20.4% at the protein level and 7.4% at the electrophoretic level (Li and Sadler, 1991). This measure of nucleotide sequence diversity is affected by errors in the information contained in DNA sequence databanks. The measure is also affected by the biased nature of the information contained in sequence databanks, which is limited to analysis of North American Caucasian populations and genes of medical importance (Li and Sadler, 1991). There is a need for accurate determination of nucleotide sequence information and direct examination of nucleotide sequences in different regions of the human nuclear genome and in different populations.

Enzymatic amplification of genomic DNA *in vitro* by the polymerase chain reaction (PCR) has simplified the analysis of genetic diversity when used in combination with several techniques for the detection of sequence differences.

One method involves detection of the RNase cleavage of a labeled RNA probe at mismatched positions with a target DNA (or RNA) sequence (Meyers et al., 1985; Winter et al., 1985). A chemical cleavage method (Cotton et al., 1988; Smooker and Cotton, 1993) cuts mismatched sites in a DNA probe following the heteroduplex formation of the probe and the target DNA (or RNA). The probe is first modified with osmium tetroxide for T and C mismatches or with hydroxylamine for C mismatches and then incubated with piperideine to cleave the probe DNA at the modified bases. Denaturing gradient gel electrophoresis subjects double-stranded DNA to an increasing gradient of denaturant and permits identification of heteroduplexes between mutant and wild-type DNA fragments (Abrams et al., 1990; Keen et al., 1991; White et al., 1992; Cariello and Skopek, 1993). The mismatching between the mutant and wild-type DNA strands results in aberrant migration due to an altered melting temperature. Hybridization of immobilized genomic DNA or PCR products with chemically synthesized oligonucleotide probes permits identification of a single nucleotide change (Conner et al., 1983). The allele-specific oligonucleotides hybridize poorly to a target molecule having a single internal mismatch. Analysis of single-strand conformation polymorphism (SSCP) is based on the principle that single-stranded DNA molecules take on specific sequence-based secondary structures (conformers) under nondenaturing conditions. Molecules differing by as little as a single base substitution may form different conformers and migrate differently in a nondenaturing polyacrylamide gel. (Orita et al., 1989a, b; Mashiyama et al., 1990; Carrington at al., 1992; Sekiya,1993). These methods permit detection of sequence differences but still represent a partial analysis of nucleotide sequence variation as they do not permit a complete description of both the number and type of sequence alterations.

The polymerase chain reaction (PCR) and direct nucleic acid sequencing permit the rapid analysis of DNA sequence differences and overcome the limitations of other techniques for the detection of nucleotide polymorphism (Gyllensten and Erlich, 1988; Hunkapiller, 1991). The difficulty associated with direct sequencing of PCR-amplified products is the poor resolution achieved with double-stranded DNA templates. Heterozygous templates give ambiguous information and purification of single-stranded templates is time and labor intensive. The cloning of PCR-generated DNA fragments permits the sequencing of single-stranded templates and the isolation of the different alleles in heterozygous individuals (Scharf et al., 1986; Ichinose et al., 1991). This approach permits the unambiguous identification of the nucleotide sequence but is time consuming and error prone. Errors made by the *Taq* polymerase (Keohavong et al., 1993) during the PCR are isolated and amplified in the cloning of the PCR product and necessitate the sequence determination of at least two independent clones. Asymmetric PCR amplifies preferentially one of the complementary DNA strands and direct sequencing of this single-stranded template avoids identification of errors generated by the *Taq* polymerase by sequencing the pool of PCR-amplified fragments (Innis et al., 1988; Liu et al., 1993). Sequencing of double-stranded PCR templates has been plagued by the rapid reassociation of the complementary DNA strands of the template and inhibition of annealing by the sequencing primer. Typically, use of sequencing primers nested in relation to the PCR primers have been used to improve the sequencing of double stranded templates (Yandell and Dryja, 1989; Engelke et al., 1988; Wong et al, 1987; Wrischnik et al., 1987).

Cycle sequencing overcomes the difficulties associated with the direct sequencing of double-stranded templates without the use of nested primers by repeating the melting, annealing and extension phases of the sequencing

reaction as many as 20 to 30 times (Adams and Blakesley, 1993). The thermophilic *Taq* polymerase used in cycle sequencing also permits the use of higher annealing and extension temperatures that assist annealing of the sequencing primer by inhibiting the rapid reannealing of short double-stranded sequencing templates (Murray, 1989; Sarkar et al., 1993). Direct sequencing of the double-stranded PCR product often requires extensive purification of the sequence template (Kretz et al., 1989; Aguilar-Cordova and Lieberman, 1991; Anderson et al., 1993; Dowton and Austin, 1993; Harvey et al., 1993; Moncany and Keller, 1993) and does not permit the identification of the allele to which a novel mutation is linked in heterozygous individuals. Direct sequencing of heterozygous templates does not permit identification of the phase of two or more novel mutations, that is whether multiple mutations are associated with the same or different alleles. Despite their limitations, the PCR-based methods for the detection of differences in nucleotide sequences enable direct examination of different regions of the human nuclear genome. Direct sequencing permits measurement of the total genetic variation and provides a description of the pattern and nature of sequence differences.

1.2 Sampling Considerations in the Study of Genetic Variation Among Humans

Differences among sequences are not randomly distributed along the sequence length (Charlesworth, 1994; Fickett et al., 1992; Hess, et al. 1994; Sharp and Lloyd, 1993). Sequence composition is affected by such genome-specific features as high mutation rates associated with methylated cytosine residues in vertebrates (Beutler et al., 1989; Bird, 1980; Ehrlich and Wang, 1981; Tasheva and Roufa, 1993) nonrandom dinucleotide to tetranucleotide

composition (Nussinov, 1980, 1981a, b, 1984; Phillips et al., 1987; Volinia et al., 1989) and biases in synonymous codon usage (Grantham et al., 1980, 1981; Kemura, 1985; Wain-Hobson et al., 1981). The extent of nucleotide variation at different positions of codons and among different functional domains is related to functional constraints of the gene product and is thus gene-specific. Therefore, the extent and location of sequence differences is determined by unequal probabilities of the occurrence of base alterations at specific sites and genome- and gene-specific constraints upon the maintenance of such alterations. Total measurement of genetic variation at different regions of the human genome is essential to refine theories that pertain to the accumulation of mutations and the maintenance of genetic variation.

Stratified random sampling has been proposed for surveying the human genome for novel nucleotide sequence variation (Cavalli-Sforza, 1990). In this process the genome is divided into layers that have a different degree of interest and demand different efforts for study (exons, introns, promoters, enhancers, repeated sequences, etc.) and sampling is randomized within each layer. Individuals within each group (subpopulations, populations, species, etc.) are selected at random since stratification of the subject groups to be screened for variability is based upon genetic relatedness. The number and the genetic relatedness of the individuals to be screened and the length and the degree of conservation of the DNA segments to be sequenced determines the type of information obtained about nucleotide sequence variation. This strategy has not yet been applied in a systematic analysis of nucleotide sequence variation in different regions of the human nuclear genome for different populations.

1.3 Two Distinct Human Populations

Different human populations are characterized by differences in geographic location, population structure and/or evolutionary history. In the present study genetic variation was examined in a random collection of individuals from two distinct populations. Southwestern Ontario, has a large human population of heterogeneous ancestry, mostly European. In contrast, the Athabaskan Dogrib population from the Northwest Territories represents a smaller and more homogenous collection of individuals whose ancestry is mainly Oriental with a limited European admixture (0.082, Szathmary and Ossenberg, 1978; Szathmary, 1978, 1993). The size of the founder populations and the number and time of founding events, migration rates and population size are some of the parameters that are unique among these two human populations. The differences in the structure and evolutionary history of the populations are useful in examining the forces responsible for the absence or maintenance of genetic variation.

1.4 Two Distinct Regions of the Human Nuclear Genome

The alcohol dehydrogenase (*Adh*) multigene family in humans is a member of the evolutionary group of ADH enzymes that have a long chain (375 amino acids), require zinc as a cofactor (Jornvall et al., 1984; Jornvall, 1985) and are highly conserved among phylogenetically diverse species (Sun and Plapp, 1992; Yokoyama and Harry, 1993; Yokoyama et al., 1990). Study of the enzyme's function have included examination of the protein's three-dimensional structure in two species, horse and human (Eklund et al., 1976; Hurley et al., 1991, respectively). Mammalian ADHs in comparison with ADHs of maize and

yeast species have an amino acid sequence similarity of 50% and 20%, respectively (Eklund et al., 1976; Jornvall et al., 1987). ADH is representative of the collection of general "house-keeping enzymes" important in detoxification, and specifically in the case of ADH, the metabolism of alcohol.

The *Adh* multigene family in humans, as in other species, is highly conserved implying that most amino acid changes in ADH would be selectively deleterious (Kreitman, 1983; Jornvall, 1985; Sun and Plapp, 1992; Yokoyama and Harry, 1993). The zinc-containing ADHs are dimeric enzymes, and in humans, the three classes (I, II and III) are distinguished by their ability to form intergenic heterodimers, randomly. Class I subunits α, β and γ are encoded by three distinct loci *Adh1*, *Adh2* and *Adh3*, respectively and these three genes are highly homologous (93 to 96% identity in amino acid sequence comparisons, Ikuta et al., 1985, 1986; von Bahr-Lindstrom et al., 1986). No variability has been identified at the *Adh1* locus but three and two alleles exist at the *Adh2* and *Adh3* loci, respectively.

Two alleles (the "typical" β_1 and "atypical" β_2 allele) at the *Adh2* locus predominate in human populations (Bosron and Li, 1986; Duester et al., 1986). The alleles differ in a single base change in the third exon (Jornvall et al., 1984; Ikuta et al., 1985; Heden et al., 1986) and the specific activity of the $\beta_2\beta_2$ enzyme is approximately 100 times higher than that of $\beta_1\beta_1$ at physiological pH (Bosron et al., 1985; Yoshida et al., 1981; Hurley et al., 1991). Most Japanese individuals have the "atypical" β_2 subunit while the "typical" β_1 subunit is predominant in English populations. The frequency of the "atypical" *Adh2* β_2 is >70% in the Japanese and <10% in the English (Stamatoyannopoulos et al., 1975). The third exon of *Adh2*, which harbors the β_1 and β_2 alleles, has been PCR-amplified and distinguished using allele-specific oligonucleotide probes directed at the single base-pair differences (Gennari, et al., 1988; Xu et al.,

1988). The well recognized and population-specific frequencies of the *ß1* and *ß2* alleles of the *Adh2* locus have been the focus of attention in studies of human population genetics, alcohol sensitivity and alcoholism for several decades (Stamatoyannopoulos et al., 1975; Thomasson et al., 1991; Agarwal and Goedde, 1987). This region of the human genome, with its established population-specific molecular markers, has the potential to provide insight into the pattern of divergence, including allele-specific mutational events. As well, intra-allelic variation, additional DNA sequence variation within allelic types detected previously by protein electrophoresis or immunogenetics, could be examined.

Human leukocyte antigen (*HLA*) loci represent a different region of the human genome from that characterized by the *Adh* multigene family both in terms of evolutionary history and genetic variation. HLA molecules are highly polymorphic glycoproteins that are normally expressed on the surface of B cells and antigen-presenting cells of the immune system (Parham and Strominger, 1982; Trowsdale and Powis, 1992). Class II HLA molecules are heterodimers composed of noncovalently associated subunits that are encoded by separate genes (α and *ß* genes, Klein et al., 1990; Parham and Strominger, 1982). There are five class II *HLA* regions (*DP*, *DN*, *DO*, *DQ* and *DR*) generally containing one or more α genes and one or more *ß* genes. The polymorphism of *HLA* loci is unique in its extent and in the genetic distance between individual alleles (Hughes and Nei, 1990; Dupont, 1990). There are at least 26 alleles at the *HLA-DQß1* locus (Bodmer et al., 1994). The highly polymorphic second exon of *HLA-DQß1* contains the antigen-binding site codons and has been PCR-amplified from genomic DNAs and restriction endonuclease digested (Trucco et al., 1989) or sequenced (Santamaria et al., 1992) for the rapid identification of *HLA-DQß1* alleles.

The polymorphism of the *HLA* gene complex is characterized by two key features: the existence of a large number of alleles in a single population that occur at appreciable frequencies and a high degree of genetic diversity between the alleles (Klein, 1986, 1987; Gyllensten et al., 1990; Hughes and Nei, 1990). The evolutionary rate at such loci is not higher than that of most other loci (Hayashida and Miyata, 1983; Klein, 1986). The maintenance of diversity is explained by mechanisms such as gene conversion (Weiss et al., 1983) that lead to exchanges of DNA from one gene to another. There is an alternative explanation for the diversity at *HLA* loci. The alleles existed prior to the inception of the species and are transmitted in a trans-species process in which a group of major alleles is passed on in the phylogeny from one species to another, with subsequent accumulation of further mutations (Figueroa et al., 1988; Gaur et al., 1992). Under this hypothesis it is expected that certain alleles from one species are more closely related to certain alleles from other species than they are to each other. There is much evidence for such a hypothesis stemming from analysis of alleles at *HLA* loci in other primates (Lawlor et al., 1988; Erlich and Gyllensten, 1989; Fan et al., 1989). No between-species diversity has been found to be greater than within-species diversity in examination of second exons of seven *DQB1* alleles from five to 13 nonhuman primate species (Otting et al., 1992). There is a preponderance of nonsynonymous nucleotide substitutions at antigen-binding site codons among alleles of different primate species. This pattern of mutation is in contrast to the pseudogene *DQB2*. The long persistence of allelic lineages, the prevalence of nonsynonymous over synonymous substitutions in the peptide-binding region and the greater variation at this locus than at a related pseudogene are indicative of balancing selection relating to antigen presentation (Satta et al., 1994). The diversity at *HLA* loci is useful in

clarification of human evolution and the historical relationships between different human populations (Klein et al., 1990; Riley and Olerup, 1992).

The major genetic difference between the *Adh* and *HLA* gene families is in the degree of polymorphism. It appears that the evolution and molecular properties of the human *Adh* and *HLA* genes are similar in that they both represent multigene families and at least some of the loci in each case are shared in different species. Also, established alleles of the two families are shared across most human populations with the frequency of the alleles being population-specific. However, *Adh2* has two major alleles and *HLA-DQß1* has 26 known alleles in human populations. What determines such differences among loci represents a fundamental question in evolutionary genetics and forms a general objective of this research. Different selection regimes have been used to explain the nature and extent of polymorphism in the two gene families. Experimentation is necessary to test directly this hypothesis but is beyond present capabilities. Thus any argument towards this end must be made by indirect and associative observations. It is also possible that the differences in polymorphism at these two loci are due to differences in mutation rates (Wu and Maeda, 1987). Given that the substitution rate involving the four bases is known to be different and that this difference may depend on sequence-specific molecular properties (Beutler et al., 1989; Blake et al., 1992) differences in primary sequence organization and associated mutational biases may account for the differences in polymorphism and may be reflected in the pattern of substitutional differences among the allele sequences. Analysis of sequence specific-mutational biases require characterization of the primary DNA sequence organization of genes and entire genomes.

1.5 Primary DNA Sequence Organization

Primary DNA sequence organization is defined by the frequency and arrangement of the four nucleotides, adenine, cytosine, guanine and thymine. The organization of nucleotides contains significant information but investigation of this phenomenon has been limited to examination of short stretches of a few tens to hundreds of nucleotides at a time and to the examination of transcription processes, regulatory signals, repetitive patterns (Lefevre and Ikeda, 1994), nucleotide frequencies (Nussinov, 1981a, b) and the distinctions between introns and exons (Mani, 1992a, b; Solovyev, 1993; Solovyev et al., 1992). At a macro level, where a gene sequence is considered in its totality including all exons and introns and even adjacent intergenic regions and neighboring genes, global characteristics of sequence organization are considered, i.e, the overall patterns to the arrangement of nucleotides. Primary organization of DNA has a global structure that is characterized by unequal frequencies of the four nucleotides and/or a nonrandom arrangement of nucleotides (Jeffrey, 1990; Rogerson, 1991; Oliver et al., 1993; Voss, 1993a, b; Solovyev, 1993).

Global structure describes a higher-order organization that is independent of the length and the functional properties of the DNA (Mani, 1992a, b) and is similar on both strands of the DNA double helix (Rogerson, 1989, 1991). Global DNA sequence organization has been examined using a variety of approaches (Nandy, 1994), such as the examination of short-sequence composition (Burge et al., 1992; Karlin and Brendel, 1993; Karlin et al., 1993; Rogerson, 1991), chaos patterns (Jeffrey, 1990; Oliver et al., 1993), fractal landscapes (Buldyrev et al 1993; Voss, 1993a, b) and long range correlations (Peng et al., 1992). DNA sequences appear to have a higher-order structure that is self-similar at numerous different length scales indicating the existence of a fractal-like nature

(Voss, 1993a, b; Nussinov, 1993; Solovyev 1993) or long-range order (Peng et al., 1992; Karlin and Brendel, 1993).

Different patterns of global DNA sequence organization have been identified in different species (Nussinov, 1984; Jeffrey, 1990; Rogerson 1991; Burge et al., 1992; Voss, 1993a b; Karlin and Brendel, 1993). The existence of global structure in DNA sequences and different structures in species representing different evolutionary categories remains questionable (Tsonis et al., 1993). The determinants and origin of global structure, its correlations with and implications in gene sequences and its evolution remain to be resolved. A step toward their elucidation in this study uses chaos game representation, determination of short-sequence representation and the analysis of a large database of DNA sequence information from numerous different species.

1.6 Chaos Game Representation

Deciphering the significance of nucleotide composition and order must begin with the recognition of still uncharacterized patterns in large DNA sequences and requires a general analytical approach. One such approach is the chaos game representation of DNA sequences (Jeffrey, 1990, 1992; Solovyev et al. 1992; Dutta and Das, 1992; Burma et al., 1992; Solovyev, 1993; Oliver et al., 1993). Chaos, the randomness generated by simple dynamic systems, can be used to reveal order within complex systems (Tsonis and Tsonis, 1989). Chaos game theory produces pictures that represent complex systems in their entirety. Chaos was first recognized in biological systems with the use of a logistic difference equation to model population dynamics (May, 1976) and seemed an appropriate means to portray the organization within the complexity of large DNA sequences.

Chaos game theory was applied to the description of the order within the complexity of nucleotide sequences by using the nucleotides of a DNA sequence rather than a series of random numbers to control the chaos game algorithm (Jeffrey, 1990, 1992). The result is a two-dimensional scatter plot depicting base composition and sequentiality. The chaos game is plotted on x and y axes, producing a square whose four vertices correspond to the four nucleotides adenine (A), cytosine (C), guanine (G) and thymine (T) of a DNA sequence. The result is a transformation of a linear sequence of nucleotides into a visual portrayal of all aspects of nucleotide composition and order in a two-dimensional scatter plot. This approach is a unique and holistic perspective to the examination of the macro structure of a nucleotide sequence.

A large sequence composed of equal frequencies of the four nucleotides and a random order of those nucleotides generates a uniform distribution of data points in the two-dimensional chaos plot (Jeffrey, 1990). DNA sequences are not composed of equivalent frequencies of the four nucleotides and the order of nucleotides is not random (Jeffrey, 1990). The nonrandom composition of DNA sequences is displayed graphically by nonuniform distributions of data points in two-dimensional chaos plots (Jeffrey, 1990, 1992; Dutta and Das, 1992; Burma et al., 1992; Oliver et al., 1993; Goldman, 1993). The unique perspective of a visual assessment of chaos patterns is the simultaneous assessment of bias in subsequence composition for single nucleotides and successively longer oligonucleotides (i.e., to tetranucleotides and perhaps longer subsequences depending on the degree of bias in composition; Goldman, 1993). The visual assessment of total sequence structure permits rapid identification of specific features that can then be characterized quantitatively and compared among different sequences. The examination of sequence organization using chaos

representation and an appropriate DNA sequence database could offer insight into the determinants and evolution of higher-order DNA sequence organization.

Chaos representation of DNA sequences is particularly suited to the rapid analysis of the primary sequence organization of large DNA sequences. The presence and absence of subsequences of all lengths are displayed simultaneously in a single visual image and represented by a data point within the chaos plot. Pattern recognition using chaos plots is not driven and thus restricted by description of known sequence structures. As a result, pattern recognition requires no preconceived notion of subsequence structure and unknown patterns in sequence organization can be revealed (Burma et al. 1992; Oliver et al., 1993).

Examination of chaos patterns has revealed several important observations pertaining to DNA sequence organization. Jeffrey (1990) showed that large DNA sequences from different species have different chaos patterns and hence have different global structures. Chaos plots of nucleotide sequences have not been described quantitatively and their biological significance (and hence the usefulness of the technique) remains to be demonstrated. Chaos game representation has not yet been applied to shorter gene sequences or strictly to the coding regions that may be under different selection regimes. Also, chaos patterns have not been generated for similar genes in the same species or for the same or similar genes occurring in phylogenetically divergent species. Demonstration of the utility of the chaos game representation requires an accumulation of results reported from different types of sequences followed by the construction of a data base of chaos patterns.

1.7 Objectives

Nucleotide diversity in the human nuclear genome is poorly characterized for different populations and for regions known to be under different selection regimes. There is also little information regarding the determinants, origin and evolution of nucleotide sequence organization at a macro level. The global structure of the nucleotide sequence has not been incorporated into previous direct analyses of genetic variation. The present study addresses these concerns in the form of three general objectives.

The first objective was to characterize DNA sequence variation at two regions of the human nuclear genome that were known to differ in the nature and extent of polymorphism. Specifically, PCR amplification of genomic DNA and direct dideoxynucleotide sequencing were used to measure the total extent of nucleotide diversity at the polymorphic third exon of the *Adh2* locus and the highly polymorphic second exon of the *HLA-DQβ1* gene. The sequence variation was examined in a random sample of individuals from two geographically distinct populations known to differ in population structure and evolutionary history, specifically individuals from Southwestern Ontario and Athabaskan Dogrib individuals form the Northwest Territories of Canada.

The second objective was to characterize the features of the organization of nucleotide sequences by identifying and defining the global organization of nucleotide sequences. A macro level to the organization of nucleotide sequences was examined using measurement of short-sequence structure and the novel graphical approach of chaos game representation. Specific objectives involved the characterization of the global organization of nucleotide sequences through examination of computer-generated and naturally-occurring DNA sequences. The chaos patterns and short-sequence representation of the

polymorphic second exon of *Adh2* and the highly polymorphic third exon of *HLA-DQB1* were compared. As well, the global structures of the exons were compared with that of the cDNAs of these human genes. The sequence organization of the *Adh* multigene family in the human genome was examined in greater detail and compared with the well-studied structure of individual human globin cDNAs and the entire human globin gene complex. The global structure of human *Adh* cDNAs was compared with those of numerous phylogenetically diverse species. Global structure was identified and defined following examination of sequence organization with different sequence lengths, at different sites within a genome and among functionally diverse regions of a genome. Sequence organization was examined in exons, introns, complete cDNA sequences, gene clusters, different regions of the same chromosome and regions of different chromosomes of the same genome.

The final objective of the present study was to define more clearly the genome-type specificity of the global organization of nucleotide sequences and to identify the biological determinants of the global structure of DNA. Characterization of the global structure of nucleotide sequences of several genomes was carried out by examination of numerous regions of the same genome and large continuous DNA sequences of phylogenetically diverse species. In total, 56 large continuous DNA sequences (>36,000 nucleotides) from 10 different species were examined. Global structure in nuclear genomes was also characterized by comparing sequence organization in mitochondrial genomes of the same species and in viral genomes capable of integration into a host nuclear genome.

This investigation is unique in that it encompasses a direct examination of nucleotide sequence diversity and a holistic analysis of sequence structure at a macro level using a large database of diverse sequence information. The

present analysis used a large database of nucleotide sequence information, which permitted a systematic evaluation of the advantages, limitations and usefulness of a novel graphical representation of nucleotide sequence organization. The findings permitted discussion pertaining to explanations for the differences in nucleotide diversity related to the region of the human nuclear genome and the history and structure of the human populations. The significance of a macro level of nucleotide sequence organization that is genome-type specific was also contemplated.

Chapter 2

MATERIALS AND METHODS

2.1 DIRECT MEASUREMENT OF NUCLEOTIDE SEQUENCE VARIATION

2.1.1 MATERIALS

2.1.1.1 Subjects

This study included 32 individuals recruited from Southwestern Ontario by Dr. S.M. Singh, University of Western Ontario, London Ontario, Canada (numbered 1 to 32) and 38 Athabaskan Dogrib individuals from Lac LaMatre, Rae Lakes and Rae of the Northwest Territories, Canada and recruited by Dr E. Szathmary, McMaster University, Hamilton Ontario (numbered 33 to 70). Of the individuals from Southwestern Ontario, number 29 was of Oriental ancestry, individual 28 was of Native North American ancestry and the remaining subjects were of European ancestry. Individuals 14 through 19 comprised three generations of a single family where 14 was the paternal grandmother, 15 was the mother, 16 was the father and 17 to 19 were daughters. Individuals 20 and 21 were father and daughter and 22 and 23 were brother and sister. Individuals 26 and 27 were father and daughter, respectively. There were 23 unrelated subjects of European ancestry in the sample of individuals from Southwestern Ontario. All Dogrib individuals were unrelated and recruited randomly. It was assumed that the 38 Dogrib individuals constituted members of a single population. This assumption was based upon measures of genetic variation at

other loci and information regarding the population dynamics of the three settlements from which the individuals were recruited (Szathmary, 1993; Szathmary et al., 1983). Approval for this research was granted by the University of Western Ontario Review Board for Research Involving Human Subjects.

2.1.1.2 Chemicals

Unless otherwise noted, molecular biology chemicals were obtained from Amersham Corporation (Arlington Heights, Illinois), BDH Chemicals Canada Ltd. (Toronto, Ontario), Boehringer Mannheim (Laval, Quebec), Fisher Scientific Company. (Philpsburg, New Jersey), Life Technologies (Gaithersburg, Maryland), Perkin Elmer (Branchburg, New Jersey), Qiagen (Chatsworth, California) or Sigma Chemical Co. (St. Louis, Missouri).

2.1.1.3 Whole Blood Samples and DNA

Fresh, whole blood samples (15 ml) in 0.10 ml of 15% solution EDTA(K_3) containing 0.016 mg potassium sorbate (antimycotic agent) were obtained by Dr. C.F.P. George from the subjects of Southwestern Ontario. Samples were stored immediately at -20°C and used for the isolation of genomic DNA. Genomic DNA from Dogrib individuals was obtained by Dr. R. Ferrell (Pittsburgh) and was stored at 4°C in TE (10 mM Tris-Cl, 1 mM EDTA, pH 8.0). Primers synthesized for use in the polymerase chain reaction (PCR) and DNA sequencing are given in Table 1. The primers were manufactured and size-selected using gel electrophoresis by Dr. G. Hammond, London Regional Cancer Center, London, Ontario.

Table 1. Oligonucleotides used as primers for PCR amplification and direct sequencing. The oligonucleotide primers and their sequences, listed 5' to 3', used to amplify the third exon of the alcohol dehydrogenase locus *Adh2* (P1 and P2) and the second exon of the human leukocyte antigen locus *HLA-DQβ1* (P3, P4). Those nucleotides in lower case represent an oligonucleotide sequence used to introduce an *Eco* RI site into the PCR product. The restriction site was not used in this analysis.

Primer (Location)	Sequence (given 5' to 3')
***Adh2* Third Exon**	
P1 (5' region)	**AATCTTTTCTGAATCTGAACAG**
P2 (3' region)	**ggaattccGTGTGAATCCTGTACCTGGTTT**
***HLA-DQβ1* Second Exon**	
P3 (5' region)	**ATTTCGTGTACCAGTTTAAGG**
P4 (3' region)	**CCACCTCGTAGTTGTGTCTGCA**

2.1.2 METHODS

2.1.2.1 Isolation of Genomic DNA

Isolation of genomic DNA from frozen samples of anticoagulated whole blood (-20°C, 5 ml) was carried out using the method of Jeanpierre (1987),with several modifications. Samples were thawed at 37°C and resuspended in 10 ml of saline solution #1 (75 mM NaCl, 25 mM EDTA). The samples were centrifuged in 15-ml conical centrifuge tubes (Fisher Scientific, Phillipsburg, New Jersey) in a bench top clinical centrifuge (International Equipment Company, Needham Heights, Massachusetts) at 690 x g for 20 min. The supernatant (3 ml) was aspirated and replaced with an equal volume of saline solution #2 (10 mM NaCl, 10 mM EDTA). The pellet was resuspended and centrifuged at 690 x g for 20 min. Again, the supernatant (10 ml) was aspirated and replaced with an equal volume of saline solution #2 and centrifuged 1200 x g for 15 min. The supernatant was aspirated to leave 0.5 ml of sample and the aspirated volume was replaced with an equal volume of saline solution #2 and centrifuged (1200 x g, 15 min). Finally, the entire supernatant was removed and the pellet resuspended in 100 µl of saline solution #2. Guanidine hydrochloride (3 ml, 6 M) was added to the resuspension and the sample was placed on a rotary wheel at 4°C for 15 min or until the pellet was completely resuspended. Ammonium acetate (216 µl, 7.5 M) was then added and the sample was spun on the rotary wheel for 3 hours at 4 °C.

Protein digestion was carried out by the addition of 216 µl of 20% sarkosyl and 35 µl proteinase K (20 mg/ml). Samples were replaced on the rotary wheel for 1 hour at room temperature and then transferred to a 65°C water bath for incubation overnight. The centrifuge tubes were inverted periodically during the

incubation. Following the overnight incubation, ethanol was added to the samples (2.5 volume of 70% ethanol, stored at -20°C) and the tubes were inverted gently until the high molecular weight DNA could be spooled into 2 ml of TE (pH 7.6). Spooled DNA was removed from the spooling rod by incubation of the sample for 5 min at 65°C. Proteins were removed from genomic DNA with the addition of 150 µl of sarkosyl (20%) and incubation on the rotary wheel for 1 hour at room temperature. Proteinase K (20 µl, 20 mg/ml) digestion was carried out at 65°C overnight. DNA was spooled from the solution after the addition of 4 ml of 95% ethanol (stored at -20°C) and rinsed in 5 ml of 70% ethanol (stored at -20°C). The DNA was then removed from the spooling rod and resuspended in 1 ml TE (pH 7.6) and incubated at 65°C for 5 min. Isolated genomic DNA samples were then stored at 4°C until required.

The concentration of the DNA solution was determined by measuring its optical density at a wavelength of 260 nm using a Shimadzu spectrophotometer (model UV160U). One microgram of undigested genomic DNA from each sample was subjected to electrophoresis in a 0.8% mini agarose gel to ensure that it was of high molecular weight and to verify the spectrophotometric quantitation. As well, 1 µg of each sample was digested with the type II restriction endonuclease, *Eco* RI and was subjected to electrophoresis through a 0.8% mini agarose gel to ensure that the genomic DNA could be digested with a restriction endonuclease. This digestion was used to assess the quality of the genomic DNA prior to its use for *in vitro* amplification.

2.1.2.2 Restriction Endonuclease Digestion of Genomic DNA

Genomic DNA equivalent to 1.0 µg was digested overnight (20 hours) with the restriction enzyme *Eco* RI (Pharmacia, Uppsala, Sweden) at 37°C and using

One-Phor-All Buffer PLUS conditions (100 mM Tris-acetate, pH 7.5, 100 mM magnesium acetate, 500 mM potassium acetate). The digestions were stopped by heat inactivation (65°C, 20 min) and by adding a one-fifth volume of gel loading buffer (50 mM EDTA, pH 8.0, 0.125% bromophenol blue, 12.5% Ficoll).

2.1.2.3 Electrophoresis of Agarose Gels

Electrophoresis of restriction enzyme digested genomic DNA was performed at room temperature in horizontal 0.8% agarose gels (8 x 9 cm), in 1 X TBE running buffer (89 mM Tris; 89 mM boric acid; 2 mM EDTA, pH 8.0). Electrophoresis was carried out at 100 V for 1 hour. The λ *Hind* III digest/ϕX174 *Hae* III digest molecular size standard (1 µg; Pharmacia) or 2 µg of a 1 kb ladder (Life Technologies) was used on all gels. Gels were stained with 10 µg/ml ethidium bromide in 1 X TBE either before or after electrophoresis. A Polaroid MP-4 Land Camera, a Kodak Wratten No. 22 gelatin filter and black-and-white Polaroid type 57 Land film were used to photograph the stained DNA fragments visualized using an Ultra Lum transilluminator (model UVB-40E; wavelength 302 nm).

2.1.2.4 Primary PCR Amplification of the Adh2 Third Exon

A 100-µl aliquot of each genomic DNA sample was heated at 95°C for 10 min prior to use as template for polymerase chain reaction (PCR) amplification as recommended by the *Taq* DNA polymerase supplier (Life Technologies, Bethesda Maryland). This heat treated aliquot of genomic DNA was stored at 4°C until required. The third exon of the alcohol dehydrogenase locus, *Adh2* was amplified from each of the genomic DNA samples using a primary PCR with

temperature and chemical conditions determined empirically to give optimal specificity and high yield of PCR product. Approximately 500 ng of genomic DNA was brought up to a 10-µl volume with TE (pH 7.6) and was incubated with 50 pmol of each of the two primers (P1 and P2, Table 1), 0.25 mM dATP, 0.25 mM dCTP, 0.25 mM dGTP, 0.25 mM dTTP, 3 mM $MgCl_2$, 1X reaction buffer (50 mM KCl, 20 mM Tris·Cl, pH 8.4) and 2.5 U *Taq* DNA Polymerase (Life Technologies, Bethesda Maryland). The reaction was brought up to 100 µl with sterile distilled water and overlaid with 100 µl of mineral oil (Sigma, St. Louis, Missouri). Thermocycling was performed in a Perkin-Elmer Cetus Thermocycler, using the following conditions. PCR reactions were prepared on ice and then placed directly into the preheated thermocycler (94°C). After an initial melting at 94°C for 2 min, 40 cycles of 94°C for 1 min, 50°C for 2 min, and 72°C for 1 min were performed. A 7-min extension period at 72°C followed the 40 cycles. The samples were then stored at 4°C until required. Ten PCR reactions were prepared simultaneously from a master mix of reactants that lacked template. Each master mixture was monitored for contamination by the use of a single negative control reaction which contained 10 µl of distilled water in place of genomic DNA template.

2.1.2.5 Primary PCR Amplification of the HLA-DQβ1 Second Exon

Prior to the primary PCR amplification of the second exon of the human leukocyte antigen locus, *HLA-DQβ1*, genomic DNA template was digested with the restriction endonuclease *Alu* I. Genomic DNA equivalent to 1.0 µg was digested overnight (20 hours at 37°C) with the restriction enzyme *Alu* I (Pharmacia, Uppsala, Sweden) and One-Phor-All Buffer PLUS. This restriction endonuclease digestion was used to enhance the specificity of the primary PCR

amplification of the *HLA-DQβ1* locus by cutting within a region of the pseudogene *HLA-DQβ2* that is homologous to the second exon of *HLA-DQβ1* and would otherwise coamplify given the primer pair selected for PCR amplification in this analysis. Genomic DNA digested by the restriction endonuclease was precipitated and resuspended in H_2O since the salt concentration of the restriction endonuclease digestion buffer was inhibitory to the *Taq* DNA polymerase. Precipitation was performed using one volume of 4 M ammonium acetate, two volumes of isopropanol (100%) and 2 µl of linear polyacrylamide (10 µg/µl) as carrier. The samples were vortexed and incubated at room temperature for 10 min prior to centrifugation for 10 min at 13,200 rpm using a MicroMax microcentrifuge (International Equipment Company, Division, Needham Heights, Massachusetts). The precipitate was washed in 70% ethanol, dried and resuspended in 20 µl of H_20.

Approximately 250 ng of this DNA was incubated with 25 pmol of each of the two primers (P3 and P4, Table 1), 0.25 mM dATP, 0.25 mM dCTP, 0.25 mM dGTP, 0.25 mM dTTP, 2 mM $MgCl_2$, 1X reaction buffer (50 mM KCl, 20 mM Tris·Cl, pH 8.4) and 2 U *Taq* DNA Polymerase (Life Technologies, Bethesda Maryland). The reaction was brought up to 100 µl with sterile distilled water and overlaid with 100 µl of mineral oil (Sigma, St. Louis, Missouri). After an initial melting at 94°C for 3 min, 30 cycles of 94°C for 1 min, 64°C for 1 min, and 72°C for 1 min were performed. A 10-min extension period at 72°C followed the 30 cycles and samples were stored immediately at 4°C until required. Master mixes and negative controls lacking genomic DNA template also were used for the amplification of this locus (section 2.1.2.4).

2.1.2.6 *Analysis and Purification of PCR Products*

The yield and specificity of all PCR reactions were assessed following electrophoresis of a one-tenth volume of the reaction through an 8% mini polyacrylamide gel (30% weight/volume of total monomer, 1 X TBE running buffer) using the Bio-Rad (Richmond, California, U.S.A) MiniPROTEAN II apparatus. A one-sixth volume of gel loading buffer (0.25% bromophenol blue, 0.25% xylene cyanol, 30% glycerol in H_2O) was added to each aliquot of the PCR reactions and the samples were subjected to electrophoresis for 45 min at 150 V. Polyacrylamide gels were stained and photographed as described for unstained agarose gels (see section 2.1.2.3).

Products of the primary PCR reaction (90 µl) were precipitated using one volume of 4 M ammonium acetate, two volumes of isopropanol (100%) and 2 µl of linear polyacrylamide (10 µg/µl) as carrier. The samples were vortexed and incubated at room temperature for 10 min prior to centrifugation for 10 min at 13,200 rpm. The precipitate was washed in 70% ethanol, dried and resuspended in 10 µl of H_2O. A one sixth volume of gel loading buffer (0.25% bromophenol blue, 0.25% xylene cyanol, 30% glycerol in H_2O) was added to the resuspension and the samples were subjected to electrophoresis in 1X TAE buffer (40 mM Tris, 20 mM sodium acetate·3 H_2O, 1 mM EDTA, pH 7.2) through a low melting point agarose gel at 90 mAmps for 2 hours. Gels were stained and photographed as described in section 2.1.2.3. The PCR products of the expected length of the third exon of the *Adh2* locus or the second exon of the *HLA-DQβ1* locus were cut out of the agarose gel and transferred to a 1.5-ml microcentrifuge tube. The gel slice containing the desired PCR product was melted at 70°C for 10 min and diluted in 1 ml of H_2O. These procedures achieved the isolation of the desired

PCR products from the genomic DNA template, primers and primer artifacts of the primary PCR reaction. The purified PCR product (10 µl) was then used as a template for secondary PCR amplification. The chemical and thermocycling conditions of the secondary PCR reactions were identical to those of the primary PCR reactions.

The product of the secondary PCR amplification of the *Adh2* third exon was purified from nucleotides, the *Taq* DNA polymerase, primers and primer-dimers through the use of the QIAquick PCR purification kit (Qiagen). The purified PCR product was eluted in 50 µl of H_2O. The product of the secondary PCR amplification of the *HLA-DQβ1* second exon was precipitated using 1 volume of 4 M ammonium acetate, 2 volumes of isopropanol (100%) and 2 µl linear polyacrylamide (10 µg/µl) as carrier. The samples were vortexed and incubated at room temperature for 10 min prior to centrifugation for 10 min at 13,200 rpm. The precipitate was washed in 70% ethanol, dried and resuspended in 10 µl H_2O. A one-sixth volume of gel loading buffer was added to 10 µl of the resuspension and the samples were subjected to electrophoresis in 1X TBE through an 8% polyacrylamide gel. The gel was stained and the DNA visualized using the methods described in section 2.1.2.3. The region of the acrylamide gel containing the amplified exon was excised, the gel slice was crushed and the DNA eluted from the gel in 250 µl of H_2O with agitation overnight at 37°C. The acrylamide was removed from the eluted DNA fragments using microcentrifugation for 15 min at 13,200 rpm in a Costar Spin-X centrifuge filter unit (Costar, Cambridge Massachusetts). The QIAquick PCR purification kit purified the amplified *HLA-DQβ1* second exon and eluted it in 50 µl of H_2O.

The purity and concentration of the PCR product was assessed by electrophoresis (45 min at 150 V) through an 8% polyacrylamide gel using a 1/10 volume of the purified PCR product and containing a one sixth volume of gel

loading buffer. The PCR product length and quantity were determined in comparison with the standards given in section 2.1.2.3. Polyacrylamide gels were stained and photographed as described in section 2.1.2.3.

2.1.2.7 Single-Strand Conformation Polymorphism Analysis

Approximately 500 ng of the QIAquick purified PCR product was mixed with an equal volume of a denaturing gel loading buffer (95% formamide, 4.6 M urea, 0.25% xylene cyanol and 0.25% bromophenol blue), heat denatured (95°C for 10 min) and cooled on ice (4 min). The single-stranded conformers and double stranded PCR-amplified products were subjected to electrophoresis in an 8% nondenaturing mini polyacrylamide gel (1X TBE running buffer at 150 V for 45 min) using the Mini PROTEAN II system. Gels were stained and photographed as described in section 2.1.2.3. The mobility of single-stranded conformers and nondenatured double-stranded PCR products were compared with a molecular size standard (section 2.1.2.3) and with samples which were not heat denatured, contained nondenaturing gel loading buffer and were run on the same gel as the denatured samples.

In an alternative SSCP protocol, lesser amounts of DNA were required and a greater separation of the different single-stranded conformers was achieved. PCR products (50 ng) were end-labeled in 1X kinase buffer (70 mM Tris-Cl, pH 7.6, 10 mM $MgCl_2$, stored a 4°C) and 10 mM DTT using 10 uCi of γ-[^{32}P]-ATP (3000 Ci/mmol; Amersham) and 30 U T_4 polynucleotide kinase (Life Technologies). End-labeling was performed in a Perkin Elmer Thermocycler at 37°C for 10 min and was immediately inactivated by incubation at 90°C for 5 min. End-labeled PCR products were used immediately or stored up to 5 days at -20°C. End-labeled PCR products were denatured as described above and were

subjected to electrophoresis in 1X TBE running buffer on 8% nondenaturing polyacrylamide gels using Bio-Rad (Richmond, California, U.S.A) SequiGene sequencing system (0.4 cm X 21 cm X 50 cm) and an LKB Bromma 2302 Multidrive XL 3.5 kV power pack (Pharmacia). The gels were prerun for 30 min at 2 W prior to loading samples, at 90 W for 2 min to run samples into the gel and finally at 4 W for 10 hr at room temperature. After electrophoresis, the gels were dried onto Sequencing Gel filter paper (Bio-Rad) for 1 hr in a Bio-Rad Model 583 gel drier. X-ray film (Kodak X-OMAT AR) was placed next to the dried gel in a film cassette for 18 to 72 hr at room temperature. The film was then developed according to the manufacturer's instructions using GBX developer and fixer.

SSCP conformers were identified without knowledge of the identity of the sample under examination. The relative mobilities of single-stranded DNA conformers upon nondenaturing polyacrylamide gel electrophoresis were described relative to a molecular size standard and to a reference sample to which all samples were compared.

2.1.2.8 Primer End-Labeling For Direct Sequencing of PCR Products

Primers (20 uM) to be used for direct sequencing were end-labeled in 1X kinase buffer (70 mM Tris·Cl, pH 7.6, 10 mM $MgCl_2$, stored at 4°C) and 10 mM DTT using 10 uCi of γ-[^{32}P]-ATP (3000 Ci/mmol; Amersham) and 30 U of T_4 polynucleotide kinase (Life Technologies). End-labeling was performed in a Perkin Elmer Thermocycler at 37°C for 10 min and was inactivated by incubation at 90°C for 5 min. End-labeled primers were used immediately or stored up to 5 days at -20°C The primer P2 was used to determine the sequence of the third exon of *Adh2* and the primers P3 and P4 were used to determine the sequence

of the second exon of the *HLA-DQß1* locus (primer sequences are contained in Table 1).

2.1.2.9 Direct DNA Sequencing of the PCR-Amplified Adh2 Third Exon

Approximately 200 to 500 ng of PCR-amplified and purified DNA was sequenced directly using the Perkin Elmer Ampli*Taq* cycle sequencing kit and a Perkin Elmer Thermocycler. The template was denatured initially at 95°C for 1 min and then cycle sequenced for 20 cycles of template denaturation at 95°C for 1 min and primer extension at 72°C for 1 min. Following the 20 cycles, a final extension step was carried out at 72°C for 7 min. The conditions for optimal sequencing were determined empirically for each primer used. The primer P2 was selected for use in direct sequencing because it produced sequence ladders with minimal background banding. Extensive sequencing was not performed using the P1 primer as no nucleotide variation was detected among the samples using SSCP analyses and direct sequencing with the P2 primer.

2.1.2.10 Direct DNA Sequencing of the PCR-amplified HLA-DQß1Second Exon

Approximately 10 fmol (0.79 ng) of purified PCR-amplified DNA was sequenced directly using the Perkin Elmer Ampli*Taq* cycle sequencing kit and a Perkin Elmer Thermocycler. The template was denatured initially at 95°C for 2 min and then cycle sequenced for 25 cycles. Using the P3 primer, the cycling conditions included template denaturation at 95°C for 1 min, primer annealing at 64°C and primer extension for 1 min at 72°C. The primer P4 required cycling conditions of template denaturation at 95°C for 1 min, primer annealing at 70°C

and primer extension for 1 min at 72°C. Both cycling conditions were followed by a final extension step carried out at 72°C for 7 min. The sequencing reactions were held for no longer than 45 min at 4°C prior to addition of the stop buffer. Sequencing reactions were used within 24 hours with storage at -20°C. The conditions for optimal sequencing were determined empirically for each of the primers tested.

