Report Number: WUCSE-2002-39

2002-11-18

# Assembly and Compositional Analysis of Human Genomic DNA - Doctoral Dissertation, August 2002

Eric C. Rouchka

In 1990, the United States Human Genome Project was initiated as a fifteen-year endeavor to sequence the approximately three billion bases making up the human genome (Vaughan, 1996).As of December 31, 2001, the public sequencing efforts have sequenced a total of 2.01 billion finished bases representing 63.0% of the human genome (http://www.ncbi.nlm.nih.gov/genome/seq/page.cgi?F=HsProgress.shtml&&ORG=Hs) to a Bermuda quality error rate of 1/10000 (Smith and Carrano, 1996). In addition, 1.11 billion bases representing 34.8% of the human genome has been sequenced to a rough-draft level. Efforts such as UCSC's GoldenPath (Kent and Haussler, 2001) and NCBI's contig assembly (Jang et al., 1999)... **Read complete abstract on page 2.**

# Assembly and Compositional Analysis of Human Genomic DNA - Doctoral Dissertation, August 2002

Eric C. Rouchka

Complete Abstract:

In 1990, the United States Human Genome Project was initiated as a fifteen-year endeavor to sequence the approximately three billion bases making up the human genome (Vaughan, 1996).As of December 31, 2001, the public sequencing efforts have sequenced a total of 2.01 billion finished bases representing 63.0% of the human genome (http://www.ncbi.nlm.nih.gov/genome/seq/page.cgi?F=HsProgress.shtml&&ORG=Hs) to a Bermuda quality error rate of 1/10000 (Smith and Carrano, 1996). In addition, 1.11 billion bases representing 34.8% of the human genome has been sequenced to a rough-draft level. Efforts such as UCSC's GoldenPath (Kent and Haussler, 2001) and NCBI's contig assembly (Jang et al., 1999) attempt to assemble the human genome by incorporating both finished and rough-draft sequence. The availability of the human genome data allows us to ask questions concerning the maintenance of specific regions of the human genome. We consider two hypotheses for maintenance of high G+C regions: the presence of specific repetitive elements and compositional mutation biases. Our results rule out the possibility of the G+C content of repetitive elements determining regions of high and low G+C regions in the human genome. We determine that there is a compositional bias for mutation rates. However, these biases are not responsible for the maintenance of high G+C regions. In addition, we show that regions of the human under less selective pressure will mutate towards a higher A+T composition, regardless of the surrounding G+C composition. We also analyze sequence organization and show that previous studies of isochore regions (Bernardi,1993) cannot be generalized within the human genome. In addition, we propose a method to assemble only those parts of the human genome that are finished into larger contigs. Analysis of the contigs can lead to the mining of meaningful biological data that can give insights into genetic variation and evolution. I suggest a method to help aid in single nucleotide polymorphism (SNP)detection, which can help to determine differences within a population. I also discuss a dynamic-programming based approach to sequence assembly validation and detection of large-scale polymorphisms within a population that is made possible through the availability of large human sequence contigs.

Department of Computer Science & Engineering

Washington University in St. Louis
School of Engineering & Applied Science

2002-39

# Assembly and Compositional Analysis of Human Genomic DNA - Doctoral Dissertation, August 2002

Authors: Rouchka, Eric C.

November 18, 2002

Abstract: In 1990, the United States Human Genome Project was initiated as a fifteen-year endeavor to sequence the approximately three billion bases making up the human genome (Vaughan, 1996). As of December 31, 2001, the public sequencing efforts have sequenced a total of 2.01 billion finished bases representing 63.0% of the human genome (http://www.ncbi.nlm.nih.gov/genome/seq/page.cgi?F=HsProgress.shtml&&ORG=Hs) to a Bermuda quality error rate of 1/10000 (Smith and Carrano, 1996). In addition, 1.11 billion bases representing 34.8% of the human genome has been sequenced to a rough-draft level. Efforts such as UCSC's GoldenPath (Kent and Haussler, 2001) and NCBI's contig assembly (Jang et al., 1999) attempt to assemble the human genome by incorporating both finished and rough-draft sequence. The availability of the human genome data allows us to ask questions concerning the maintenance of specific regions of the human genome. We consider two hypotheses for maintenance of high G+C regions: the presence of specific repetitive elements and compositional mutation biases. Our results rule out the possibility of the G+C content of repetitive elements determining regions of high and low G+C regions in the human genome. We determine that there is a compositional bias for mutation rates. However, these biases are not responsible for the maintenance of high G+C regions. In addition, we show that regions of the human under less selective pressure will mutate towards a higher A+T composition, regardless of the surrounding G+C composition. We also analyze sequence organization and show that previous studies of isochore regions