2.1.2.11 General DNA Sequencing Gel Protocols

Sequencing reactions were separated on 8% polyacrylamide, 8 M urea gels using a Bio-Rad (Richmond, California, U.S.A.) SequiGene sequencing system (0.4 cm X 21 cm X 50 cm) or a Model S2 BRL sequencing apparatus (0.4 cm X 39 cm X 33 cm) and an LKB Bromma 2302 Multidrive XL 3.5 kV power pack (Pharmacia). Electrophoresis conditions were 50°C and a constant power of 60 W for 2 or 5 hr. After electrophoresis, the gels were dried on Sequencing Gel filter paper (Bio-Rad) for 1 hr in a Bio-Rad Model 583 gel drier. X-ray film (Kodak X-Omat AR) was placed next to the dried gel in a film cassette and exposed 24 to 48 hr at room temperature. The film was processed according to manufacturer's instructions using GBX developer and fixer.

2.1.2.12 Typing the Banding Patterns of DNA Sequencing Reactions

The two alleles at a single locus in heterozygous individuals were not isolated prior to DNA sequencing. Thus, heterozygosity in DNA sequencing reactions was identified as two single-stranded DNA sequences of identical length yet terminating with two different ddNTPs. In the set of four termination

reactions two bands of approximately equal intensity are seen to migrate to the same horizontal position upon denaturing polyacrylamide gel electrophoresis.

A single nucleotide sequence in homozygous individuals and the two sequences of heterozygotes were compared with known allele sequences in order to identify individual alleles. Three allele sequences were considered in the case of the third exon of the *Adh2* locus. Twenty-five known allele sequences (Appendix 1) at the second exon of the *HLA-DQβ1* locus were compared with the sequences determined in this analysis in order to identify the allele or alleles at this locus in individual samples.

2.2 COMPUTER ANALYSIS OF NUCLEOTIDE SEQUENCE ORGANIZATION AND VARIATION

2.2.1 MATERIALS

2.2.1.1 Nucleotide Sequences

Nucleotide sequences were obtained from release 69.0 of the Genbank DNA sequence data base (Bilofsky and Burks, 1988) and release 28.0 of the EMBL data base. Table 2 lists the 29 alcohol dehydrogenase (*Adh*) sequences selected for analysis of the organization of nucleotides. Alleles at the human leukocyte antigen locus, *HLA-DQβ1* second exon and the human, pig and sheep *DQβ1* cDNAs were included in the analysis (cDNA accession numbers, M24364, M31497 and L08792, respectively). The organization of nucleotides in large continuous DNA sequences (>36,000 nts) was examined using 56 sequences ranked among the 100 longest in the Genbank and EMBL sequence databanks (releases 68.0 and 27.0, respectively; Table 3).

Table 2. The 29 alcohol dehydrogenase nucleotide sequences used in an analysis of nucleotide sequence organization.

GenBank Accession No.	*Species* Label (*Locus*)
I. Iron-binding (IB)	
M15394	*Zymomonas mobilis* **Zym**
X05992	*Sacharomyces cerevisiae* **Ys4** (*Adh4*)
II. Nonmetal-binding (NB)	
Z00033/X00603	*Drosophila mauritiana* **Dma**
X04672/Z00045	*Drosophila sechellia*
X00607/X03226/Z00031	*Drosophila simulans*
III. Zinc-binding (ZB)	
J03362	*Alcaligenes eutrophus* **Alc**
X02764	*Aspergillus nidulans* **Asn**
J01313	*Sacharomyces cerevisiae* **Ys1** (*Adh1*)
K03292	*Sacharomyces cerevisiae* (*Adh3*)
J01314/M13475	*Sacharomyces cerevisiae* (*Adh2*)
J01341	*Schizosaccharomyces pombe* **Ysp**
M12196	*Arabidopsis thaliana* **Ath**
X07774	*Hordeum vulgare* **Bly**
X00580	*Zea mays* (*Adh1-F*)
X01965	*Zea mays* (*Adh1-N*)
M25154	*Solanum tuberosum* (*Adh1*)
M25153	*Solanum tuberosum* (*Adh2*)
M25152	*Solanum tuberosum* (*Adh3*)
X14826	*Trifolium repens*
M11307	*Mus musculus* **Mus**
M15327	*Rattus norvegicus*
M25035	*Papio hamadryas anubis* **Bab**
M12963	*Homo sapiens* (*Adh1*)
M12271	*Homo sapiens* (*Adh1*)
D00137	*Homo sapiens* **Hum** (*Adh2*)
M24317/M11831-M11839 /K01883	*Homo sapiens* (*Adh2*)
M21692	*Homo sapiens* (*Adh2*)
X04299	*Homo sapiens* (*Adh3*)
M12272	*Homo sapiens* (*Adh3*)

Table 3. Large continuous DNA sequences (>36,000 nts) from 10 phylogenetically diverse species used for analysis of nucleotide sequence organization. Sequences are listed on two pages.

Species (Total No. of nts. examined) LOCUS	Gene Region or Location	Length (nts)
Escherichia coli (687,631)		
ECO110K	0-2.4 min.	111,401
ECOU47	47-48 min.	75,888
ECOUW82	81.5-84.5 min.	136,254
ECOUW85U	84.5-86.5 min.	91,408
ECOUW87	87.2-89.2 min.	96,484
ECOUW89	89.2-92.8 min.	176,196
Rhodobacter capsulatus		
RCPHSYNG	Photosynthetic gene cluster	45,959
Saccharomyces cerevisiae (432,583)		
SCPEKGAI	Ornithine decarboxylase chromosome XI	38,467
YSCCHRIII	Complete chromosome III	315,357
YSCCHROMI	Centromeric region chromosome I	41,987
YSCSYGP2	CYC7 region chromosome V	36,772
Caenorhabditis elegans (1,008,914)		
CEB0464	Aspartyl-tRNA synthetase	40,909
CEF58A4	RNA polymerase I and III	38,000
CEF59B2	Cosmid F59B2	43,782
CELC08C3	Homeobox DNA-binding proteins	44,025
CELC14B9	Met-tRNA; alpha-B-crystallin	43,492
CELC29E4	Adenylate kinase	40,050
CELC50C3	Calmodulin	44,733
CELF09G8	Cuticle collagen protein	41,645
CELF44B9	A1 accessory protein	36,327
CELF54F2	DNAJ protein; protease	39,573
CELR05D3	DNA topoisomerase II	38,810
CELTWIMUSC	Unknown open reading frame	54,962
CELZC21	Breakpoint cluster region protein	36,087
CELZK112	Cadherin	38,269
CELZK370	Mariner transposition protein	37,675
CELZK652	60S ribosomal protein L35	36,052
CELZK688	N-acetylgalactosaminyltransferase	36,977
CER107	CosmidR107	40,970
CEUNC22	Unc-22	47,081
CEZC84	Serine protease inhibitor repeats	38,955
CEZK1098	GCN3	37,310
CEZK512	RNA helicase	36,997
CEZK632	ADP ribosylation factor	36,000
CEZK637	Cosmid ZK637	40,699
CEZK643	Cosmid ZK643	39,534

Species (Total No. of nts. examined) LOCUS	Gene Region or Location	Length (nts)
Drosophila melanogaster		
DROABDB	Abdominal-B	80,423
Rattus Norvegicus		
RATCRYG	γ-crystallin	54,670
Mus musculus		
MUSTCRA	T-cell receptor	94,647
Oryctolagus cuniculus		
RABBGLOB	β-like globin	44,594
Galago crassicaudatus		
GCRHBEGEB	β-globin	41,101
Homo sapiens (965,092)		
HSMHCAPG	Major histocompatibility complex Class II	66,109
HUMADAG	Adenosine deaminase	36,741
HUMDYSTROP	Dystrophin	38,770
HUMFIXG	Factor IX	38,059
HUMGHCSA	Growth hormone	66,495
HUMHABCD	4p16.3	58,864
HUMHBB	β-globin	73,326
HUMHDA	Cosmid HDAC	40,289
HUMHP2HPR	Haptoglobin	38,524
HUMMDA	19q13.3	37,314
HUMNEUROF	Neurofibromatosis-1	100,849
HUMRETBLAS	Retinoblastoma	180,388
HUMTCRADCV	Tcr-C-delta	97,634
HUMTPA	Tissue plasminogen activator	36,594
HUMVITDBP	Vitamin D binding protein	55,136

Nucleotide sequence organization in the yeast *Saccharomyces cerevisiae* genome was examined in greater detail using large regions of four different chromosomes and numerous regions of chromosome III. Regions of heterogeneous functional significance (i.e., that contained both translatable and nontranslatable sequences) were obtained by dividing chromosome III into 10 consecutive and nonoverlapping regions, each 30,000 nts in length, and 20 consecutive and nonoverlapping regions, each 15,000 nts in length. Subdivision of the chromosome into regions always began with the first nucleotide at the left telomere. The remaining 15,357 nts of the chromosome comprised a final region.

Sequence organization was also compared for strictly translatable and entirely nontranslatable sequences. In the case of the yeast chromosome III, a 28-kb sequence composed entirely of translatable sequences was created by assembling end to end the 22 putative open reading frames located closest to the left telomere. The relative order of the reading frames in the chromosome sequence was maintained in the newly assembled sequence. A 28-kb sequence entirely composed of nontranslatable sequences was created by assembling end to end the sequences dispersed amongst the translatable sequences. These sequences were obtained beginning at the left telomere and their relative order in the chromosome sequence was maintained in the newly assembled sequence. The coding and noncoding regions of the human globin gene complex and neurofibromatosis-1 gene regions were analyzed in a similar manner.

Nucleotide sequence organization was also examined in twenty-eight complete mitochondrial DNA genomes and short, adjacent and nonoverlapping regions of mitochondrial genomes (Table 4) The regions of mitochondrial genomes were 16, 30 and 4 kb in size from *Saccharomyces cerevisiae*, *Marchantia polymorpha* and *Homo sapiens*, respectively.

Table 4. The complete mitochondrial genome sequences for 28 phylogenetically diverse species used for analysis of nucleotide sequence organization.

Evolutionary Category *Genus species*	Accession No. /Keyword	Sequence Length (nts)
Protozoan		
Paramecium aurelia	X15917/PAUR	40,469
Fungus		
Podospora anserina	M61734/PANS	100,314
Alga		
Prototheca wickerhamii	U02970/PWIC	55,328
Yeast		
Schizosaccharomyces pombe	X54421/SPOM	19,431
Saccharomyces cerevisiae	M62622/SCER	78,294
Plant		
Marchantia polymorpha	M68929/MPOL	186,608
Invertebrate		
Ascaris suum	X54253/ASUU	14,283
Caenorhabditis elegans	X54242/CELE	13,794
Artimiidae franciscana	X69067/AFRA	15,770
Paracentrotus lividus	J04815/PLIV	15,679
Strongylocentrotus purpuratus	X12631/SPUR	15,650
Drosophila yakuba	X03240/DYAK	16,019
Apis melifera	L06178/AMEL	16,343
Anopheles gambiae	L20934/AGAM	15,363
Vertebrate		
Gallus gallus	X52392/GGAL	16,775
Xenopus laevis	X01601/XLAE	17,553
Cyprinus carpio	X61010/CCAR	16,364
Crossostoma lacustre	M91245/CLAC	16,558
Oncorhynchus mykiss	L29771/OMYK	16,660
Dipelphis virginiana	Z29573/DVIR	17,084
Balaenoptera physalus	X61145/BPHY	16,398
Balaenopteridae musculus	X72204/BMUS	16,402
Halichoerus grypus	X72004/HGRY	16,797
Phoca vitulina	X63726/PVIT	16,826
Bos taurus	J01394/BTAU	16,338
Rattus norvegicus	X14848/RNOR	16,298
Mus musculus	V00711/MMUS	16,295
Homo sapiens	J01415/HSAP	16,569

Table 5 contains the names of the complete viral genome sequences available from sequence databanks. Only complete genome sequences were selected due to the small size of the partial genome sequences and the difficulty in characterizing genome-specific sequence structures from small and select regions of a genome.

Nucleotide sequences having simple global structures were constructed using the Microsoft Excel spreadsheet program (version 2.2; Microsoft Corporation, Bellevue, WA) and the random number generator of the Systat statistical package (version 5.1; Wilkinson, 1991). A column of cells in the spreadsheet were each assigned a nucleotide and the cells were sorted using the random number generator. The number of cells used was equal to the number of nucleotides in the sequence and the number of cells assigned a particular nucleotide was dependent upon the single nucleotide composition to be examined. Table 6 lists the types of simple sequence structures constructed through alteration of single nucleotide frequencies. All sequences were 16,295 nts in length, a length comparable to a vertebrate mitochondrial genome and a length that generated a sufficient number of data points for the visual recognition of nonrandom patterns in the distribution of data points. Nucleotides were ordered randomly in all of the computer-generated nucleotide sequences.

2.2.1.2 Computer Hardware and Software Packages

Chaos game representation of short nucleotide sequences (shorter than 2 kb) was performed using a HyperBasic program according to the method of Jeffrey (1990, 1992). Chaos Generator/Plot programs were run on the Macintosh Classic microcomputer (4 megabytes of memory and operated under system software version 6.07; Cupertino, CA). Chaos generator and plotting programs

Table 5. Complete viral genome sequences used for analysis of nucleotide sequence organization.

Viral Genome	Accession Keyword	Sequence Length (nts)
Adenovirus type 2	ADRCG	35,937
Bacteriophage L2	BLZCG	11,965
Hog cholera	HCVCGSA	12,285
Human immunodeficiency	HIVANT70C	9,754
Human immunodeficiency	HVMVP5180	9,793
Kunjin	KUNCG	10,664
Sindbis	SINCG	11,703
Simian immunodeficiency	SIVMM32H	10,277
Sonchus yelow net	SYENMGC	13,720
Variola major	VARCG	186,102
Vesicular stomatitis	VSVCG	11,161
West Nile Virus	WNTCG	10,960
Visna lentivirus	VLVCG	9,203
Hepatitis C	HPCJCG	9,413
Woodchuck hepatitis	OHVCGD	3,323
Visna	VLVLVIA	9,221
Heron Hepatitis B	HPUCG	3,027
Duck Hepatitis B	HPUGA	3,024
Borna disease	BDVSEQ	8,903
Moloney murine leukemia	MLMCG	8,332
Rous sarcoma	ALRCG	9,625
Tobacco mosaic	MTVVCG	6,395
Nurine Minute	MVMPCG	5,149
Hepatitis A	HPA	7,478
Feline immunodeficiency	FIVCG	9,474
Bovine Leukemia	BLVCG	8,714
Moloney murine sarcoma	MLMPROCG	5,828
Yellow fever	YFVCG	10,862
JC polyomavirus	PLYCG	5,130
Potato A	PVCGA	9,585
Dengue	DENZCGA	10,723

Table 6. Computer-generated nucleotide sequences having known and simple structures. Each of the sequences differs in single nucleotide composition. All sequences are 16,295 nucleotides in length and the nucleotides are ordered randomly.

Sequence Description	Nucleotide Composition
Equivalent frequencies of nucleotides	A=C=G=T=25%
Under-representation of a single nucleotide	A=40%, C=G=T
	A=30%, C=G=T
	A=20%, C=G=T
	A=10%, C=G=T
Over-representation of a single nucleotide	C=90%, A=G=T
	C=80%, A=G=T
	C=70%, A=G=T
	C=60%, A=G=T
Under-representation of two nucleotides	A=T 10%, C=G 90%
	A=T 20%, C=G 80%
	A=T 30%, C=G 70%
	A=T 40%, C=G 60%
	C=T 10%, A=G 90%
	C=T 20%, A=G 80%
	C=T 30%, A=G 70%
	C=T 40%, A=G 60%

based upon the algorithm of Jeffrey (1990, 1992) were also written in Fortran 77 and operated in the VAX/VMS operating system in order to analyze nucleotide sequences longer than 2 kb (Appendix 2). Chaos generator and plotting programs utilized the SAS Information System (SAS Institute Inc, Cary NC). Another Fortran 77 program, Coordinate Frequency was employed to count repetitions of (x,y) coordinates in chaos plots (Appendix 3). This program determines the frequency of superimposed data points in individual chaos plots and identifies the nucleotides or subsequences that are repeated in the entire nucleotide sequence. Other Macintosh productivity software packages including The DNA Inspector program (version 3.13; Gross 1986), Systat and Excel were used in the data analysis. Sequence alignment and related genetic distance calculations were performed using The University of Wisconsin Genetics Computer Group Software (UWGCG, version 8; Devereux et al. 1981).

2.2.2 METHODS

2.2.2.1 Manipulation of Nucleotide Sequence Information

Coding regions were edited from genomic sequences using the software program DNA Inspector IIe. Amino acid sequences were translated from nucleotide sequences using either the DNA Inspector package or the 'Translate' program of the UWGCG package.

2.2.2.2. Multiple Sequence Alignment and Genetic Distance Calculations

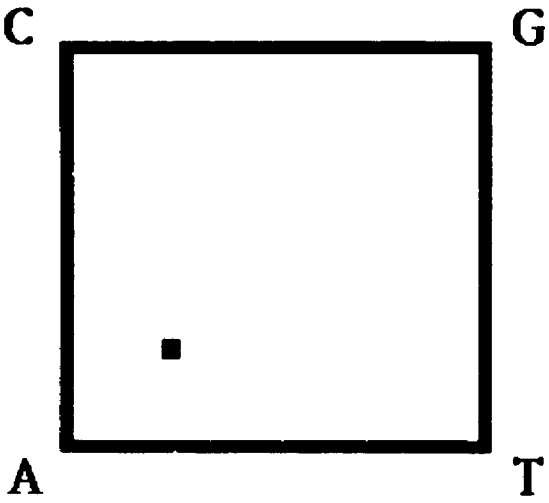
The UWGCG program 'Gap', which uses the algorithm of Needleman and Wunsch (1970), was used to make an optimal alignment between two nucleotide

sequences. Default values for gap weight and gap length were used (5.0 and 0.3, respectively). The scoring matrix used the default match threshold of 0.5. All possible pairwise comparisons of sequences were made using the 'Gap' program prior to sequence alignment. Multiple sequence alignment was performed using the UWGCG program 'Lineup', which used the alignments generated by the 'Gap' program. 'Pileup' was also used to create a multiple sequence alignment from a group of related sequences using progressive pairwise alignments (Needleman and Wunsch, 1970). The UWGCG program 'Distances' generated a matrix of pairwise genetic distances between each sequence in the multiple sequence alignment. For closely related sequences, the observed distance between sequences was computed with no correction for multiple substitutions. This uncorrected distance measure was defined as the difference between one and the number of matches between each sequence pair divided by their length (Swofford & Olsen, 1990).

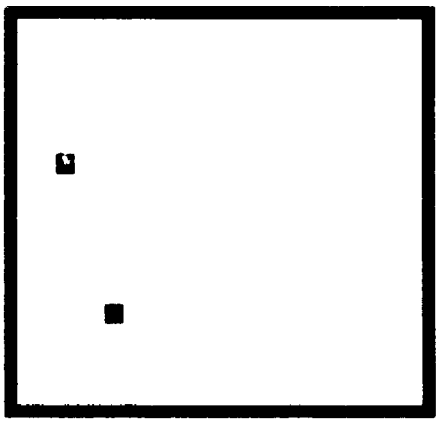
2.2.2.3 Chaos Game Representation of DNA Sequences

Chaos game representation of DNA sequences (Jeffrey 1990) was used to obtain a graphic display of the primary organization of large nucleotide sequences. The chaos game (Figure 1) is plotted on x and y axes, producing a square whose four vertices correspond to the four nucleotides of a DNA sequence, adenine (A), cytosine (C), guanine (G) and thymine (T). All chaos plots were constructed such that x and y were integer values and the x and y axes were set at a minimum value of 0 and a maximum value of 1000. The standard format for chaos plots sets the (0,0) coordinate as the vertex representing A, (0,1000) as C, (1000,1000) as G and (1000,0) as the T vertex. The first point (i.e., x,y coordinate) is plotted halfway between the center of the

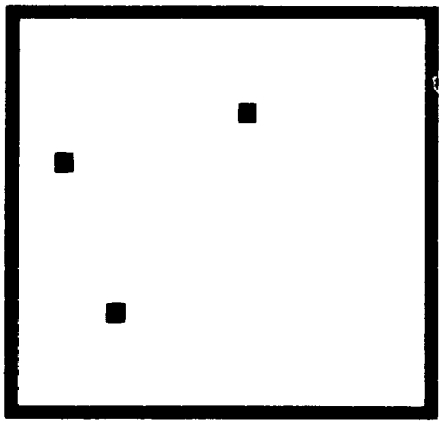
Figure 1. Chaos representation of DNA sequences. The chaos game is plotted on a square whose four vertices correspond to the four nucleotides adenine (A), cytosine (C), guanine (G) and thymine (T). For the short sequence "ACGT" adenine, the first nucleotide of the sequence is plotted halfway between the center of the square and the 'A' vertex. The second nucleotide, cytosine is plotted halfway between the previous point plotted and the 'C' vertex. Plotting continues with the remaining nucleotides in the DNA sequence being plotted half-distance between the previous point plotted and the vertex representing the current nucleotide being plotted.



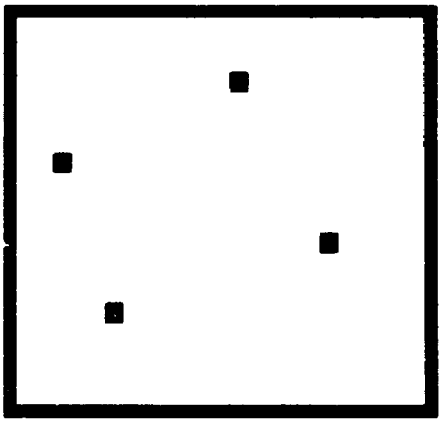
"A"



"AC"



"ACG"



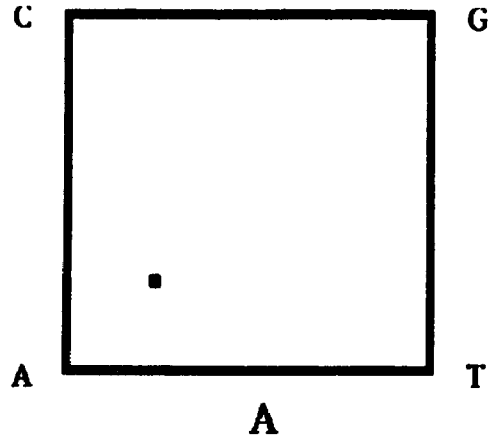
"ACGT"

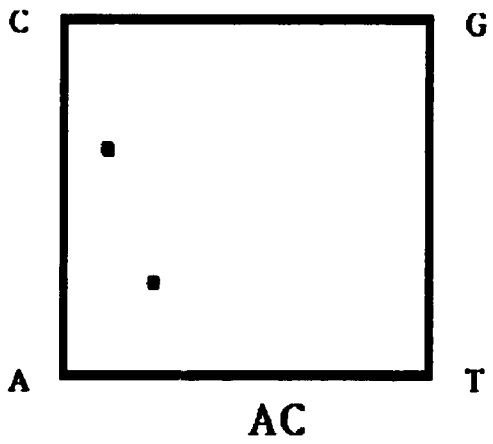
square and the vertex representing the first nucleotide of the sequence. Successive nucleotides in the sequence are plotted halfway between the previous point plotted and the vertex representing the nucleotide being plotted. The result is a transformation of a linear sequence of nucleotides into a graphic presentation of nucleotide composition and order in a two-dimensional scatter plot.

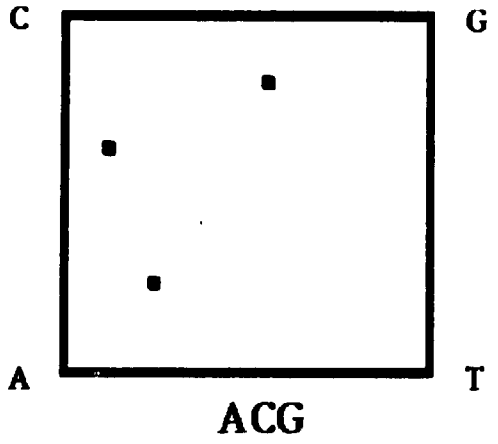
A rough method of quantification of data point distribution in chaos plots involves the division of the chaos plot into smaller quadrants, followed by determination of the frequency of data points within each of the subquadrants (Figure 2). Subquadrants of chaos plots representing particular subsequences were defined in terms of (x,y) coordinates and subsequence composition was determined through calculation of data point frequency within the representative subquadrants. Division of the chaos plot into four quadrants of equivalent size and determination of the data point frequency within each subquadrant revealed the single nucleotide composition of the sequence and a further division of the plot into 16 subquadrants of equivalent size provided information regarding the dinucleotide composition of the sequence (Jeffrey, 1990, 1992). Trinucleotide frequencies were obtained from examination of 64 equal-sized subquadrants and further subdivision of chaos plots was used to analyze the frequency of subsequences of increasing length. Coordinate data obtained from the chaos generator programs were divided into 4, 16 or 64 subquadrants and the number of points in each subquadrant was determined using the data base functions of Excel.

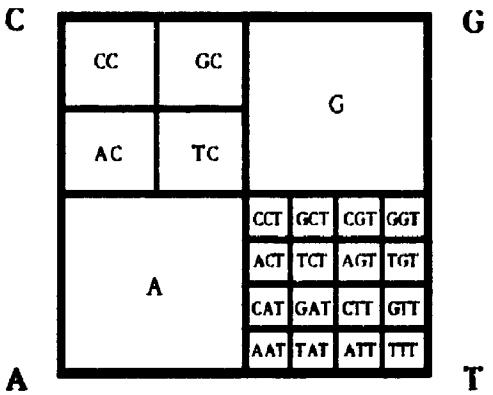
Subdivision of chaos plots and determination of short sequence frequencies was used to identify the biases in short sequence composition that best represented the biases observed in the overall chaos pattern. The $\chi 2$ statistic was employed to determine the uniformity of the data point distribution in

Figure 2. Subdivision of the chaos plot permits calculation of oligonucleotide frequencies. The chaos plot can be subdivided into quadrants and subquadrants etc., where the data point frequency in each subdivision represents the frequency of a particular oligonucleotide in the DNA sequence. This figure indicates the size and location of different quadrants and subquadrants in which the density of data points would correspond to either single, di- or trinucleotide frequencies. Further subdivision of the chaos plot would permit determination of the frequencies of longer subsequences.


C
G
A
T
A


C
G
A
T
AC


C
G
A
T
ACG


C
G
CC
GC
G
AC
TC
A
CCT
GCT
CGT
GGT
ACT
TCT
AGT
TGT
CAT
GAT
CTT
GTT
AAT
TAT
ATT
TTT
A
T

each chaos pattern among 4, 16 or 64 subquadrants. As well, standard genetic distances (Nei, 1975) between pairs of nucleotide sequences were calculated using the 4, 16 or 64 subquadrant data point frequencies representing single nucleotide, dinucleotide and trinucleotide composition.

Similarities and differences among chaos patterns were visually ascertained for same-length sequences and were described as relative differences in data point density between the four axes and four vertices of the chaos plot and among similar-sized subquadrants of the plot. Comparisons of chaos patterns for sequences of different lengths were made only after it was demonstrated that the major features of chaos patterns were not altered with increased sequence length. Chaos patterns generated for biological sequences were compared with the chaos patterns produced using computer-generated sequences that had known and simple sequence structures.

In contrast to the subdivision of a chaos plot into its constituent parts in order to describe quantitatively the overall pattern in data point distribution, two-dimensional contour plots were used to describe regional densities of data points within the entire plot. Contour lines were drawn around regions of a chaos plot that contained similar densities of data points. Different contour lines were used to represent different data point densities. Contour plots distinguish dense and sparse regions of chaos plots without consideration of the subquadrants of a chaos plot. Relative frequencies of subsequences of all possible combinations of length and composition are indicated. Contour plots were generated using the bivariate nonparametric kernel density estimators (default parameters) available in the Systat package (Wilkinson, 1991).

The resolution of a chaos plot with the dimensions used in this analysis was such that (x,y) coordinates repeated for identical subsequences ten nucleotides in length and of random composition and order. A biased

composition and/or a nonrandom order of nucleotides would tend to generate repetition of an (x,y) coordinate for a shorter subsequence length. Repeated (x,y) coordinates in a two-dimensional chaos plot are superimposed and thus not considered in a visual evaluation of chaos scatter plots. The "Chaos Frequency" program assigned an (x,y,z) coordinate to each nucleotide where the z value indicated the frequency of that particular (x,y) coordinate up to that nucleotide position in the plotting of the entire DNA sequence (Appendix 3). The (x,y,z) coordinates were used to generate three-dimensional chaos plots that display the composition of the repeated subsequence and the frequency of its repetition within the entire DNA sequence plotted. Repeated (x,y) coordinates are considered in contour plots and in the subdivided chaos plots used to portray the relative frequencies of subsequences that differ in composition but are of identical length.

2.2.2.4 Short-Sequence Representation

The UWGCG program 'Composition' was another method used to determine nucleotide, dinucleotide and trinucleotide content of nucleotide sequences. Single nucleotide, dinucleotide and trinucleotide frequencies were used to obtain measures of strand-symmetric relative abundances of short sequences (i.e., short-sequence representation; Cardon et al., 1994; Burge et al., 1992). The frequency of a nucleotide X was expressed as f_X. The representation of A and T ($f^*_{A/T}$) and C and G ($f^*_{C/G}$), considering the DNA sequence and the complementary strand, was determined as $f^*_{A/T} = 1/2(f_T + f_A)$ and $f^*_{C/G} = 1/2(f_C + f_G)$. Relative over- and under-representation of A and T versus C and G nucleotides were used to measure the degree of nonrandomness in single nucleotide composition. Standard assessment of oligonucleotide bias in

2.2.2.5 *General Statistical Analyses*

Distance measures between pairs of sequences based on short sequence representation, codon usage, and sequence alignment (nucleotide substitutions, etc) were compared using the nonparametric statistic, the Spearman rank-order correlation coefficient, r_S. Correlation coefficients were assumed to vary monotonically with genetic distance and thus were equated with similarity coefficients (Rogerson, 1991). These values were used to generate a cluster diagram (Wilkinson, 1991) describing phylogenetic relationships based upon analysis of primary sequence organization, specifically in terms of single nucleotide, dinucleotide and trinucleotide representation. The cluster diagrams generated for the three types of representation values were assessed in relation to the known phylogenies of the species used in this analysis.

The polymorphic index of Hamrick and Allard (1972) or the measure of gene diversity (Nei, 1975) were used to determine the heterozygosity at the *HLA-DQB1* locus in the two populations examined. The χ^2 statistic was used to determine if the genotypic frequencies were representative of a random association of alleles.

Chapter 3

RESULTS

3.1 DIRECT MEASUREMENT OF NUCLEOTIDE SEQUENCE VARIATION

3.1.1 The Third Exon of the *Adh2* locus

3.1.1.1 PCR-Amplification

The third exon of the human alcohol dehydrogenase gene encoding the ß subunit (i.e., the *Adh2* locus) was PCR-amplified from the genomic DNA of 29 individuals from Southwestern Ontario, Canada and 37 Athabaskan Dogrib individuals from the Northwest Territories of Canada (Table 7). Figure 3 contains the nucleotide sequence of the third exon of the *Adh2* locus *ß1* allele and indicates the positions of the PCR primers P1 and P2 that border this exon sequence. Three of the samples from Southwestern Ontario (individuals 30, 31 and 32) were not available at the time of the screening for nucleotide sequence variation at the third exon of the *Adh2* locus. Amplification using the P1 and P2 primer set was unsuccessful for the genomic DNA sample of one Dogrib individual (individual 64). It was possible to achieve PCR amplification of another region of the genome of individual number 64 using a different primer set (section 2.1.1.4). The failure of the PCR reaction using the P1 and P2 primer set and the PCR conditions given in section 2.1.2.4 may be due to a failure or inhibition of primer annealing possibly due to an alteration in either or both of the primer binding sequences bordering the exon sequence. For all other samples, the PCR

Table 7. A summary of the single strand polymorphism (SSCP) and DNA sequence analyses performed to determine the nature and extent of nucleotide sequence variation at the third exon of the *Adh2* locus and the second exon of the *HLA-DQβ1* locus. Human subjects included 32 individuals from Southwestern Ontario and 38 Athabaskan Dogrib individuals from the Northwest Territories, of Canada.

Types of analyses performed (+), or not performed (np)

	Adh2 SSCP/Sequence	*HLA-DQB1* Sequence		*Adh2* SSCP/Sequence	*HLA-DQB1* Sequence
Southwestern Ontario, Canada Individual No.			Northwest Territories, Canada Individual No.		
1	np/np	+	33	+/+	+
2	+/+	+	34	+/+	np
3	+/+	+	35	+/+	np
4	+/+	+	36	+/+	+
5	+/+	+	37	+/+	+
6	+/+	+	38	+/+	+
7	+/+	+	39	+/+	np
8	+/+	np	40	+/+	+
9	+/+	+	41	+/+	+
10	+/+	np	42	+/+	+
11	+/+	np	43	+/+	np
12	+/+	+	44	+/+	+
13	+/+	+	45	+/+	+
14	+/+	+	46	+/+	+
15	+/+	+	47	+/+	+
16	+/+	np	48	+/+	np
17	+/+	+	49	+/+	+
18	+/+	+	50	+/+	np
19	+/+	+	51	+/+	+
20	+/+	+	52	+/+	+
21	+/+		53	+/+	np
22	+/+	np	54	+/+	np
23	+/+	+	55	+/+	np
24	+/+	+	56	+/+	np
25	+/+	+	57	+/+	+
26	+/+	np	58	+/+	+
27	+/+	np	59	+/+	+
28	+/+	+	60	+/+	+
29	+/+	+	61	+/+	+
30	np/np	+	62	+/+	np
31	np/np	+	63		+
32	np/np	+	64	+/np	+
			65	+/+	+
			66	+/+	np
			67	+/+	+
			68	+/+	+
			69	+/+	+
			70	+/+	+

Figure 3. The third exon of the alcohol dehydrogenase gene encoding the *B* subunit, *Adh2*. The amino acid encoding nucleotide sequence is presented in the 5' to 3' orientation. The oligonucleotide primers P1 and P2 were used to amplify this exon. The exon sequence is written in triplets, the intron sequences are given as a continuous stretch of nucleotides and the primer sequences are in bold print. P2' represents the complementary sequence to P2, the actual primer. P2 was the primer used in the direct sequencing of the PCR product. The single codon altered between typical *B1* (CGC) and atypical *B2* (CAC) alleles at this locus is underlined.

P1

AATCTTTTCTGAATCTGAACAGCTTCTCTTTATTCTGTAG ATG GTG (I GTA GGA

ATC TGT CGC ACA GAT GAC CAC GTG GTT AGT GGC AAC CTG GTG ACC

CCC CTT CCT GTG ATT TTA GGC CAT GAG GCA GCC GGC ATC GTG GAG

AGT GTT GGA GAA GCG GTG ACT ACA GTC **AAA CCA GGTACAGGATTCACAC**

P2'

reactions were specific for the exon sequence, yielding a single fragment, 194 nts in length (Figure 4). All PCR reactions were free from nonspecific products and primer and product artifacts.

3.1.1.2 Single-Strand Conformation Polymorphism Analysis

The PCR-amplified and purified exon 3 fragments of *Adh2* were initially examined for sequence differences using two protocols for the analysis of single-strand conformation polymorphism (SSCP). The single-strand conformers of the amplified third exon of the *Adh2* locus were of three pattern-types (Figure 5). Pattern A consisted of two single-strand conformers migrating between nucleotide fragments of the double stranded DNA in the molecular size standards that were 1,018 and 517 nts in length. Pattern A was evident in the *Adh2* exon 3 sequences amplified from 27 of the 29 individuals of Southwestern Ontario. Pattern B consisted of two single-strand conformers that displayed slightly slower migration relative to both Pattern A conformers. Pattern B was in the single individual from Southwestern Ontario who was of Oriental ancestry (individual number 29). Pattern C was a four-band pattern indicative of the coexistence of two different alleles at the *Adh2* third exon. The migration of the four single-strand conformers in Pattern C was consistent with the existence of the single-strand conformers of both Pattern A and Pattern B (Figure 5c). The PCR-amplified exon 3 of the *Adh2* locus of a single individual from Southwestern Ontario (number 27) displayed the SSCP Pattern C. No evidence of heteroduplex formation by the PCR-amplified exon sequence was observed in any of the samples. Unlike the individuals from Southwestern Ontario, all Dogrib subjects showed a SSCP pattern consistent with the A type described above. The results of the SSCP analysis were identical in the two different SSCP

Figure 4. A representative set of PCR-amplified fragments (194 nts in length) of the *Adh2* third exon from the genomic DNA of a random sample of individuals from Southwestern Ontario and the Northwest Territories (lanes A to I). The fragments had been purified from the reactants and artifacts of the PCR reaction and were subjected to electrophoresis through a 2.0% agarose gel. The size of the fragment was calculated using the molecular size marker (M; 1 kb ladder, Life Technologies) as a standard.

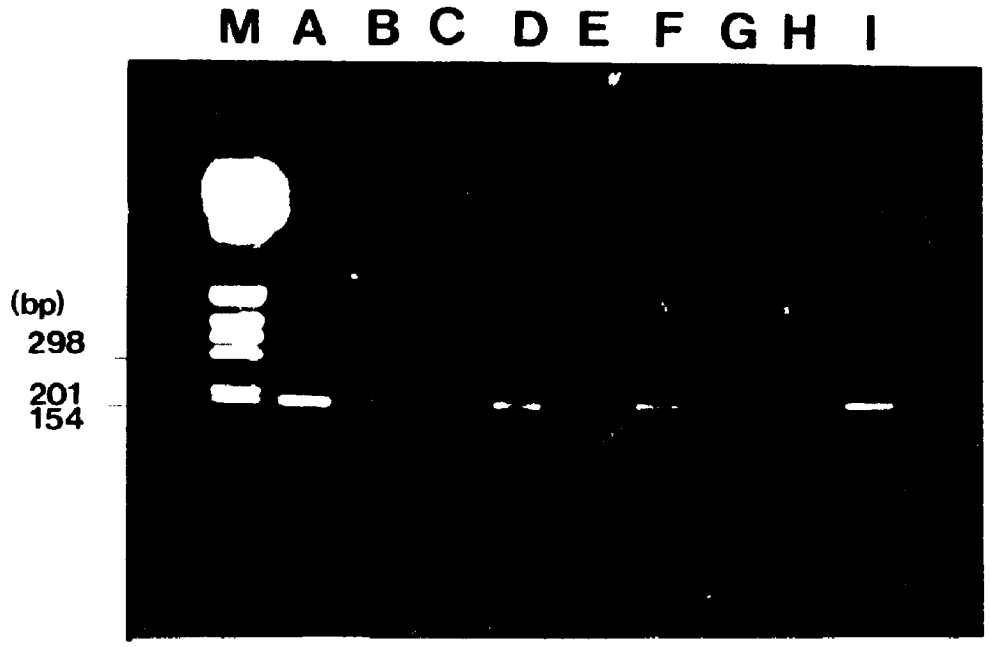
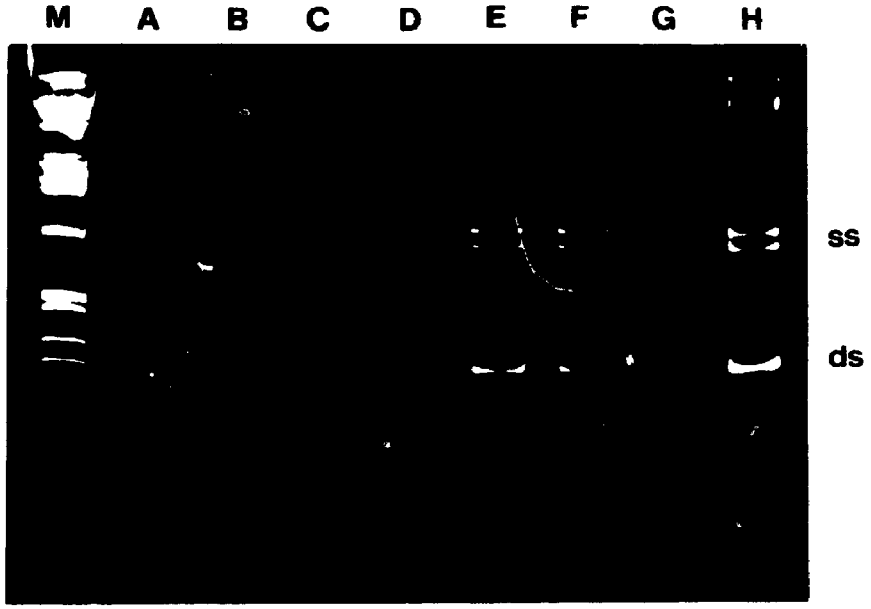

M A B C D E F G H I
(bp)
298
201
154

Figure 5. a) A representative set of the single-strand conformers (ss; lanes A to H) generated by the PCR-amplified fragments (ds; 194 nts) of the *Adh2* third exon from the genomic DNA of a random sample of individuals from Southwestern Ontario and the Northwest Territories (Dogrib individuals). The fragments had been purified from the reactants and artifacts of the PCR reaction and were subjected to electrophoresis through an 8% polyacrylamide gel. The mobility of the single-strand conformers was interpreted relative to the molecular size marker (M) λ DNA-*Hind* III/ϕX-174 RF DNA-*Hae* II, (Pharmacia). The single-strand polymorphism (SSCP) banding pattern of each of these samples is Pattern A

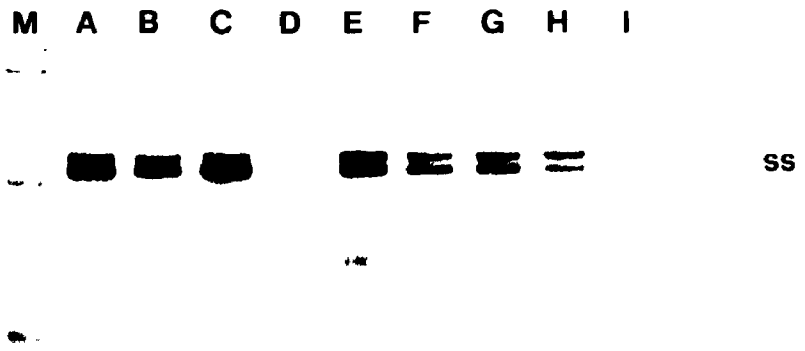
(b) A different representative set of the single-strand conformers generated by the PCR-amplified fragments (194 nts) of the *Adh2* third exon. Each of the SSCP banding patterns presented is that of Pattern A. The electrophoretic conditions were identical to those used in (a) except that the PCR amplified fragment was visualized through the use of γ-[^{32}P]-ATP end-labeling and the molecular size standard (M) used was the 1 kb ladder (Life Technologies).

(c) A different representative set of the single-strand conformers generated by the PCR-amplified fragments (194 nts) of the *Adh2* third exon. Three different SSCP banding patterns were identified (labelled A, B and C). Pattern C is a composite of patterns A and B. The electrophoretic conditions, molecular size standard (M) and visualization technique were identical to those used in (a).

a



b



c

protocols described in section 2.1.2.7.

3.1.1.3 *Nucleotide Sequence Variation in Individuals from Southwestern Ontario and Dogrib Individuals*

Nucleotide sequence variation among the *Adh2* exon 3 sequences of a sample of 28 individuals from Southwestern Ontario and 37 Dogrib individuals was ultimately described through the use of direct DNA sequencing of the PCR-amplified fragments containing the exon sequence. The complete *Adh2* exon 3 sequence is given in Figure 3. The exon 3 sequences from 26 individuals from Southwestern Ontario were identical to the nucleotide sequence of ß1, the 'typical' allele of the *Adh2* locus (Figure 6; Jornvall et al., 1984; Xu et al., 1988). One subject from Southwestern Ontario (individual 29) was homozygous for the *ß2*, 'atypical' allele sequence at exon 3. The nucleotide sequence *ß2* differs from that of the *ß1* by a single nucleotide substitution at the second nucleotide position of codon 47. One individual of the Southwestern Ontario sample (individual 27) was heterozygous at the *Adh2* locus and possessed both the *ß1* and *ß2* alleles (Figure 6). The difference between the alleles was identified due to the presence of two bands at the same horizontal position in a single set of sequencing reactions. The DNA sequence information was used to interpret the SSCP banding patterns (section 3.1.1.2) as follows: Pattern A is consistent with the presence of *ß1ß1* alleles, Pattern B with the presence of *ß2ß2* and Pattern C with the heterozygous condition, *ß1ß2*. DNA sequence information for the exon 3 of the *Adh2* gene for each of the 37 Dogrib individuals indicates that they are all homozygous for the *ß1* allele. No novel or intra-allelic nucleotide sequence variation was detected at the *Adh2* exon 3 among the individuals screened in this analysis.

Figure 6. The direct sequencing of the PCR-amplified third exon of the *Adh2* locus. The same region of the sequence is shown following 2 hrs (a to c) and 5 hrs (d and e) of electrophoresis. The four dideoxynucleotide termination reactions are labelled at the top of each lane (C, T, A and G for ddCTP, ddTTP, ddATP and ddGTP, respectively). a) The exon sequence derived from an individual who was heterozygous at this locus and possessed both the "typical" *B1* and "atypical" *B2* alleles. b) and c) The exon sequence from individuals who were homozygous for the "typical" *B1* allele. The heterozygous position is marked with an arrowhead. The two alleles differ by only one point mutation (G:C to A:T transition) and thus the heterozygous nucleotide position resulted in two bands occurring at the same horizontal position in two lanes (G and A) of the single set of sequencing reactions.


CTAG
a
b
c
d
e

Duplicate DNA sequence information was obtained using another primer (P1) in order to assess the reliability of the sequencing results. Sequencing was considered reliable after sequencing using primer P1 and only four samples for the following reasons. Preliminary sequence information obtained using the second primer (P1) had not revealed any inconsistencies in sequence information obtained using the primer P2. The results of the SSCP analysis were consistent with the sequence information derived from the use of one sequencing primer. No heteroduplex formation was observed that would otherwise have indicated heterozygosity in the PCR fragment not detected upon sequencing with only one primer.

It should be noted that sequence information was obtained within three nucleotides of the sequencing primer (Figure 3). Thus, the total number of nucleotide positions analyzed for sequence variation in a single allele were 147, of which the exon sequence comprised 129 nts. Thus, this study examines a total of 147 nucleotides at the *Adh2* locus in each of 65 diploid individuals from two different populations (28 individuals from Southwestern Ontario and 37 Dogrib individuals). This represents a screening of a total of 19,110 nucleotides. The only variants observed are consistent with the existence of the well known alleles of this locus.

3.1.2 The Second Exon of the *HLA-DQβ1* Locus

3.1.2.1 PCR-Amplification

A region of the second exon of the human leukocyte antigen locus *HLA-DQβ1* was PCR-amplified from genomic DNA of 25 individuals from Southwestern Ontario and 26 Dogrib individuals. The samples were selected

randomly for this analysis. Sequence information was obtained in duplicate for all samples examined. Figure 7 contains a known representative nucleotide sequence of the second exon of the *HLA-DQβ1* locus (Bodmer et al., 1994) and indicates the positions of the PCR primers P3 and P4 bordering this exon sequence. The second exon of *HLA-DQβ1* includes codons 6 to 85. The PCR-amplified region of the second exon includes codons 13 to 78. In all of the samples selected for analysis of nucleotide sequence variation, amplification of *HLA-DQβ1* exon 2 using P3 and P4 produced a single fragment of the expected length of 240 nts (Figure 8).

A 237-nucleotide region of the pseudogene *HLA-DQβ2* is known to coamplify with the amplification of the second exon of the *HLA-DQβ1* using the primer set P3 and P4 (Trucco et al., 1989). The agarose and polyacrylamide gel electrophoresis conditions used in this analysis were not sensitive enough to detect the presence of the homologous region of the pseudogene as it differs in length from the *HLA-DQβ1* exon 2 by a three-base pair deletion of the third position of codon 58 and the first and second codon positions of codon 59. SSCP protocols (section 2.1.2.7) were used to determine crudely the number of single-strand conformers produced by purified PCR fragments in eight samples. The presence of more than four single-strand conformers was taken as evidence of the presence of a contaminant in the PCR reaction, probably a region of the pseudogene *HLA-DQβ2.* Preliminary sequencing reactions using contaminated PCR templates readily displayed the presence of the shorter pseudogene template. The coamplification of the highly homologous pseudogene locus *HLA-DQβ2* in this study was eliminated by using the *Alu* I restriction digestion of genomic DNA and the primary PCR product. *Alu* I restriction sites are not within the second exon of the *HLA-DQβ1* gene.

PCR amplification of the *HLA-DQβ1* second exon also produced primer

Figure 7. The second exon of the human leukocyte antigen locus *HLA-DQβ1*. The amino acid encoding nucleotide sequence is presented in the 5' to 3' orientation. The oligonucleotide primers P3 and P4 were used to amplify the exon. The exon sequence is written in triplets, the primers used for PCR amplification are in bold print. P4' represents the complementary sequence to P4, the actual primer. Both P3 and P4 were used as primers in the direct sequencing of amplified exon sequences.

P3

AT TTC GTG TAC CAG TTT AAG GGC CTG TGC TAC TTC ACC AAC GGG

ACG GAG CGC GTG CGG GGT GTG ACC AGA CAC ATC TAT AAC CGA GAG

GAG TAC GTG CGC TTC GAC AGC GAC GTG GGG GTG TAC CGG GCA GTG

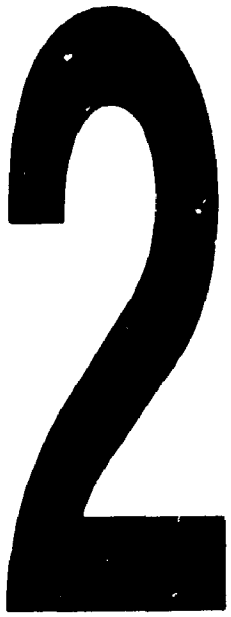
ACG CCG CAG GGG CGG CCT GTT GCC GAG TAC TGG AAC AGC CAG AAG

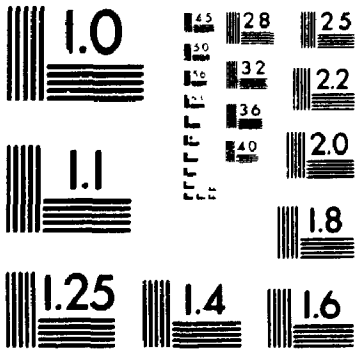
GAA GTC CTG GAG GGG GCC CGG GCG TCG GTG GAC AGG GTG **TGC AGA**

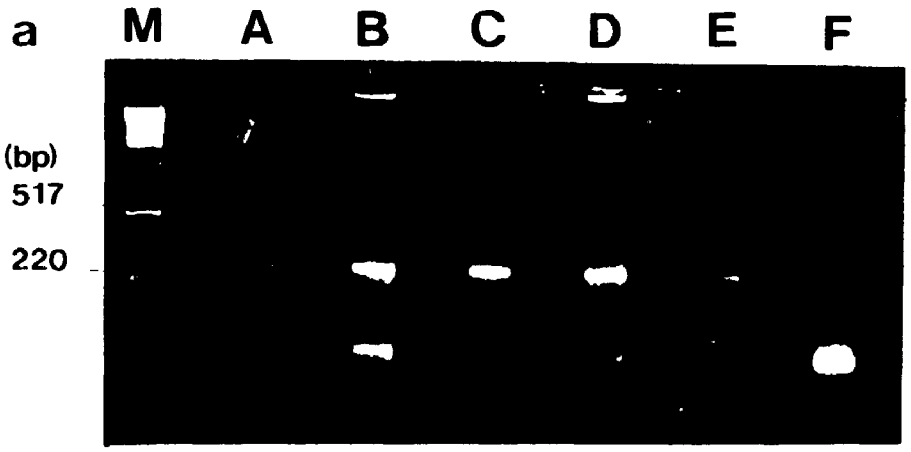
CAC AAC TAC GAG GTG G

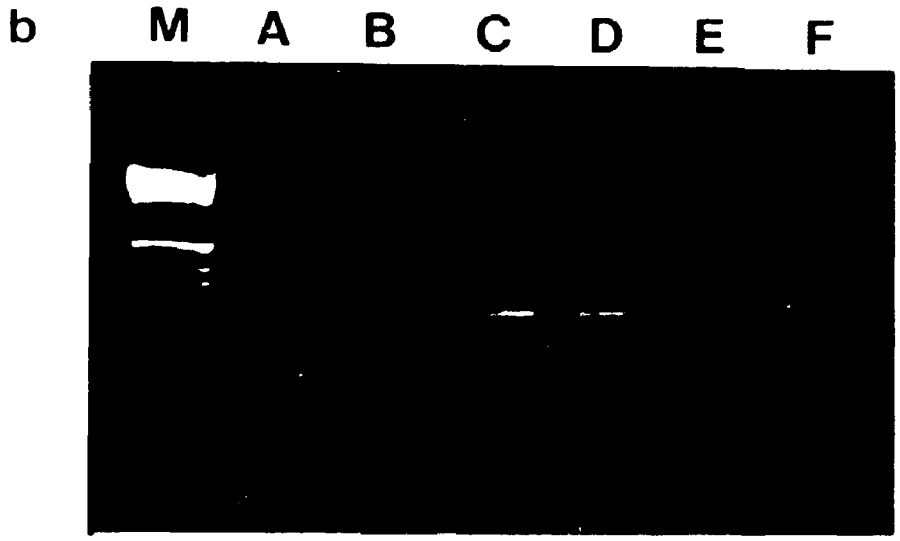
P4'

Figure 8. (a) A representative set of PCR-amplified fragments (240 nts) of the second exon of the *HLA-DQB1* locus (lanes A to F). The exon was amplified from the genomic DNA of a random sample of individuals from Southwestern Ontario and the Northwest Territories. The fragments in (b) were purified from the reactants and artifacts of the PCR reaction (lanes A to F). In both (a) and (b) fragments were electrophoresed through a 2.0% agarose gel. The size of the fragment was calculated using the molecular size marker (M; 1 kb ladder, Life Technologies) as a standard.






a
M
A
B
C
D
E
F
(bp)
517
220


b
M
A
B
C
D
E
F

and product artifacts necessitating the use of extensive purification protocols (section 2.1.2.6). Primer artifacts included primer-dimer formation and concatemers of primer sequences. Product artifacts were identified by their presence in preliminary DNA sequencing reactions and were the result of the formation of hairpin structures in single-stranded conformers of the PCR product during the PCR reaction. The hairpin structures created new priming sites in subsequent rounds of amplification in PCR reactions. Longer PCR products were produced that comigrated with double stranded molecular size standards of approximately 500 nts.

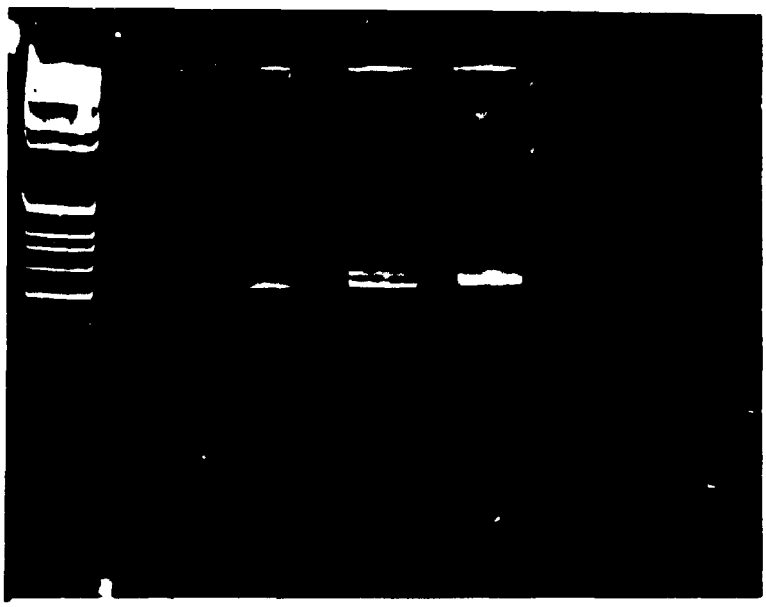
Heteroduplex formation was a frequent occurrence among the PCR products and was seen as a doublet banding pattern upon electrophoresis in 8% polyacrylamide gels (Figure 9a). The homoduplex migrated to a position consistent with the length of the double stranded DNA (240 bp). The heteroduplex which contained mismatches and bulges in the DNA duplex migrated at a slightly slower rate than the homoduplex form of the PCR product. The presence of heteroduplex formation was interpreted as representing the heterozygous condition of the amplified region.

Preliminary SSCP analysis of eight purified PCR reactions revealed seven unique SSCP pattern types (three pattern types are shown in Figure 9b). SSCP analyses were not pursued as the limitations of the SSCP protocols employed in this analysis offered poor resolution of the single-strand conformers of this exon sequence and the limitations of the SSCP technique, in comparison with sequencing techniques, are well documented (Sarkar et al., 1992). As a result, all nucleotide sequence variation was identified using the direct sequencing of the PCR-amplified second exon of the *HLA-DQβ1* locus.

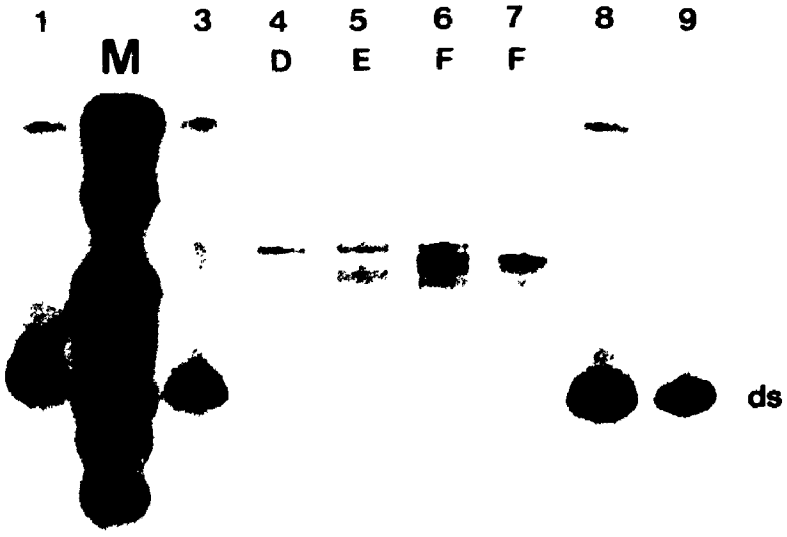
Figure 9. (a) A representative set of homo- and heteroduplex conformers of the PCR-amplified fragments (240 nts) of the second exon of the *HLA-DQβ1* locus. The exon was amplified from the genomic DNA of a random sample of individuals from Southwestern Ontario and the Northwest Territories. The doublet band pattern in lanes A to D is interpreted as resulting from heterozygosity at this exon sequence. Annealing of the complementary DNA single strands from two different alleles forms an imperfect double-stranded heteroduplex. The heteroduplex conformer of the double-stranded DNA migrates slower than the homoduplex form.

In (b) lanes 1,3, 8 and 9 contained nondenatured PCR-amplified fragments of this exon (ds, double-stranded DNA). These PCR fragments were melted and the migration of the single-stranded conformers was examined in lanes 4 to 7. Three different single-strand conformation polymorphism (SSCP) pattern-types were observed (labeled D, E and F). DNA was visualized using γ-[^{32}P]-ATP end labeling. The fragments in (a) and (b) had been purified from the reactants and artifacts of the PCR reaction and were electrophoresed through an 8% polyacrylamide gel. The size of the fragment was calculated using the molecular size marker (M; 1 kb ladder, Life Technologies) as a standard.

a



b



3.1.2.2 Nucleotide Sequence Variation in Individuals From Southwestern Ontario

The nucleotide sequence of a region of the second exon of the *HLA-DQβ1* locus was determined for 25 individuals of Southwestern Ontario using the techniques of PCR amplification and direct dideoxynucleotide DNA sequencing of the PCR products. Representative examples of the sequence information obtained are given in Figure 10. It is apparent from these data that most individuals are heterogeneous with bands occurring at the same horizontal position in two lanes of a set of sequencing reactions. Such results are expected since the two alleles were not separated prior to sequencing. Sequence information was obtained in duplicate for each sample using both of the PCR primers P3 and P4. The complete sequence information for all individuals is contained in Figure 11. The nucleotide sequences were compared with the 26 known allele sequences at this exon (Appendix 1) in order to interpret the alleles possessed by each individual. The first digit of the numeric designations for the alleles indicates the allele subtype. Given that the sequencing protocol was not preceded by the isolation of the two alleles in heterozygous individuals, it is not possible to present the individual allele sequences and to identify directly the sequence variability inherent to each allele. As well, these techniques do not permit determination of the phase of multiple novel mutations (i.e., which allele contains each novel mutation). Such determinations can only be made following the separation of the two alleles and the determination of each nucleotide sequence. The observed sequence variation were attributed to the existence of known allele sequences without introducing any novel sequence differences.

The alleles at the *HLA-DQβ1* locus in the sample of 25 individuals from

Figure 10. The direct sequencing of the PCR-amplified second exon of the *HLA-DQβ1* locus. The four dideoxynucleotide termination reactions are labeled at the top of each lane (C, T, A and G for ddCTP, ddTTP, ddATP and ddGTP, respectively). Panels a) and b) show the nature of the results for sequence differences among individuals and heterozygosity within a single individual. In a) three positions that differed among five different individuals are marked with arrowheads. At position 1 the nucleotides are G, G/A, G/A, A and G, in individuals 1 to 5 respectively. At position 2 the nucleotides are C, C/T C/T, C and C. At position 3 the nucleotides are C, C, C/T C and C. In b) the position labeled 4 contains C, A, and C/A in individuals 1 to 3, respectively.


b
1
2
3
4>

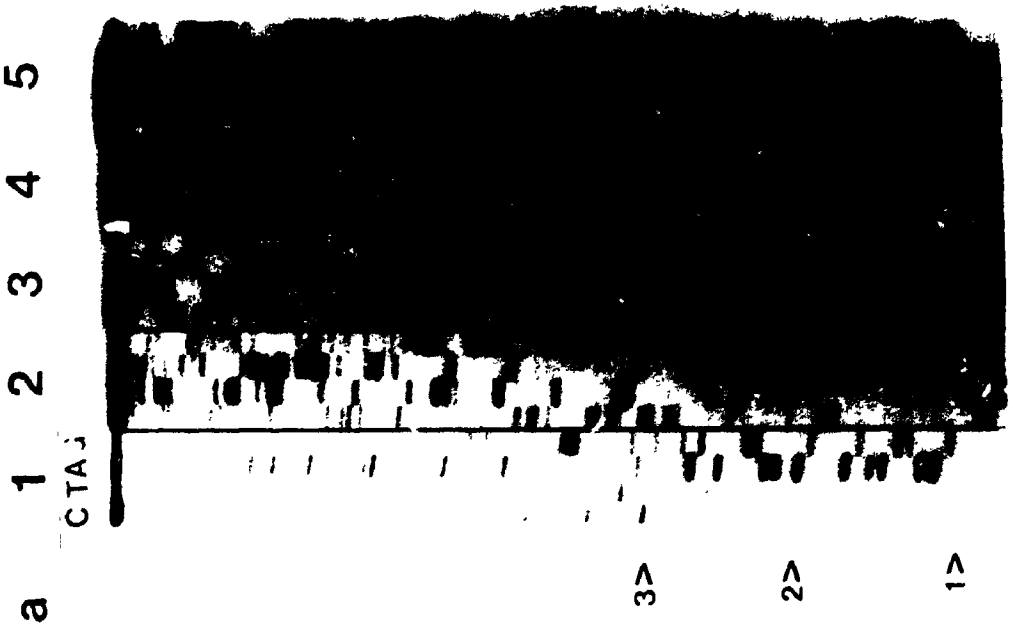
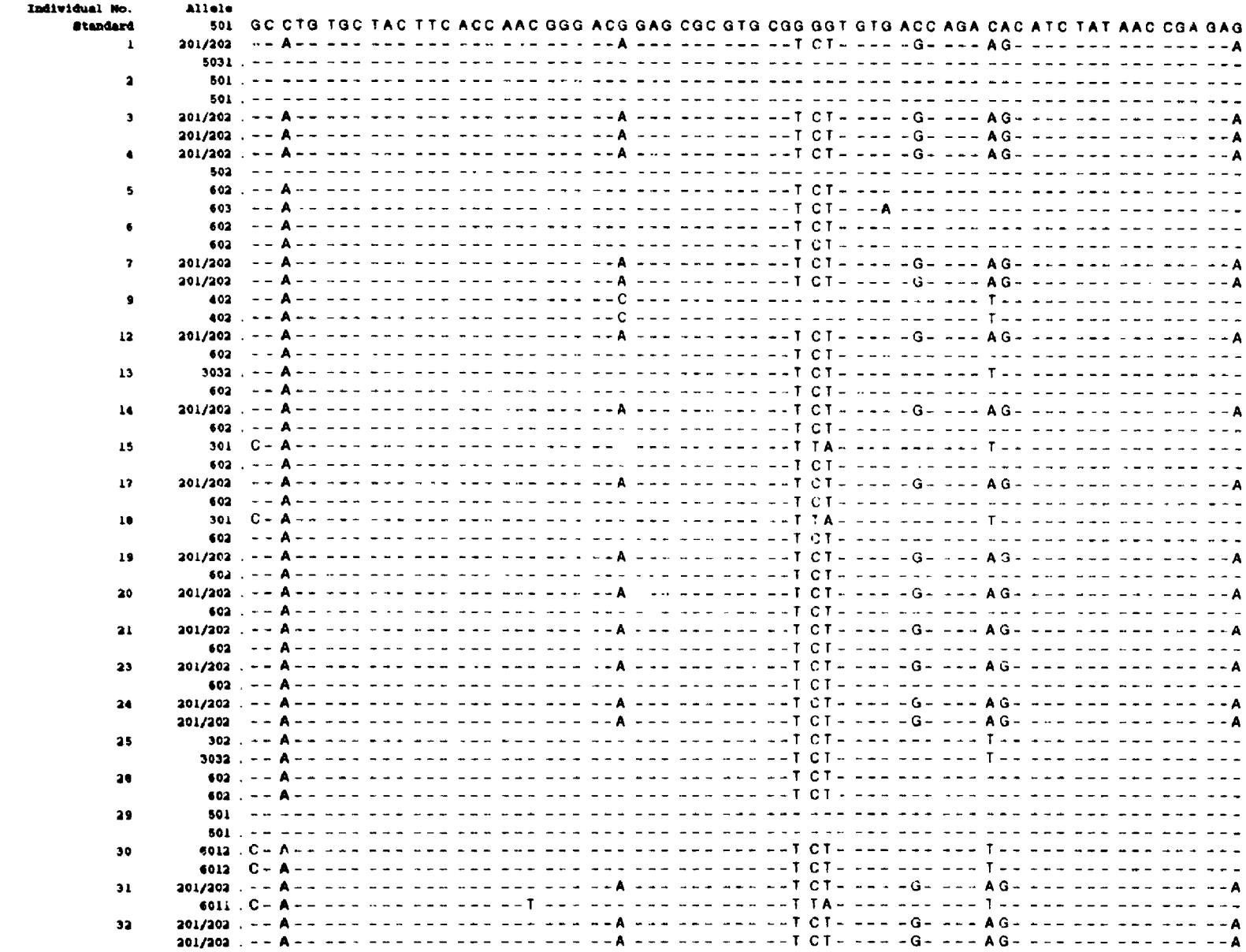
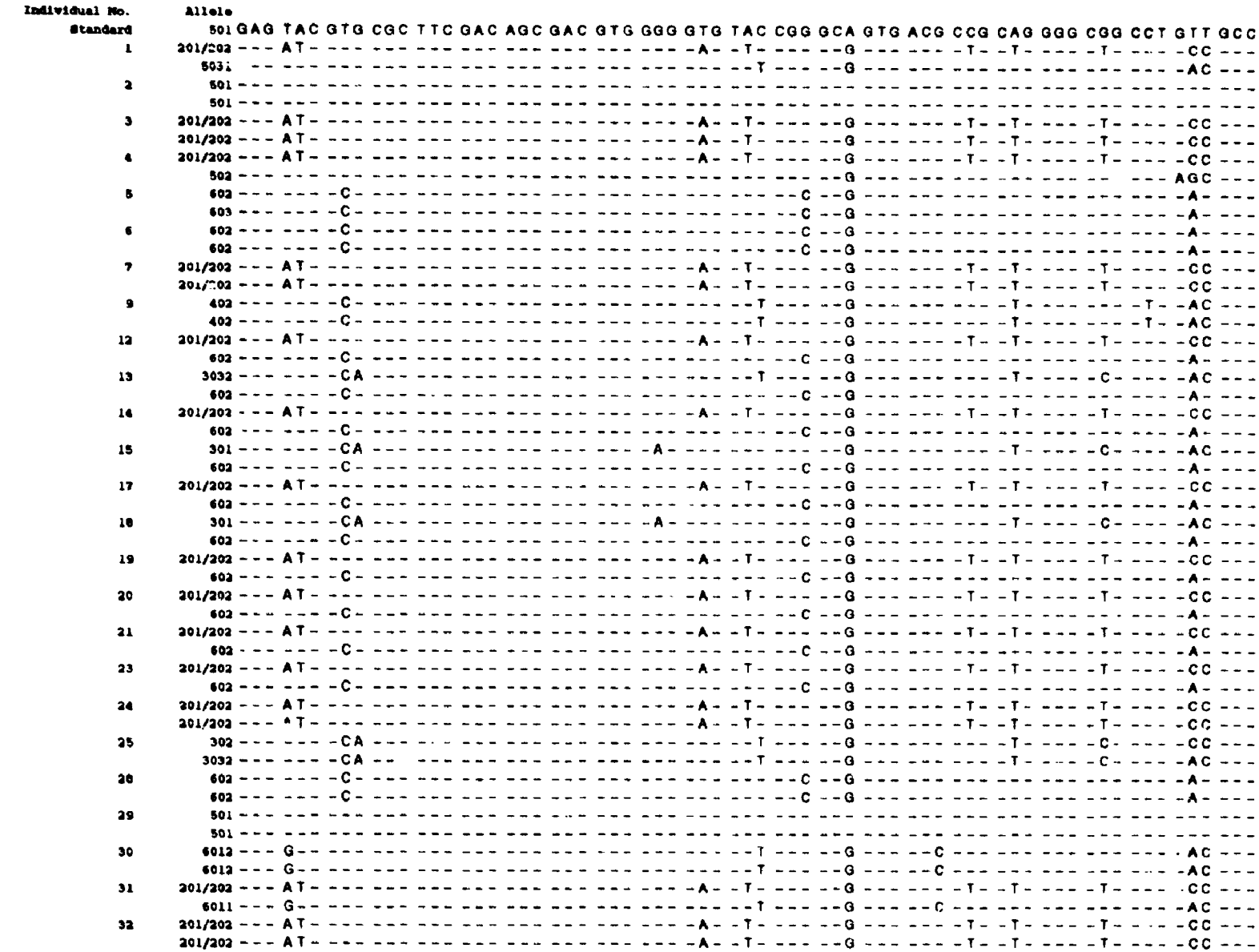
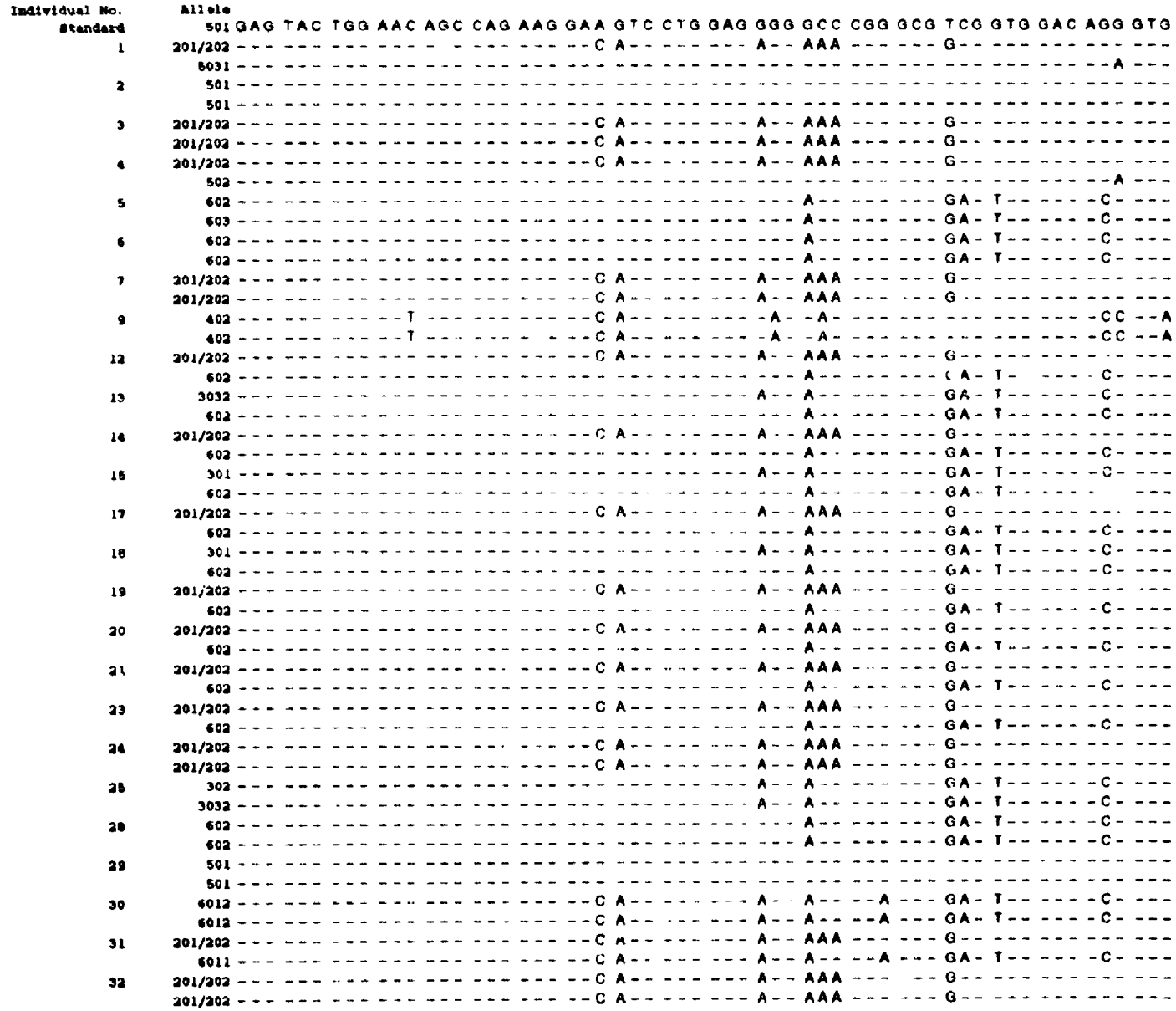

a
1
2
3
4
5
CTAJ
3>
2>
1>

Figure 11. An alignment of the nucleotide sequences found in a large region of the second exon of the *HLA-DQβ1* locus in a sample of 25 individuals from Southwestern Ontario. The allele designations for the nucleotide sequences are listed to the left of each sequence. The alleles 201 and 202 do not differ in the region examined in this analysis and are thus considered as a single allele 201/202. The primer sites used in the PCR amplification of this region of *HLA-DQβ1* exon 2 are located immediately 5' and 3', respectively of the sequence presented. Nucleotides are identified at those positions where differences occurred in comparison with allele 501, a standard commonly used for comparison. A dash is used to indicate no difference from allele 501


Individual No.
Allele
Standard
501 GC CTG TGC TAC TTC ACC AAC GGG ACG GAG CGC GTG CGG GGT GTG ACC AGA CAC ATC TAT AAC CGA GAG
1 201/202
5031
2 501
501
3 201/202
201/202
4 201/202
503
5 602
603
6 602
602
7 201/202
201/202
9 402
402
12 201/202
602
13 3032
602
14 201/202
602
15 301
602
17 201/202
602
18 301
602
19 201/202
602
20 201/202
602
21 201/202
602
23 201/202
602
24 201/202
201/202
25 302
3032
26 602
602
29 501
501
30 6012
6012
31 201/202
6011
32 201/202
201/202


Individual No.
Allele
Standard
501 GAG TAC GTG CGC TTC GAC AGC GAC GTG GGG GTG TAC CGG GCA GTG ACG CCG CAG GGG CGG CCT GTT GCC
1 201/202 --- AT- --- --- --- --- --- --- --- --- --- --- --- -A- -T- --- --G --- --- -T- -T- --- -T- --- -CC ---
5031 -- --- --- --- --- --- --- --- --- --- --- --- --- --T --- --G --- --- --- --- --- --- --- --- -AC ---
2 501 ---
501 ---
3 201/202 --- AT- --- --- --- --- --- --- --- --- --- --- --- -A- -T- --- --G --- --- -T- -T- --- -T- --- -CC ---
201/202 --- AT- --- --- --- --- --- --- --- --- --- --- --- -A- -T- --- --G --- --- -T- -T- --- -T- --- -CC ---
4 201/202 --- AT- --- --- --- --- --- --- --- --- --- --- --- -A- -T- --- --G --- --- -T- -T- --- -T- --- -CC ---
502 --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --G --- --- --- --- --- --- --- AGC ---
5 602 --- --- --- -C- --- --- --- --- --- --- --- --- --- --- --- --- --C --G --- --- --- --- --- --- --- -A- ---
603 --- --- --- -C- --- --- --- --- --- --- --- --- --- --- --- --- --C --G --- --- --- --- --- --- --- -A- ---
6 602 --- --- --- -C- --- --- --- --- --- --- --- --- --- --- --- --- --C --G --- --- --- --- --- --- --- -A- ---
602 --- --- --- -C- --- --- --- --- --- --- --- --- --- --- --- --- --C --G --- --- --- --- --- --- --- -A- ---
7 201/202 --- AT- --- --- --- --- --- --- --- --- --- --- --- -A- -T- --- --G --- --- -T- -T- --- -T- --- -CC ---
201/202 --- AT- --- --- --- --- --- --- --- --- --- --- --- -A- -T- --- --G --- --- -T- -T- --- -T- --- -CC ---
9 402 --- --- --- -C- --- --- --- --- --- --- --- --- --- --- --T --- --G --- --- --- --- -T- --- --- -T- -AC ---
402 --- --- --- -C- --- --- --- --- --- --- --- --- --- --- --T --- --G --- --- --- --- -T- --- --- -T- -AC ---
12 201/202 --- AT- --- --- --- --- --- --- --- --- --- --- --- -A- -T- --- --G --- --- -T- -T- --- -T- --- -CC ---
602 --- --- --- -C- --- --- --- --- --- --- --- --- --- --- --- --- --C --G --- --- --- --- --- --- --- -A- ---
13 3032 --- --- --- -CA --- --- --- --- --- --- --- --- --- --- --- --T --- --G --- --- --- --- -T- --- -C- --- -AC ---
602 --- --- --- -C- --- --- --- --- --- --- --- --- --- --- --- --- --C --G --- --- --- --- --- --- --- -A- ---
14 201/202 --- AT- --- --- --- --- --- --- --- --- --- --- --- -A- -T- --- --G --- --- -T- -T- --- -T- --- -CC ---
602 --- --- --- -C- --- --- --- --- --- --- --- --- --- --- --- --- --C --G --- --- --- --- --- --- --- -A- ---
15 301 --- --- --- -CA --- --- --- --- --- --- --- -A- --- --- --- --- --G --- --- --- --- -T- --- --- -C- --- -AC ---
602 --- --- --- -C- --- --- --- --- --- --- --- --- --- --- --- --- --C --G --- --- --- --- --- --- --- -A- ---
17 201/202 --- AT- --- --- --- --- --- --- --- --- --- --- --- -A- -T- --- --G --- --- -T- -T- --- -T- --- -CC ---
602 --- --- --- -C- --- --- --- --- --- --- --- --- --- --- --- --- --C --G --- --- --- --- --- --- --- -A- ---
18 301 --- --- --- -CA --- --- --- --- --- --- --- --- -A- --- --- --- --G --- --- --- --- -T- --- --- -C- --- -AC ---
602 --- --- --- -C- --- --- --- --- --- --- --- --- --- --- --- --- --C --G --- --- --- --- --- --- --- -A- ---
19 201/202 --- AT- --- --- --- --- --- --- --- --- --- --- --- -A- -T- --- --G --- --- -T- -T- --- -T- --- -CC ---
602 --- --- --- -C- --- --- --- --- --- --- --- --- --- --- --- --- --C --G --- --- --- --- --- --- --- -A- ---
20 201/202 --- AT- --- --- --- --- --- --- --- --- --- --- --- -A- -T- --- --G --- --- -T- -T- --- -T- --- -CC ---
602 --- --- --- -C- --- --- --- --- --- --- --- --- --- --- --- --- --C --G --- --- --- --- --- --- --- -A- ---
21 201/202 --- AT- --- --- --- --- --- --- --- --- --- --- --- -A- -T- --- --G --- --- -T- -T- --- -T- --- -CC ---
602 --- --- --- -C- --- --- --- --- --- --- --- --- --- --- --- --- --C --G --- --- --- --- --- --- --- -A- ---
23 201/202 --- AT- --- --- --- --- --- --- --- --- --- --- --- -A- -T- --- --G --- --- -T- -T- --- -T- --- -CC ---
602 --- --- --- -C- --- --- --- --- --- --- --- --- --- --- --- --- --C --G --- --- --- --- --- --- --- -A- ---
24 201/202 --- AT- --- --- --- --- --- --- --- --- --- --- --- -A- -T- --- --G --- --- -T- -T- --- -T- --- -CC ---
201/202 --- AT- --- --- --- --- --- --- --- --- --- --- --- -A- -T- --- --G --- --- -T- -T- --- -T- --- -CC ---
25 302 --- --- --- -CA --- --- --- --- --- --- --- --- --- --- --T --- --G --- --- --- --- -T- --- -C- --- -CC ---
3032 --- --- --- -CA --- --- --- --- --- --- --- --- --- --- --T --- --G --- --- --- --- -T- --- -C- --- -AC ---
26 602 --- --- --- -C- --- --- --- --- --- --- --- --- --- --- --- --- --C --G --- --- --- --- --- --- --- -A- ---
602 --- --- --- -C- --- --- --- --- --- --- --- --- --- --- --- --- --C --G --- --- --- --- --- --- --- -A- ---
29 501 ---
501 ---
30 6012 --- G-- --- --- --- --- --- --- --- --- --- --- --- --- --T --- --G --- --C --- --- --- --- --- . AC ---
6012 --- G-- --- --- --- --- --- --- --- --- --- --- --- --- --T --- --G --- --C --- --- --- --- --- -AC ---
31 201/202 --- AT- --- --- --- --- --- --- --- --- --- --- --- -A- -T- --- --G --- --- -T- -T- --- -T- --- CC ---
6011 --- G-- --- --- --- --- --- --- --- --- --- --- --- --- --T --- --G --- --C --- --- --- --- --- -AC ---
32 201/202 --- AT- --- --- --- --- --- --- --- --- --- --- --- -A- -T- --- --G --- --- -T- -T- --- -T- --- -CC ---
201/202 --- AT- --- --- --- --- --- --- --- --- --- --- --- -A- -T- --- --G --- --- -T- -T- --- -T- --- -CC ---


Individual No.
Allele
Standard
501 GAG TAC TGG AAC AGC CAG AAG GAA GTC CTG GAG GGG GCC CGG GCG TCG GTG GAC AGG GTG
1
201/202
5031
2
501
501
3
201/202
201/202
4
201/202
502
5
602
603
6
602
602
7
201/202
201/202
9
402
402
12
201/202
602
13
3032
602
14
201/202
602
15
301
602
17
201/202
602
18
301
602
19
201/202
602
20
201/202
602
21
201/202
602
23
201/202
602
24
201/202
201/202
25
302
3032
28
602
602
29
501
501
30
6012
6012
31
201/202
6011
32
201/202
201/202

Southwestern Ontario are listed in Table 8. Individuals 17, 18, 19 and 21 were related to another member of the study and were not included in analysis of the extent or nature of the genetic variation in the sample of individuals from Southwestern Ontario. It should be noted that alleles 201 and 202 are identical at the region of the second exon examined in this study and are thus considered as a single allele 201/202. There were 10 cases of homozygosity among the 21 unrelated individuals and the theoretical heterozygosity (Hamrick and Allard, 1972; Nei, 1975) calculated from the observed genotypes was 0.78. Twelve different alleles exist among this group of individuals. The alleles were of the five known subtypes and ranged in frequency from 0.02 to 0.36 (Table 9). The pairs of alleles observed in the 13 unique genotypes differed significantly from that expected of a random association of the alleles ($p<0.01$). All observed sequence variation was compatible with the sequences of existing alleles of the *HLA-DQβ1* locus and no intra-allelic nucleotide substitutions (i.e., no additional sequence differences from those known to exist for alleles identified previously) were observed in this analysis.

Of 197 nucleotide positions of the PCR-amplified regions 44 were variable among the different alleles observed. At the 44 variable sites, nucleotide substitutions are equally frequent at second and third codon positions (39%) and are only slightly less frequent at the first codon position (23%). Amino acid replacement occurred at 71% of the nucleotide sites with substitutions. At the 22 amino acid replacement sites among the alleles there are 10 conservative and 34 nonconservative amino acid replacements. Transversions are only slightly more frequent than transitions (52% and 48%, respectively). At CpG sites, the frequency of transitions and transversions does not differ significantly given the relative frequencies of all possible transition and transversion events ($p<0.01$). The variable sites were compatible with the existence of known allele sequences

Table 8. A summary of the alleles determined to exist at the *HLADQβ1* locus in a random sample of 25 individuals of Southwestern Ontario. The allele designations were based on the nucleotide sequencing of a region of the second exon of the *HLADQβ1* locus in each individual.

Southwestern Ontario Individual No.	*HLA-DQB1* Second Exon Allele	Allele
1	201/202	5031
2	501	501
3	201/202	201/202
4	201/202	502
5	602	603
6	602	602
7	201/202	201/202
9	402	402
12	201/202	602
13	3032	602
14	201/202	602
15	301	602
17*	201/202	602
18*	301	602
19*	201/202	602
20	201/202	602
21*	201/202	602
23	201/202	602
24	201/202	201/202
25	302	3032
28	602	602
29	501	501
30	6012	6012
31	201/202	6011
32	201/202	201/202

* an individual related to another sample member(s)

Table 9. The frequencies of 12 alleles at the *HLA-DQß1* locus identified in a sample of 25 individuals from Southwestern Ontario. Twenty-one of the individuals represent a random sample of unrelated individuals (corrected). The alleles were identified from PCR amplification and direct nucleotide sequence determination of a region of the second exon of the *HLA-DQß1* locus.

HLA-DQB1 **Locus in Individuals of Southwestern Ontario**

Allele	Allele Frequency	
	N = 25	N = 21 (corrected)
201/202	0.36	0.36
301	0.04	0.02
302	0.02	0.02
3032	0.04	0.05
402	0.04	0.05
501	0.08	0.10
502	0.02	0.02
5031	0.02	0.02
6011	0.02	0.02
6012	0.04	0.05
602	0.30	0.26
603	0.02	0.02

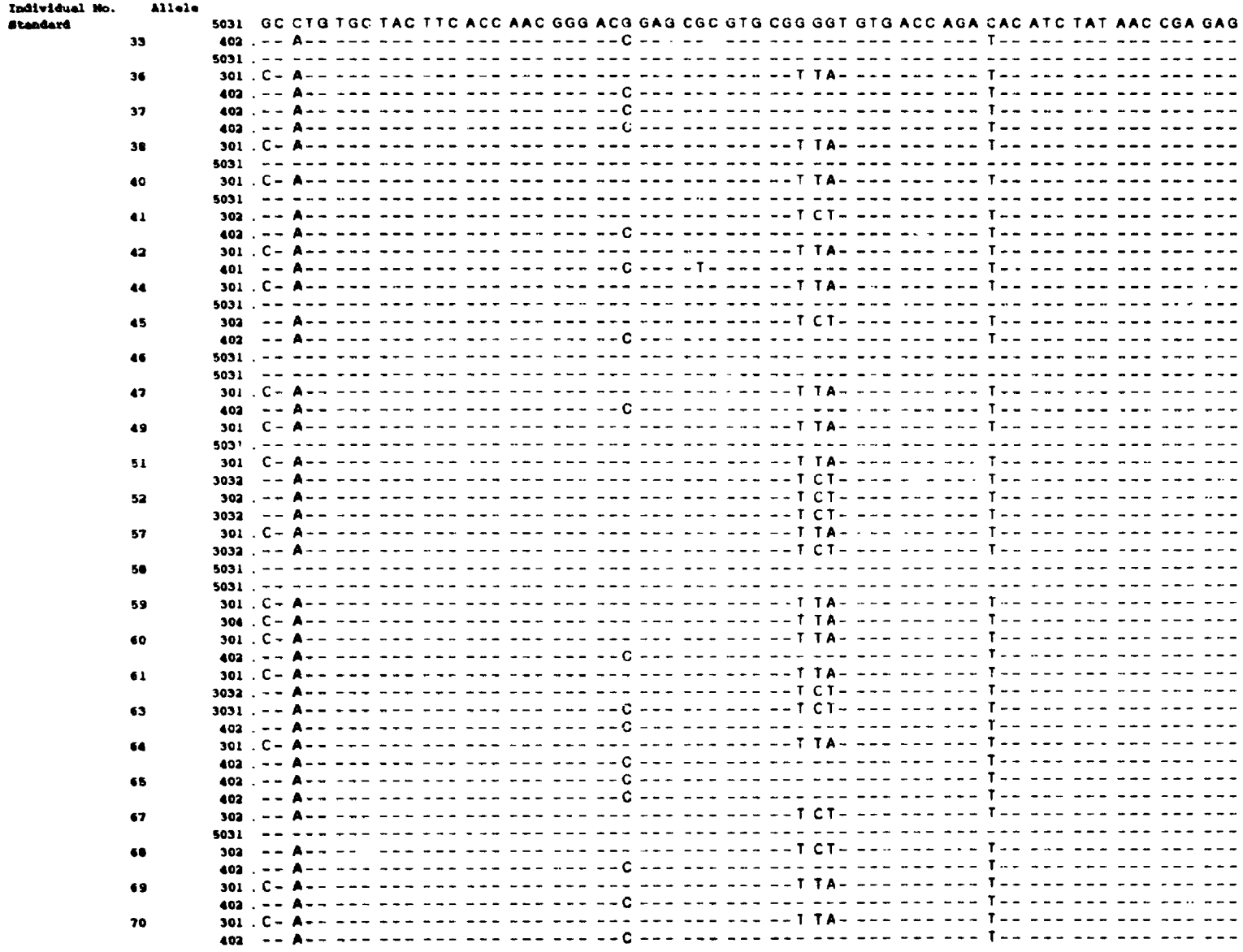
and the absence of any novel or intra-allelic variability.