WASHINGTON UNIVERSITY
SEVER INSTITUTE OF TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE

---

# Assembly and Compositional Analysis of Human Genomic Data

by
Eric C. Rouchka

Prepared under the direction of Professors
David J. States
Warren Gish
Ron Cytron

---

A dissertation presented to the Sever Institute of
Washington University in partial fulfillment
of the requirements for the degree of

DOCTOR OF SCIENCE

August, 2002
Saint Louis, Missouri

WASHINGTON UNIVERSITY
SEVER INSTITUTE OF TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE

---

ABSTRACT

---

Assembly and Compositional Analysis
of Human Genomic Data
by Eric C. Rouchka

---

ADVISOR: David J. States

---

AUGUST, 2002
Saint Louis, Missouri

---

In 1990, the United States Human Genome Project was initiated as a fifteen-year endeavor to sequence the approximately three billion bases making up the human genome (Vaughan, 1996). As of December 31, 2001, the public sequencing efforts have sequenced a total of 2.01 billion finished bases representing 63.0% of the human genome (`http://www.ncbi.nlm.nih.gov/genome/seq/page.cgi?F=HsProgress.shtml&&ORG=Hs`) to a Bermuda quality error rate of 1/10000 (Smith and Carrano, 1996). In addition, 1.11 billion bases representing 34.8% of the human genome has been sequenced to a rough-draft level. Efforts such as UCSC's `GoldenPath` (Kent and Haussler, 2001) and NCBI's contig assembly (Jang *et al*., 1999) attempt to assemble the human genome by incorporating both finished and rough-draft sequence. The availability of the human genome data allows us to ask questions concerning the maintenance of specific regions of the human genome. We consider two hypotheses for maintenance of high G+C regions: the presence of specific repetitive elements and compositional mutation biases. Our results rule out the possibility of the G+C content of repetitive elements determining regions of high and low G+C regions in the human genome. We determine that there is a compositional bias for mutation rates. However, these biases are not responsible for the maintenance of high G+C regions. In addition, we show that regions of the human under less selective pressure will mutate towards a higher A+T composition, regardless of the surrounding G+C composition. We also analyze sequence organization and show that previous studies of isochore regions (Bernardi, 1993) cannot be generalized within the human genome. In addition, we propose a method to assemble only those parts of the human genome that are finished into larger contigs. Analysis of the contigs can lead to the mining of meaningful biological data that can give insights into genetic variation and evolution. I suggest a method to help aid in single nucleotide polymorphism (SNP) detection, which can help to determine differences within a population. I also discuss a dynamic-programming based approach to sequence assembly validation and detection of large-scale polymorphisms within a population that is made possible through the availability of large human sequence contigs.

TO MY FAMILY AND FRIENDS whom have been very supportive while anxiously awaiting the day when I would reach this milestone. Yes, the day is Monday and the month is August, but now you know the year -- 2002.

# Contents

# Tables

# Figures

# Equations

# List of Abbreviations

**BAC -** bacterial artificial chromosome
**BP -** base pair
**DFA -** deterministic finite automaton
**DNA -** deoxyribonucleic acid
**DOE -** Department of Energy
**EST -** expressed sequence tag
**HERV** – human endogenous retrovirus
**HTGS -** high throughput genome sequence
**IBC -** Washington University Institute for Biomedical Computing
**indel -** insertion or deletion
**KB** - kilobase
**LINE -** long interspersed element
**LTR** – long terminal repeat
**MB** - megabase
**NIH -** National Institute for Health
**NT -** nucleotide
**SINE -** short interspersed element
**SNP** - single nucleotide polymorphism
**STS -** sequence tagged site
**TIGR -** The Institute for Genome Research
**UCSC -** University of California at Santa Cruz
**WUSTL** - Washington University in St. Louis
**YAC -** yeast artificial chromosome

# Glossary[1]

**ALU** - An interspersed DNA sequence, approximately 300 bp long, found in the genome of primates that is cleaved by the restriction enzyme ALU I.