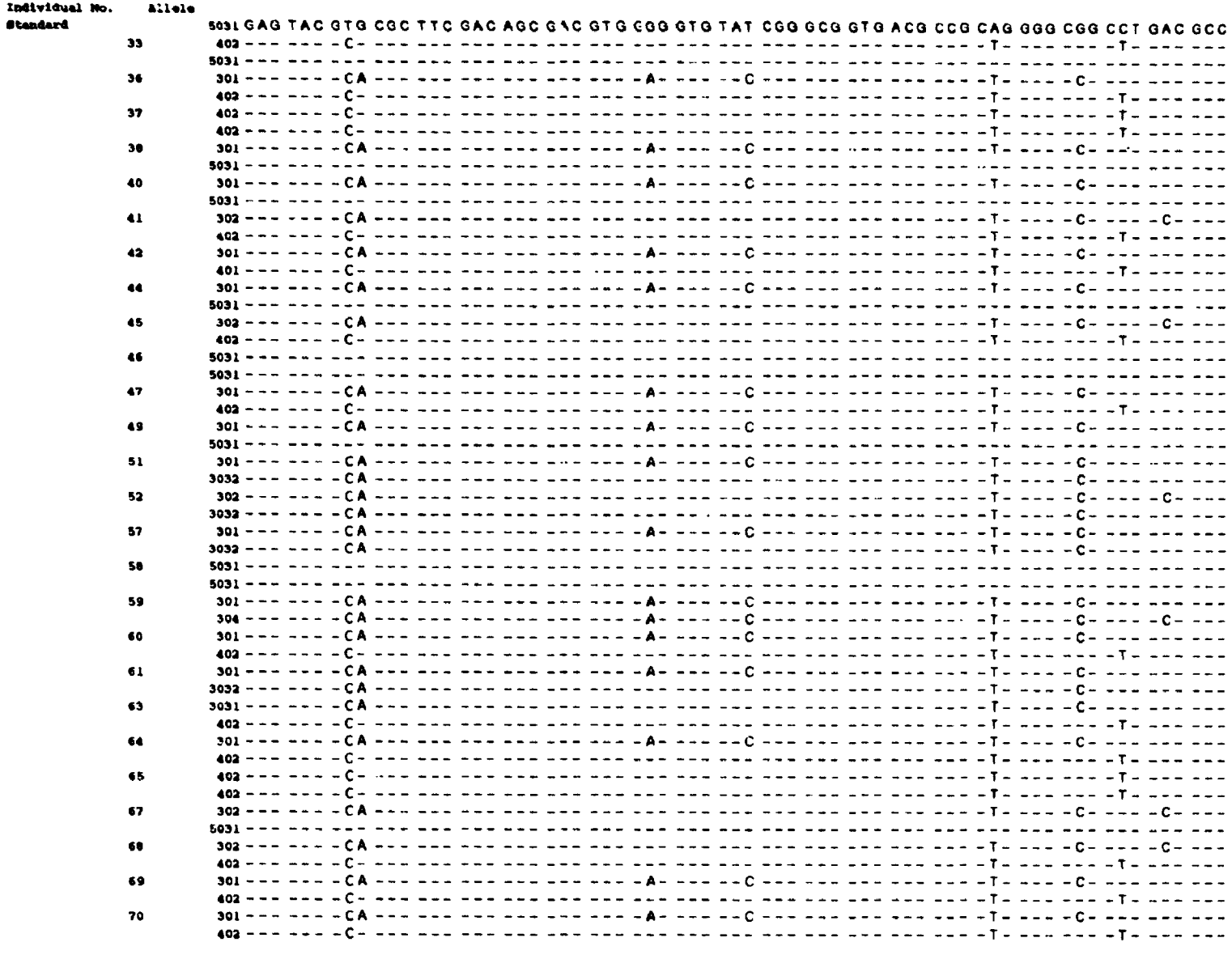
3.1.2.3 Nucleotide Sequence Variation in a Sample of Dogrib Individuals

The nucleotide sequence of a region of the second exon of the *HLA-DQβ1* locus was determined for 26 Dogrib individuals using the techniques of PCR amplification and direct dideoxynucleotide DNA sequencing. Sequence information was obtained in both directions for each sample using the PCR primers P3 and P4 (Figure 7). The sequence determined for each Dogrib individual is contained in Figure 12. Twenty-two of 26 individuals show heterozygosity in this genomic region. These individuals show two bands at the same horizontal position in a single set of sequencing reactions at a number of positions in the sequence. The observed sequence variation can be attributed to the existence of known allele sequences without introducing any novel sequence differences. The alleles at the *HLA-DQβ1* locus for the 26 Dogrib individuals surveyed are listed in Table 10. Alleles were assigned following comparison of sequence information with the 26 known alleles at this locus given in Appendix 1. Alleles 201 and 202 are identical in the sequenced gene region but differ in in other regions of the gene. Eight different alleles of three different subtypes exist among this group of individuals. There were four cases of homozygosity among the 26 individuals and the theoretical heterozygosity calculated from the observed genotypes was 0.82 The frequencies of the eight alleles observed in this population ranged in frequency from 0.02 to 0.29 (Table 11). The pairs of alleles observed in the form of 12 unique genotypes did not differ significantly from that expected of a random association of the alleles ($p<0.01$).

Alignment of the different allele sequences at this exon revealed that 29 of 197 nucleotide positions contain single nucleotide substitutions. Transversions

Figure 12. An alignment of the nucleotide sequences determined from examination of a large region of the second exon of the *HLA-DQβ1* locus in a sample of 25 individuals from Southwestern Ontario. The allele designations for the nucleotide sequences are listed to the left of each sequence. The primer sites used in the PCR amplification of this region of *HLA-DQβ1* exon 2 are in bold print. Nucleotides are identified at those positions where differences occur in comparison with the allele 501. A dash is used to indicate no difference from allele 501.


Individual No.
Allele
Standard
5031 GC CTG TGC TAC TTC ACC AAC GGG ACG GAG CGC GTG CGG GGT GTG ACC AGA CAC ATC TAT AAC CGA GAG
33 402 . -- A-- --- --- --- --- --- --- --C -- - -- --- --- --- --- --- --- T-- --- --- --- --- ---
5031 . -- ---
36 301 . C- A-- --- --- --- --- --- --- --- --- --- --- --- --T TA- --- --- --- T-- --- --- --- --- ---
402 . -- A-- --- --- --- --- --- --- --- --C --- --- --- --- --- --- --- --- T-- --- --- --- --- ---
37 402 . -- A-- --- --- --- --- --- --- --- --C --- --- --- --- --- --- --- --- T-- --- --- --- --- ---
402 . -- A-- --- --- --- --- --- --- --- --C --- --- --- --- --- --- --- --- T-- --- --- --- --- ---
38 301 . C- A-- --- --- --- --- --- --- --- --- --- --- --- --T TA- --- --- --- T-- --- --- --- --- ---
5031 -- ---
40 301 . C- A-- --- --- --- --- --- --- --- --- --- --- --- --T TA- --- --- --- T-- --- --- --- --- ---
5031 -- ---
41 302 . -- A-- --- --- --- --- --- --- --- --- --- --- --- --T CT- --- --- --- T-- --- --- --- --- ---
402 . -- A-- --- --- --- --- --- --- --- --C --- --- --- --- --- --- --- --- T-- --- --- --- --- ---
42 301 . C- A-- --- --- --- --- --- --- --- --- --- --- --- --T TA- --- --- --- T-- --- --- --- --- ---
401 -- A-- --- --- --- --- --- --- --- --C --- -T- --- --- --- --- --- --- T-- --- --- --- --- ---
44 301 . C- A-- --- --- --- --- --- --- --- --- --- --- --- --T TA- --- --- --- T-- --- --- --- --- ---
5031 . -- ---
45 302 -- A-- --- --- --- --- --- --- --- --- --- --- --- --T CT- --- --- --- T-- --- --- --- --- ---
402 -- A-- --- --- --- --- --- --- --- --C --- --- --- --- --- --- --- --- T-- --- --- --- --- ---
46 5031 . -- ---
5031 -- ---
47 301 . C- A-- --- --- --- --- --- --- --- --- --- --- --- --T TA- --- --- --- T-- --- --- --- --- ---
402 -- A-- --- --- --- --- --- --- --- --C --- --- --- --- --- --- --- --- T-- --- --- --- --- ---
49 301 C- A-- --- --- --- --- --- --- --- --- --- --- --- --T TA- --- --- --- T-- --- --- --- --- ---
503' . -- ---
51 301 C- A-- --- --- --- --- --- --- --- --- --- --- --- --T TA- --- --- --- T-- --- --- --- --- ---
3032 -- A-- --- --- --- --- --- --- --- --- --- --- --- --T CT- --- --- --- T-- --- --- --- --- ---
52 302 . -- A-- --- --- --- --- --- --- --- --- --- --- --- --T CT- --- --- --- T-- --- --- --- --- ---
3032 -- A-- --- --- --- --- --- --- --- --- --- --- --- --T CT- --- --- --- T-- --- --- --- --- ---
57 301 . C- A-- --- --- --- --- --- --- --- --- --- --- --- --T TA- --- --- --- T-- --- --- --- --- ---
3032 . -- A-- --- --- --- --- --- --- --- --- --- --- --- --T CT- --- --- --- T-- --- --- --- --- ---
58 5031 . -- ---
5031 . -- ---
59 301 . C- A-- --- --- --- --- --- --- --- --- --- --- --- --T TA- --- --- --- T-- --- --- --- --- ---
304 . C- A-- --- --- --- --- --- --- --- --- --- --- --- --T TA- --- --- --- T-- --- --- --- --- ---
60 301 . C- A-- --- --- --- --- --- --- --- --- --- --- --- --T TA- --- --- --- T-- --- --- --- --- ---
402 . -- A-- --- --- --- --- --- --- --- --C --- --- --- --- --- --- --- --- T-- --- --- --- --- ---
61 301 . C- A-- --- --- --- --- --- --- --- --- --- --- --- --T TA- --- --- --- T-- --- --- --- --- ---
3032 . -- A-- --- --- --- --- --- --- --- --- --- --- --- --T CT- --- --- --- T-- --- --- --- --- ---
63 3031 . -- A-- --- --- --- --- --- --- --- --C --- --- --- --T CT- --- --- --- T-- --- --- --- --- ---
402 . -- A-- --- --- --- --- --- --- --- --C --- --- --- --- --- --- --- --- T-- --- --- --- --- ---
64 301 . C- A-- --- --- --- --- --- --- --- --- --- --- --- --T TA- --- --- --- T-- --- --- --- --- ---
402 . -- A-- --- --- --- --- --- --- --- --C --- --- --- --- --- --- --- --- T-- --- --- --- --- ---
65 402 . -- A-- --- --- --- --- --- --- --- --C --- --- --- --- --- --- --- --- T-- --- --- --- --- ---
402 -- A-- --- --- --- --- --- --- --- --C --- --- --- --- --- --- --- --- T-- --- --- --- --- ---
67 302 . -- A-- --- --- --- --- --- --- --- --- --- --- --- --T CT- --- --- --- T-- --- --- --- --- ---
5031 -- ---
68 302 -- A-- -- --- --- --- --- --- --- --- --- --- --- --T CT- --- --- --- T-- --- --- --- --- ---
402 . -- A-- --- --- --- --- --- --- --- --C --- --- --- --- --- --- --- --- T-- --- --- --- --- ---
69 301 . C- A-- --- --- --- --- --- --- --- --- --- --- --- --T TA- --- --- --- T-- --- --- --- --- ---
402 . -- A-- --- --- --- --- --- --- --- --C --- --- --- --- --- --- --- --- T-- --- --- --- --- ---
70 301 . C- A-- --- --- --- --- --- --- --- --- --- --- --- --T TA- --- --- --- T-- --- --- --- --- ---
402 -- A-- --- --- --- --- --- --- --- --C --- --- --- --- --- --- --- --- T-- --- --- --- --- ---


Individual No.
Allele
Standard
5031 GAG TAC GTG CGC TTC GAC AGC GAC GTG CGG GTG TAT CGG GCG GTG ACG CCG CAG GGG CGG CCT GAC GCC
33 402 C T T
5031
36 301 CA A C C T C T
402 C T T T
37 402 C T T T
402 C T T T
38 301 CA A C C T C T
5031
40 301 CA A C C T C T
5031
41 302 CA C C T C C
402 C T T T
42 301 CA A C C T C T
401 C T T T
44 301 CA A C C T C T
5031
45 302 CA C C T C C
402 C T T T
46 5031
5031
47 301 CA A C C T C T
402 C T T T
49 301 CA A C C T C T
5031
51 301 CA A C C T C T
3032 CA T C T
52 302 CA C C T C C
3032 CA T C T
57 301 CA A C C T C T
3032 CA T C T
58 5031
5031
59 301 CA A C C T C T
304 CA A C C T C C
60 301 CA A C C T C T
402 C T T T
61 301 CA A C C T C T
3032 CA T C T
63 3031 CA T C T
402 C T T T
64 301 CA A C C T C T
402 C T T T
65 402 C T T T
402 C T T T
67 302 CA C C T C C
5031
68 302 CA C C T C C
402 C T T T
69 301 CA A C C T C T
402 C T T T
70 301 CA A C C T C T
402 C T T T

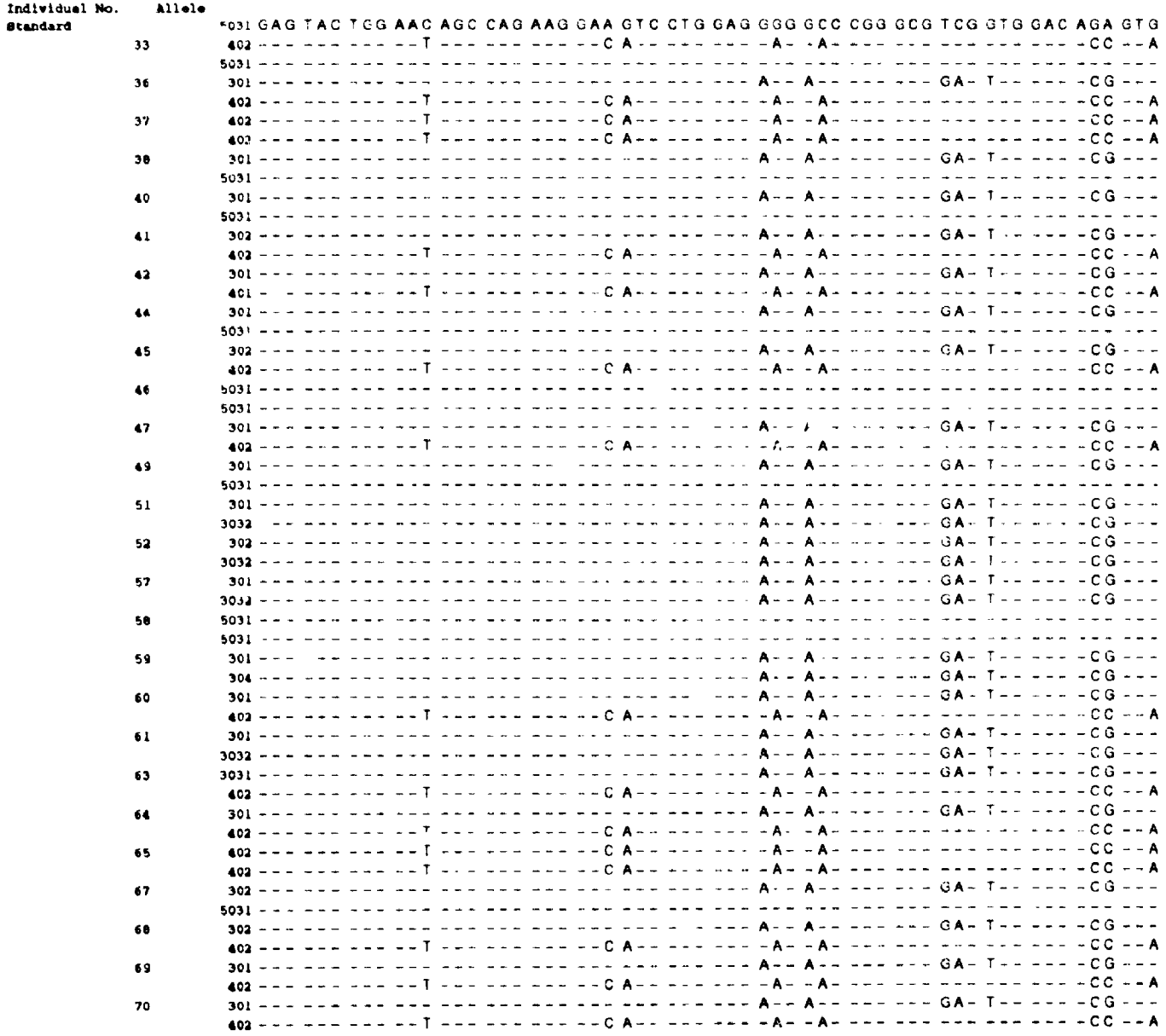

Individual No.
Allele
Standard
5031 GAG TAC TCG AAC AGC CAG AAG GAA GTC CTG GAG GGG GCC CGG GCG TCG GTG GAC AGA GTG
33 402
36 301
402
37 402
402
38 301
5031
40 301
5031
41 302
402
42 301
401
44 301
5031
45 302
402
46 5031
5031
47 301
402
49 301
5031
51 301
3032
52 302
3032
57 301
3032
58 5031
5031
59 301
304
60 301
402
61 301
3032
63 3031
402
64 301
402
65 402
402
67 302
5031
68 302
402
69 301
402
70 301
402

Table 10. A summary of the alleles determined to exist at the *HLADQB1* locus in a random sample of 26 Dogrib individuals. The alleles were identified from determination of the nucleotide sequence of a large region of the second exon of the *HLADQB1* locus in each individual.

Dogrib Individual No.	*HLA-DQB1* Second Exon Allele	Allele
33	402	5031
36	301	402
37	402	402
38	301	5031
40	301	5031
41	302	402
42	301	401
44	301	5031
45	302	402
46	5031	5031
47	301	402
49	301	5031
51	301	3032
52	302	3032
57	301	3032
58	5031	5031
59	301	304
60	301	402
61	301	3032
63	3031	402
64*	301	402
65	402	402
67	302	5031
68	302	402
69	301	402
70	301	402

* sequence information derived from sequencing in one direction only

Table 11. The frequencies of eight alleles identified in 25 Dogrib individuals from determination of the nucleotide sequences of a large region of the second exon of the *HLA-DQβ1* locus.

HLA-DQB1 Locus in Dogrib Individuals

Allele	Frequency
301	0.29
302	0.09
3031	0.02
3032	0.08
304	0.02
401	0.02
402	0.29
5031	0.19

were only slightly more frequent than transitions (54 and 45%, respectively). The pattern of nucleotide substitution at CpG sites did not differ significantly from that expected, given the ratio of four transversion events to two transition events ($p<0.01$). Nucleotide substitutions were most frequent at the second position of codons (44%). Substitutions at the first and third positions of codons were equally frequent (28%). Seventy-two percent of the nucleotide substitutions among the alleles resulted in amino acid replacements. Of 79 amino acid sites examined 21 contain amino acid replacements among the alleles present in the Dogrib sample. At these 21 sites there are eight conservative and 19 nonconservative amino acid replacements. It is possible to attribute all of the differences among the nucleotide sequences obtained in the Dogrib individuals to nucleotide variation known to exist among the alleles at the *HLA-DQβ1* locus and thus it was concluded that no intra-allelic nucleotide substitutions were observed in this analysis.

The total number of nucleotide positions analyzed for sequence variation in a single allele were 197. The survey examines a total of 197 nucleotides of the second exon of *HLA-DQβ1* locus in each of 51 diploid individuals from two different populations. This represents a screening of a total of 20,094 nucleotides. The variation observed is consistent with the existence of previously known alleles of this locus.

3.2 COMPUTER ANALYSES OF NUCLEOTIDE SEQUENCE ORGANIZATION AND VARIATION

3.2.1 Short-Sequence Representation in a Single Exon and The Complete cDNA Sequence of *Adh2* and *DQβ1* Loci

The nucleotide sequence organization of the third exon of the human *Adh2* locus (*ß1* allele) and the second exon of the *HLA-DQß1* locus was described in terms of single nucleotide, dinucleotide and trinucleotide composition. The *Adh2* exon 3 (*ß1* allele) sequence was composed of equivalent frequencies of the four nucleotides ($p<0.01$) but biases existed in dinucleotide and trinucleotide composition. The dinucleotide and its strand-symmetric partner, TG/CA were significantly over-represented ($\rho_{XY}>1.23$) while CG/GC and TA showed significant under-representation ($\rho_{XY}<0.78$). Representation of trinucleotides ranged from $\gamma_{XY} = 0.29$ to 1.59. The trinucleotides ATT/AAT and GCC/GGC were the two extremes in over-representation and ATA/TAT and TTG/CAA were the two extremes in under-representation.

Description of the nucleotide sequence organization at the second exon of the *HLA-DQß1* locus involved examination of each of the 26 known allele sequences. The %G+C content of the sequences ranged from 59.2 to 62.9. The various allele sequences did not differ significantly in single nucleotide composition ($p<0.01$). All of the alleles tended to have a bias in single nucleotide frequencies with an over-representation of guanine (%G of 36) and an under-representation of thymine in all of the allele sequences (%T of 18). Biases in dinucleotide composition were significant in alleles 501 to 504 and 6051 to 606 ($p<0.05$). The dinucleotide and its strand symmetric partner GT/AC were over-represented significantly ($\rho_{XY}>1.23$) in all but alleles 201/202. Dinucleotides TA, GC, AT and TT/AA were under-represented significantly in at least one of the 26 alleles ($\rho_{XY}<0.78$). Spearman rank-order correlation coefficients, r_S were used as similarity coefficients to describe the similarities in dinucleotide composition among the 26 known alleles. Alleles 201/202 differed in dinucleotide composition in comparisons with certain other alleles ($r_S = 0.29$ to 0.73 for comparisons of 201/202 with all other alleles). All other alleles had similar dinucleotide

compositions (r_S = 0.54 to 0.95). Representation of trinucleotides (γXYZ) ranged from 0.54 to 1.87. Those trinucleotides representing the extremes in over-representation were AAT/ATT, TAG/CTA, GCA/TGC, TTT/AAA, CAT/ATG, TTG/CCA, TGA/TCA and GAT/ATC. The two most under-represented trinucleotides in the 26 allele sequences included AAG/CTT, CCA/TGG, TTA/TAA, GCT/AGC, CGA/TCG, GAA/TTC, TCC/GGA, GTT/AAC and CGC/GCG. Similarities in dinucleotide and trinucleotide composition were greater within allele subtypes than between different subtypes.

Dinucleotide and trinucleotide representation at the *Adh2* third exon (*B1* allele) and the *HLA-DQB1* second exon (26 alleles) were compared using the Spearman rank-order correlation coefficient. Dinucleotide representation in the exons of the two different genes differed significantly (r_S = 0.05 to 0.54) except in the case of the *Adh2* exon 3 and alleles 201/202 of *HLA-DQB1* where r_S was 0.70. No significant correlation exists in trinucleotide representation values for the exon sequences of the two different genes (r_S = -0.35 to 0.35). The *Adh* exon 3 and *HLA-DQB1* exon 2 possess different short-sequence compositions.

The nucleotide sequence organization of the third exon of the human *Adh2* locus was compared with the full length cDNA sequence of this human gene. The cDNA sequence was composed of equivalent frequencies of the four nucleotides ($p<0.01$). Significant biases in dinucleotide representation in the cDNA included the over-representation of TG/CA and CC/GG and the under-representation of CG and TA. Four extremes in trinucleotide representation were CGC/GCG, TTG/CAA, AAA/TTT and TCG/CGA (γXYZ = 0.74, 0.77, 1.33 and 1.22, respectively). The dinucleotide representations of the exon and the full length cDNA sequence were similar (r_S = 0.68) but differences in trinucleotide representation were evident (r_S = 0.24).

The nucleotide sequence organization of the second exon of the *HLA-*

3.2.2 Graphic Portrayal and Measurement of Short-Sequence Representation

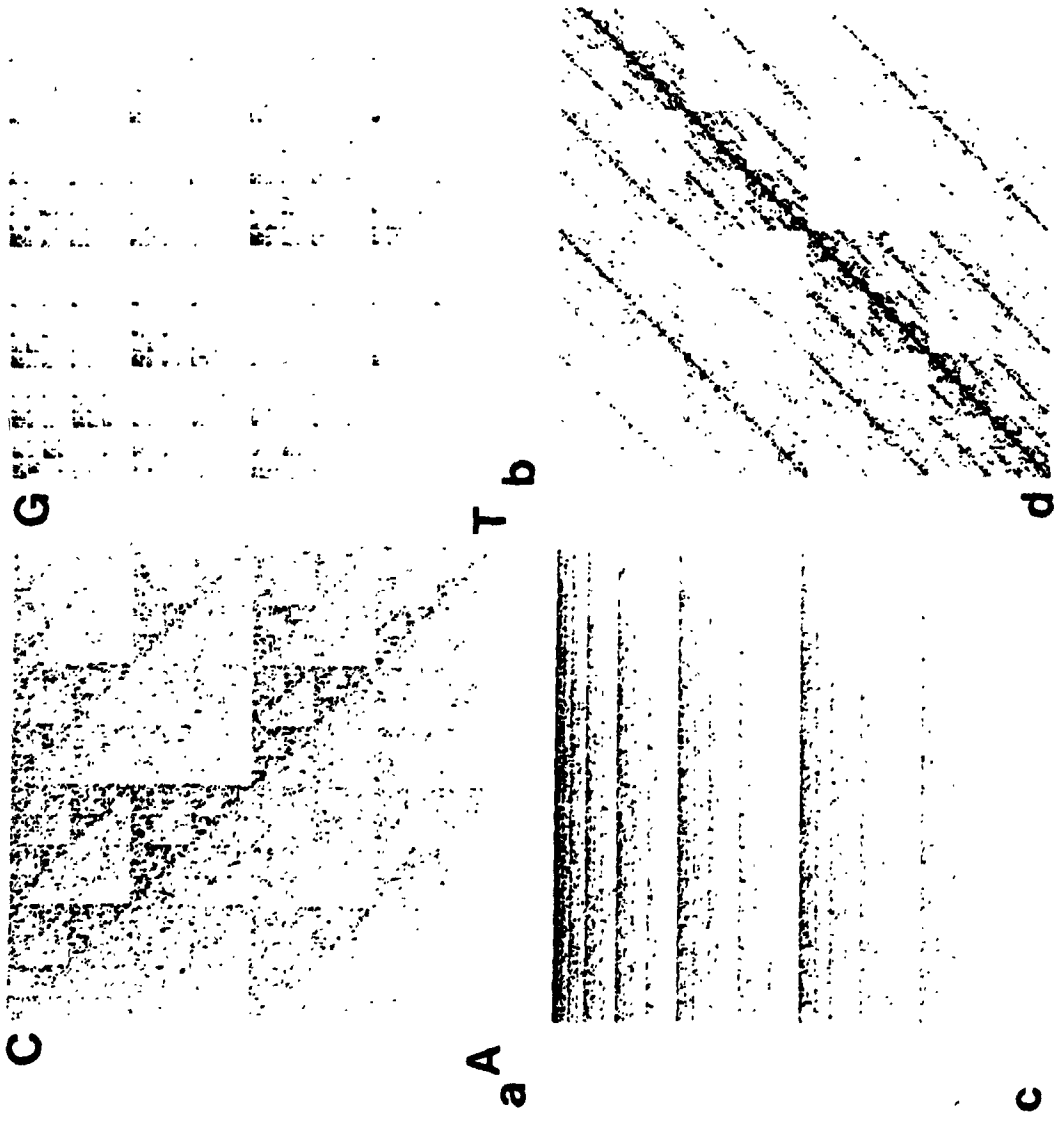
The examination of the sequence organization at an exon and the complete cDNA sequences of *Adh2* and *DQβ1* loci revealed interesting similarities and differences. The examination of sequence organization of numerous and much larger sequences necessitated the use of a graphic portrayal of sequence structure. Chaos game representation was selected for use in this study, but due to its novelty and limited application, it was initially applied to the examination of computer-generated sequences with known and simple structures. These preliminary analyses formed the basis for comparisons of the complexity presented by chaos patterns of naturally-occurring DNA sequences.

Chaos game representation has been used to portray the nucleotide sequence organization of large biological DNA sequences, but the major sequence determinants of the nonrandom distribution of data points in chaos patterns are generally not known. In this study, computer-generated sequences having known and simple structures were used to identify the major determinants of data point distribution in chaos plots. Sequences were constructed to differ in single nucleotide composition. All of the computer-generated sequences were 16,295 nts in length and the nucleotides were ordered randomly. The differences in the data point distribution of chaos plots generated by these sequences were compared with a sequence having an identical length, a random order of nucleotides and equivalent frequencies of the four nucleotides. A random order and equivalent frequencies of the four nucleotides produced a uniform distribution of data points within the chaos plot. The under-representation of a single

nucleotide relative to the other three nucleotides produced decreased data point frequency in regions of the chaos plot representing that nucleotide. Under-representation of adenine (10%) and equivalent frequencies of the other three nucleotides produced low data point density in regions of the plot near the vertex representing adenine (Figure 13a). Over-representation of cytosine (80%) and equivalent frequencies of the other three nucleotides resulted in increased data point density towards the C vertex (Figure 13b). Under-representation of adenine and thymine (20%) produced horizontal striations in the chaos pattern with increased data point density towards the axis joining C and G vertices (Figure 13c). Over-representation of two nucleotides at diagonally opposed vertices produced increased data point density along the diagonal line joining the two vertices of the over-represented nucleotides (in Figure 13d, the frequency of A and G is 80% and that of C and T is 20%). The nature and extent of departures from equivalent nucleotide frequencies and random associations of nucleotides were portrayed by the data point distribution of chaos patterns. Biases in single nucleotide frequencies produced patterns in data point distribution in chaos plots that were readily recognized and interpreted in terms of single nucleotide composition from a visual examination of chaos patterns.

Self-similarity was evident in each of the chaos patterns produced by the computer-generated sequences (Figure 13). It should be noted that in all chaos patterns the major features of data point distribution in the entire plot are repeated in subquadrants of the plot. For a given nucleotide, or oligonucleotide that was in high or low frequency, subsequences of successively longer length ending in that nucleotide or oligonucleotide were also in high or low frequency. In the chaos pattern this was seen as high or low data point frequency in smaller and smaller subquadrants that represented longer and longer subsequences ending in that nucleotide or dinucleotide. Self-similarity occurs as a result of plot

Figure 13. Chaos game representation of computer-generated nucleotide sequences having known and simple structures. All sequences were 16,295 nts in length and the nucleotides were ordered randomly. a) Adenine was under-represented (10%) and there were equivalent frequencies of the other three nucleotides. b) Cytosine was over-represented (80%) and the other three nucleotides had equivalent frequencies. c) Adenine and thymine were under-represented (20%) relative to cytosine and guanine and each pair of nucleotides had equivalent frequencies. d) Adenine and guanine were under-represented (80%) relative to cytosine and thymine and each pair of nucleotides had equivalent frequencies.


C
G
T
A
a
b
d
c

construction and is thus inherent to all chaos patterns. The features of the entire chaos pattern that are repeated in subquadrants of the plot are important because they are characteristic of the structure of the DNA sequence being plotted. Self-similarity produces a repetition of patterns in data point distribution that assists the visual recognition and interpretation of chaos patterns.

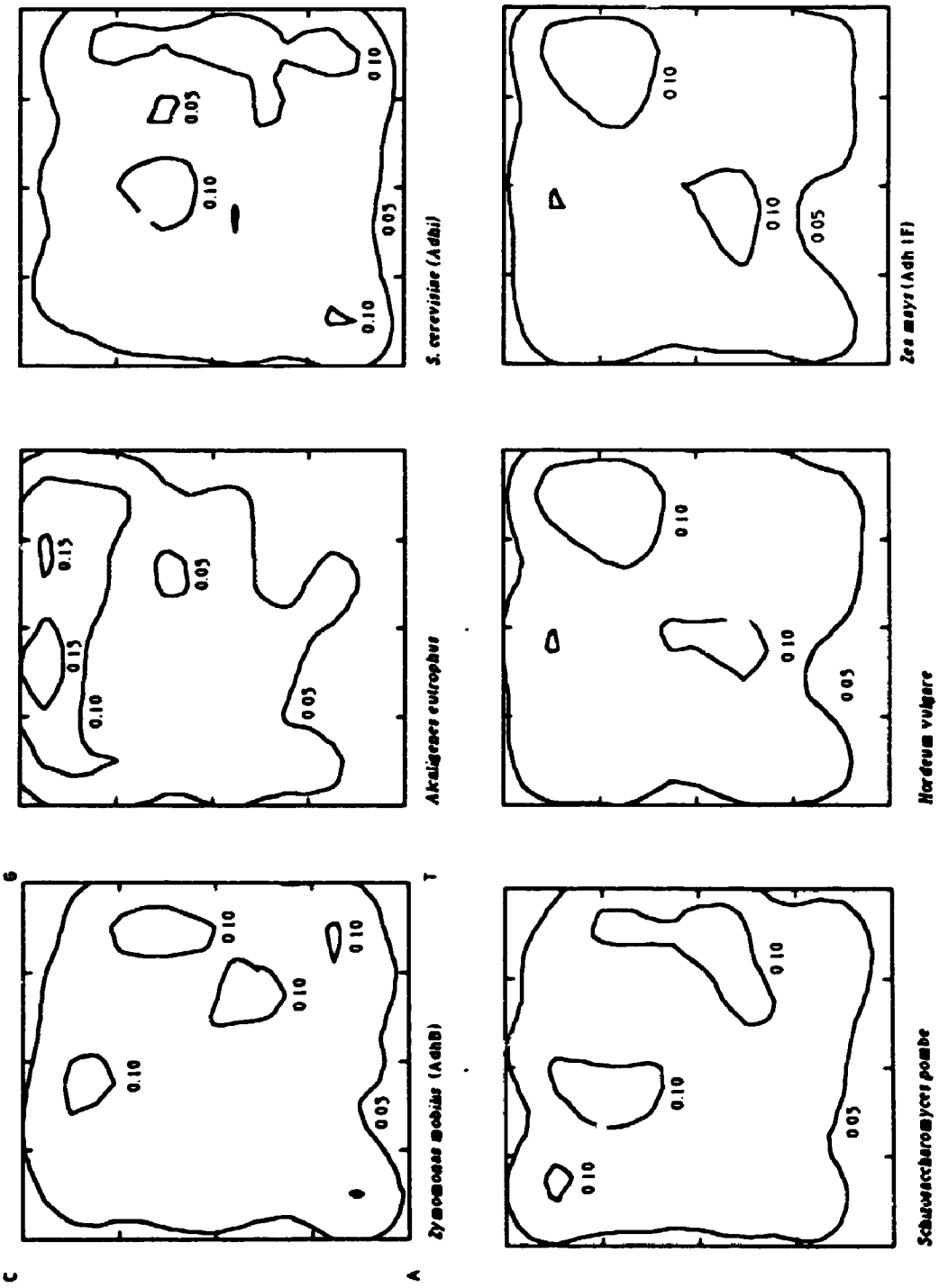
Biases in single nucleotides were used to interpret the effect of biases in dinucleotides and longer subsequences. In the computer-generated sequence, a deficiency in a single nucleotide produced deficiencies in subsequences of all lengths that contained that nucleotide. Biases in the frequencies of single nucleotides and oligonucleotides have a more pronounced effect on data point distribution in the chaos plot than do biases in frequencies of longer subsequences. The occurrence of single nucleotides and oligonucleotides are represented by larger subquadrant areas in the chaos plot. It is important to note that there is no single determinant of chaos patterns but rather the major determinants of data point distribution are nonequivalent single nucleotide frequencies and/or nonrandom occurrences of oligonucleotides and thus are DNA sequence dependent. The major determinant of data point distribution in any chaos pattern is the shortest subsequence at which nonrandomness in composition occurs. The ease with which biases in long subsequence composition can be identified visually from chaos patterns depends upon the degree to which the sequence has biases in the frequencies of shorter subsequences. Thus, chaos representation is best suited, but not limited to, the portrayal of short-sequence representation since shorter sequences are represented by larger areas of the chaos plot.

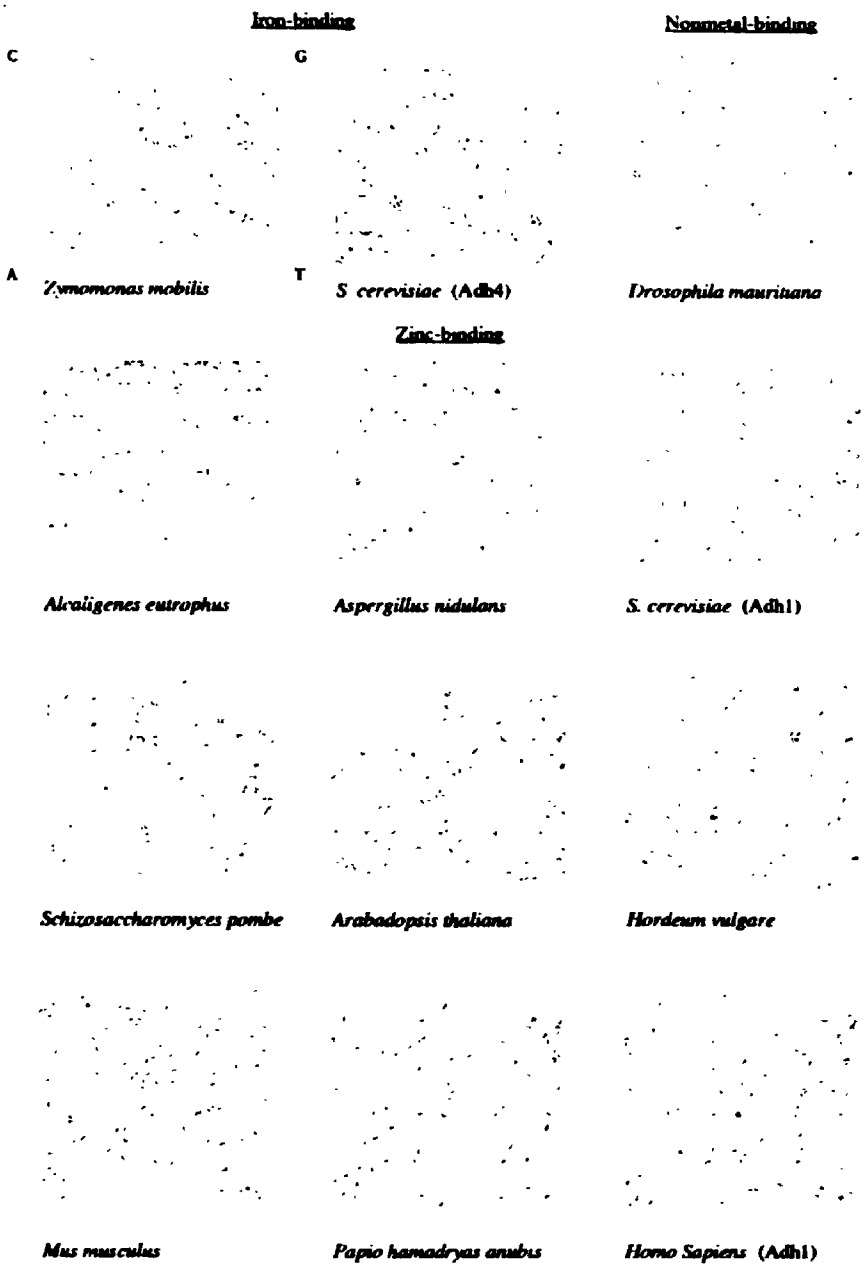
Visual assessment of data point distribution in chaos plots is subjective and mathematical characterization of the total distribution of data points in chaos plots is lacking. Deviation of sequence structure from a random order of

nucleotides can be quantified partially through an analysis of data point frequency in larger subquadrants (i.e., measurement of short-sequence representation). An objective and quantitative measure of nonrandom sequence organization is the calculation of strand-symmetric relative abundance of subsequences (Burge et al., 1992) seen in chaos patterns as a nonrandom distribution of data points among subquadrants of a chaos plot representing subsequences of similar length.

3.2.3 Short-Sequence Representation in *Adh* cDNAs of Phylogenetically Divergent Species

Chaos patterns were generated for 29 cDNA sequences representing *Adh* coding regions from phylogenetically divergent species (sequences are identified in Table 2). A sample of 12 such plots is given in Figure 14. A visual evaluation of the 29 chaos patterns was subjective but identified major similarities and distinctions. All of the chaos patterns appeared to have a nonuniform distribution of data points. The two iron-binding sequences had distinct chaos patterns. The three *Drosophila* chaos patterns were almost identical. The patterns for the different *Adh* genes of yeast were each distinct. The three potato sequences (*Adh1*, *Adh2* and *Adh3* coding sequences) had nearly identical patterns. All of the human and the baboon chaos patterns were almost superimposable, whereas mouse and rat chaos patterns had regional similarities. A sparseness of CG dinucleotides was particularly evident in the chaos patterns of human and baboon cDNAs. In general, only *Adh* coding sequences of the same species or those of closely related species produced recognizable similarities. A consistent feature in all of the chaos plots that was characteristic of the *Adh* cDNAs of diverse species was not identified.


C
A
G
T
Zymomonas mobilis (AdhB)
Saccharomyces cerevisiae (Adh1)
Saccharomyces pombe
Acaligenes eutrophus
Hordeum vulgare
Zea mays (Adh 1F)
0.10
0.05
0.15


Iron-binding
Nonmetal-binding
C
G
A
T
Zymomonas mobilis
S. cerevisiae (Adh4)
Drosophila mauritiana
Zinc-binding
Alcaligenes eutrophus
Aspergillus nidulans
S. cerevisiae (Adh1)
Schizosaccharomyces pombe
Arabadopsis thaliana
Hordeum vulgare
Mus musculus
Papio hamadryas anubis
Homo Sapiens (Adh1)

Dinucleotide composition of 12 representative sequences was determined from calculation of the percentage of nucleotides plotted in each of the 16 subquadrants of the chaos plots given in Figure 14 (Table 12). Data points among the 16 subquadrants of *Adh* chaos patterns were not uniformly distributed. The relative low frequency of the CG dinucleotides and high frequency of TG dinucleotides was evident, particularly in human and baboon sequences. Further, there was a relative low frequency of TA dinucleotides in all cDNA sequences. In general, the dinucleotide compositions of mammalian and plant sequences were not significantly different from one another, whereas bacteria, yeast and *Drosophila* sequences showed significant differences in the frequencies of the dinucleotides. There was no correlation between distance measures based on dinucleotide frequencies and the distance estimates based upon codon usage ($r = 0.39$). Distance measures based upon dinucleotide frequencies and multiple sequence alignment were also not correlated ($r = 0.06$).

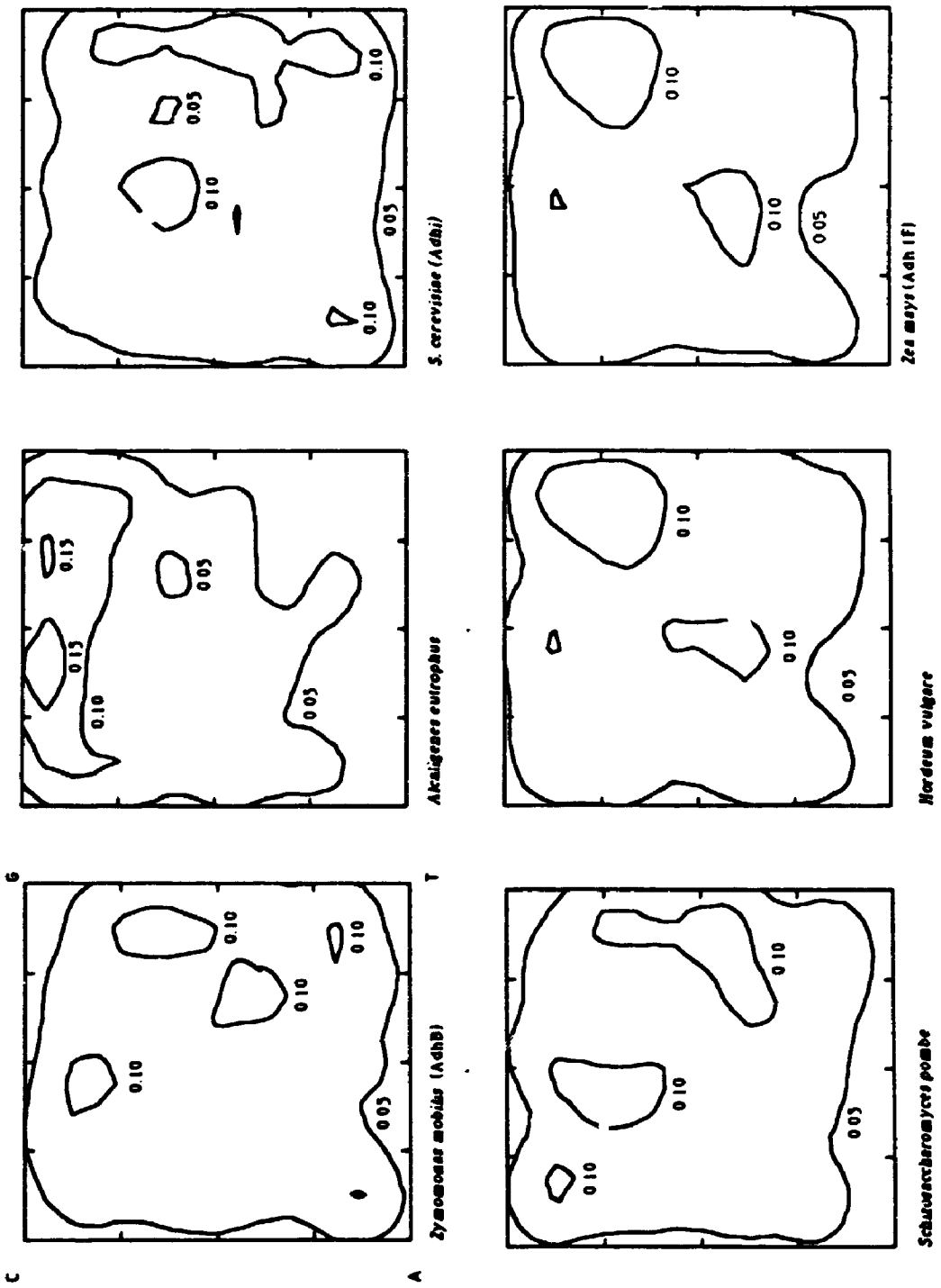
A visual comparison of chaos plots containing fewer than 4000 data points proved difficult. A more objective comparison of data point distribution in chaos plots was achieved through use of contour plots of chaos patterns. Kernel density estimation was used to draw contour lines around regions of the chaos plot that contained similar data point densities. The contour plots of chaos patterns generated by *Adh* cDNAs were different for different species (Figure 15). Some similarities among chaos patterns exist within human, primate, rodent and monocot plant groups. Generally, the chaos patterns of mammalian sequences were similar to one another. The *Adh* genes for more distantly related species had extensive homologies yet generated different chaos patterns. Such results did not follow the expectation that the chaos patterns of cDNAs might possess visually discernible features characteristic of the polypeptide for which they code. The major features of the chaos patterns and contour plots do not appear to be

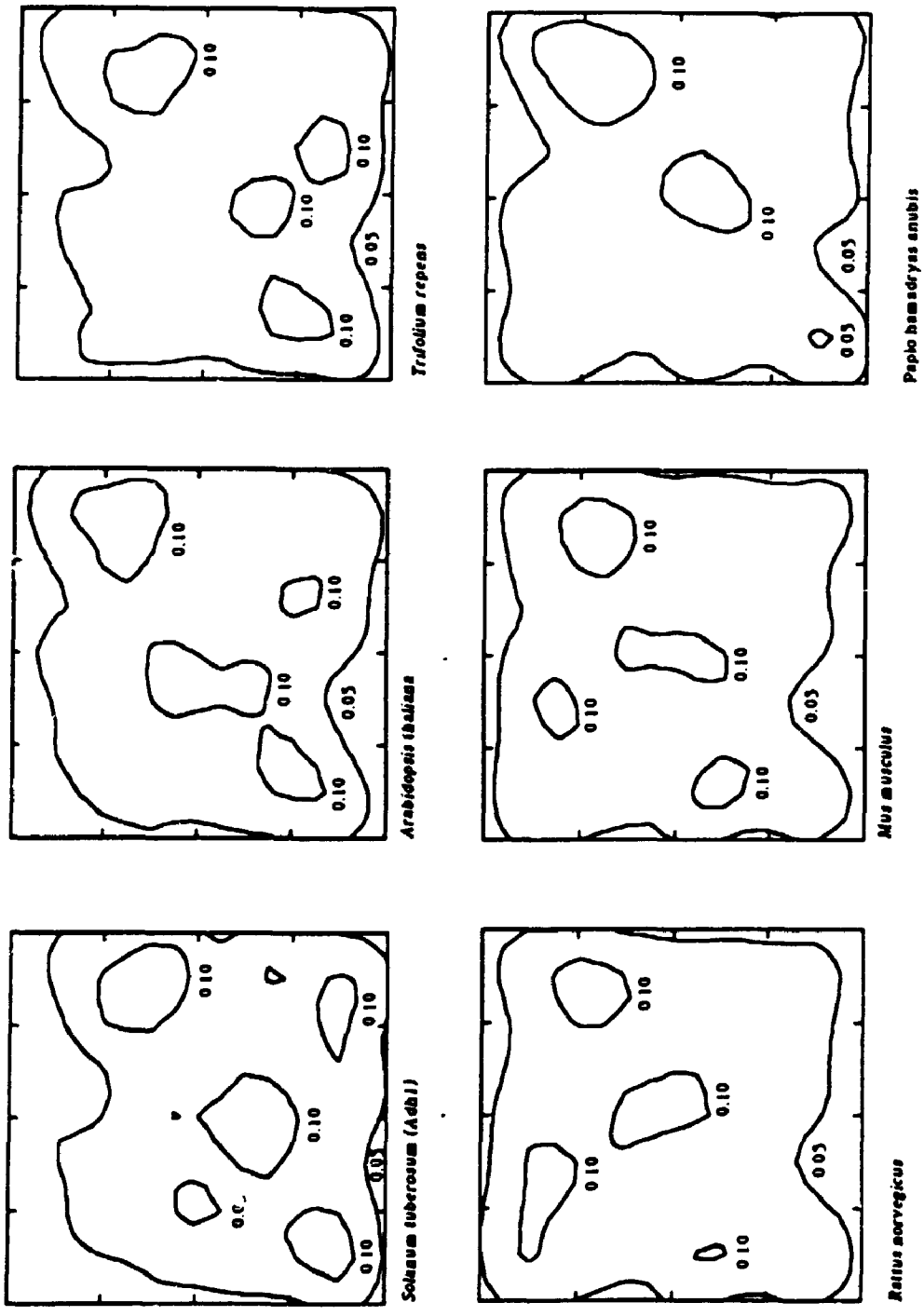
Table 12. The percentage of each dinucleotide determined for 12 representative *Adh* cDNA sequences. The chaos patterns for the sequences are presented in Figure 14 and the identity of nucleotide sequences is presented in Table 2.

Dinucleotide	Sequence Source*											
	Zym IB	Ys4 IB	Dma NB	Alc ZB	Asn ZB	Ys1 ZB	Ysp ZB	Ath ZB	Bly ZB	Mus ZB	Bab ZB	Hum ZB
CC	5.7	4.0	10.9	9.3	7.6	6.4	8.7	3.5	6.3	7.4	6.3	5.9
GC	7.3	5.3	7.8	13.2	8.6	6.5	6.0	4.2	7.6	6.8	6.2	5.4
CG	**6.1**	**2.2**	**5.7**	**12.8**	**7.4**	**2.2**	**5.9**	**3.6**	**4.8**	**3.7**	**2.4**	**1.6**
GG	5.1	4.2	6.8	10.2	7.6	8.7	6.2	7.5	9.0	6.9	8.2	7.8
AC	4.8	6.3	7.8	6.6	4.4	4.4	5.7	5.1	5.0	5.0	4.4	4.8
TC	7.2	5.4	6.0	5.2	7.3	5.1	8.2	7.0	5.5	6.1	5.0	5.1
AG	4.2	5.2	5.1	4.2	6.6	7.4	4.7	6.9	6.9	6.6	7.0	6.8
TG	**9.4**	**8.5**	**8.1**	**6.5**	**6.0**	**8.8**	**9.0**	**9.8**	**10.5**	**8.7**	**9.9**	**10.0**
CA	5.4	7.7	8.8	7.3	7.1	7.4	6.0	7.1	7.8	8.0	6.8	7.6
GA	6.2	5.6	7.2	6.3	6.8	6.8	5.5	8.8	7.6	6.7	6.6	6.3
CT	8.5	7.2	7.4	4.4	5.7	6.7	7.8	5.8	5.7	7.7	6.9	6.6
GT	5.8	5.2	4.2	4.0	5.1	6.7	8.6	7.4	5.9	6.4	6.5	6.5
AA	7.5	10.9	4.9	3.3	6.2	7.9	4.1	7.1	5.0	6.6	8.2	8.9
TA	**3.4**	**5.4**	**1.4**	**1.2**	**2.7**	**3.4**	**3.2**	**2.7**	**1.8**	**2.7**	**3.5**	**4.0**
AT	6.2	7.1	4.2	4.4	5.5	5.5	4.4	6.4	5.4	4.8	5.3	6.0
TT	6.9	9.8	3.8	1.4	5.4	6.1	6.0	7.1	5.0	5.9	6.6	6.6
Total No. of Bases	1155	1397	769	1098	1060	1034	1050	1131	1119	1130	1130	1130

*Abbreviations are defined in Table 2.

Figure 15. Contour plots for chaos patterns for 12 *Adh* cDNA sequences from phylogenetically divergent species. The contour plots are contained on two pages.


G
T
C
A
Zymomonas mobilis (AdhB)
Alcaligenes eutrophus
S. cerevisiae (Adh1)
Schizosaccharomyces pombe
Hordeum vulgare
Zea mays (Adh 1F)

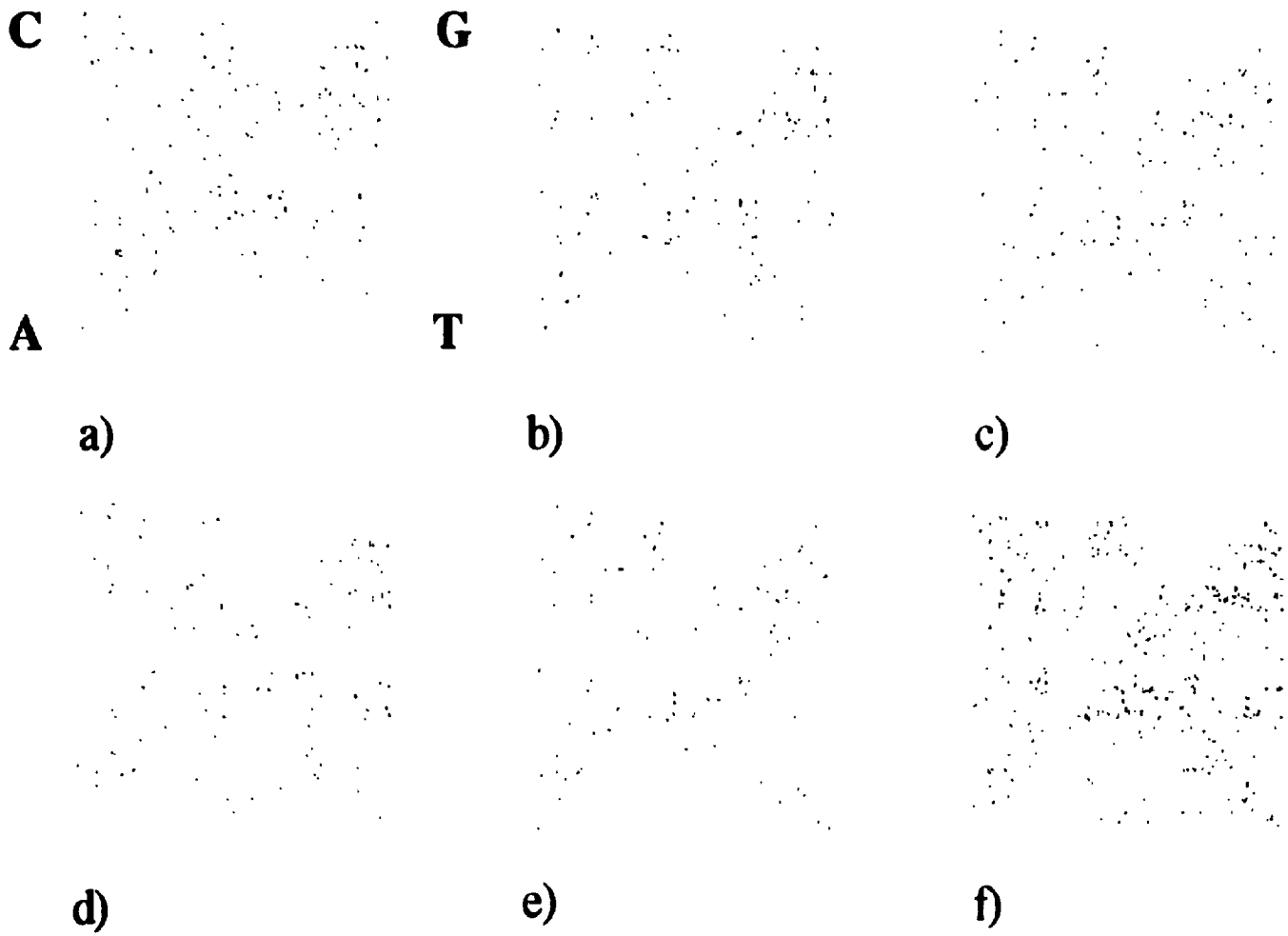

Trifolium repens
Papio hamadryas anubis
Arabidopsis thaliana
Mus musculus
Solanum tuberosum (Adbl)
Rattus norvegicus

gene-specific. The predominant features of chaos patterns represent biases in dinucleotide composition. This feature of nucleotide sequence organization appears to indicate a genome-type specificity in nucleotide sequence organization.

3.2.4 Short-Sequence Representation in Two Different Human Multigene Families

Human globin genes contained within a known 73 kb continuous stretch of DNA sequence information generated chaos patterns that were similar to those of human *Adh* genes. Figure 16 (a to e) contains five representative chaos plots of globin coding sequences and a composite plot of seven globin cDNAs superimposed upon one another (Figure 16f). The chaos patterns generated from individual human globin cDNA sequences (roughly 280 nts) were similar to the chaos patterns of the entire human globin gene complex that contained both coding and noncoding sequences (73,357 nts; Jeffrey, 1990). There was a sparseness of data points in the CG subquadrant and in the upper left portion of the GC subquadrant (i.e., where the subsequences plotted end in the trinucleotide CGC). This feature of a chaos pattern has been described as the "double scoop" (Jeffrey, 1990). There was a sparseness of data points along the horizontal midline of the chaos patterns, where the "double scoop" was repeated in both the A and T quadrants (i.e., where the subsequences plotted ended in the trinucleotides CGA and CGT). Also there was a paucity of data points in the TA subquadrant and a dense collection of points in the TG subquadrant. These features were also discernible in the composite plot containing roughly 2000 data points. The self-similarity that was characteristic of the chaos plot of the entire human globin gene complex (Jeffrey, 1990) was not a readily apparent feature in

Figure 16. Chaos patterns for five of seven coding sequences (each roughly 280 nts) found within the human globin gene complex (a, Aγ; b, β; c δ; d, ε; e, Gγ). The sixth plot (f) represents a composite of the seven globin cDNA chaos plots superimposed upon one another.


C
A
G
T
a)
b)
c)
d)
e)
f)

chaos patterns of only 2000 nts.

Significantly different frequencies of the 16 dinucleotides ($p<0.05$) contributed to the nonuniform data point distribution seen in the chaos patterns of the coding regions (Figure 17) and the entire complex of globin genes (Jeffrey, 1990). It was evident in Figure 18 that the globin coding sequences had a relatively low frequency of CG and TA dinucleotides and a high occurrence of TG dinucleotides. The dinucleotide compositions of the seven globin coding sequences did not differ significantly (a heterogeneity χ^2 test, $p<0.05$), which may account for the similarities in chaos patterns. The contour plots of A-γ, ß and ε globin cDNAs are given in Figure 18a. Overall the plots were similar but each plot could be distinguished visually from the other two. Contour plots of human globin chaos patterns were also compared with plots generated by human *Adh* cDNAs (Figure 18). It was evident that each contour plot was distinct and the plots of one gene family were more similar relative to one another than they were to the plots of the other gene family. The overall number, shape and position of contours in both gene families were strikingly similar. The major determinants of the contour lines in all of the chaos plots were the under-representation of CG and TA dinucleotides and the over-representation of the TG dinucleotide. Despite the fact that these two gene families are not homologous, they have similar biases in short-sequence composition.

3.2.5 Short-Sequence Representation in Large Continuous DNA Sequences of Different Genomes

A visual comparison of chaos patterns, generated for 56 large DNA sequences from 10 species, identified 6 different types of patterns. *Rhodobacter capsulatus*, *Escherichia coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*,

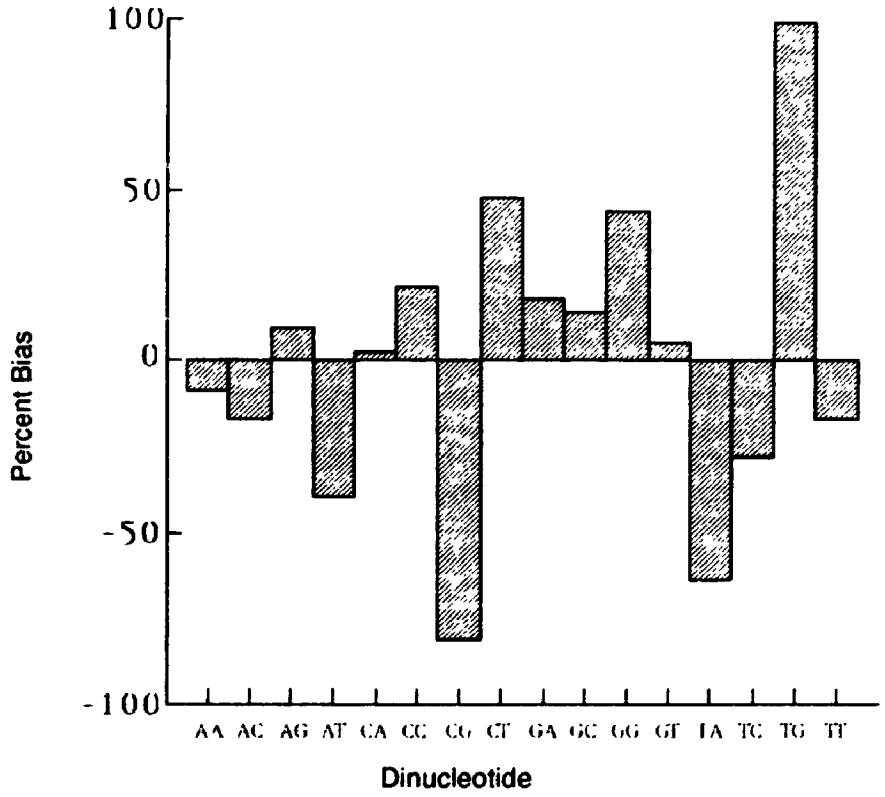
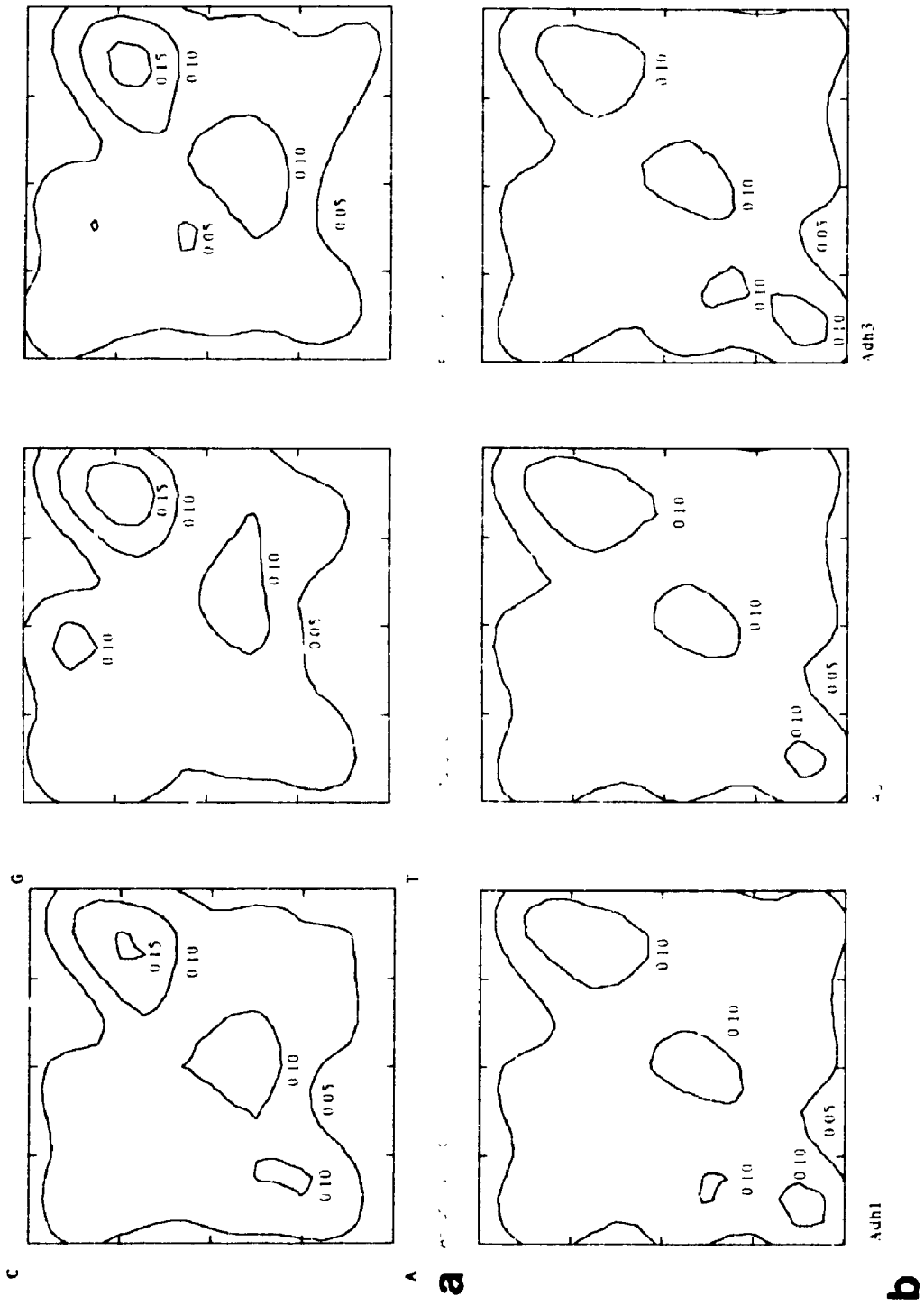

100
50
0
-50
-100
Percent Bias
AA AC AG AT CA CC CG CT GA GC GG GT TA TC TG TT
Dinucleotide

Figure 18. Contour plots of chaos patterns for a) three human globin cDNAs and b) three human *Adh* cDNAs.


a
b
Adh1
Adh3
C
G
A
T

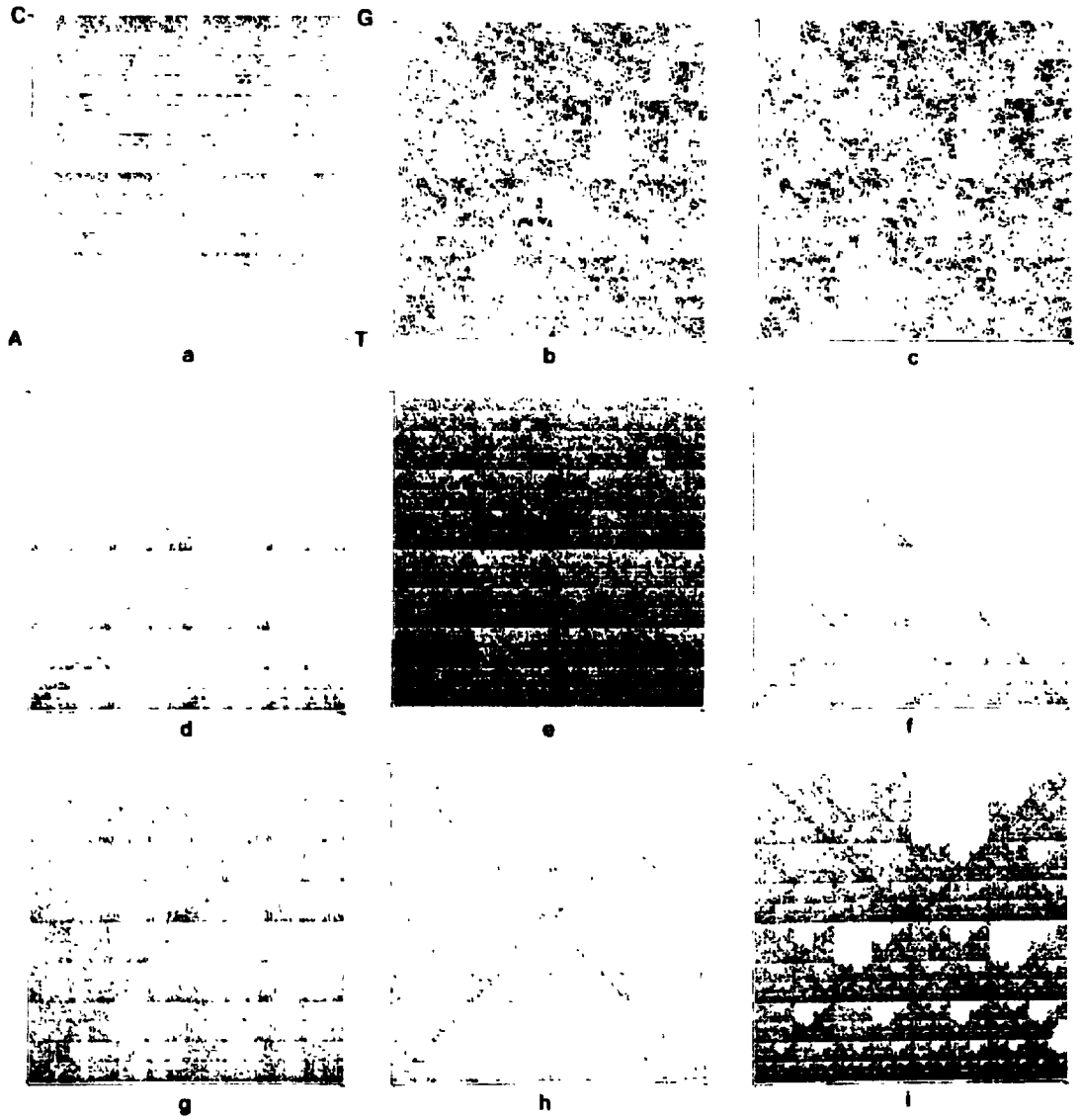
Drosophila melanogaster and mammalian genomes each had a distinct global sequence organization as assessed by visual examination of major features of data point distribution in two-dimensional chaos patterns (Figure 19). Sequence organization was examined in more than one region of *E. coli, S. cerevisiae, C. elegans* and *H. sapiens* genomes (Table 3). Different regions of the same genome had a similar global structure defined as similar major features in data point distribution in two-dimensional chaos plots.

DNA sequences of *R. capsulatus* and *E. coli* genomes have distinct chaos patterns (Figure 19a and 19b, respectively) but both display a paucity of TA dinucleotides, a feature shared by all genomes and both have a paucity of the tetranucleotide CTAG. The chaos pattern generated by the single region of the *R. capsulatus* genome (45,595 nts; Figure 19a) has increased data point density towards the C and G axis of the plot and in particular towards the C vertex representing nonequivalent single nucleotide frequencies (66% G+C and 33% C, respectively). Subquadrants of this chaos plot representing TA dinucleotides and CTAG and TTAG tetranucleotides contain few data points indicative of the rare occurrence of these subsequences. Six different regions of the *E. coli* genome (Table 3) each have a chaos pattern that is uniformly filled with data points except for a paucity of data points in subquadrants representing the occurrence of TA and CTAG subsequences (representative patterns, Figure 19b and 19c).

Chaos patterns of 25 different regions of the *C. elegans* genome were similar. A representative chaos pattern for these sequences (Figure 19d) has a high data point density between A and G vertices and between C and T vertices. Data point density also increases towards the A and T axis. The DNA sequence organization in the *C. elegans* genome as portrayed in chaos patterns appears to be unique among the species' genomes examined in this investigation.

The two-dimensional chaos scatter plot of the complete *Saccharomyces*

Figure 19. Representative chaos patterns of large continuous DNA sequences (>36,000 nts) from the genomes of phylogenetically divergent species. *Rhodobacter capsulatus* a) RCPHSYNG, 45,959 nts; *Escherichia coli* b) ECO110K, 111,401 nts and c) ECOUW85U, 91,408 nts; *Caenorhabditus elegans* d) CELCO8C3 44,025 nts; *Saccharomyces cerevisiae* e) YSCCHRIII, 315,357 nts and f) YSCCHROMI, 41,987 nts; *Drosophila melanogaster* g) DROABDB, 80, 423 nts; *Rattus norvegicus* h) RATCRYG 54,670 nts and *Homo sapiens* i) HUMRETBLAS, 180,388 nts.


C-
G
A
T
a
b
c
d
e
f
g
h
i

cerevisiae chromosome III DNA sequence (315,357 nts) has horizontal striations and increasing data point density towards the A and T axis of the plot (Figure 19e). These features reflect the near equivalent frequencies of adenine (31%) and thymine (30%) and a high A+T content. Diagonal lines joining A and G and T and C vertices of the plot represent a preponderance of purines in some regions of the chromosome sequence and pyrimidines in other regions, respectively. Evident from the variation in data point frequency in different subquadrants containing different dinucleotides is the higher frequency of TT, AA, CA and TG and lower frequency of TA and CG dinucleotides. Similar chaos patterns were generated by 3 other regions of the *S. cerevisiae* genome including the centromeric region (41,987 nts) of chromosome I (Figure 19f).

The Abdominal-B gene region (80,423 nts) of the *D. melanogaster* genome has a unique chaos pattern that is relatively filled with data points in comparison with other chaos patterns. Data point density is increased towards the A and T axis and in particular towards the A and T vertices (Figure 19g). Only a single large region (>36,000 nts) of the *Drosophila* genome was available for analysis and thus the general occurrence of this sequence structure in the genome of this species cannot be evaluated.

The chaos pattern representative of mammalian sequences examined thus far (Jeffrey, 1990; Hill et al., 1992), including the 15 sequences analyzed in this study (Table 3) is characterized by a paucity of data points in subquadrants of the plot that represent the occurrence of the CG dinucleotide (Figures 19h, rat γ-crystallin gene and 19i, human retinoblastoma gene). The pattern also contains dense accumulations of data points along diagonal lines joining A and T vertices and T and C vertices.

Over- and under- representation of short subsequences such as all dinucleotides and trinucleotides were determined for a representative sample of

sequences from species encompassing a broad phylogenetic range (Table 13). In general, the dinucleotide TA appears to be deficient in all of the genomes examined and the CG dinucleotide is under-represented in *S. cerevisiae* and mammalian genomes. The degree of nonrandomness in dinucleotide and trinucleotide composition differs among the different species, with the greater extremes in representation existing in sequences from mammalian species. This was seen in chaos patterns as more uniform data point distribution in sequences of nonmammalian genomes and the deficiency of data points in subquadrants of the chaos plots representing the occurrence of CG dinucleotides in mammalian genomes.

Entropy is a measure of nonrandomness in sequence organization and is defined as the degree of deviation from equivalent frequencies of single nucleotides, dinucleotides and trinucleotides and successively longer subsequences. In this investigation entropy was defined specifically as the coefficient of variation (%) among strand-symmetric relative abundances (i.e., representation values) for same-length subsequences. Entropy values based upon subsequence representation measure the degree of nonrandomness at single nucleotide, dinucleotide and trinucleotide levels of sequence organization, independently. Table 14 presents the coefficient of variation (%) for single nucleotide frequencies and dinucleotide and trinucleotide representation values for 17 sequences representative of the diverse genome-types examined. Entropy profiles were generally similar within a genome-type with the exception of sequences from the human genome which displayed two different profile-types. Nonrandomness in single nucleotide composition varied from 2.8 to 36% for sequences of diverse organisms. Bias in single nucleotide frequencies was greatest in the sequence from the *R. capsulatus* genome and least in all sequences of *E. coli* and a sequence from the *H. sapiens* genome. Generally,

Table 13. Percent C and G content and extremes in representation of dinucleotides and trinucleotides in representative examples of large continuous DNA sequences from phylogenetically diverse species.

Species LOCUS*	Sequence Length (nts)	%C+G Content	Dinucleotides Under-representation		Dinucleotides Over-representation		Trinucleotides Under-representation		Trinucleotides Over-representation	
E. coli										
ECO110K	111,401	53	TA	0.72**	GC	1.27	CTA/TAG	0.67	CCA/TGG	1.29
			AG/CT	0.82	AA/TT	1.22	ACA/TGT	0.77	CAG/CTG	1.21
ECOUW85U	91,408	53	TA	0.74	GC	1.29	CTA/TAG	0.68	CCA/TGG	1.33
			AG/CT	0.83	AA/TT	1.21	ACA/TGT	0.77	CAG/CTG	1.22
R. capsulatus										
RCPHSYNG	45,959	66	TA	0.32	AT	1.47	TAA/TTA	0.72	AAG/CTT	1.33
			GT/AC	0.78	AA/TT	1.26	GGA/TCC	0.74	ATA/TAT	1.30
S. cerevisiae										
YSCCHROMI	41,987	38	TA	0.77	AA/TT	1.11	CTA/TAG	0.88	CCA/TGG	1.12
			CG	0.79	CA/TG	1.10	GCA/TGC	0.89	ACC/GGT	1.10
YSCSYGP2	36,772	40	TA	0.76	AA/TT	1.14	CCC/GGG	0.88	ACC/GGT	1.12
			CG	0.82	CA/TG	1.09	CTA/TAG	0.90	CCA/TGG	1.09
C. elegans										
CELC50C3	44,733	36	TA	0.59	AA/TT	1.22	CCC/GGG	0.84	GCC/GGC	1.24
			AC/GT	0.86	GA/TC	1.20	CGC/GCG	0.86	CCA/TGG	1.14
CELC08C3	44,025	34	TA	0.67	AA/TT	1.25	ACA/TGT	0.91	TGA/TCA	1.14
			[illegible]	0.85	CC/GG	1.12	GGA/TCC	0.92	CCG/CGG	1.10
D. melanogaster										
DROABDB	80,423	42	TA	0.77	CG	1.25	CTA/TAG	0.78	AGC/GCT	1.15
			AC/GT	0.84	AA/TT	1.24	CGC/GCG	0.87	CGA/TCG	1.15
R. norvegicus										
RATCRYG	54,670	44	CG	0.28	CA/TG	1.22	CCC/GGG	0.89	CCA/TGG	1.17
			TA	0.74	AG/CT	1.21	AAG/CTT	0.92	CCC/GGG	1.11
H. sapiens										
HUMRETBLAS	180,388	37	CG	0.24	CC/GG	1.25	CCC/GGG	0.89	CCA/TGG	1.13
			TA	0.78	CA/TG	1.16	CGA/TCG	0.90	CGC/GCG	1.13
HUMHDABCD	58,864	52	CG	0.44	CA/TG	1.22	AAG/CTT	0.80	CCA/TGG	1.19
			TA	0.66	AG/CT	1.21	CCC/GGG	0.83	AAA/TTT	1.16

*LOCUS is identified [illegible]ly in Table 3.

**strand-symmetric odds ratio calculation ρ^*_{ij} and γ_{XYZ} (section 2.2.2.5 and Burge et al., 1992).

The two highest and two lowest dinucleotide and trinucleotide representation values are listed for the oligonucleotide and its inverted complement.

Table 14. Coefficient of variation (%) for single nucleotide, dinucleotide and trinucleotide representation among DNA sequences of the same genome and genomes of diverse species.

Species LOCUS*	Coefficient of Variation (%)		
	Single Nucleotide Frequencies	Dinucleotide Representation	Trinucleotide Representation
E. coli			
ECO110K	4.8	17.2	12.9
ECOUW85U	3.6	16.9	13.3
R. capsulatus			
RCPHSYNG	36.0	27.4	15.5
S. cerevisiae			
YSCCHROMI	22.8	11.2	6.7
SCPEKGAI	28.4	11.5	6.1
YSCSYGP2	26.4	11.1	6.0
C. elegans			
CELC08C3	30.8	15.7	5.7
CELC50C3	31.2	17.5	9.3
D. melanogaster			
DROABDB	19.6	16.3	8.5
R. norvegicus			
RATCRYG	14.4	24.2	6.2
M. musculus			
MUSTCRA	12.0	27.4	6.9
H. sapiens			
HUMGHCSA	2.8	28.2	7.3
HUMHABCD	6.8	23.7	10.5
HUMNEUROF	28.8	25.6	8.8
HUMFIXG	26.0	25.8	4.7
HUMRETBLAS	30.0	25.0	6.0
HUMVITDBP	30.4	27.1	6.9

*Locus is identified fully in Table 3

maximum entropy or nonrandomness occurred at single nucleotide or dinucleotide levels.

Spearman rank-order correlation coefficients that measured the similarity of single nucleotide, dinucleotide and trinucleotide frequencies for the DNA sequence and its complementary strand were highly significant for all sequences examined ($p<0.01$) . This observation is indicative of a similar organization on both strands rather than a complementary organization between the two strands.

Spearman rank-order correlation coefficients of single, dinucleotide and trinucleotide strand symmetric relative abundances for sequences of the same and different genomes identified greater similarity in short-sequence composition within the same genome or genomes of closely related species than among the genomes of diverse species (Table 15). Correlation coefficients for dinucleotide representation were applied as a similarity matrix and generally clustered sequences in a phenogram in a pattern consistent with the genome-type specificity observed in chaos patterns of global sequence organization (Figure 20). The same pattern was not evident at the trinucleotide level (phenogram not shown). It appears that among the genome-types examined in this study that nonrandom structure is greater at single and dinucleotide levels of organization and these levels are represented by the largest subquadrants in a chaos plot. The genome-specific features of global sequence organization discerned by visual examination of chaos patterns are for the most part due to nonequivalent nucleotide frequencies and biases in association of two nucleotides, the most basic elements of sequence organization.

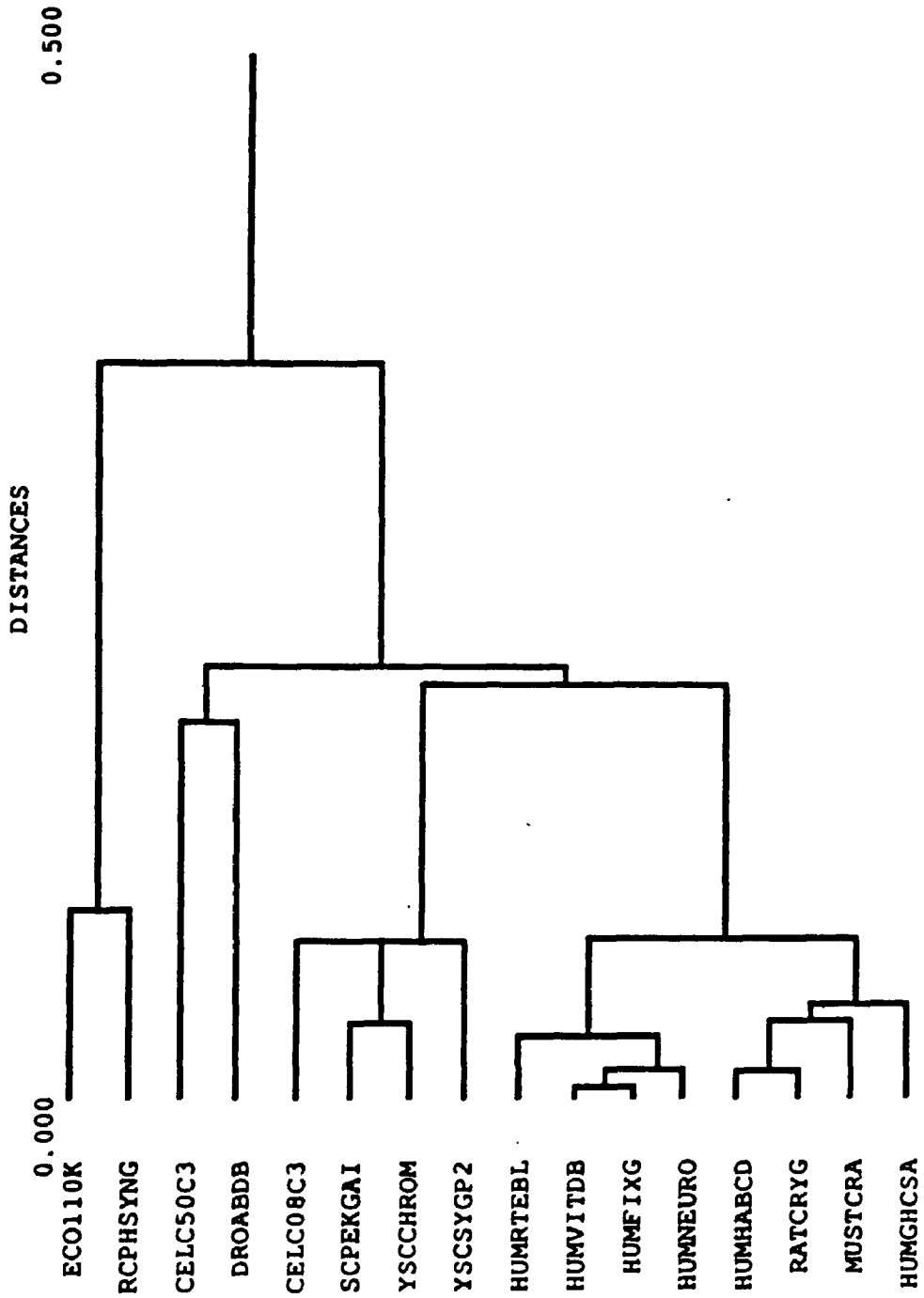
Table 15. Spearman rank-order correlation coefficients for pairwise comparisons of dinucleotide (above diagonal) and trinucleotide (below diagonal) representation for large DNA sequences from diverse species.

LOCUS*	ECO110K	[illegible]	YSCCHROMI	YSCSYGP2	CELC09C3	CELC50C3	DROARDB	MUSTCRA	RATCRYG	[illegible]	[illegible]	[illegible]
ECO110K	-----**	0.769	0.618	0.651	0.389	0.682	0.706	-0.184	-0.077	-0.092	-0.246	0.071
[illegible]	0.565	-----	0.606	0.657	0.642	0.726	0.635	-0.131	-0.003	-0.092	-0.074	0.107
YSCCHROMI	0.606	0.606	-----	0.967	0.969	0.960	0.605	0.632	0.703	0.647	0.593	0.751
YSCSYGP2	0.642	0.386	0.725	-----	0.902	0.789	0.745	0.623	0.706	0.733	0.608	0.849
CELC09C3	0.088	-0.129	0.310	0.681	-----	0.819	0.730	0.510	0.646	0.656	0.569	0.784
CELC50C3	0.326	0.021	0.661	0.832	0.636	-----	0.674	0.282	0.353	0.303	0.263	0.630
DROARDB	0.095	-0.173	0.307	-0.041	0.136	0.576	-----	0.260	0.312	0.687	0.297	0.638
MUSTCRA	0.057	-0.398	0.182	-0.122	0.217	0.659	0.654	-----	0.976	0.896	0.967	0.816
RATCRYG	0.201	-0.246	0.294	0.148	0.183	0.625	0.630	0.624	-----	0.920	0.920	0.866
[illegible]	-0.057	0.126	-0.049	-0.048	0.334	0.095	-0.114	0.052	0.109	-----	0.911	0.967
[illegible]	0.544	-0.046	0.539	0.371	0.297	0.497	0.376	0.542	0.702	0.121	-----	0.869
[illegible]	0.497	0.086	0.871	0.549	0.693	0.822	0.298	0.352	0.634	0.322	0.811	-----

* LOCUS is identified fully in Table 3

** Spearman rank correlation coefficients for comparison of dimucleotide and trinucleotide representation are presented above and below the diagonal (-----), respectively.

Figure 20. A comparison of the dinucleotide representation in large continuous DNA sequences (>36,000 nts) from eight different species. A dendogram produced by assuming that the Spearman rank-order correlation coefficients for pairwise comparisons of dinucleotide representation constitute a similarity matrix. Sequence labels are defined in Table 3. The plot is produced using a clustering method based on Euclidean distance (Wilkinson 1991).


DISTANCES
0.500
0.000
ECO110K
RCPHSYNG
CELC50C3
DROABDB
CELC08C3
SCPEKGAI
YSCCHROM
YSCSYGP2
HUMRTEBL
HUMVITDB
HUMFIXG
HUMNEURO
HUMHABCD
RATCRYG
MUSTCRA
HUMGHCSA

3.2.6 Short-Sequence Representation and the Length, Location and Function of the Nucleotide Sequence

Local DNA sequence organization was examined in nonoverlapping segments (30 and 15 kb in length) covering the entire sequence of the yeast *S. cerevisiae* chromosome III. These regions have a structure that is similar to the organization of the entire *S. cerevisiae* chromosome III sequence (see Figure 19e) and sequences from other chromosomes in the *S. cerevisiae* genome (see Figure 19f). In all of the chaos plots generated, plotting was monitored and the final pattern in data point distribution was established early in the course of plotting and did not alter with the subsequent addition of data points (personal observation of plotting). The chaos pattern generated by nucleotides 1 to 30,000 of the chromosome III sequence is shown in Figure 21a and is representative of the chaos patterns for the 30 kb and 15 kb regions examined. Twenty-two putative open reading frames encompassing 28 kb and stretching from the left telomere of the chromosome also produce the chaos pattern typical of *S. cerevisiae* nuclear genome sequences (Figure 21b). Nontranslatable sequences (28 kb) have a chaos pattern similar to that of translatable sequences and all sequences of the *S. cerevisiae* nuclear genome examined thus far (Figure 21c).

Biases in dinucleotide and trinucleotide representation do not appear to vary in regions of the chromosome III sequence or in strictly translatable or entirely nontranslatable chromosome III sequences (Table 16). Significantly high correlations exist in short-sequence representation in heterogeneous 15 and 30 kb regions and translatable and nontranslatable sequences of chromosome III ($p<0.05$). A *S. cerevisiae*-specific global sequence structure is evident in regions of different chromosomes, among different regions of a single chromosome, in

Figure 21. Chaos representation of DNA sequence organization for different regions of chromosome III of the yeast *Saccharomyces cerevisiae*. Chaos patterns were generated for a) a 30 kb region originating at the left telomere, b) a 28 kb sequence composed of 22 putative translatable elements and c) a 28 kb sequence composed of entirely non-translatable elements of the *S. cerevisiae* chromosome III. Translatable and non-tranlatable sequences were assembled from the left telomere of the *S. cerevisiae* chromosome III.