**BAC (bacterial artificial chromosome)** - A type of cloning vector use to clone DNA fragments.

**Biocluster** - A set of 25 4-cpu machines set up by Compaq Corporation for computational biology applications.

**BLAST (Basic Local Alignment Statistics Tool)** - A tool which reports the score for aligning two sequences using Karlin-Altschul statistics.

**chimera** - A clone composed of pieces derived from two or more distinct organisms.

**clone** - A DNA segment which has been inserted into a cloning vector and replicated to form many copies.

**codon** - A sequence of three nucleotides that specifies a particular amino acid during protein synthesis.

**complement** - Refers to the base which can pair with a reference base via a hydrogen bond. The complement of adenine (A) is thymine (T); the complement of cytosine (C) is guanine (G).

**contig** - Long stretches of continuous DNA sequence, represented by the concatenation of two or more shorter sequences.

**cosmid** - A type of cloning vector used to clone DNA fragments by packaging the DNA to be cloned into lambda phage viruses which then infect *E. coli*. When the *E. coli* reproduce, so does the DNA fragment of interest.

**cytogenetic** - Pertaining to chromosomes.

**deamination** - The process through which amino groups are stripped off of nucleic acids which results in base pair mismatches.

**density gradient centrifugation** - A technique for separating macromolecules using centrifugal force and solvents of varying density.

[1] Many of the definitions are adapted from three sources: an online BioTech Life Science Dictionary (`http://biotech.icmb.utexas.edu/search`), Molecular Cell Biology (Lodish *et al.*, 1995) and Concepts of Genetics (Klug and Cummings, 1991)

**dinucleotide** - A sequence of two consecutive nucleotides.

**electrophoresis** - A technique for separating DNA molecules based on their migration in a gel. The migration is based on the molecule size.

**euchromatin** - Less condensed chromosomal regions containing most transcribed regions. Euchromatin is the sequence target of the human genome project.

**exon** - The portion of a primary transcript which reaches the cytoplasm as part of the mature mRNA.

**expressed sequence tag** - DNA sequence derived by sequencing an end of a cDNA molecule.

**fasta** - A program which aligns two sequences. Fasta format is the sequence format that is used. Generally, fasta format requires the first line to be a header line beginning with '>' and each subsequent line contains the actual sequence data.

**fingerprint** - The resulting DNA fragment pattern generated by one of several methods, including electrophoresis.

**GenBank** - Database collection of all publicly available DNA sequences maintained by NCBI.

**gene conversion –** The process in which the allele of one gene is converted to another during recombination. Biased gene conversion implies that in regions of high G+C, the conversion is more likely to be to a G or C nucleotide.

**genic** - Referring to regions of a genome in which genes occur.

**genome** - The total genetic information contained within an organism.

**GoldenPath** - Assembly of human genomic DNA using both finished and unfinished clones as well as various mapping information maintained by the University of California-Santa Cruz.

**haplotype** - The set of alleles from closely linked loci carried by an individual and normally inherited as a unit.

**HERV (human endogenous retrovirus) –** One class of LTR retroviruses that have become integrated into the human germline cells and thus fixed within the population.

**heterochromatin** - Highly condensed and transcriptionally inactive portions of the genome which are typically not targeted to be sequenced.

**homologous** - Pertains to two DNA sequences sharing a common ancestor and having both sequence and functional similarity. Note that sequence homology refers only to those sequences that share sequence similarity regardless of their function.

**homologous recombination –** The process by which DNA sequences on maternal and paternal chromatids are exchanged, resulting in new sequence combinations.

**intron** - The portion of a primary transcript which is removed by splicing and is not included as a part of the mature mRNA.

**isochore** - A large scale region of relatively constant G+C composition within a vertebrate genome. According to the theories of Bernardi, there are five different isochore classification schemes depending on the G+C content.

**LINEs** - Long interspersed elements that are non-viral retrotransposons, about 6-7 kb long, which are found abundantly in mammals.