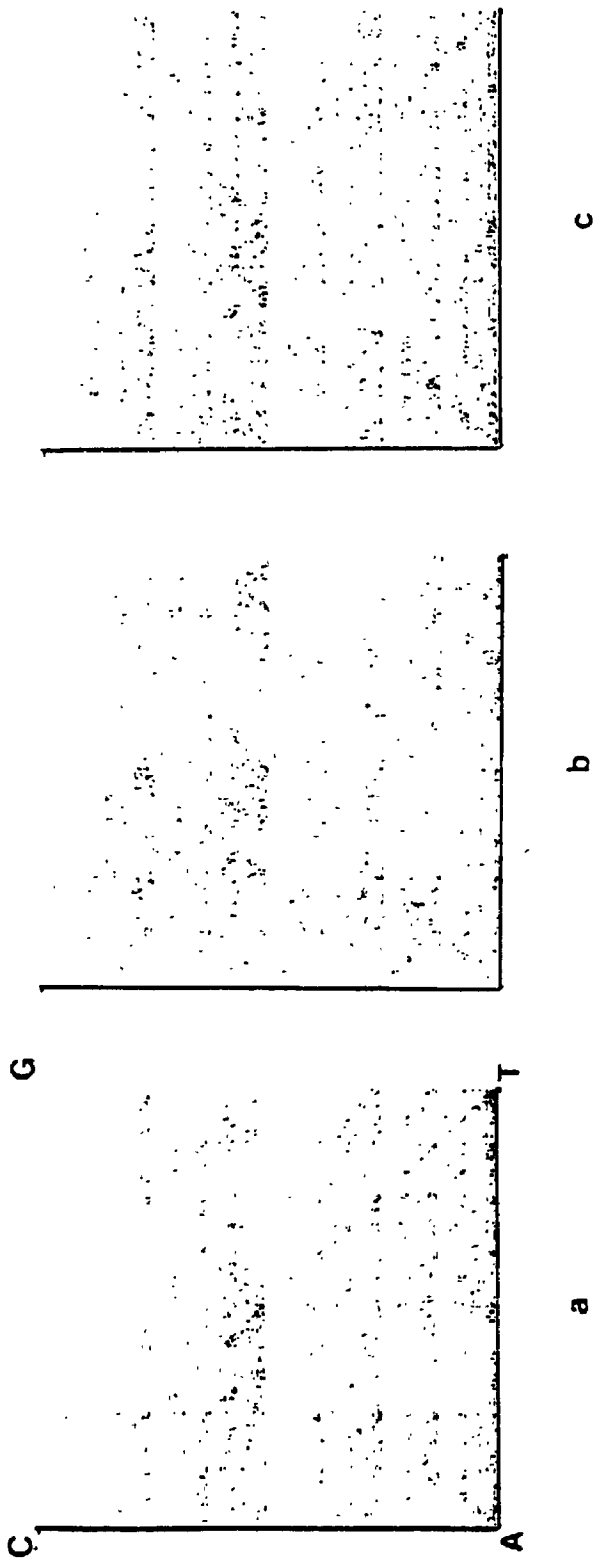

C
G
T
A
a
b
c

Table 16. Dinucleotide and trinucleotide representation values for the complete *Saccharomyces cerevisiae* chromosome III DNA sequence (315,357 nts) and 11 consecutive and non overlapping regions of that sequence (30,000 nts each) and 28 kb sequences composed of strictly translatable and entirely non-translatable sequences.

Region	%G+C	Dinucleotides and values		Trinucleotides and values	
		Under-represented	Over-represented	Under-represented	Over-represented
Total Chromosome	38	TA 0.78*	AA/TT 1.13	CCC/GGG 0.90	CCA/TGG 1.12
		CG 0.80	CA/TG 1.11	CTA/TAG 0.90	ATA/TAT 1.08
I	38	CG 0.77	CA/TG 1.12	CCC/GGG 0.87	CCA/TGG 1.11
		TA 0.80	AA/TT 1.10	CTA/TAG 0.90	ACC/GGT 1.08
II	40	TA 0.72	AA/TT 1.17	CTA/TAG 0.87	CCA/TGG 1.10
		CG 0.82	CA/TG 1.10	CCC/GGG 0.89	GTA/TAC 1.09
III	40	TA 0.78	CA/TG 1.14	CCC/GGG 0.86	ATA/TAT 1.12
		CG 0.82	AA/TT 1.10	CTA/TAG 0.87	CCA/TGG 1.11
IV	38	CG 0.77	AA/TT 1.12	CCC/GGG 0.87	ACC/GGT 1.13
		TA 0.79	CA/TG 1.10	ACA/TGT 0.89	CCA/TGG 1.11
V	37	TA 0.81	AA/TT 1.15	CTA/TAG 0.87	CCA/TGG 1.12
		CG 0.82	GC 1.08	AGG/CCT 0.92	AAG/CTT 1.09
VI	37	TA 0.81	AA/TT 1.12	GCA/TGC 0.89	CCA/TGG 1.12
		CG 0.82	CA/TG 1.08	AGG/CCT 0.90	TCA/TGA 1.08
VII	38	TA 0.77	AA/TT 1.15	CTA/TAG 0.87	CCA/TGG 1.12
		CG 0.84	CA/TG 1.08	CCC/GGG 0.91	ACC/GGT 1.10
VIII	42	TA 0.72	AA/TT 1.13	ACA/TGT 0.88	CCA/TGG 1.14
		CG 0.82	CA/TG 1.13	CCC/GGG 0.88	CAC/GTG 1.10
IX	41	TA 0.73	CA/TG 1.14	ACA, TGT 0 89	ATA/TAT 1.11
		CG 0.75	AA/TT 1.12	CTA/TAG 0.89	CCA/TGG 1.11
X	35	CG 0.76	CA/TG 1.13	CTA/TAG 0.91	CGC/GCG 1.15
		TA 0.80	AA/TT 1.11	GCA/TGC 0.91	CCA/TGG 1.08
XI**	38	CG 0.81	AA/TT 1.10	CCC/GGG 0.86	CCA/TGG 1.15
		TA 0.82	CA/TG 1.09	ACA/TGT 0.87	GCC/GGC 1.13
Non-translatable	36	CG 0.77	AA/TT 1.11	CCC/GGG 0.88	GTG/CAC 1.10
		TA 0.81	CA/TG 1.11	ACG/CGT 0.92	CCA/TGG 1.08
Translatable	42	TA 0.70	TG/CA 1.14	CTA/TAG 0.84	CCA/TGG 1.11
		CG 0.81	AA/TT 1.13	CCC/GGG 0.89	GTA/TAC 1.09

* strand-symmetric odds ratio calculation p^*_{ij} and r^*_{XYZ} (section 2.2.2.5 and Burge et al., 1992). The two highest and two lowest dinucleotide and trinucleotide representation values are listed for the oligonucleotide and its inverted complement.

** region 11 contains 15,357 nts.

strictly translatable sequences and in entirely nontranslatable sequences. Global DNA sequence structure in the *S. cerevisiae* nuclear genome did not appear to be a composite of different patterns but appears to be independent of location within the genome and the length and function of the sequence. Additionally, comparisons of chaos patterns and short-sequence representation were made for translatable and nontranslatable sequences of the neurofibromatosis-1 gene and globin gene cluster and these functionally distinct sequence elements have similar short-sequence structures ($p<0.05$) in the human genome.

3.2.7 Representation of Longer Subsequences

One aspect of sequence organization that was regionally specific was the location of highly repeated complex subsequences. These sequence structures are not visible in two-dimensional chaos plots or measured in the examination of short-sequence representation but were identified through analysis of three-dimensional chaos patterns that portray the degree of repetition of each individual (x,y) coordinate (i.e., the repeated occurrence of specific subsequences). For a single subsequence that occurs more than once in the total DNA sequence the (x,y) coordinates defining each nucleotide of the subsequence will repeat and thus indicate the reoccurrence of the oligonucleotide. In two-dimensional chaos plots such data points are superimposed and not a part of a visual interpretation of the plots.

For the *S. cerevisiae* chromosome III complete DNA sequence the construction of chaos plots in this analysis was such that 53,726 different (x,y) coordinates are repeated at least once in various locations along the *S. cerevisiae* chromosome III sequence. The most frequent (x,y) coordinates are (0,0), (667, 0), (333,0) and (1000,0) and they occur 120, 83, 82 and 64 times,

respectively. These coordinates identify polyA, $(CA)_N$ and polyT repeats. Repeated subsequences were also described for 10 consecutive 30 kb regions and a terminal 15 kb region of the chromosome. Regions of the chromosome differ in the number of different (x,y) coordinates that are repeated (837 to 1094 coordinates) and the degree to which the most frequent coordinate occurs (5 to 47 occurrences). Among the 30 kb regions, there are from 7 to 51 different (x,y) coordinates that occur at least four times and these coordinates have an average repetition of 4 to 9 times. There appear to be differences in the type and degree of repetition of short sequences along the length of the *S. cerevisiae* chromosome III.

In each of the 11 nonoverlapping regions (30 kb) of this chromosome, those subsequences that generated at least 4 repetitions of an (x,y) coordinate were determined in order to limit analysis to the most highly repeated subsequences (Table 17). Of the 39 subsequences identified, only 5 occurred at least four times in multiple regions and these oligonucleotides were simple repeats of single (polyA or polyT), dinucleotides (TA repeats) or trinucleotides (CAA and CAT repeats). The remaining 34 oligonucleotides were repeated at least four times but in a single region only.

Complex repeated subsequences were defined as containing at least 3 of the 4 nucleotides and being at least 9 nucleotides in length. The 18 repetitive oligonucleotides identified as complex (Table 18) had highly regional occurrences (i.e., primarily occurring in a single 30 kb segment). The complex repeated oligonucleotides were classified as: 1) tandem, occurring consecutively with no intervening sequences; 2) dispersed, occurring at distant locations at no regular intervals and/or 3) mixed, occurring both in tandem and at dispersed location. Most of the complex repeated oligonucleotides were dispersed in occurrence in the entire chromosome sequence. Four of the complex repeats were tandemly

Table 17. Thirty-nine oligonucleotide sequences that are repeated a minimum of four times within at least one of the 10 consecutive 30 regions of *Saccharomyces cerevisiae* chromosome III.

Label	Oligonucleotide	Regional location
1	$(A)_{7-23}$	I,II,IV,V,VI,VIII,X
2	$C_{2-3}A(CA)_{1-3}$	I
3	G_2A_6	I
4	GACGCTGAGTCTTTACC	I
5	$(TA)_{4-6}$	I,II,III,IV,VII,VIII,IX,X
6	TTGTTTATTTA	I
7	ATTTGTTTGTA	I
8	TTTTGTTTAT	I
9	T_{7-15}	I,IV,V,VI,VII,VIII,IX
10	GAATAAAAATCAA	III
11	TTTACGTTAC	III
12	$(AGGTCA)_4$	III
13	GATGATAATA	III
14	CTAGTATAT	III
15	AAAAAAATAAAA	IV
16	ATAAGAAAAA	IV
17	$(TTC)_7$	V
18	$(GTT)_8$	V
19	TGATAATATA	V
20	TTTGGAAGAA	VI
21	TTATTATTT	VI
22	TTTTGATTT	VI
23	$(CAA)_{3-5}$	VI,VIII,X
24	TAAACAAAA	VII
25	TTCCTCTTC	VII
26	$(TTC)_{3,4}$	VII
27	$(GAA)_{4-11}$	VIII
28	TAGTGGGTTCC	VIII
29	TTTTCTTTTTT	VIII
30	$(TCA)_{3-5}$	IX
31	AAACCAAATCA	IX
32	$(CATA)_{2,5}$	IX
33	CTGCTTCTGCT	IX
34	TTGCTGTTG	IX
35	$(CAT)_6$	IX,X
36	GAAATAT	X
37	TCAACTTCA	X
38	$(TTA)_{3,6}$	X
39	$(TTC)_3$	X

Table 18. Oligonucleotide sequences repeated along the length of the *Saccharomyces cerevisiae* chromosome III.

Label	Oligonucleotide	Locations in 315 kb sequence				Occurrence of inverted sequence
	Tandem repetitions					
4	GACGCTGAGTCTTTACC	6,226	6,243	6,260	6,277	0
6	TTGTTTATTTA	29,442	29,454	29,474	29,49.	5
30	AAACCAAATCA	247,693 267,899*	247,707	264,467*	267,572*	1
33	CTGCTTCTGCT	262,970*	262,982*	262,994*	263,000*	1
	Dispersed repetitions					
7	ATTGTTGTA	5,874 196,949*	12,133 198,036	19,646* 303,508*	29,786	2
8	TTTTGTTTAT	12,638* 115,780 261,967*	13,953* 197,275* 292,176*	19,749* 198,541* 298,268	29,439 199,856*	6
10	GAATAAAAATCAA	83,694	83,920	84,436	90,066	0
11	TTTACGTTAC	65,788 90,096	83,724	83,950	84,466	2
13	GATGATAATA	43,965* 84,602	77,047* 90,232	83,860 186,552*	84,087*	1
14	CTAGTATAT	83,732 84,678 149,247 168,516	83,959 90,104 149,445 247,535*	84,163* 90,308 151,486 290,162	84,474 142,292 153,466* 293,991	0
16	ATAAGAAAAA	45,147* 111,435* 290,533	55,536* 112,787	100,882 141,909	107,071 258,751	4
17	TGATAATATA	10,283* 136,443*	81,512* 142,737	124,018 142,748	124,029	2
20	TTTGGAAGAA	115,728 171,063*	144,165* 174,407*	153,216* 305,498	155,219	1
22	TTTTGATTT	11,246 30,084 161,614* 182,055* 292,022*	12,484* 35,966* 163,239* 198,387*	20,739* 48,008 163,273* 204,424	29,333 83,303 174,169* 263,357	12
25	TAAACAAAA	23,700* 186,670* 205,556	29,084 200,750* 274,882*	29,520 201,428* 286,756	112,159 205,127*	8
34	TTGCTGTTG	58,146* 174,091* 260,887*	58,336* 260,602* 263,245	69,452* 260,632* 282,199*	101,890* 260,851* 297,782	1
37	TCAACTTCA	66,920* 270,993 279,785*	153,816* 272,632 289,541	180,735* 273,483* 295,861*	266,663* 276,620*	6
	Mixed repetitions					
27	TAGTGGGTTCC	232,477*	232,507*	232,537*	239,932*	0

* The oligonucleotide is located in a putative open reading frame.

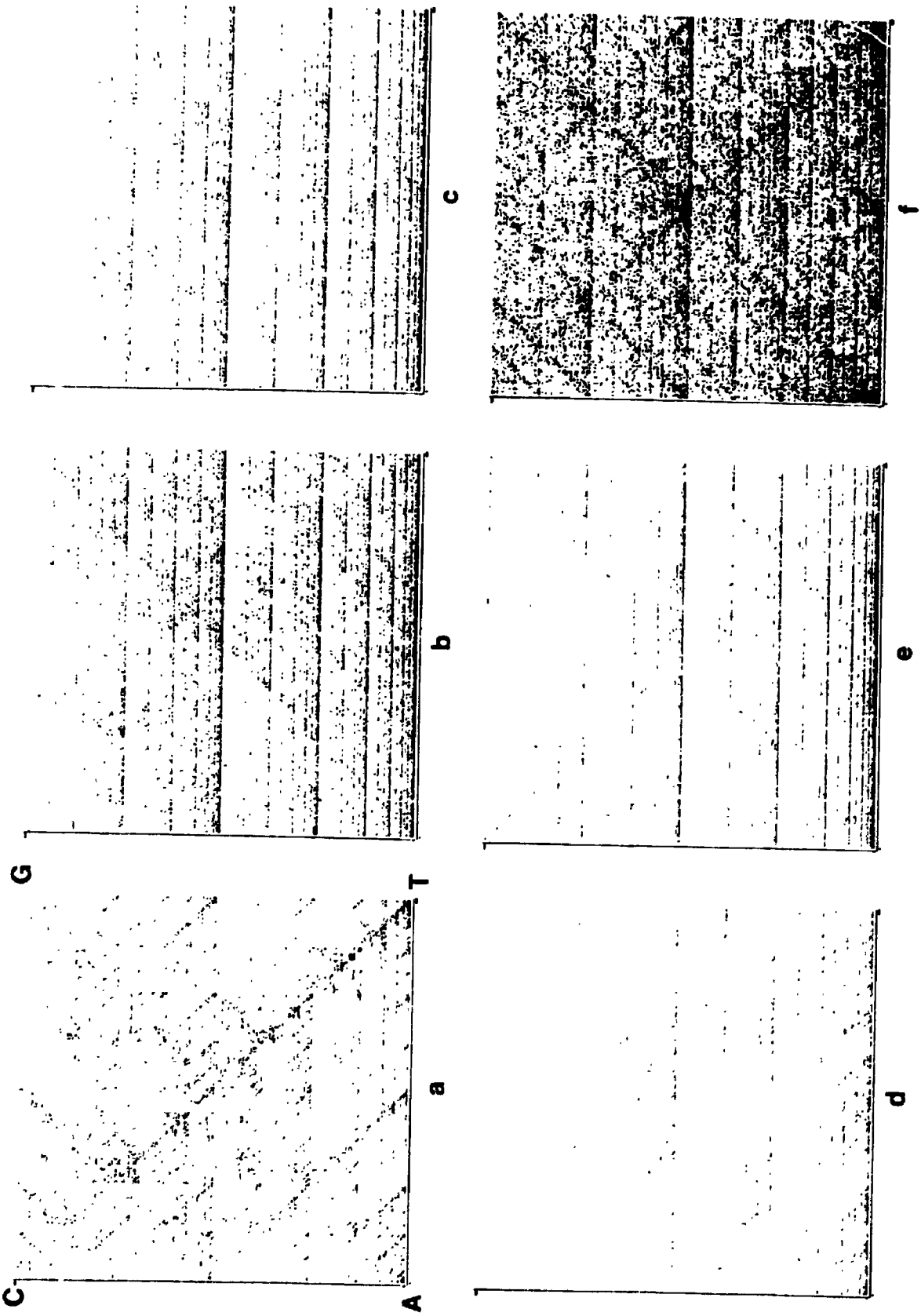
repeated and one oligonucleotide was both tandem and dispersed in occurrence. The occurrences of the inverted sequence of each of the 18 repeated oligonucleotides either do not occur at all or only in low frequency. The repetition of certain larger subsequences was not distributed randomly within the entire chromosome sequence but was region-specific.

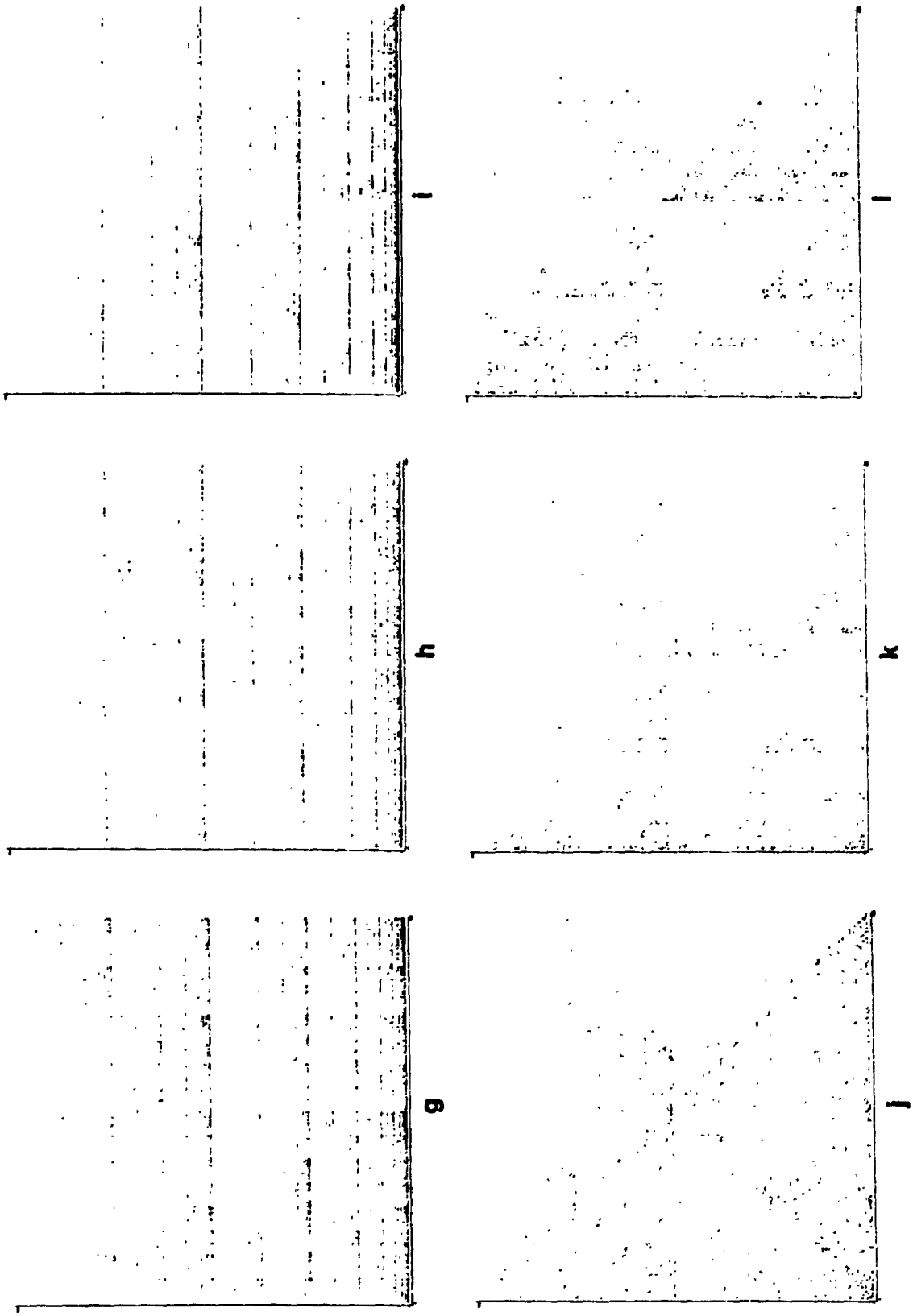
3.2.8 Sequence Organization in Mitochondrial Genomes

A visual examination of major features of data point distribution in chaos plots for the mitochondrial genomes (mtDNAs) of 28 diverse species (12 representative patterns) identified seven different types of patterns. These distinct chaos patterns were exhibited by 1) *Paramecium aurelia* and *Marchantia polymorpha*, 2) yeast and other fungi, 3) *Prototheca wickerhamii*, 4) nematode, 5) insect, 6) echinoderm and 7) vertebrate species (Figure 22). All mtDNA sequences are characterized by an over-representation of adenine and thymine. This is seen in chaos plots as increased datapoint density towards the A and T axis. The chaos patterns of mtDNAs of more distantly-related species differ in major features of data point distribution but chaos patterns of mtDNAs of more closely-related species have similar chaos patterns. Sequence organization displayed in chaos patterns is similar within consecutive and nonoverlapping regions of mitochondrial genomes as apparent from an analysis of 4, 16 and 30 kb fragments of the human, *S. cerevisiae* and *M. polymorpha* mtDNAs, respectively (chaos plots not given).

The chaos pattern of the mtDNA of *P. aurelia* (Figure 22a) is characterized by high data point density along a diagonal line joining C and T vertices. Data point density also increases towards the A and T axis and in particular towards the T vertex. Subquadrants of the plot that represent CG

Figure 22. Representative chaos patterns of complete mitochondrial genome sequences from phylogenetically divergent species. a) *Paramecium aurelia*, 40,469 nts; b) *Podospora anserina*, 100,314 nts; c) *Prototheca wickerhamii*, 55,328 nts; d) *Schizosaccharomyces pombe*, 19,431 nts; e) *Saccharomyces cerevisiae*, 78,294 nts; f) *Marchantia polymorpha*, 186,608 nts; g) *Caenorhabditis elegans*, 13,794 nts; h) *Artimiidae franciscana*, 15,770 nts; i) *Paracentrotus lividus*, 15,679 nts; j) *Drosophila yakuba*, 16,019 nts; k) *Phoca vitulina*, 16,826 nts; l) *Homo sapiens*, 16,569 nts. Chaos patterns are contained on two pages.


C
A
a
G
T
b
c
d
e
f


i
l
h
k
g
j

dinucleotides contain relatively fewer data points and subquadrants representing the CT dinucleotide, CTT trinucleotide and CTTC tetranucleotide contain a greater number of data points than other similar-sized subquadrants.

The chaos pattern of the fungus *Podospora anserini* mtDNA (Figure 22b) contains horizontal striations with increasing data point density towards the A and T axis. The higher frequency of T relative to A, C and G is represented in the chaos plot by higher data point density towards the T vertex and below the diagonal line stretching from the A and G vertices. The alga species *P. wickerhamii* mtDNA (Figure 22c) has a chaos pattern dominated by horizontal striations with increasing data point density towards the A and T axes but there is a relatively uniform data point distribution between A and T vertices. The mtDNA chaos patterns of the two yeast species, *Schizosaccharomyces pombe* (Figure 22d) and *S. cerevisiae* (Figure 22e) have features most like that of *P. anserinii* with the exception of high data point density along diagonal line between C and T and A and G vertices.

The relatively large mtDNA of *Marchantia polymorpha* (186 kb; Figure 22f) produces a chaos pattern with the highest overall data point frequency. Like the chaos pattern for the mitochondrial genome of *P. aurelia*, data point density is greatest towards A and T vertices, the A and T axis and along the diagonal line joining T and C vertices. As well, subquadrants representing subsequences with the dinucleotide suffix, CT contain a greater number of data points than other similar-sized subquadrants. Unlike the pattern for *P. aurelia*, data point density is also high along the diagonal line joining A and G vertices.

The mtDNA sequences of invertebrate species have three types of chaos patterns and the vertebrate species examined have a single type of chaos pattern. The mtDNAs of the nematode species *Ascaris suum* and *Caenorhabditis elegans* have similar chaos patterns characterized by increased data point

frequency toward the A and T axis and in particular towards the T vertex (*C. elegans*, Figure 22g) . Of the two patterns, that of *A. suum* has the greater bias for A and T nucleotides over C and G nucleotides. *Drosophila yakuba, Apis melifera* and *Anopheles gambiae* mtDNAs have similar chaos patterns and the greatest extremes in bias for A and T nucleotides (*D. yakuba*, Figure 22h and *A. melifera*, 22i). In these chaos patterns, subquadrants representing CG and GC dinucleotides contain relatively fewer data points compared with similar-sized subquadrants. The mtDNAs of *Artimiidae franciscana*, *Strongylocentrotus purpuratus* and *Paracentrotus lividus* have chaos patterns typified by increased data point density along the diagonal lines between A and G and T and C vertices (*P. lividus*, Figure 22j). Data point density in these plots increases towards the A and T axis and subquadrants representing CG and GC dinucleotides contain fewer data points compared with similar-sized subquadrants. Vertebrate mtDNAs have a chaos pattern typified by a paucity of G and the CG dinucleotide (representative plots are shown for *Phoca vitulina*, Figure 22k and *Homo sapiens*, Figure 22l). The degree to which G and the CG dinucleotide are under-represented in these genomes varies among the different vertebrate species.

The major features of mitochondrial DNA sequence organization portrayed in chaos patterns can be quantified in part through determination of the over- and under-representation of short subsequences such as single nucleotides, dinucleotides and trinucleotides. The major determinants of data point distribution in chaos patterns are nonequivalent occurrences of single nucleotides and over- and under-representation of successively longer oligonucleotides. Representation values for short sequences in mtDNAs indicate a greater degree of bias in single nucleotide and dinucleotide composition than in trinucleotide composition (Table 19). All mtDNAs have an over-representation of A and T, a

Table 19. Single nucleotide, dinucleotide and trinucleotide representation values for mitochondrial sequences from phylogenetically diverse species and for a prokaryote genome sequence and nuclear genome sequences.

Evolutionary Category SEQUENCE**		Single nucleotide representation	Dinucleotides and values Under-represented	Dinucleotides and values Over-represented	Trinucleotides and values Under-represented	Trinucleotides and values Over-represented
Mitochondrial Genomes						
Protozoan						
PAUR	A/T	1.18*	AT 0.65	AA/TT 1.37	CGC/GCG 0.76	CGA/TCG 1.33
	C/G	0.82	CA/TG 0.76	AG/CT 1.22	AGA/TCT 0.85	GTA/TAC 1.28
Fungus						
PANS	A/T	1.39	CG 0.84	GC 1.29	GCC/GGC 0.88	ACC/GGT 1.13
	C/G	0.60	CA/TG 0.84	CC/GG 1.25	AAC/GTT 0.91	CGC/GCG 1.10
Alga						
PWIC	A/T	1.48	GA/TC 0.87	GC 1.38	CCG/CGG 0.74	ACG/CGT 1.21
	C/G	0.52	CG 0.91	CC/GG 1.20	CGC/GCG 0.87	CCA/TGG 1.17
Yeast						
SPOM	A/T	1.40	CG 0.54	CC/GG 1.31	CCC/GGG 0.83	CCA/TGG 1.17
	C/G	0.60	AC/GT 0.85	AG/CT 1.13	GCC/GGC 0.85	ACC/GGT 1.15
SCER	A/T	1.65	TG/CA 0.65	CC/GG 3.11	CCC/GGG 0.54	CAA/TTG 1.34
	C/G	0.35	GT/AC 0.68	CG 1.48	CGC/GCG 0.56	AAG/CTT 1.33
Plant						
MPOL	A/T	1.15	AC/GT 0.82	AA/TT 1.24	AGA/TCT 0.89	TCA/TGA 1.15
	C/G	0.85	TA 0.85	CC/GG 1.22	ACA/TGT 0.92	AGC/GCT 1.10
Invertebrate						
ASUU	A/T	1.44	CG 0.36	CC/GG 1.61	CCC/GGG 0.79	CCG/CGG 1.27
	C/G	0.56	GC 0.72	AA/TT 1.23	CAC/GTG 0.83	CTC/GAG 1.26
CELE	A/T	1.52	CG 0.56	CC/GG 1.52	CGA/TCG 0.72	ACG/CGT 1.29
	C/G	0.47	GA/TC 0.83	AA/TT 1.11	CCC/GGG 0.85	CCG/CGG 1.10
AFRA	A/T	1.29	CG 0.66	CC/GG 1.37	CGC/GCG 0.79	GTA/TAC 1.14
	C/G	0.71	AC/GT 0.80	AG/CT 1.12	GGA/TCC 0.88	CCG/CGG 1.11
PLIV	A/T	1.21	CG 0.58	CC/GG 1.31	CGC/GCG 0.84	CCG/CGG 1.20
	C/G	0.79	AC/GT 0.82	AG/CT 1.19	ATC/GAT 0.91	AGC/GCT 1.09
SPUR	A/T	1.08	CG 0.56	TT/AA 1.86	CCC/GGG 0.89	CAC/GTG 1.11
	C/G	0.82	AC/GT 0.80	CC/GG 1.33	CGC/GCG 0.91	CCG/CGG 1.09
DYAK	A/T	1.57	CG 0.68	CC/GG 1.68	CGC/GCG 0.65	AGC/GCT 1.42
	C/G	0.43	AC/GT 0.80	GC 1.34	GCC/GGC 0.65	CGA/TCG 1.28
AMEL	A/T	1.70	AC/GT 0.71	CC/GG 1.94	CCC/GGG 0.41	GAC/GTC 1.29
	C/G	0.30	CG 0.81	GA/TC 1.12	GCC/GGC 0.44	AGG/CCT 1.26
AGAM	A/T	1.55	CG 0.68	CC/GG 1.57	CGC/GCG 0.63	CGA/TCG 1.36
	C/G	0.45	AC/GT 0.80	GC 1.40	CCG/GGC 0.71	AGC/GCT 1.28

Evolutionary Category SEQUENCE**	Single nucleotide representation		Dinucleotides and values Under-represented		Over-represented		Trinucleotides and values Under-represented		Over-represented	
Vertebrate										
GGAL	A/T	1.08	CG	0.46	CC/GG	1.37	AAG/CTT	0.85	CCG/CGG	1.24
	C/G	0.92	GC	0.82	AG/CT	1.12	GAC/GTC	0.86	GCC/GGC	1.17
CCAR	A/T	1.14	CG	0.62	CC/GG	1.30	AGA/TCT	0.88	CCG/CGG	1.16
	C/G	0.86	GA/TC	0.86	AG/CT	1.07	CCC/GGG	0.88	AGG/CCT	1.11
XLAE	A/T	1.26	CG	0.63	CC/GG	1.28	CGC/GCG	0.78	GCC/GGC	1.12
	C/G	0.74	AC/GT	0.89	CT/AG	1.06	CCC/GGG	0.87	CCG/CGG	1.12
CLAC	A/T	1.09	CG	0.60	CC/GG	1.3[illegible]	CGC/GCG	0.91	CCG/CGG	1.15
	C/G	0.91	AC/GT	0.87	AA/TT	1.0[illegible]	CCA/TGG	0.92	GAA/TTC	1.08
OMYK	A/T	1.08	CG	0.64	CC/GG	1.[illegible]	CGC/GCG	0.89	GCC/GGC	1.08
	C/G	0.92	AC/GT	0.89	AG/CT	1	GAC/GTC	0.90	CCG/CGG	1.08
DVIR	A/T	1.34	CG	0.55	CC/GG	1.43	CGC/GCG	0.86	AGG/CCT	1.10
	C/G	0.66	AC/GT	0.86	TA	1.09	AGA/TCT	0.86	TCA/TGA	1.10
BPHY	A/T	1.19	CG	0.54	CC/GG	1.31	CGC/GCG	0.83	GCC/GGC	1.17
	C/G	0.81	GA/TC	0.90	AG/CT	1.10	GAC/GTC	0.85	CCG/CGG	1.16
BMUS	A/T	1.21	CG	0.54	CC/GG	1.31	AGA/TCT	0.85	CCG/CGG	1.22
	C/G	0.79	GC	0.90	AG/CT	1.11	GAC/GTC	0.85	GCC/GGC	1.18
HGRY	A/T	1.16	CG	0.64	CC/GG	1.24	CGC/GCG	0.83	ACG/CGT	1.18
	C/G	0.84	GC	0.89	TA	1.10	AAG/CTT	0.83	GCC/GGC	1.16
PVIT	A/T	1.[illegible]7	CG	0.65	CC/GG	1.24	CGC/GCG	0.78	ACG/CGT	1.22
	C/G	0.83	GC	0.87	TA	1.09	AAG/CTT	0.83	GCC/GGC	1.20
BTAU	A/T	1.21	CG	0.56	CC/GG	1.31	CGC/GCG	0.79	GCC/GGC	1.18
	C/G	0.79	AC/GT	0.9[illegible]	AG/CT	1.10	AAG/CTT	0.86	AGC/GCT	1.14
RNOR	A/T	1.23	CG	0.53	CC/GG	1.39	GCA/TGC	0.82	CCG/CGG	1.25
	C/G	0.77	GC	0.88	AG/CT	1.05	CGC/GCG	0.83	GCC/GGC	1.15
MMUS	A/T	1.26	CG	0.52	CC/GG	1.36	AAG/CTT	0.85	GAA/TCC	1.13
	C/G	0.74	GC	0.90	AG/CT	1.07	CCC/GGG	0.86	GCC/GGC	1.12
HSAP	A/T	1.11	CG	0.53	CC/GG	1.35	AAG/CTT	0.84	GCC/GGC	1.16
	C/G	0.89	GC	0.87	AG/CT	1.09	GTC/GAC	0.86	CCG/CGG	1.15
Prokaryote Genome										
MCAP	A/T	1.43	CG	0.36	AA/TT	1.19	CGG/CCG	0.33	AGC/GCT	3.30
	C/G	0.57	TA	0.78	GC	1.24	CCC/GGG	0.43	ACT/AGT	3.02
Nuclear Genomes										
Yeast										
SCERC	A/T	1.21	TA	0.76	AA/TT	1.13	TAG/CTA	0.86	CCA/TGG	1.34
	C/G	0.79	CG	0.80	TG/CA	1.12	CCC/GGG	0.88	TAT/ATA	1.33
Invertebrate										
DMELN	A/T	1.19	TA	0.83	GC	1.23	TAG/CTA	0.79	CCA/TGG	1.17
	C/G	0.81	AC/GT	0.86	AA/TT	1.11	CCC/GGG	0.84	GTG/CAC	1.15
CELEN	A/T	1.35	TA	0.64	AA/TT	1.10	TCC/GGA	0.88	CCG/CGG	1.17
	C/G	0.65	AC/GT	0.86	GA/TC	1.26	TGT/ACA	0.89	TGA/TCA	1.14
Vertebrate										
HSAPT	A/T	1.16	CG	0.48	CC/GG	1.25	CGA/TCG	0.82	CCA/TGG	1.15
	C/G	0.84	GC	0.77	AA/TT	1.20	CCC/GGG	0.87	CGC/GCG	1.13

* strand-symmetric odds ratio calculation ρ^*_{ij} and γ^*_{XYZ} (see *Methods* and Burge et al., 1992). Representation values are given for single nucleotides and the two highest and two lowest dinucleotide and trinucleotide representation values. The oligonucleotide and its inverted complement are given.

**SEQUENCE is fully identified in Table 4.

deficiency of CG dinucleotides and an over-representation of CC/GG and AG/CT dinucleotides. The greatest extremes in short-sequence representation occur in *S. cerevisiae*, *S. pombe*, *A. suum*, *D. melanogaster* and *A. melifera* mtDNAs. The type and extent of bias in short-sequence representation are generally more similar in closely-related species than among distantly-related species, although more closely-related species, such as *S. cerevisiae* and *S. pombe*, *D. melanogaster* and *A. melifera* and *C. elegans* and *A. suum* do differ in the extent of bias in short-sequence representation.

Short-sequence representation among smaller regions of the mitochondrial genomes of *H. sapiens*, *S. cerevisiae*, *M. polymorpha* and the entire mitochondrial sequence of the same species are significantly correlated ($p<0.05$). The higher-order sequence organization of mitochondrial genomes portrayed in chaos patterns can be attributed in large part to biases in short-sequence representation, in particular single nucleotide composition. Patterns in the short-sequence organization of mitochondrial genomes also appear to be correlated with host species-type.

Spearman rank-order correlation coefficients were generated for all pairwise comparisons of 28 mitochondrial genomes based on single nucleotide composition and dinucleotide and trinucleotide representation (see Table 20 for representative sequence comparisons where comparisons between more closely related species pairs are in bold type). Single nucleotide composition differs among the mtDNAs and does not appear to be species-type specific or correlated with the similarities and differences seen in chaos representations of these sequences. Dinucleotide representation considers both single nucleotide composition and nearest-neighbor nucleotide associations and for mtDNAs was similar among closely-related species and different among distantly-related species. The similarities and differences observed in chaos patterns of

Table 20. Spearman rank-order correlation coefficients for pairwise comparisons of the relative proportions of single nucleotides and the representation of dinucleotides and trinucleotides for mitochondrial genomes of phylogenetically diverse species.

SEQUENCE*	PHER	PHRB	PHIC	SYCH	SCER	MYOL	ABRU	CELE	APRA	PLIV	SPUR	DYAK	AMEL	CHAL	XLAE	CCAR	HSAP	PVIT	MMUS	RNOR
PHER	-----	0.600[a]	0.000	0.000	0.600	0.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000	0.200	0.000	0.400	0.400	0.400	0.000	0.200
	-----	0.630[b]	0.596	0.772	0.320	0.916	0.193	0.602	0.709	0.001	0.025	0.700	0.699	0.513	0.696	0.672	0.264	-0.021	0.036	0.395
	-----	0.290[c]	0.224	0.260	-0.165	0.509	0.193	0.336	0.250	0.076	0.100	0.317	0.349	0.117	-0.001	-0.070	0.039	0.244	-0.051	0.000
PHRB		-----	0.000	0.000	1.000	0.000	0.000	0.000	0.600	0.000	0.600	0.000	0.000	-0.200	0.000	0.600	0.600	0.600	0.000	-0.200
		-----	0.820	0.617	0.493	0.671	0.660	0.001	0.896	0.710	0.742	0.004	0.362	0.591	0.001	0.776	0.569	0.610	0.721	0.620
		-----	0.279	0.626	0.031	0.383	-0.171	0.000	-0.007	-0.013	0.067	0.611	0.627	0.069	-0.126	0.095	-0.019	0.096	-0.030	0.129
PHIC			-----	0.600	0.000	0.600	0.000	0.600	0.000	1.000	0.000	1.000	1.000	0.600	1.000	0.000	0.000	0.000	1.000	0.600
			-----	0.726	0.564	0.617	0.217	0.748	0.769	0.031	0.795	0.916	0.651	0.525	0.819	0.605	0.670	0.601	0.733	0.561
			-----	0.350	0.339	0.602	0.091	0.605	0.002	0.156	0.600	0.587	0.606	-0.199	0.371	-0.049	-0.160	0.213	0.207	-0.211
SYCH				-----	0.000	1.000	1.000	1.000	0.000	0.600	0.000	0.600	0.600	-0.600	0.600	0.000	0.000	0.000	0.600	-0.600
				-----	0.276	0.022	0.651	0.677	0.955	0.982	0.950	0.020	0.656	0.043	0.007	0.736	0.651	0.651	0.760	0.712
				-----	0.210	0.163	-0.339	0.163	0.004	0.096	0.110	0.523	0.350	-0.203	-0.007	-0.2?3	-0.262	-0.226	-0.230	-0.296
SCER					-- --	0.000	0.000	0.000	0.600	0.000	0.600	0.000	0.000	-0.200	0.000	0.600	0.600	0.600	0.000	-0.200
					-----	0.670	-0.024	0.320	0.368	0.371	0.610	0.505	0.507	0.119	0.347	0.237	0.053	0.040	0.250	0.160
						0.174	-0.094	-0.156	-0.036	0.356	0.122	0.327	0.253	-0.099	0.093	-0.056	-0.075	-0.231	-0.077	-0.220
MYOL						-----	1.000	1.000	0.000	0.600	0.000	0.600	0.600	-0.600	0.600	0.000	0.000	0.000	0.600	-0.600
						-----	0.315	0.667	0.748	0.020	0.007	0.810	0.660	0.575	0.733	0.607	0.205	0.006	0.693	0.604
						-----	0.200	0.119	0.265	0.177	0.230	0.375	0.233	-0.016	0.202	0.071	0.176	0.216	0.006	0.056
ABRU							-----	1.000	0.000	0.600	0.000	0.600	0.600	-0.600	0.600	0.000	0.000	0.000	0.600	-0.600
							-----	0.745	0.626	0.621	0.693	0.306	0.231	0.660	0.816	0.632	0.650	0.667	0.660	0.697
							-----	0.335	0.066	0.117	0.013	-0.101	0.000	0.062	0.163	-0.001	0.037	0.106	0.223	0.151
CELE								-----	0.000	0.600	0.000	0.600	0.600	-0.600	0.600	0.000	0.000	0.000	0.600	-0.600
								-----	0.638	0.742	0.777	0.795	0.315	0.650	0.813	0.769	0.596	0.660	0.632	0.691
								-----	0.333	-0.062	0.191	-0.076	0.076	0.030	0.129	0.050	-0.063	0.342	0.342	0.050
APRA									-----	0.000	1.000	0.000	0.000	0.200	0.000	0.600	0.600	0.600	0.000	0.200
									-----	0.955	0.952	0.843	0.641	0.852	0.272	0.769	0.605	0.519	0.040	0.769
									-----	-0.099	0.100	0.265	0.010	0.100	0.161	0.060	0.003	0.229	0.168	0.075
PLIV										--- -	0.000	1.000	1.000	0.600	1.000	0.000	0.000	0.000	1.000	0.600
										-----	0.976	0.995	0.626	0.013	0.929	0.777	0 653	0.663	0.795	0.726
										-----	0.500	0.140	0.329	0.116	0.230	0.250	0 209	0 173	0.332	0.106
SPUR											-- --	0.000	0.000	0.200	0.000	0.600	0.600	0.600	0.000	0.200
											-----	0.917	0.709	0.709	0.905	0.762	0.593	0.300	0.760	0.600
											-----	0.305	0.230	0.679	0.705	0.647	0.690	0.332	0.670	0.323

SEQUENCE*	PMUR	PAUB	PWIC	PPCM	SCER	MPOL	AGUU	CELE	APRA	PLIV	SPUR	DYAK	AMEL	CHAL	ILBA	CCAR	BMUS	PVIT	MMUS	BMAP
DYAK												....	1.000 0.600 0.690	0.400 0.671 0.132	1.000 0.893 0.274	0.000 0.730 0.329	0.000 0.532 0.100	0.000 0.360 0.202	1.000 0.703 0.159	0.600 0.617 0.112
AMEL													....	0.400 0.433 -0.013	1.000 0.404 0.001	0.000 0.279 0.026	0.000 0.202 -0.029	0.000 0.077 0.106	1.000 0.615 0.124	0.600 0.249 -0.074
CHAL														....	0.400 0.884 0.949	0.000 0.917 0.820	0.000 0.933 0.839	0.000 0.769 0.604	0.400 0.899 0.662	1.000 0.967 0.929
ILBA															....	0.000 0.932 0.724	0.000 0.807 0.710	0.000 0.641 0.320	1.000 0.916 0.747	0.600 0.855 0.652
CCAR																....	1.000 0.947 0.937	1.000 0.846 0.721	0.000 0.947 0.800	0.800 0.970 0.775
BMUS																	....	1.000 0.947 0.723	0.800 0.893 0.710	0.000 0.976 0.761
PVIT																		....	0.800 0.810 0.835	0.800 0.807 0.626
MMUS																			....	0.400 0.917 0.653
BMAP																				.....

*SEQUENCE is identified fully in Table 5.

Spearman rank correlation coefficients for comparison of single nucleotide composition[a] and dinucleotide[b] and trinucleotide[c] representation ($p=0.05$, r_s(single nucleotides)=0.878, r_s(dinucleotides)=0.482 and r_s(trinucleotides)=0.250).

mitochondrial genomes correlate with patterns in dinucleotide representation. Trinucleotide representation among mitochondrial genomes is also species-type specific but has a consistent low extent of bias in comparison with the biases in single nucleotide composition and dinucleotide representation. Thus, species-type specific patterns in DNA sequence organization for mitochondrial genomes occur primarily in dinucleotide composition. Cluster diagrams relating the various mitochondrial genomes based upon short-sequence representation were correlated with a known phylogeny of these species (Gray, 1989; 1992) and the clustering based on dinucleotide representation best portrayed the known phylogenetic relatedness of these species (Figure 23).

3.2.9 Short-Sequence Representation in Nuclear and Mitochondrial Genomes

The mitochondrial genomes of *S. cerevisiae*, *C. elegans*, *D. yakuba* and *H. sapiens* and same-length nuclear sequences of the same or a closely-related species generate different chaos patterns (compare Figure 22 e, g, j, l for mitochondrial sequences with Figure 24 a, b, c, d for nuclear sequences of the same species). The mtDNA of *S. cerevisiae* has a greater over-representation of A and T compared with a representative region of the nuclear genome. The *C. elegans* mtDNA compared with a representative nuclear region has a higher relative frequency of T and a greater over-representation of A and T. The *D. yakuba* mtDNA has an over-representation of A and T and a deficiency of CG dinucleotides not visible in the chaos pattern of an 80 kb region of the *D. melanogaster* nuclear genome. Vertebrate nuclear genomes have a greater deficiency of CG dinucleotides than do mitochondrial genomes of these species. Vertebrate mitochondrial genomes are deficient in guanine, unlike nuclear

Figure 23. The phenogram produced by assuming that the Spearman rank-order correlation coefficients for pairwise comparisons of dinucleotide representation constitute a similarity matrix. The phenogram was produced using a clustering method based on Euclidean distance (Wilkinson 1991). Sequence labels for mtDNAs are defined in Table 4. Sequences from nuclear genomes include SCERN (*Saccharomyces cerevisiae*, 78,295 nts of chromosome III, accession No. X59720), CELEN (*Caenorhabditis elegans* 13,794 nts of a homeobox DNA binding gene region, accession No. L15201), DMELN (*Drosophila melanogaster* 16,019 nts of the abdominal-B gene region, accession No. L07835) and HSAPN (*Homo sapiens* 16,569 nts of the retinoblastoma gene region, accession No. L11910). The prokaryote genomes examined included MCAP (*Mycoplasma capricolum* 12,971 nts of the ribosomal protein gene cluster; accession No. X06414), ECOL (*Escherichia coli*, ECO110K and ECOUW85U, 111,401 and 91,408 nts, respectively) and RCAP (*Rhodobacter capsulatus* 45,959 nts of the photosynthetic gene cluster, accession No. Z11165).

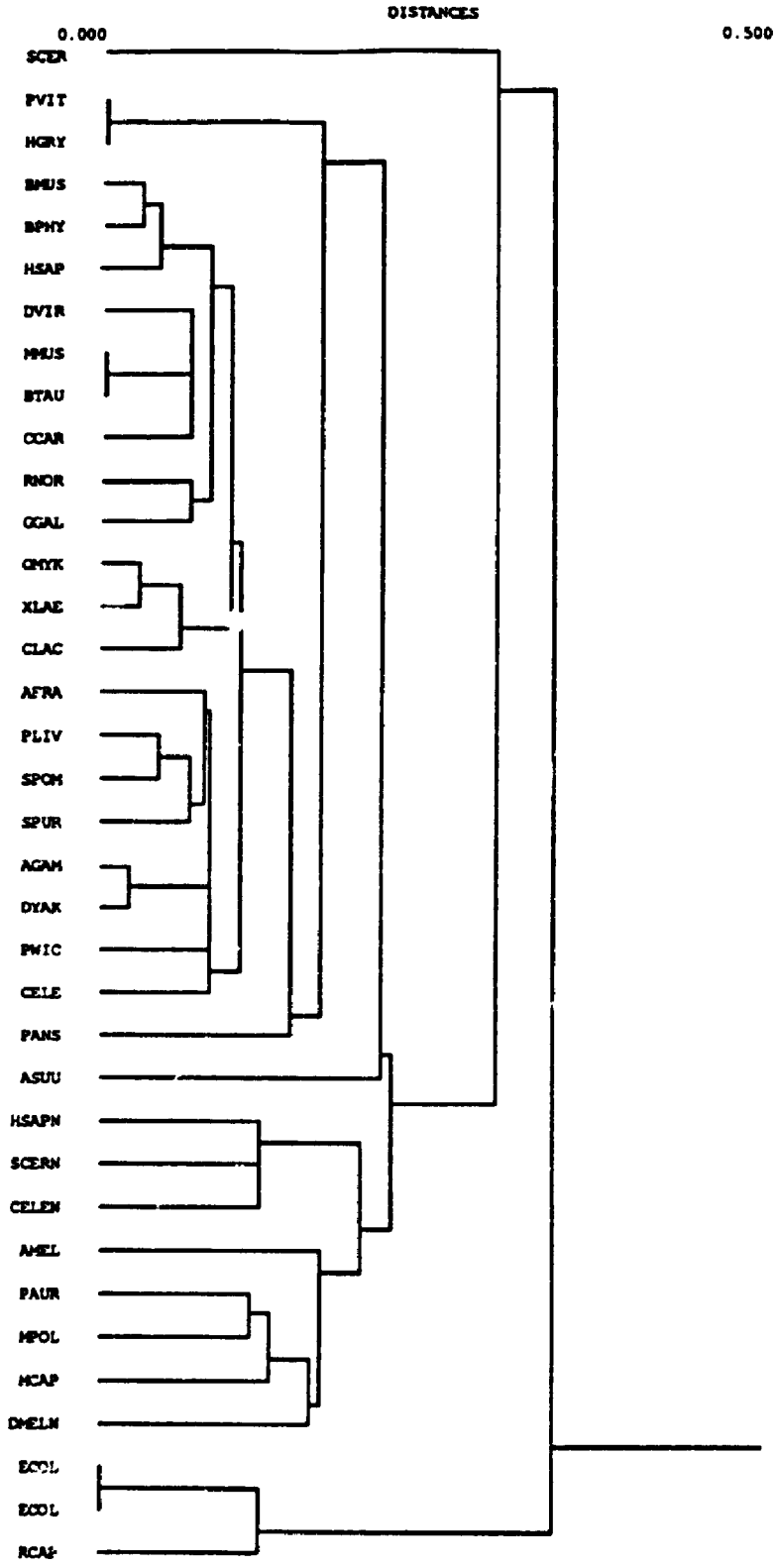
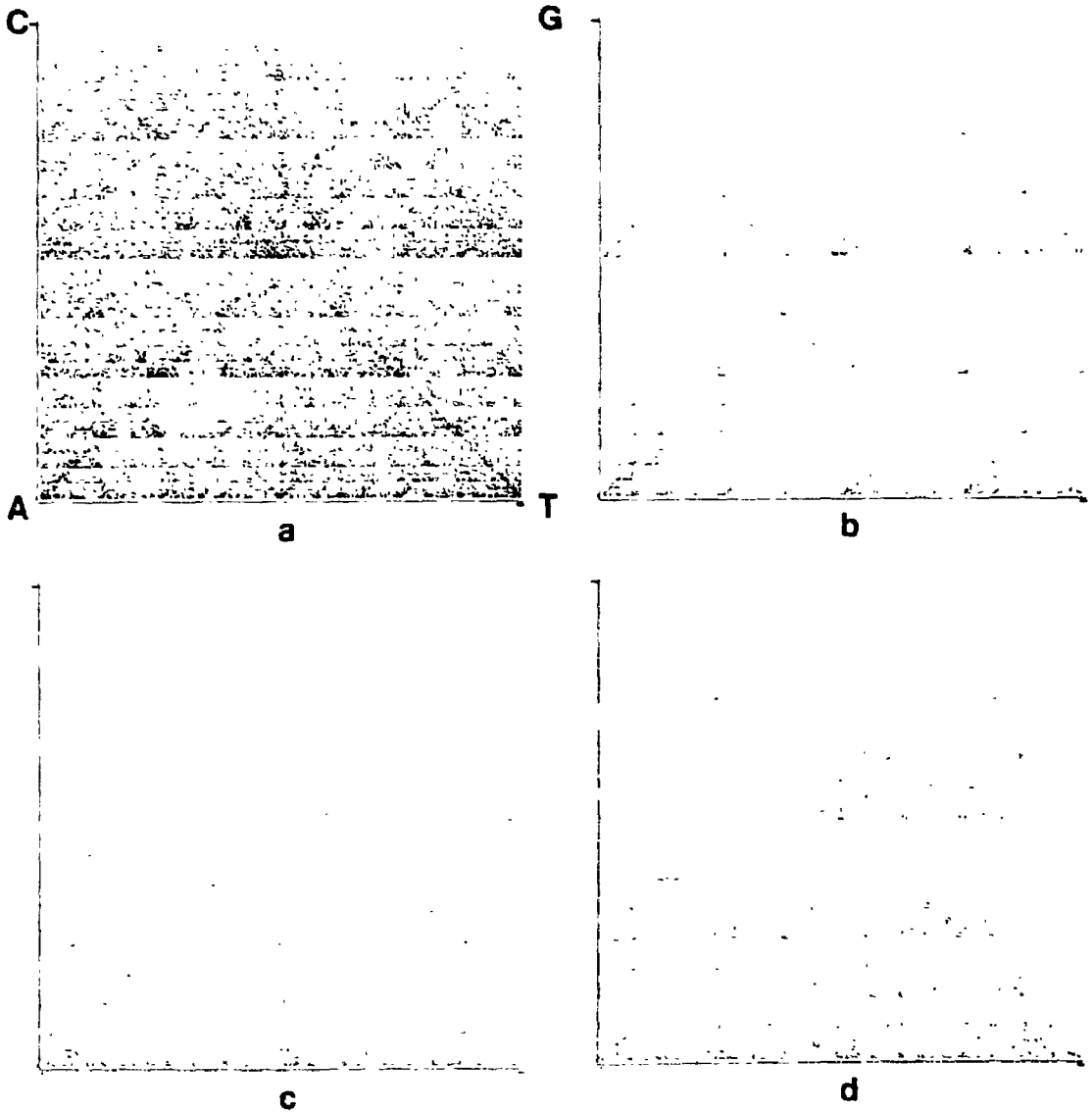

DISTANCES
0.000
0.500
SCER
PVIT
HGRY
BMUS
BPNY
HSAP
DVIR
MMUS
BTAU
CCAR
RNOR
OGAL
OMYK
XLAE
CLAC
AFRA
PLIV
SPOM
SPUR
AGAM
DYAK
PWIC
CELE
PANS
ASUU
HSAPN
SCERN
CELEN
AMEL
PAUR
MPOL
MCAP
DMELN
ECOL
ECOL
RCAS

Figure 24. Chaos patterns of the DNA sequence organization representative of the nuclear genomes of a) *Saccharomyces cerevisiae* (78,295 nts of chromosome III, accession No. X59720); b) *Caenorhabditis elegans* (13,794 nts of a homeobox DNA binding gene region, accession No. L15201); c) *Drosophila melanogaster* (16,019 nts of the abdominal-B gene region, accession No. L07835); d) *Homo sapiens* (16,569 nts of the retinoblastoma gene region, accession No. L11910).


C
A
a
G
T
b
c
d

sequences of these species. Short-sequence composition is more similar among diverse mitochondrial genomes than it is between mitochondrial and nuclear genomes of the same species.

The sequence organization of a region of the *Mycoplasma capricolum* genome has a unique chaos pattern (Figure 25) but a dinucleotide and trinucleotide composition like that of the mtDNAs of *P. aurelia*, *M. polymorpha*, and *A. melifera* and the nuclear genome of *D. melanogaster* (Figure 23). The chaos patterns and short-sequence representation for representative regions of *E. coli* and *R. capsulatus* genomes differed from each other and from the various types of chaos patterns generated by the 28 mitochondrial genomes (Figure 23).

3.2.10 Short-Sequence Representation in Viral Genomes

Chaos patterns generated for viral genomes had a fairly uniform distribution of data points except for the viral genomes capable of integration within a vertebrate nuclear genome which generated chaos patterns similar to regions of the host genome. The "double-scoop" pattern was evident in these chaos plots (Figure 26). Dinucleotide representation in viral genomes capable of integration in a vertebrate genome was similar to that of the host genome ($p<0.05$) and unlike that of viral genomes replicating in the host cytoplasm or nucleus without integration into the host genome ($p>0.05$).

Figure 25. Chaos representation of the ribosomal protein gene cluster of *Mycoplasma capricolum* (12,971 nts; accession No. X06414).

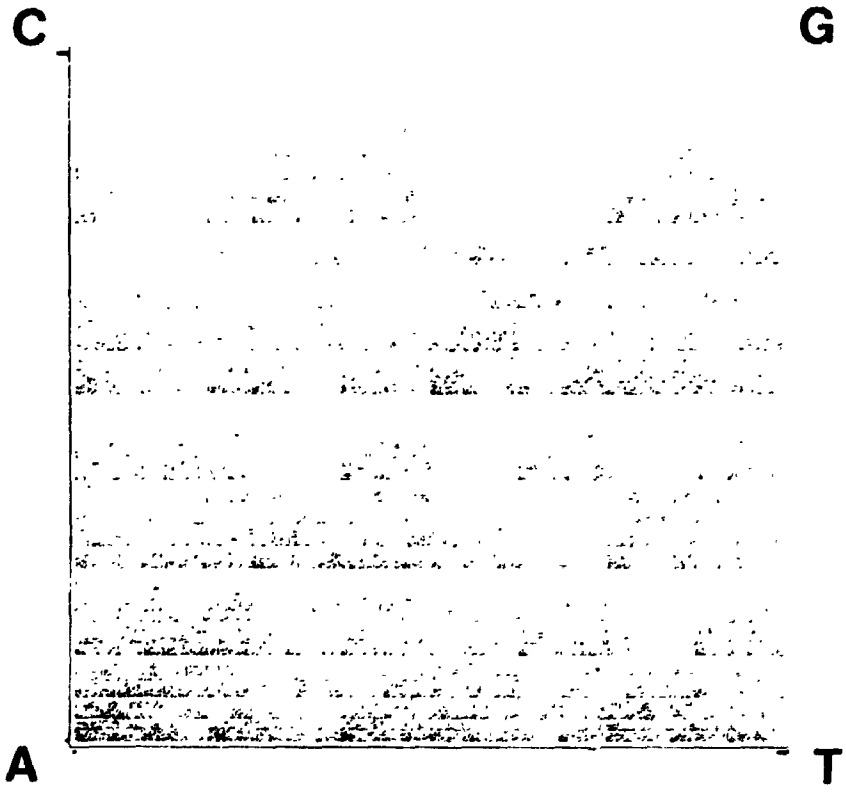
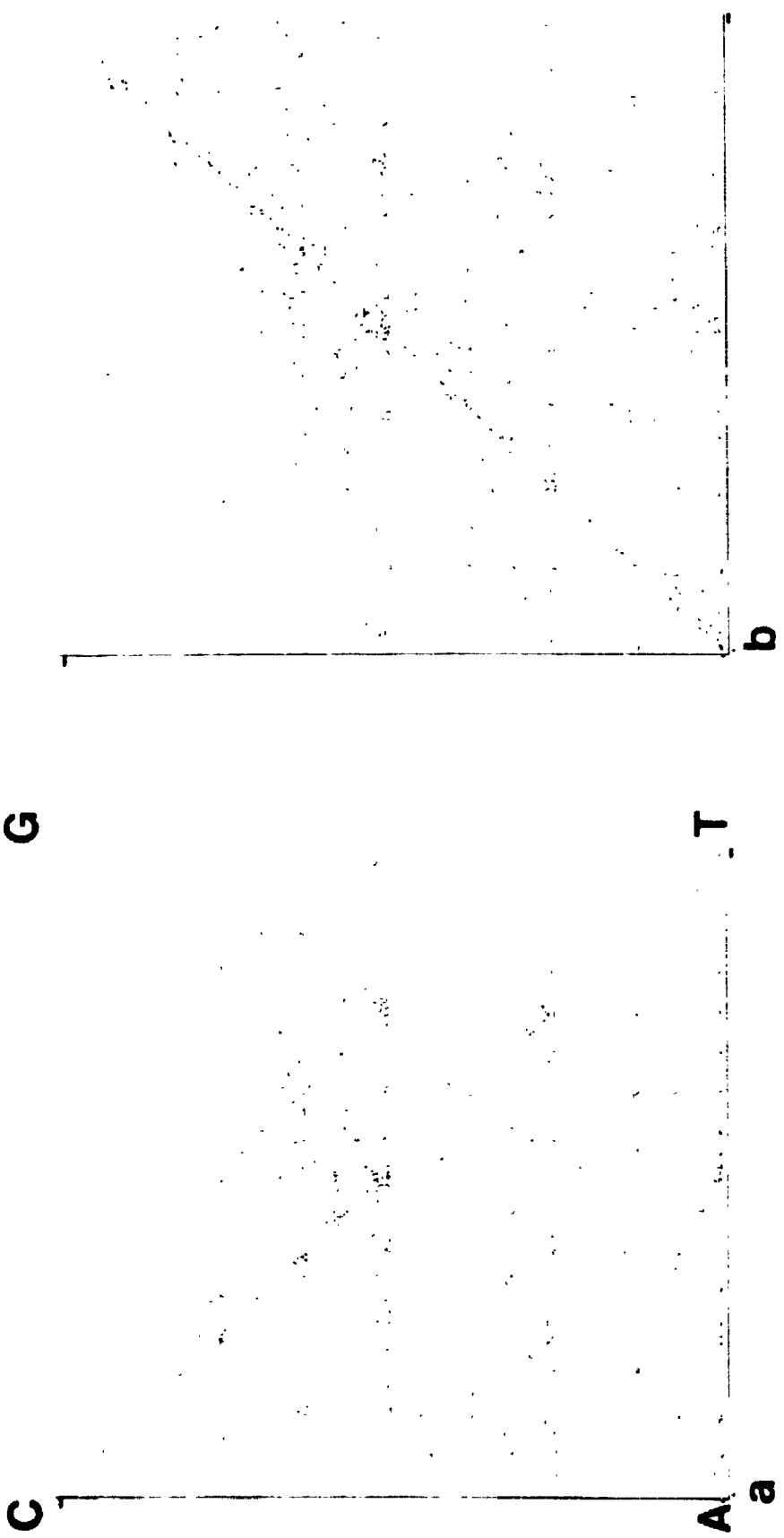

C
G
A
T

Figure 26. Chaos representation of the DNA sequence organization for two viral genomes capable of integration into a primate nuclear genome. The chaos patterns were generated by the a) human and b) simian immunodeficiency viruses (9,793 and 10,277 nts, respectively).


G
T
C
A
b
a

Chapter 4

DISCUSSION

It is through the examination of genetic variation that insight into the dynamics of evolution is gained. Estimates provided by examination of phenotypes and a partial analysis of DNA sequences are of limited value. Measurement of the extent of genetic variation is essential to elucidate its origin and maintenance. The best measure is obtained from the direct examination of nucleotide sequence diversity. DNA sequencing provides high resolution and simple interpretation by measuring the total extent of genetic variation and describing the pattern of nucleotide sequence diversity and organization. Sequence determination of any region of a genome for any species is now possible using the polymerase chain reaction. Knowledge of both the pattern and extent of nucleotide sequence diversity is essential to refine theories of the origin, pattern and maintenance of genetic variation that were based on examination of phenotypic differences.

A direct examination of nucleotide sequence differences has revealed a large number of previously hidden substitutions. Kreitman (1983) compared 11 *Adh* genes from five geographically distinct populations of *Drosophila melanogaster* and revealed the presence of 42 silent substitutions in exons and introns but only one amino acid replacement, providing strong evidence that most amino acid replacement mutations in this gene have been selectively deleterious and that this gene has a high level of silent nucleotide variation within a species. It is not known if the level of purifying selection and rate of silent substitution are representative of other genes nor how much of this variation occurs in single

populations. The examination of *Drosophila Adh* genes has raised several questions that can be addressed only by comparison of additional sequences.

There is limited knowledge of the extent and nature of the diversity in nucleotide sequence information in human populations. This diversity is defined as the number of nucleotide differences per site between two randomly chosen sequences from a population (Li and Sadler, 1991). The highest degree occurs at fourfold degenerate sites and is 0.11% in the human genome (Li and Sadler, 1991), much less than that observed in *Drosophila* (2.2%; Aquadro, 1991; Li and Sadler, 1991). The nature of the sequence data used by Li and Sadler (1991) has sampling biases both for loci of medical importance and for Caucasian North Americans and it is not known if it is representative of the entire human genome and of different human populations. Direct examination of numerous and functionally diverse loci in different human populations is necessary to obtain a representative measure of nucleotide sequence diversity in humans.

Descriptions of the origin and maintenance of genetic variation generally have not accounted for the constraints placed upon nucleotide sequence organization which are not associated with gene-specific features, but are related to the necessity of a higher-order sequence organization. The existence of any genome-specific constraints to sequence organization would be expected to affect the level and nature of nucleotide sequence diversity. However, DNA sequences may have a fractal nature which is defined as a global structure to nucleotide sequence organization (Tsonis et al., 1993) and appears to differ in different species; any genome-type specificity has not been defined (Jeffrey, 1990; Rogerson, 1991; Voss, 1993a b). Few determinants of a macrolevel of organization of DNA sequences are known (Burge et al., 1992; Karlin and Brendel, 1993). Given our limited knowledge of the global structure of DNA

sequences, its existence has not been integrated with the examination of the origin and maintenance of nucleotide sequence diversity in any genome.

The present study addressed these concerns with three general objectives: 1) to characterize the nucleotide sequence variation at two regions of the human nuclear genome that were known to differ in the nature and extent of polymorphism and in two populations with different evolutionary histories and population structures; 2) to describe the features of the primary organization of nucleotide sequences and identify the global structure of DNA; 3) to define the genome-type specificity of the global structure of nucleotide sequences and identify the biological determinants of the global structure of DNA.

4.1 Nucleotide Sequence Variation at a Polymorphic Locus in Two Different Human Populations

The direct sequencing of the polymorphic third exon of the *Adh2* locus in two different human populations did not reveal any intra-allelic variation in a total of 19,110 nucleotides examined. Instead the two predominant alleles at this locus, the "typical" *B1* and the "atypical" *B2* alleles were observed. These observations are consistent with the selection regime affecting the *Adh2* gene product and the structures of the two human populations.

A lack of intra-allelic variation is not unexpected given the functional constraints and thus the highly conserved nature of *Adh* genes. Eleven clones of the entire *Adh* genes from five natural populations of *Drosophila* contained 43 previously undetected polymorphisms that are considered intra-allelic (Kreitman, 1983). Only one of these polymorphisms resulted in an amino acid change, the change responsible for the known electrophoretic variants present in all natural populations of *Drosophila*, the "fast "allele, *Adh-f* and the "slow" allele *Adh-s*.

These data imply that most amino acid changes in *Adh* would be selectively deleterious.

The *Adh* gene product is a general "housekeeping" enzyme of critical importance in the metabolism of alcohol. Human *Adh1*, *Adh2* and *Adh3* genes encode amino acid sequences with 93 to 96% identity (Ikuta et al., 1985; 1986; von Bahr-Lindstrom et al., 1986). Mammalian ADH enzymes, in comparison with ADHs of maize and yeast species, have an amino acid sequence similarity of 50% and 20%, respectively (Eklund et al., 1976; Jornvall et al., 1987). In this study, nucleotide sequence diversity was examined in the *Adh2* third exon which encodes the catalytic domain of the enzyme (Hurley et al., 1991). This exon contains nucleotide sites encoding amino acid residues critical for substrate binding and other residues essential for maintaining the three-dimensional structure of the active site of the enzyme (Eklund et al., 1976; Hurley et al., 1991). Certain of these amino acid sites are invariant across numerous diverse species (Sun and Plapp, 1992), so that nucleotide substitutions at the third exon of *Adh2* would not be expected at a large number of sites as they would likely alter amino acids subject to functional constraints.

In an examination of 49 human genes Li and Sadler (1991) observed no sequence differences between two randomly chosen sequences for 33 human genes. The 16 remaining loci contained only one to four nucleotide differences. *Adh* loci were not among the 49 loci examined but other general "housekeeping" enzymes were and little to no nucleotide sequence diversity was observed in the sequences analyzed.

Substitutions at synonymous codon sites are not affected by the selective influences upon the *Adh2* gene product. However, synonymous codon usage is biased and genome-specific (Grantham et al., 1980). Synonymous codon usage biases in prokaryotes and lower eukaryotes are associated with tRNA content

and gene expression, particularly in highly abundant proteins (Grantham et al., 1981; Ikemura, 1985; Lloyd and Sharp, 1992). In higher eukaryotes synonymous codon usage is associated with the nucleotide composition of the region in which the gene is located (Sharp and Matasi, 1994; Zhang and Chou, 1993; D'Onofrio et al., 1991; Lipman and Wilbur, 1983). There is a selective constraint upon nucleotide sequence organization at synonymous sites that is associated with the organization of the nucleotide sequence in the region of the genome containing the gene of interest. Thus in genes there are selective constraints even upon synonymous sites and these constraints are species-type specific in nature and relate to sequence composition.

In this study there was a sequence difference detected at the third exon of the *Adh2* locus. It is not unexpected that the substitution occurs at a CG dinucleotide, a known "hot spot" for mutation (Beutler et al., 1989; Bird, 1980; Ehrlich and Wang, 1981). Transitions at CG dinucleotides are elevated approximately 24-fold relative to transitions at non-CpG dinucleotides (Sommer 1992). The dinucleotide mutation rate produces a bias against G and C nucleotides that has been hypothesized in the case of the human factor IX gene to be sufficient to maintain the G+C content at its evolutionarily conserved level of 40% (Sommer 1992).

The sequence difference observed in the third exon of the *Adh2* locus represents the occurrence of the two alleles *β1* and *β2* that predominate in human populations. The existence of the two alleles in different populations is not hypothesized to be due to recurrent mutation but to the single occurrence of the mutation and a founder effect in multiple populations (Stamatoyannopoulos et al., 1975). The distribution of the *β2* allele is >70% in Japanese populations and <10% in English populations (Stamatoyannopoulos et al., 1975) so that the single occurrence of the *β2* allele in Southwestern Ontario is consistent with the

European ancestry of these people. The individuals of Southwestern Ontario represent a relatively large population derived primarily from Western European countries.

The Dogrib population differs from that of Southwestern Ontario population in many respects. All Dogrib individuals were homozygous for the *B1* allele of the *Adh2* locus. Archeological evidence on ancestral American Indian migrations is ambiguous but analyses of linguistic diversity and mitochondrial DNA have placed them at 5,200 to 10,500 years before present and indicate only one or two separate waves of migration (Wallace et al., 1985; Szathmary, 1993; Torroni et al., 1992; Wallace and Torroni, 1992). Traditional anthropological investigations have confirmed that American Indians came from Asia, probably crossing the Bering land bridge when it was exposed during an episode of glaciation (Crawford and Enisco, 1983). The nearest living Asian relatives of present day Dogrib individuals are Siberians and little if anything is known of genetic similarities and differences between Dogrib and Siberian individuals (Posukh et al., 1990). The nature of the alleles at the *Adh2* locus in the Siberians is not known but would be of interest to this study. Considering their Asian ancestry it might be assumed that the *B2* allele would predominate in the Dogrib population as it does in the more closely-related Japanese populations.

The structure of the Dogrib population is unlike that of the Europeans or the Japanese and is the result of perhaps two founding events with a limited number of founders (Crawford, 1992). It has been subjected to severe bottlenecks associated with disease related to colonization by Western Europeans (Crawford, 1992). The population is small and geographically isolated. European admixture is limited (Szathmary, 1978). The population is subject to genetic drift and the effect is the eventual fixation of one allele at a particular locus. It is possible that the original founding population contained only

the *β1* allele or alternatively both *β1* and *β2* alleles with frequencies similar to that seen in present-day Japanese populations. Genetic drift may have been a significant force in eliminating *β2* allele from a polymorphic founding population. The specific absence of the *β2* allele could also be consistent with the existence of selective pressure against the "atypical" *β2* allele. The *β2* allele does produce a "super active" form of the ADH enzyme that would be expected to cause a build up of the highly toxic metabolite, acetaldehyde, an obvious deleterious result (Bosron and Li, 1988).

4.2 Nucleotide Sequence Variation at a Highly Polymorphic Locus in Two Different Human Populations

The direct sequencing of the highly polymorphic second exon of the *HLA-DQβ1* locus in two different human populations did not reveal any intra-allelic differences in a total of 20,094 nucleotides examined. Instead the sequence information determined at the second exon of this locus could be attributed to known alleles at this locus. The Southwestern Ontario and Dogrib populations differ in the number and type of alleles at this locus. The average heterozygosity at this locus in the individuals of Southwestern Ontario was 0.78. The individuals of Southwestern Ontario contained 12 alleles and all five allele subtypes. The most frequent alleles were 201/202 and 602, the two alleles with the greatest nucleotide diversity at the second exon. The Dogrib populations had an average heterozygosity of 0.82 but contained four fewer alleles and two fewer allele subtypes than the individuals of Southwestern Ontario. These observations are consistent with the evolutionary histories and population structures of the two human populations and a selection regime affecting the *HLA-DQβ1* gene product that differs from that affecting the *Adh2* gene product.

The fewer alleles and allele subtypes in the Dogrib population would be expected given the structure and evolutionary history of the population and the significance of genetic drift as a force affecting the maintenance of genetic variation in this population. The small size of the population indicates the influence of genetic drift in reducing the level of genetic variation. An opposing force affecting the level of genetic variation is the selection regime. Although genetic variation at *HLA* loci is critical to the survival of the population, there is no evidence that it is maintained by an increased mutation rate (Hayashida and Miyata, 1983; Klein, 1986) but rather by an overdominant selection regime (Figueroa et al., 1988; Fan et al., 1989; Gaur et al., 1992; Lawlor et al., 1988). The fewer alleles and allele subtypes and the increased heterozygosity in the Dogrib population in comparison with the Southwestern Ontario population is consistent with the coexistence of genetic drift and balancing selection as forces affecting the maintenance of genetic variation.

The lack of intra-allelic variation at the second exon of the *HLA-DQB1* locus is consistent with evidence (Hayashida and Miyata, 1983; Klein, 1986) that genetic variation is not maintained by an increased mutation rate. It is reasonable to assume that there is a limit to the extent of nucleotide diversity that can occur without disrupting the function of the *HLA-DQB1* molecule. Thus, the type and number of mutations, even within the antigen recognition domain of the *HLA-DQB1* molecule, are still limited by functional constraints.

The lack of intra-allelic variation at exons of the polymorphic and the highly polymorphic loci argues for no difference in mutation rates as a significant force in the maintenance of the observed levels of genetic variation. The different levels of genetic variation are hypothesized to be due to different selection regimes acting upon the gene products of these loci. The different levels of genetic variation between the two human populations are hypothesized to be the

result of differences in population size and ancestry. Genetic drift is proposed as a significant force affecting the heterozygosity within the Dogrib population.

4.3 Evidence for the Global Structure of Nucleotide Sequence Organization

This investigation defines global DNA sequence organization as a higher order structure that is depicted in chaos patterns of DNA sequences and appears to be determined largely by single nucleotide frequencies and nearest-neighbor associations of nucleotides. It is similar in both translatable and non-translatable elements of a DNA sequence. Global structure is similar on both strands of the DNA double helix and is scale and region invariant. Its specific features appear to be related to the genome-type in which the DNA sequence occurs.

Different regions of the same genome have a common structure, defined as similar major features in data point distribution in two-dimensional chaos plots and similar short-sequence representation (i.e., single nucleotide frequencies, dinucleotide and trinucleotide representation). Sequence organization was examined in more than one region of the *Escherichia coli* genome and in different regions of the nuclear genomes of *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Homo sapiens*. The short-sequence composition of each DNA sequence and its complementary strand were significantly correlated as has been reported previously in examination of tetranucleotide compositions of sequences from diverse species (Rogerson, 1991). In one particular genome, that of the yeast *S. cerevisiae*, the characteristic sequence organization was seen in regions of four different chromosomes and in adjacent 15 and 30 kb regions of chromosome III. The nucleotide sequence organization in the nuclear genome of the yeast *S. cerevisiae* (a total of 432 kb examined) has a common structure

regardless of the size of the DNA sequence examined, its location within a chromosome sequence and the specific chromosome examined. Similar conclusions are reached from analysis of numerous sequences from different chromosomes in *C. elegans* (a total of 1,008 kb) and human nuclear (965 kb) genomes. Sequence structure was also similar among different regions of the mitochondrial genomes of *Marchantia polymorpha* and humans. These observations are consistent with the hypothesis that a higher-order organization exists in DNA sequences. Sequence structure is said to be global as it is similar on both strands of the duplex DNA and is independent of the length of the nucleotide sequence and the location of the sequence within the genome.

The global nature of sequence structure also includes an independence from the function of the DNA sequence. Translatable or non-translatable elements have a similar organization of nucleotides in the chromosome III sequence of *S. cerevisiae* and the neurofibromatosis gene and globin gene cluster of humans.

The chaos patterns and contour plots of unrelated genes from a single species appear to be more similar than those of a relatively conserved gene from unrelated species. For example, in humans the *Adh* and globin gene families and the *HLA-DQβ1* gene have similar chaos patterns, contour plots and biases in short-sequence composition. In contrast, the highly conserved *Adh* genes of phylogenetically diverse species have different chaos patterns and short-sequence compositions. Sequence structure in Adh genes of different species were similar only among closely-related species. These observations are consistent with a structure that is not gene-specific but genome-type specific. Structures depicted in two-dimensional chaos patterns and roughly quantified in terms of biases in short-sequence composition portray the existence of a highly

conserved and species-type specific pattern that appears to affect both DNA strands and underlie all other genetic information.

Not all features of nucleotide sequence organization have a global nature. Three-dimensional chaos patterns portray the frequency and nature of highly repeated longer subsequences and have a region-specific pattern of occurrence in the *S. cerevisiae* chromosome III. These sequence structures are not measured in the examination of two-dimensional chaos patterns and the analysis of short sequence composition and were not examined further in the present study. The chaos method has the potential to identify all repeating subsequences, including imperfect repeats, but algorithms for their identification have not been developed. It is more expedient to determine the nature and location of all repeats using other analysis techniques (Leung et al., 1991). Thus only a limited analysis of subsequence repetition was undertaken in the present study. As well, repetition of longer subsequences did not appear to be a global feature of biological organization, at least in the *S. cerevisiae* genome. Future analyses could determine if this is the case in other species' genomes.

4.4 Evidence for a Global DNA Sequence Structure that is Genome-Type Specific

The genome-type specific nature of global DNA structure, although identified previously, remains poorly defined (Jeffrey, 1990; Rogerson, 1991; Voss, 1993 b; Burge et al., 1992; Karlin and Brendel, 1993). In the present study the chaos patterns of unrelated genes from a single species appear to be more similar than chaos patterns for a relatively conserved gene sequence from unrelated species. Two different multigene families, *Adh* and globin in the human genome, have similar global sequence structures and the same gene family, *Adh*,

known to be highly conserved in numerous distantly-related species, has greater variability in global structure in different species. Similar sequence structures were only identified among the *Adh* genes of more closely-related species (i.e., among mammals). These observations argue for the existence of species-type specific influences upon global sequence structure.

Chaos representation of 56 large DNA sequences (> 36,000 nts) from 10 phylogenetically diverse species revealed six different chaos patterns or global structures. *Rhodobacter capsulatus*, *E. coli*, *S. cerevisiae*, *C. elegans*, *Drosophila melanogaster* and mammalian genomes each have a unique chaos pattern and global structure. Data point distribution in chaos patterns was quantified partially by examination of single nucleotide frequencies and biases in di- and trinucleotide frequencies. Spearman rank correlation coefficients for all pairwise comparisons of short-sequence representation were considered as similarity coefficients. A similarity matrix clustered most sequences in a phenogram in a pattern consistent with the genome-type specificity suggested from the visual comparison of chaos patterns. The different global sequence structures in different prokaryote DNAs and eukaryote nuclear genomes suggest the existence of constraints upon nucleotide sequence organization which are species-type specific.

It should be noted that genome-type specificity remains poorly defined due to the limited number of different species' genomes represented in the data set of this study. To refine the definition of genome-type specificity in global sequence organization it is necessary to examine numerous other sequences from a greater variety of species. Perhaps subtle differences between more-closely related species and the determinants of the structures can then be identified.

4.5 Evolution of a Global Sequence Organization That is Genome-Type Specific

The monophyletic and endosymbiotic origin of mitochondria permits examination of the influence of the ancestral species-type, evolutionary time and host species-type on the evolution of global structure. Global DNA sequence organization was characterized for the complete mitochondrial genomes of 28 phylogenetically diverse species using chaos representation and measures of short-sequence representation. This approach revealed 7 different patterns of global sequence organization in mitochondrial genomes of 1) protozoan and plant, 2) yeast and fungus, 3) alga, 4) nematode, 5) echinoderm, 6) insect and 7) vertebrate species. A cluster diagram based on dinucleotide representation for select prokaryote, eukaryote and mitochondrial sequences revealed 1) similarities in the sequence organization of mitochondrial genomes of more closely-related host species, 2) similarities in the sequence organization of mitochondrial genomes and that of *Mycoplasma capricolum,* the proposed prokaryote progenitor of present day mitochondrial genomes (Cardon et al., 1994) and 3) more similarity among mitochondrial genomes of diverse species than between the mitochondrial genomes and the nuclear genome of the same or closely related species.

Examination of mitochondrial DNA sequences in this analysis has permitted a further characterization of species-type specificity, at least in terms of the close and distant relationships among the 28 species for which complete mitochondrial sequences are available. Mitochondrial genomes were also selected as a biological model for monitoring the relationship between species-type and DNA sequence organization due to their unique ancestry. It is assumed that mitochondrial sequence organization is a product of the multiple influences of

a monophyletic and prokaryote origin and symbiosis with the constraints imposed upon the host nuclear sequence organization. The endosymbiotic relationship with the host cell could maintain features specific to the prokaryote ancestry and unique codon usage patterns and cellular location of the mitochondrial genomes. It is apparent from the chaos patterns generated in this investigation that mitochondrial genomes differ in global structure from nuclear genomes of the same species. Mitochondrial sequence organization also differs from the presumed ancestral sequence, which itself may have altered with evolutionary time. Also, the eukaryote host and evolutionary time are expected to have altered the mitochondrial sequence organization from the ancestral form. It has been observed that certain features of sequence organization are common to all mitochondrial genomes (mitochondrial-specific). Certain other features are similar among closely related-species and different among distantly-related species (i.e., host species-type specific). Certain features are similar in the proposed prokaryote ancestor of present day mitochondrial genomes, *Mycoplasma capricolum* genome (specific to the species of origin). Such results are compatible with two hypotheses: 1) that insufficient evolutionary time has passed for mitochondrial genomes to have been influenced by the same host-specific constraints as nuclear genomes or 2) different and yet species-type specific constraints account for the global structures of nuclear and mitochondrial genomes. The global structure of nucleotide sequences does not appear to be simply species-type specific but since it differs between the mitochondrial and nuclear genomes of the same species it is more properly defined as genome-type specific.

Why the global sequence organization of the mitochondrial and nuclear genomes is species-type specific but different from one another is not known. It may be that the nuclear and mitochondrial genomes of a cell, although differently

located in the cell, are under same/similar evolutionary constraints for higher-order organization. In such a hypothesis, the observed differences in sequence organization represent a transitional state and, given time, the two sequences will attain a similar (global sequence) organization where the time required could be species-type dependent. In a second hypothesis, mitochondrial and nuclear genomes could accommodate differently to the same/similar global sequence organization for mitochondrial and nuclear genomes. This explanation suggests mitochondrial and nuclear genomes could accommodate the same/similar species-type specific constraints on DNA sequentiality differently given their differences in subcellular location, mode of replication, transmission, codon usage and genome organization. The differences in the global sequence organization of the mitochondrial and nuclear genomes of a species are representative of the many differences in mitochondrial and nuclear genomes such as their different cellular locations and codon usage patterns. The second explanation suggests that the observed differences may represent relatively stable species-type specific mitochondrial and nuclear features of higher-order sequence organization. Identifying which of the proposed mechanisms account for the differences between nuclear and mitochondrial genomes could form the basis for identifying particular determinant(s) of global sequence organization describing the evolution of nucleotide composition and order in DNA sequences.

The genome specificity of the chaos pattern at the present level of our understanding could be attributed to differences in mutation rates associated with a particular base, dinucleotide, etc. One example of this is the relatively high mutation rate associated with CG dinucleotides to TG dinucleotides in vertebrates due primarily to the methylation specificity of CG (Russell et al. 1976; Coulondre et al., 1978; Bird 1980; Ehrlich and Wang 1981). Sequence-specific methylation and mutation rates in other genome-types remains poorly understood.

Explanation of some of these observations may be critical for complete elucidation of the patterns observed in chaos patterns. The global structure of all mitochondrial genomes is characterized by an over-representation of A and T, a deficiency of CG dinucleotides and an over-abundance of CC/GG and AG/CT dinucleotides. It should be noted that the global structure of nuclear genomes is determined primarily by biases in dinucleotide composition, while the determinants of the global structure of mitochondrial genomes are biases in single nucleotide and dinucleotide composition. There is a need to examine further genome-specific mutation rates for single nucleotides and oligonucleotides and/or genome-specific constraints upon sequence organization to explain the differences in global structure between mitochondrial and nuclear genomes.

Viral genomes capable of integration within the host genome represent a means to test indirectly the hypothesis that genome-type specific influences determine global sequence organization. The global structures of different viral genomes are consistent with this hypothesis. Only those viral genomes requiring integration into the human genome for replication have a global structure similar to that seen in the human nuclear genome (also seen by Jeffrey, 1990). Specifically the under-representation of CG dinucleotides is present in these integrative viral genomes. Other viral genomes replicating in the cytoplasm or the nuclear matrix but not requiring integration into the host genome for replication have different global structures. These observations also point out that the genome-specific constraints upon global sequence structure require integration into the host genome. These analyses were limited by sequence availability to the examination of global sequence structure in mammals. In order to test the pervasiveness of this observation, future analyses should examine other integrative viruses and host species pairs, in particular those host

genomes having a different global structure from that of mammals. The pattern and extent of mutation in transgenic sequences such as the prokaryote lactose operon sequences in a murine nuclear genome (Kohler et al., 1991; Myhr, 1991) could be examined to monitor directly the transition of global sequence structure from one genome-type to another.

4.6 Summary

1. There is no evidence for higher substitution rates for a gene with extensive polymorphism (*HLA-DQβ1*) as compared to a gene with relatively low polymorphism (*Adh2*) in human populations. No intra-allelic nucleotide sequence variation was observed following DNA sequence determination of roughly 39,000 base pairs at these loci. A lack of novel nucleotide diversity argues for a major role for selective forces in maintaining a gene-specific degree of polymorphism in most populations.

2. In the evolutionary history of a population, admixture and stochastic processes such as founder effects and bottlenecks are important determinants of the observed genetic variation in human populations. These processes can account for the similarities and differences between the relatively homogeneous Dogrib population and the heterogeneous Southwestern Ontario population at the *Adh2* and *HLA-DQβ1* loci.

The difference in the number of alleles and heterozygosity at the two loci however could not be explained by stochastic processes alone. These differences are best explained by a hypothesis involving balancing selection and heterozygote advantage at the *HLA-DQβ1* locus in both populations.

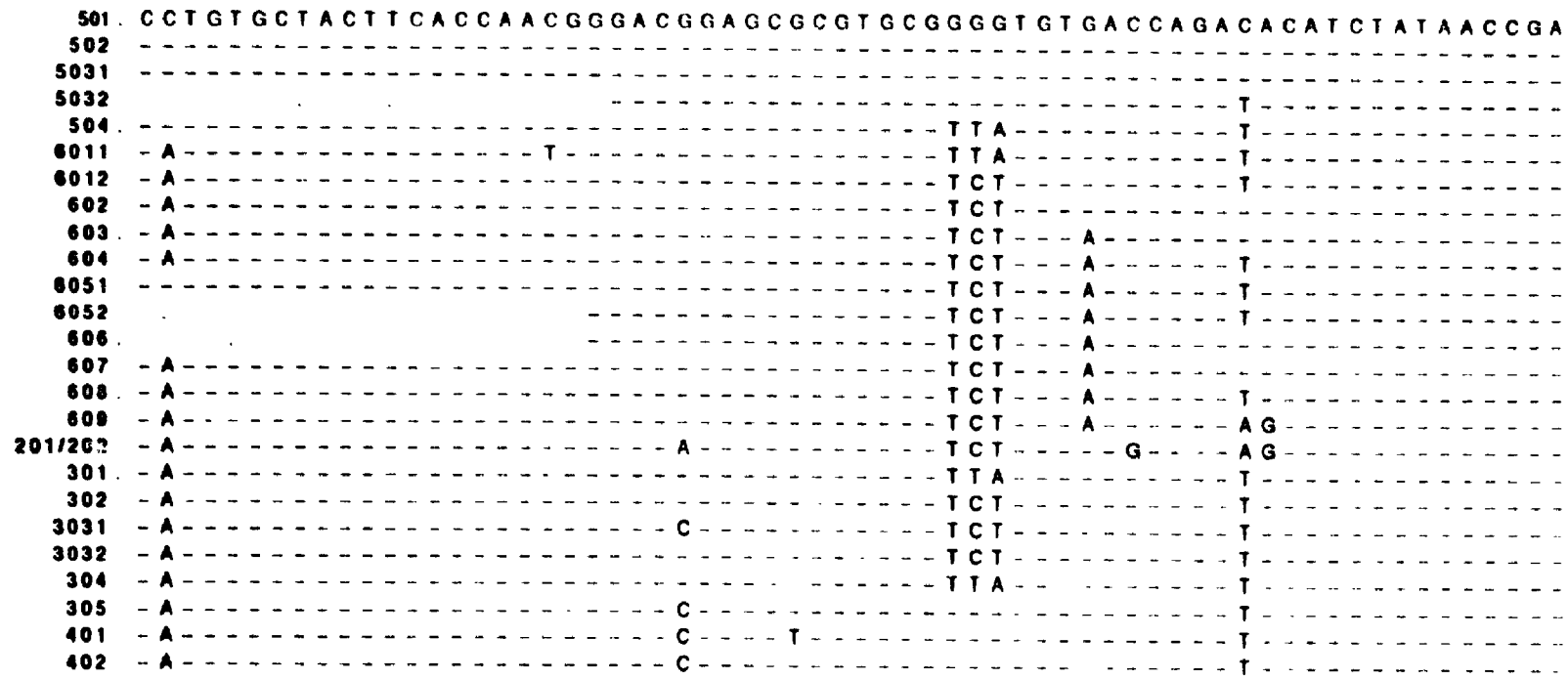
3. Examination of substitutional events at the *Adh2* and *HLA-DQβ1* loci argue for a nonrandom nature of mutational events and emphasize the significance of nearest-neighbor nucleotide associations. Dinucleotide representation is a major determinant of sequence organization in prokaryote sequences and the nuclear and mitochondrial genomes of eukaryotes.

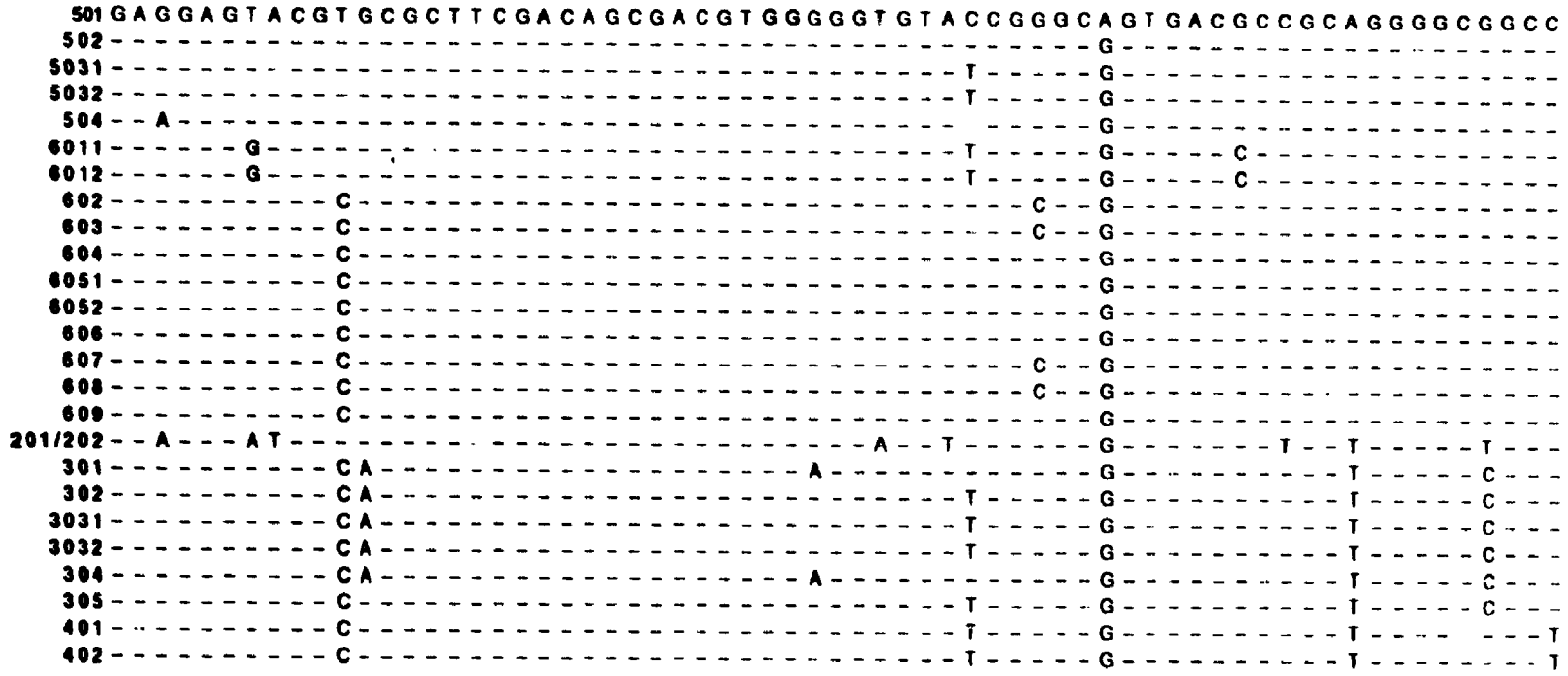
An extensive analysis of a variety of large continuous DNA sequences from sequence databanks using the novel chaos representation of DNA sequences suggests that short-sequence representation, in particular for dinucleotides appears to be genome-type specific rather than a gene-specific property. Such features would predict a global structure to substitutions associated with individual nucleotides in the DNA of a given species.

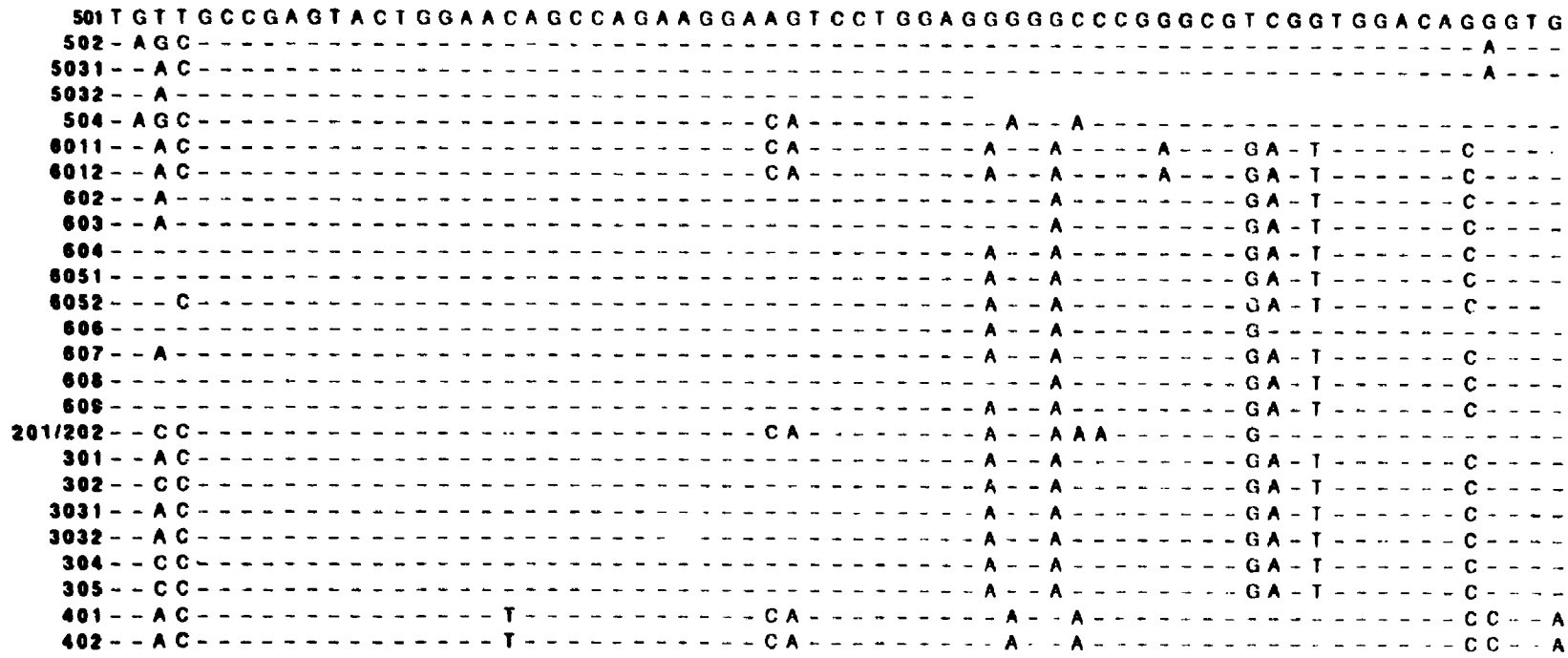
4. It was hypothesized that the genome-type specific global structure of a DNA is attained under the constraint of an unknown force(s) or factor(s). This hypothesis was evaluated by analysis of 28 mitochondrial sequences assuming a monophyletic origin for mitochondria. Extensive chaos and representational analysis revealed some similarities in the global sequence organization of mitochondrial and nuclear sequences of species. Viral genomes capable of integrating in the human nuclear genome showed greater similarity with the host nuclear sequence organization than did mitochondrial sequence of the same host.

5. The results and interpretations included in this thesis suggest that the evolution of DNA sequentiality of genes should be viewed in the dual context of constraints related to gene function and to the genome-type specificity of higher-order organization.

Appendix 1. An alignment of the nucleotide sequences of the 26 known alleles of the *HLA-DQB1* locus (for review see Bodmer, 1994). Nucleotides are identified at those positions where differences from allele 501 occur otherwise a dash is used. A period indicates sites for which the nucleotide sequence is not yet determined. The alignments for the entire exon sequence are contained on three pages.


501 CCTGTGCTACTTCACCAACGGGACGGAGCGCGTGCGGGGTGTGACCAGACACATCTATAACCGA
502
5031
5032
504
6011
6012
602
603
604
6051
6052
606
607
608
609
201/202
301
302
3031
3032
304
305
401
402


501 GAGGAGTACGTGCGCTTCGACAGCGACGTGGGGGGTGTACCGGGCAGTGACGCCGCAGGGGCGGCC
502
5031
5032
504
6011
6012
602
603
604
6051
6052
606
607
608
609
201/202
301
302
3031
3032
304
305
401
402


501 TGTTGCCGAGTACTGGAACAGCCAGAAGGAAGTCCTGGAGGGGGCCCGGGCGTCGGTGGACAGGGTG
502 -AGC----------A---
5031 --AC----------A---
5032 --A-----------
504 -AGC----------CA---------A--A----------
6011 --AC----------CA---------A--A----A---GA-T-------C----
6012 --AC----------CA---------A--A----A---GA-T-------C----
602 --A----------A---------GA-T-------C----
603 --A----------A---------GA-T-------C----
604 ----------A--A---------GA-T-------C----
6051 ----------A--A---------GA-T-------C----
6052 ---C----------A--A---------GA-T-------C---
606 ----------A--A---------G----------
607 --A----------A--A---------GA-T-------C----
608 ----------A---------GA-T-------C----
609 ----------A--A---------GA-T-------C----
201/202 --CC----------CA---------A--AAA-------G----------
301 --AC----------A--A---------GA-T-------C----
302 --CC----------A--A---------GA-T-------C----
3031 --AC----------A--A---------GA-T-------C----
3032 --AC----------A--A---------GA-T-------C----
304 --CC----------A--A---------GA-T-------C----
305 --CC----------A--A---------GA-T-------C----
401 --AC----------T----------CA---------A--A----------CC--A
402 --AC----------T----------CA---------A--A----------CC--A

Appendix 2

The Fortran 77 program used to generate chaos (x,y) coordinate data from nucleotide sequences.

1. Generation of (x,y) coordinates

```
      program xypairs
      character*1 la,lt,lc,lg,ua,ut,uc,ug,un,ln
      character string(50)*1
      integer case, base
      real x, y
      open(unit=5,file= infile ',status='old')
      open(unit=6,file='outfile',status='new')
      la='a'
      lt='t'
      lc='c'
      lg='g'
      ln='n'
      ua='A'
      ut='T'
      uc='C'
      ug='G'
      un='N'
      x=500.0
      y=500.0
      case=0
      base=0
      no let=50
10    continue
c
c     Reads a line of 50 characters
c
      read(5,20,end=999) (string(i),i=1,nolet)
20    format(50a1)
c
c     for each letter calculates the x,y values and base number
c
      do 50 i=1,nolet
         if (string(i) .eq. lc .or. string(i) .eq. uc)then
            base = 4
            x = (x+0.0)/2.0
            y = (y+1000.0)/2.0
         else if (string(i) .eq. lg .or. string(i) .eq. ug)then
            base = 3
            x = (x+1000.0)/2.0
            y = (y+1000.0)/2.0
         else if (string(i) .eq. la .or. string(i) .eq. ua) then
            base = 1
```

```
            x = (x+0)/2.0
            y = (y+0)/2.0
         else if (string(i) .eq. lt .or. string(i) .eq. ut) then
            base = 2
            x = (x+1000)/2.0
            y = (y+0)/2.0
         else if (string(i) .eq. ln .or. string(i) .eq. un) then
            go to 50
         else
            go to 50
         end if
         case = case+1
         write(6,40) case,base,x,y
40       format(1x,i6,1x,i1,f6.0,f6.0)
50    continue
      go to 10
999   stop
      end
```

2. Plotting of (x,y) coordinates

```
filename con1 "filename.ps";
symbol1 v=point;
axis1 major=none minor=none label=none length=7in order = (0 1000);
axis2 major=none minor=none label=none length=7.125 in order = (0 1000);
*goptions device=tek4010 gepilog='18'X rotate;

goptions device=ps300 gsfname=con1 gsfmode=replace gsfmode=replace
      gsflen=132 gprolog='2521'X nodisplay ftext=swissb htitle=0.5 htext=0.05
      gunit=in;
options linesize=80 nodate nonumber;
title 'filename';
data;
      drop cases bases ;
infile "[12010_3700.chaos]filename.chaos" ;
      input cases bases x y ;
proc gplot;
      plot y*x=1 /haxis=axis1 vaxis=axis2;
```

Appendix 3

Determination of a z value that represents the frequency of repetition of (x,y) coordinates in the chaos representation of DNA sequences.

1. Calculation of the frequency of repetition of (x,y) coordinates

```
options linesize=80;
title 'filename';
data;
        drop cases bases;
        infile "[12010_3700]filename.chaos" ;
        input cases bases x y '
proc sort;
        by x y;
proc means noprint;
        by x y;
        var x;
        output out=hum n=z;
proc sort data=hum(WHERE=(7>1));
        by z x y'
proc print data=hum(WHERE=(Z>1));
        var x y z;
```

2. Three-dimensional plotting of chaos coordinates where z values represent the frequency of (x,y) coordinates

```
filename con1 "filename.ps";
*goptions device=tek4010 gepilog='18'X rotate;
goptions device=ps300 gsfname=con1 gsfmode=replace gsflen=132
        gprolog='2521'X nodisplay ftext=swissb htitle=6 htext=3 gunit=pct;
options linesize=80 nodate nonumber;
title 'freq-hummt';
data;
        drop cases bases;
        infile '[12010_3700]filename.chaos" dlm=',';
        input cases bases x y ;
proc sort;
        by x y;
proc means noprint;
        by x y;
        var x;
        output out=hum n=z;
proc print data=hum(WHERE=(Z>1));
        var x y z;
proc g3d data=hum(WHERE=(Z>1));
        scatter x*y=z/grid;
```

REFERENCES

Abrams, E. S., Murdaugh, S. E. and Lerman, L. S. 1990. Comprehensive detection of single base changes in human genomic DNA using denaturing gradient gel electrophoresis and a GC clamp. *Genomics*. 7: 463-475.

Adams, S. M. and Blackesley, R. W. 1993. Sequencing a PCR-amplified DNA with the dsDNA cycle sequencing system. *Focus*. 14: 31-32.

Agarwal, D. P. and Goedde, H. W. 1987. Genetic variation in alcohol metabolizing enzymes: Implications in alcohol use and abuse. In *Progress in Clinical Biological Research*. Vol. 241. Alan R. Liss.

Aguilar-Cordova, E. and Lieberman, M. W. 1991. Direct directional sequencing of PCR-amplified genomic DNA. *Biotech*. 11: 63-65.

Anderson, R. D., Bao, C. Y., Minnick, D. T., Baoder, J., Veigl, M. L. and Sedwick, W. D. 1993. Sequencing of double-stranded polymerase chain reaction products for mutation analysis. *Mut Res.*. 288: 181-185.

Aquadro, C. F. 1991. Molecular population genetics of Drosophila. In J. Oakeshotl and M. Whitten (Eds.), *Molecular approaches to Pure and Applied Entomology*. Springer-Verlag: New York.

Beutler, E., Gelbart, T., Han, J., Koziol, J. A. and Beutler, B. 1989. Evolution of the genome and the genetic code: Selection at the dinucleotide level by methylation and polyribonucleotide cleavage. *Proc Natl Acad Sci, USA*. 86: 192-196.

Bilofsky, H. S. and Burks, C. 1988. The GenBank genetic sequence data bank. *Nucl Acids Res*. 16: 1861-1863.

Bird, A. P. 1980. DNA methylation and the frequency of CpG in animal DNA. *Nucl Acids Res*. 8: 1499-1504.

Blake, R. D., Hess, S. T. and Nicholson-Truell, J. 1992. The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *J Mol Evol.* 34: 189-200.

Bodmer, J. et al. 1994. Nomenclature for factors of the HLA system, 1994. *Hum Immunol.* 41: 1-20.

Bosron, W. F., Yin, S-J. and Li, T-K. 1985. Purification and characterization of human liver ß1ß1, *ß2ß2* and *ßIndßInd* alcohol dehydrogenase isoenzymes. In *Enzymology of Carbonyl Metabolism 2: Aldehyde Dehydrogenase, Aldo-Keto Reductase and Alcohol Dehydrogenase* (pp. 193-206). New York: Alan R. Liss, Inc.

Bosron, W. F. and Li, T 1986. Genetic polymorphism of human liver alcohol and aldehyde dehydrogenases, and their relationship to alcohol metabolism and alcoholism. *Hepatology.* 6: 502-510.

Buldryrev, S. V., Goldberger, A. L., Havlin, S., Peng, C-K., Stanley, H. E., Stanley, M. H. R. and Simons, M. 1993. Fractal landscapes and molecular evolution: Modeling the myosin heavy chain gene family. *Biophys. J.* 65: 2673-2679.

Burge, C., Campbell, A. M. and Karlin, S. 1992. Over- and under-representation of short oligonucleotides in DNA sequences. *Proc Natl Acad Sci. USA.* 89: 1358-1362.

Burma, P. K., Raj, A., Deb, J. K. and Brahmachari, S. K. 1992. Genome analysis - A new approach for visualization of sequence organization in genomes. *J Biosci.* 17: 395-411.

Cardon, L., Burge, C., Clayton, D. A. and Karlin, S. 1994. Pervasive CpG suppression in animal mitochondrial genomes. *Proc Natl Acad Sci U.S.A.* 91: 3799-3803.

Cariello, N. R. and Skopek, T. R. 1993. Mutational analysis using denaturing gradient gel electrophoresis and PCR. *Mut Res.* 288: 103-112.

Carrington, M., Miller, T., White, M., Gerrard, B., Stewart, C., Dean, M. and Mann, D. 1992. Typing of *HLA-DQA1* and *DQB1* using DNA single-strand conformation polymorphism. *Hum Immun.* 33: 208-212.

Cavalli-Sforza, L. 1990. Opinions: How can one study individual variation for 3 billion nucleotides of the human genome. *Am J Hum Gen.* 46: 649-651.

Charlesworth, B. 1994. Patterns in the genome. *Curr Biol.* 4: 182-184.

Conner, B. J., Reyes, A. A., Morin K., Itakura, K., Teplitz, R. L. and Wallace, R. B. 1983. Detection of sickle cell β^S-globin allele by hybridization with synthetic oligonucleotides. *Proc Natl Acad Sci USA.* 80: 278-282.

Cotton, R. G. H., Rodrigues, N. R. and Campbell, R. D. 1988. Reactivity of cytosine and thymine in single-base-pair mismatches with hydroxylamine and sodium tetroxide and its application to the study of mutations. *Proc Natl Acad Sci USA.* 85: 4397-4401.

Coulondre, C., Miller, J. H., Farabaugh, P. J. and Gilbert, W. 1978. Molecular basis of base substitution hot spots in *Escherichia coli. Nature.* 274: 775-780.

Crawford, M. H. 1992. When two worlds collide. *Hum Biol.* 64: 271-279.

Crawford, M. H. and Enisco, V. B. 1983. Population structure of circumpolar groups of Siberia, Alaska, Canada and Greenland. In M. H. Crawford and J. H. Mielke (eds.). *Current Developments in Anthropological Genetics, vol. 2, Ecology and Population Structure*, (pp. 51-91). New York: Plenum Press.

Devereux, J., Haeberli, P. and Smithies, O. 1981. A comprehensive set of sequence analysis programs for the VAX. *Nucl Acids Res.* 12: 387-395.

Dobzhansky, T. 1941. *Genetics and the origin of species.* New York: Columbia University Press.

D'Onofrio, G., Mouchiroud, D., Aissani, B., Gautier, C. Bernardi, G. 1991. Correlations between the compositional properties of human genes, codon usage and amino acid composition of proteins. *J Mol Evol.* 32: 504-510.

Dowton, M. and Austin, A. D. 1993. Direct sequencing of double-stranded PCR products without intermediate fragment purification; digestion with mung bean nuclease. *Nucl Acids Res.* 21: 3599-3600.

Duester, G., Smith, M., Bilanchone, V. and Hatfield, G. W. 1986. Molecular analysis of the human class I alcohol dehydrogenase gene family and nucleotide sequence of the gene encoding the ß subunit. *J Biol Chem.* 261: 2027-2033.

Dupont, B. 1990. *Immunobiology of HLA.* New York: Springer-Verlag.

Dutta C. and Das, J. 1992. Mathematical characterization of chaos game representation - New algorithms for nucleotide sequence analysis. *J Mol Biol.* 228: 715-719.

Ehrlich, M. and Wang, RY-H. 1981. 5' Methylcytosine in eukaryotic DNA. *Science.* 212: 1350-1357.

Eklund, H., Nordstrom, B., Zeppezauer, E., Soderlund, G., Ohlsson, I., Boiwe, T., Soderberg, B., Tapia, O., Branden, C-I. and Åkeson, Å. 1976. Three-dimensional structure of horse liver alcohol dehydrogenase at 2.4 Å resolution. *J Mol Biol.* 102: 27-59.

Engelke, D., Hoener, P. and Collins, F. 1988. Direct sequencing of enzymatically amplified human genomic DNA. *Proc Natl Acad Sci, USA.* 85: 544-548.

Erlich, H. and Gyllensten, U. B. 1989. The evolution of allelic diversity at the primate major histocompatibility complex class II loci. *Hum Immun.* 30: 110-118.

Fan, W., Kasahara, M., Gutknecht, J., Klein, D., Mayer, W. E., Jonker, M. and Klein, J. 1989. Shared class II MHC polymorphisms between human and chimpanzees. *Hum Immun.* 26: 107-121.

Ferris, S. D., Brown, W. M., Davidson, W. S. and Wilson, A. C. 1981. Extensive polymorphism in the mitochondrial DNA of apes. *Proc Natl Acad Sci, USA.* 78: 6319-6323.

Fickett, F. W., Torney, D. C. and Wolf, D. R. 1992. Base compositional structure of genomes. *Genomics.* 13: 1056-1064.

Figueroa, F., Gunther, E. and Klein, J. 1988. MHC polymorphism pre-dating speciation. *Nature.* 335: 265-2647.

Ford, E. B. 1940. Polymorphism and taxonomy. Pp. 493-513. In J. Huxley (ed.), *The New Systematics.* Oxford: Clarendon Press.

Gaur, L. K., Hughes, A. L., Heise, E. R. and Gutknecht, J. 1992. Maintenance of *DQβ1* polymorphisms in primates. *Mol Biol Evol.* 9: 599-609.

Gennari, K., Wermuth, B., Muellener, D., Ehrig, T. and Wartburg, J. 1988. Genotyping of human class I alcohol dehydrogenase: Analysis of enzymatically amplified DNA with allele-specific oligonucleotides. *Fed Euro Biochem Soc.* 228: 305-309.

Goldman, N. 1993. Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. *Nucl Acids Res.* 21: 2487-2491.

Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pavw, A. 1980. Codon catalog usage and the genome hypothesis. *Nucl Acids Res.* 8: R49-R62.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucl Acids Res.* 9: r43.

Gray, M. W. 1989. The evolutionary origins of organelles. *Trends Genet.* 5: 294-299.

Gray, M. W. 1992. The endosymbiont hypothesis revisited. *Int Rev Cytol.* 141: 233-357.

Gross, R. H. 1986. A DNA sequence analysis program for the Apple Macintosh *Nucl Acids Res.* 14: 591-596.

Gyllensten, U. B. and Erlich, H. A. 1988. Generation of single-stranded DNA by the polymerase chain reaction and its application to direct sequencing of the *HLA-DQA* locus. *Proc Natl Acad Sci USA.* 85: 7652-7656.

Gyllensten, U. B., Lashkari, D. and Erlich. 1990. Allelic diversification at the class II *DQß* locus of the mammalian major histocompatibility complex. *Proc Natl Acad Sci.* 87: 1835-1839.

Hamrick, J. L. and Allard, R. W. 1972. Microgeographic variation in allozyme frequencies in *Avena barbata*. *Proc Natl Acad Sci, USA.* 69: 2100-2104.

Harris, H. 1966. Enzyme polymorphisms in man. *Proc Roy Soc London Series B.* 164: 198-310.

Harris, H., Hopkinson, D. A. and Edwards, Y. H. 1977. Polymorphism and the subunit structure of enzymes: A contribution to the neutralist-selectionist controversy. *Proc Natl Acad Sci, USA.* 74: 698-701.

Harvey, M., Brisson, I. and Guerin, S. 1993. A simple apparatus for fast and inexpensive recovery of DNA from polyacrylamide gels. *Biotech.* 14: 942-948.

Hayashida, H. and Miyata, T. 1983. Unusual evolutionary conservation and frequent DNA segment exchange in class I genes of the major histocompatibility complex. *Proc Natl Acad Sci USA.* 80: 2671-2675.

Haden, L., Hoog, J., Larsson, K., Lake, M., Lagerholm, E., Holmgren, A., Vallee, B., Jornvall, H. and Lindstrom, H. 1986. cDNA clones coding for the ß-subunit of human liver alcohol dehydrogenase have differently sized 3'-noncoding regions. *Fed Euro Biochem Soc.* 3210 : 327-332.

Hess, S. T., Blake, J. D. and Blake, R. D. 1994. Wide variations in neighbor-dependent substitution rates. *J Mol Biol.* 236: 1022-1033.

Hill, K. A., Schisler, N. J. and Singh, S. M. 1992. Chaos game representation of coding regions of human globin genes and alcohol dehydrogenase genes of phylogenetically divergent species. *J Mol Evol.* 35: 261-269.

Hong, J. 1990. Prediction of oligonucleotide frequencies based upon dinucleotide frequencies obtained from the nearest-neighbor analysis. *Nucl Acids Res.* 18: 1625-1629.

Hughes, A., and Nei, M. 1990. Evolutionary relationships of class II major-histocompatibility-complex genes in mammals. *Mol Biol Evol.* 7: 491-514.

Hunkapiller, M. W. 1991. Advances in DNA sequencing technology. *Curr Opin Gen Dev.* 1: 88-92.

Hurley, T., Bosron, W. F., Hamilton, J. A. and Amzel, L. M. 1991. Structure of human $\beta_1\beta_1$ alcohol dehydrogenase: Catalytic effects of non-active site substitutions. *Proc Natl Acad Sci. USA* 88, 8149-8153.

Ichinose, A., Espling, E. S., Takamatsu, J., Saito, H., Shinmyozu, K., Maruyama, I., Petersen, T. and Davie, E. W. 1991. Two types of abnormal genes for plasminogen in families with a predisposition for thrombosis. *Proc Natl Acad Sci USA.* 88: 115-119.

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

Ikuta, T., Fujiyoshi, T., Kurachi, K. and Yoshida, A. 1985. Molecular cloning of a full-length cDNA for human alcohol dehydrogenase. *Proc Natl Acad Sci USA.* 82: 2703-2707.

Ikuta, T., Szeto, S. and Yoshida, A. 1986. Three human alcohol dehydrogenase subunits: cDNA structure and molecular and evolutionary divergence. *Proc Natl Acad Sci USA.* 83: 634-638.

Innis, M., Myambo, K., Gelfand, D. and Brow, M. 1988. DNA sequencing with Thermus aquatica DNA polymerase and direct sequencing of polymerase chain reaction-amplified DNA. *Proc Natl Acad Sci. USA.* 85: 9436-9440.

JeanPierre, M. 1987. A rapid method for the purification of DNA from blood. *Nucl Acids Res.* 15: 9611.

Jeffrey, H. J. 1990. Chaos game representation of gene structure. *Nucl Acids Res.* 18: 2163-2170.

Jeffrey, H. J. 1992. Chaos game visualization of sequences. *Comput & Graphics.* 16: 25-33.

Jornvall, H. 1985. Alcohol dehydrogenase, aldehyde dehydrogenase, and related enzymes. *Alcohol.* 2: 61-66.

Jornvall, H., Persson, M. and Jeffrey, J. 1987. Characteristics of alcohol/polyol dehydrogenases: The zinc-containing long-chain alcohol dehydrogenases. *Eur J Biochem.* 167: 195-201.

Jornvall, H., Hempel, J., Vallee, B., Bosron, W. and Li, T. 1984. Human liver alcohol dehydrogenase: Amino acid substitution in the *β2β2* Oriental isozyme explains functional properties, establishes an active site structure, and paralleles mutational exchanges in the yeast enzyme. *Proc Natl Acad Sci, USA.* 81: 3024-3028.

Karlin S. and Brendel, V. 1993. Patchiness and correlations in DNA sequences. *Science*, 259: 677-683.

Karlin, S., Blaisdell, B. E., Sapolsky, R. J., Cardon, L. and Burge,C. 1993. Assessments of DNA inhomogeneities in yeast chromosome III. *Nucl Acids Res.* 21: 703-711.

Keen, J., Lester, D., Inglehearn, C., Curtis, A. and Bhattacharya, S. 1991. Rapid detection of single base mismatches as heteroduplexes on hydrolink gels. *Trends Genet.* 7: 5.

Keohavong, P., Ling, L., Dias, C. and Thilly, W. G. 1993. Predominant mutations induced by the *Thermococcus litoralis,* Vent DNA polymerase during DNA amplification. *PCR Met Appl.* 2: 288-292.

Klein, J. 1986. *Natural history of the major histocompatibility complex* New York: John Wiley.

Klein, J. 1987. Origin of major histocompatibility complex polymorphism: The trans-species hypothesis. *Human Imm.* 19: 155-162.

Klein, J., Gutknecht, J. and Fischer, N. 1990. The major histocompatibility complex and human evolution. *Trends Gen.* 6: 7-11.

Kohler, S. W., Provost, G S., Fieck, A., Kretz, P. L., Bullock, W. O., Sorge, J. A., Putman, D. L., Short, J. M. (1991). Spectra of spontaneous and mutagen-induced mutations in the *lacI* gene in transgenic mice. *Proc Natl Acad Sci USA.* 8: 7958-7962.

Kreitman, M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster. Nature.* 304: 412-417.

Kretz, K. A., Carson, G. S. and O'Brien, J. S. 1989. Direct sequencing from low-melt agarose with sequenase.® *Nucl Acids Res.* 17: 5864.

Lawlor, D. A., Ward, F. E., Ennis, P. D., Jackson, A. and Parham, P. 1988. HLA-A and B polymorphisms predate the divergence of humans and chimpanzees. *Nature*. 335: 268-271.

Lefevre C. and Ikeda, J-E. 1994. A fast word search algorithm for the representation of sequence similarity in genomic DNA. *Nucl Acids Res*. 22: 1514.

Leung, M. Y., Blaisdell, B. E., Burge, C. and Karlin, S. 1991. An efficient algorithm for identifying matches with errors in multiple long molecular sequences. *J Mol Biol*. 221: 1367-1378.

Lewontin, R. C. 1974. *The genetic basis of evolutionary change*. New York: Columbia University Press.

Lewontin, R. C. and Hubby, J. L. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura Genetics*. 54: 595-609.

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

Lipman, D. J. and Wilbur, W. J. 1983. Contextual constraints on synonymous codon choice. *J Mol Biol*. 163: 363-376.

Liu, Y. G., Mitsukawa, N. and Whittier, R. F. 1993. Rapid sequencing of unpurified PCR products by thermal asymmetric PCR cycle sequencing using unlabeled sequencing primers. *Nucl Acids Res*. 21: 3333-3334.

Lloyd and Sharp 1992. Evolution of codon usage patterns - The extent and nature of divergence between *Candida-albicans* and *Saccharomyces-cerevisiae*. *Nucl Acids Res*. 20: 5289-5295.

Mani, G. S. 1992a. Correlations between the coding and non-coding regions in DNA. *J Theor Biol*. 158: 429-445.

Mani, G. S. 1992b. Long-range doublet correlations in DNA and the coding regions. *J. Theor. Biol.* 158: 447-464.

Mashiyama, S., Murakami, S., Yoshimoto, T., Sekiya, R. and Hayashi, K. 1990. Detection of p53 gene mutations in human brain tumors by single-strand conformation polymorphism analysis of polymerase chain reaction products. *Oncogene*. 6: 1313-1318.

May, R. 1976. Simple mathematical models with very complicated dynamics. *Nature*. 261: 459-467.

Meyers, R. M., Larin, Z. and Maniatis, T. 1985. Detection of single base substitutions by ribonuclease cleavage at mismatches in RNA:DNA duplexes. *Science*. 230: 1242-1246.

Moncany, M. L. J. and Keller, R. 1993. Comparison of some procedures to recover the PCR-amplified products for a direct sequencing. *GATA*. 10: 24-26.

Murray, V. 1989. Improved double-stranded DNA sequencing using the linear polymerase chain reaction. *Nucl Acids Res.* 17: 8889.

Myhr, B. C. 1991. Validation studies with Muta™ Mouse: A transgenic mouse model for detecting mutations *in vivo*. *Environ Mol Mutagen*. 18: 308-315.

Nandy, A. 1994. Recent investigations into global characteristics of long DNA sequences. *Indian J Biochem & Biophys.* 31: 149-155.

Needleman, S. B. and Wunsch, C. D. 1970. A general method applicable to search for similarities in the amino acid sequence of two proteins. *J Mol Biol.* 48: 443-453.

Nei, M. 1975. *Molecular population genetics and evolution*. New York: American Elsevier.

Nei, M. and Graur, D. 1984. Extent of protein polymorphism and the neutral mutation theory. *Evol Biol.* 17: 73-118.

Nussinov, R. 1980. Some rules in the ordering of nucleotides in the DNA. *Nucl Acids Res.* 8: 4545-4561.

Nussinov, R. 1981a. Nearest neighbor nucleotide patterns: Structural and biological implications. *J Biol Chem.* 256: 8458-8462.

Nussinov, R. 1981b. Eukaryote dinucleotide preference rules and their implications for degenerate codon usage. *J Mol Biol.* 149: 125-131.

Nussinov, R. 1984. Doublet frequencies in evolutionary distinct groups. *Nucl Acids Res.* 12: 1749-1763.

Nussinov, R. 1993. Fractal graphical representation and analysis of DNA and protein sequences. *BioSystems.* 30: 137-160.

Oliver, J. L., Bernaola-Galvan, P., Guerrero-Garcia, J. and Roman-Roldan, R. 1993. Entropic Profiles of DNA sequences through chaos-game-derived images. *J Theor Biol.* 160: 457-470.

Orita, M. H., Iwahana, H., Kanazawa, K., Hayashi, K. and Sekiya, T. 1989a. Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proc Natl Acad Sci USA.* 86: 2766-2770.

Orita, M. Y., Suzuki, T., Sekiya, T. and Hayashi, K. 1989b. Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction. *Genomics.* 5: 874-879.

Otting, N., Kenter, M., van Weeren, P., Jonker, M. and Bontrop, R. E. 1992. MHC-DQß repertoire variation in hominoid and old world primate species. *J Immun.* 149: 461-470.

Parham, P. and Strominger, J. 1982. *Histocompatibility antigens: structure and function (receptors and recogni* Series B, vol 14). Chapman and Hall: London.

Peng, C. -K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simmons, M. and Stanley, H. E. 1992. Long-range correlations in nucleotide sequences. *Nature*. 356: 168-170.

Phillips, G. J., Arnold, J. and Ivarie, R. 1987. Mono- through hexanucleotide composition of the *Escherichia coli* genome: A Markov chain analysis. *Nucl Acids Res.* 15: 2611-2626.

Posukh, O. L., Weibe, V. P., Sukernick, R. I., Osipova, L. P., Karaphet, T. M. and Schanfield, M. S. 1990. Genetic studies of the Evens, an ancient human population in eastern Siberia. *Hum Biol.* 62: 457-465.

Riley, E. and Olerup, O. 1992. HLA polymorphisms and evolution. *Imm Today*. 13: 333-335.

Rogerson, A. C. 1989. The sequence asymmetry of the *Escherichia coli* chromosome appears to be independent of strand or function and may be evolutionarily conserved. *Nucl Acids Res.* 17: 5547-5563.

Rogerson, A. C. 1991. There appear to be conserved constraints on the distribution of nucleotide sequences in cellular genomes. *J Mol Evol.* 32: 24-30.

Russell, G. J., Walker, P. M. B., Elton, R. A., Subad-Sharpe, J. H. 1976. Doublet frequency analysis of fractionated vertebrate nuclear DNA. *J Mol Biol.* 108: 1-23.

Santamaria, P., Boyce-Jacino, M. T., Lindstrom, A. L., Barbosa, J. J., Faras, A. J. and Rich, S. S. 1992. HLA class II typing: Direct sequencing of DRB, DQB and DQA genes. *Hum Immun.* 33: 69-81.

Sarich, V. M. and Wilson, A. C. 1966. Quantitative immunochemistry and the evolution of primate albumins: Micro-complement fixation. *Science*. 154: 1563-1566.

Sarkar, G., Yoon, H. and Sommer, S. 1992. Dideoxy fingerprinting (ddF): A rapid and efficient screen for the presence of mutations. *Genomics*. 13: 441-443.

Sarkar, F. H., Li, Y-W. and Crissman, J. D. 1993. A method for PCR sequencing of the p53 gene from a single 10 μm frozen or paraffin-embedded tissue section. *BioTech*. 15: 36-38.

Satta, Y., O'HUigin, C., Takahata, N. and Klein, J. (1994). Intensity of natural selection at the major histocompatibility complex loci. *Proc Natl Acad Sci USA*. 91: 7184--7188.

Sekiya 1993. Detection of mutant sequences by single-strand conformation polymorphism analysis. *Mut Res*. 288: 79-83.

Scharf, S. J., Horn, G. T. and Erlich, H. A. 1986. Direct cloning and sequence analysis of enzymatically amplified genomic sequences. *Science*. 233: 1076-1078.

Sharp, P. M. and Lloyd, A. T. 1993. Regional base composition variation along yeast chromosome III: Evolution of chromosome primary structure. *Nucl Acids Res*. 21: 179-183.

Sharp, P. M. and Matassi, G. 1994. Codon usage and genome evolution. *Curr Opin Gen Dev*. 4: 851-860.

Smooker, P. M. and Cotton, R. G. H. 1993. The use of chemical reagents in the detection of DNA mutations. *Mut Res*. 288: 65-77.

Solovyev, V. V. 1993. Fractal graphical representation and analysis of DNA and protein sequences. *BioSystems*. 30: 137-160.

Solovyev, V. V., Korolev, S. V. and Lim, H. A. 1992. A new approach for the classification of functional regions of DNA sequences based on fractal representation. Report FSCU-SCRI-91-40 Supercomputer Computations Research Institute, Tallahassee, Forida, USA.

Sommer, S. S. 1992. Assessing the underlying pattern of human germline mutations: Lessons from the factor IX gene. *FASEB J.* 6: 2767-2774.

Stamatoyannopoulos, G., Chen, S. H. and Jukui, M. 1975. Liver alcohol dehydrogenase in Japanese: High population frequency of atypical form and its possible role in alcohol sensitivity. *Am J Hum Gen.* 27: 780-797.

Sun, H-W. and Plapp, B. V. 1992. Progressive sequence alignment and molecular evolution of the Zn-containing alcohol dehydrogenase family. *J Mol Evol.* 34: 522-535.

Swofford, and Olsen 1990. Phylogeny Reconstruction. In D. M. Hillis and C. Moritz (Ed.), *Molecular Systematics*, (ch-11) New York: Sinauer Associates.

Szathmary, E. J. E. 1978. Peopling of North America: Clues from genetic studies. In W. S. Laughlin (Ed.), *Origins and Affinities of the First Americans.* New York: Gustav Fischer.

Szathmary, E. J. E. 1984. Peopling of northern North America: clues from genetic studies. *Acta Anthropol.* 8: 79-109.

Szathmary, E. J. E. 1993. MtDNA and the peopling of the Americas. *Am J Hum Genet.* 53: 793-799.

Szathmary, E. J. E. and Ossenberg, N. S. 1978. Are the biological differences between North American Indians and Eskimos truly profound? *Curr Anthropol.* 19: 673-701.

Szathmary, E. J. E. Ferrall, R. E. and Gershowitz, H. 1983. Genetic differentiation in Dogrib Indians: Serum protein and erythrocyte enzyme variation. *Am J Phys Anthropol.* 62: 249-254.

Tasheva, E. S. and Roufa, D. J. 1993. Deoxycytidine methylation and the origin of spontaneous transition mutations in mammalian cells. *Som Cell Mol Gen.* 19: 275-283.

Thomasson, H. R., Edenberg, H. J., Crabb, D. W., Mai, X-L, Jerome, R. E., Li, T-K., Wang, W-P., Lion, Y-T., Lu, R-B and Yin, S-J. 1991. Alcohol and aldehyde dehydogenase genotypes and alcoholism in Chinese men. *Am J Hum Genet.* 48: 677-681.

Trowsdale, J. and Powis, S. H. 1992. The MHC: Relationship between linkage and function. *Curr Opin Gen Dev.* 2: 492-497.

Thatcher, D. R. 1980. The complete amino acid sequence of three alcohol dehydrogenase alleloenzymes (AdhN-11, AdhS and Adh UF) from the fruitfly *Drosophila melanogaster*. *Biochem J.* 187: 875-886.

Torroni, A., Schurr, T. G., Yang, C. -C., Szathmary, E. J. E., Williams, R. C., Schanfield, M. S., Troup, G. A., Knowler, W.C., Lawerence, D. N., Weiss, K. M. and Wallace, D. C. 1992. Native American mitochondrial DNA analysis indicates that the Amerind and Nadene populations were founded by two independent migrations. *Genetics.* 130: 153-162.

Trucco, G., Fritsch, R., Giorda, R. and Trucco, M. 1989. Rapid detection of IDDM susceptibility with HLA-DQß -alleles as markers. *Diabetes.* 38: 1617-1622.

Tsonis P. A. and Tsonis, A. A. 1989. Chaos: Principles and implications in biology. *Cabios.* 5: 27-32.

Tsonis, A. A., Elsner J. B. and Tsonis P. A. 1993. On the existence of scaling in DNA sequences. *Biochem Biophys Res Com.* 197: 1288-1295.

Volinia, S., Gambari, R., Bernardi, F. and Barrai I. 1989. The frequency of oligonucleotides in mammalian genic regions. *Cabios.* 5: 33-40.

Von Bahr-Lindstrom, H., Hoog, J-O., Kaiser, R., Fleetwood, L., Larsson, K., Lake, M., Holmquist, B., Holmgren, A., Hempel, J., Vallee, B. L. and Jornvall, H. 1986. cDNA and protein structure for the α subunit of human liver alcohol dehydrogenase. *Biochem.* 25: 2465-2470.

Voss, R. F. 1993a. *1/f* noise and fractals in DNA-base sequences. In: Crilly AJ, Earnshaw RA, Jones H (Eds.), *Applications of Fractals and Chaos: The Shape of tThings*. pp. 7-20. New York: Springer-Verlag.

Voss, R. F. 1993b. 1/f noise and fractals in DNA-base sequences. *Phys Rev Lett.* 68: 3805-3808.

Wain-Hobson J. M., Nussinov R., Brown R. J. and Sussman, J. L. (1981) Preferential codon usage in genes. *Gene*. 13: 335-364.

Wallace, D. C. and Torroni, A. 1992. American Indian prehistory as written in the mitochondrial DNA: A review. *Hum Biol*. 64: 403-416.

Wallace, D. C., Garrison, K. and Knowler, W. C. 1985. Dramatic founder effects in Amerindian mitochondrial DNAs. *Am J Phys Anthropol*. 68: 149-155.

Weiss, E. H., Mellor, A. L., Fahrner, K., Simpson, E., Hurst, J. and Flavell, R. A. 1983. The structure of a mutant H-2 gene suggests that the generation of polymorphism in *H-2* genes may occur be gene conversion-like events. Nature. 301: 671.

White, M. B., Carvalho, M., Derse, D., O'Brien, S. J. and Dean, M. 1992. Detecting single base substitutions as heteroduplex polymorphisms. *Genomics*. 12: 301-306.

Wilkinson, L. 1991. *Systat: The system for statistics*. Systat Inc., Evanston, IL.

Winter, E., Yamanoto, F., Almoguera, C. and Peruchio, M. A. 1985. Methods to detect and characterize point mutations in transcribed genes: Amplification and overexpression of the mutant cKi-ras allele in human tumor cells. *Proc Natl Acad Sci USA*. 82: 7575-7579.

Wong, C., Dowling, C., Saiki, R., Higuchi, R., Erlich, H. and Kazazian, H. 1987. Characterization of ß-thalassaemia mutations using direct genomic sequencing of amplified single copy DNA. *Nature*. 370: 384-386.

Wrischnik, L. A., Higuchi, R. G., Stoneking, M., Erlich, H. A., Arnheim, N. and Wilson, A. C. (1987). Length mutations in human mitochondrial DNA: Direct sequencing of enzymatically amplified DNA. *Nucl Acids Res.* 15: 529-542.

Wu, C-I. and Maeda, N. 1987. Inequality in mutation rates of the two strands of DNA. *Nature*. 327: 169-170.

Xu, Y., Carr, L., Bosron, W., Li, R. and Edenberg, H. 1988. Genotyping of human alcohol dehydrogenases at the *Adh2* and *Adh3* loci following DNA sequence amplification. *Genomics*. 2: 209-214.

Yandell, D. and Dryja, T. 1989. Detection of DNA sequence polymorphisms by enzymatic amplification and direct genomic sequencing. *Am J Hum Gen.* 45: 547-555.

Yokoyama, S. and Harry, D. E. 1993. Molecular phylogeny and evolutionary rates of alcohol dehydrogenases in vetebrates and plants. *Mol Biol Evol.* 10: 1215-1226.

Yokoyama, S., Yokoyama, R., Kinlaw, C. S. and Harry, D. E. 1990. Molecular evolution of zinc-containing long-chain alcohol dehydrogenase genes. *Mol Biol Evol.* 7: 143-154.

Yoshida, A., Impraim, C. C. and Huang, J. Y. 1981. Enzymatic and structural differences between usual and atypical human liver alcohol dehydrogenases. *J Biol Chem.* 256: 12430-12436.

Zhang, C-T and Chou, K-C. 1993. Graphic analysis of codon usage strategy in 1490 human proteins. *J Prot Chem.* 12: 329-335.