**locus** - A specific location within a chromosome.

**LTR (long terminal repeat)** – A sequence directly repeated at both ends of a defined sequence, typically found in retroviruses (such as the HERV elements).

**methylation** - The process by which a methyl group is added to a nucleotide base thereby modifying it. In humans, general cytosine methylation occurs frequently.

**nucleotide** - one of the four bases, adenine (A), cytosine (C), guanine (G) or thymine (T) composing genomic DNA.

**oligomer** - A short polymer consisting of short stretches of amino acids or nucleic acids.

**oligonucleotide** - A short stretch of nucleic acids.

**orthologous** - DNA sequences from two different species which arose from a common ancestral gene which may or may not have functional conservation.

**paralogous** - DNA sequences within a single genome which are similar to one another and arise from a duplication event.

**physical map** - A map of the location of identifiable landmarks within a nucleotide sequence, including sequence tagged sites and restriction sites.

**pseudogene** - A duplicated gene copy which has become non-functional.

**purine** - One of two nucleic acids, either adenine (A) or guanine (G).

**pyrimidine** - One of two nucleic acids, either cystine (C) or thymine (T).

**RefSeq -** A database for NCBI's reference sequence project, containing transcript and protein coding data among others.

**RepBase** - A database of prototypic sequences representing repetitive elements in eukaryotes. The database is maintained and curated by the Genome Research Institute.

**RepeatMasker** - A program developed by Arian Smit that locates and masks out various repeats, including SINEs, LINEs and simple tandem repeats within a genomic sequence.

**repetitive element** - Any nucleotide sequence that is repeated many times within a genome. SINEs, LINEs, and simple tandem repeats are instances of repetitive elements.

**restriction enzyme** - An enzyme that recognizes and cleaves a specific short sequence.

**restriction site** - A specific short sequence which is recognized by a restriction enzyme.

**single nucleotide polymorphism** - A mutation that occurs at a single point.

**shotgun sequencing** - A technique in which a genome is sequenced by cloning randomly created DNA fragments.

**SINEs** - Short interspersed elements, approximately 300 bp long, which occur abundantly throughout mammalian genomes.

**synonymous** – Referring to a mutation in a codon that does not affect the resulting amino acid.

**transcription** - The process in which one strand of DNA is used as a template to produce a single strand of complementary RNA.

**transition** - A mutational event in which one purine is replaced by another or in which one pyrimidine is replaced by another.

**transversion** - A mutational event in which one purine is replaced by a pyrimidine or a pyrimidine is replaced by a purine.

**wobble base** - The third nucleotide position in a codon. Due to the degeneracy of the genetic code, the wobble base can be mutated and still code for the same amino acid.

**YAC** - A vector used to clone DNA fragments up to 400 kb in length. It is constructed from the replication origin regions needed for replication in yeast cells.

# Acknowledgements

With any work that takes years to complete, there are many people who have been extremely helpful throughout my graduate studies. First of all, I would like to thank Dr. Chip Lawrence of the Wadsworth Institute in Albany, New York who took my interests in computer science and biology and introduced me into the field of computational biology. He has also instrumental in setting up contacts making it possible for me to join Washington University. I which to thank my advisor, Dr. David States, for his support and guidance.

**Institute for Biomedical Computing (Center for Computational Biology)** I would like to thank the staff at the Institute for Biomedical Computing, including Janice Cole, Ken Kaiser and Debbie Peterson, for all their help in dealing with various administrative issues including mailing manuscripts, contacting speakers, setting travel arrangements, making sure I was reimbursed for expenses and helping me to deal with tuition hassles. I am greatly indebted to the computer staff at the IBC, including Gerald Johns, Stan Phillips, Carlos Santos and Brian Dunford-Shore for all of their help in maintaining a stable computing environment.

**States Lab** I would like to thank all of the people formerly in the States lab which have been helpful in countless ways. George Kan has been vitally helpful in providing insight and critique into various aspects as well as providing productive collaboration. Thomas Blackwell has provided clarification and direction with problems of statistical nature, and

has proven to be an excellent guide into what exists in the literature. Volker Nowotny has been a point of contact for obtaining restriction digest data from the Center of Genetics in Medicine. I would also like to thank Ron Liu, David Pollitte, David Maffit, Lisa Gu, and Bill Reisdorf for countless hours of discussion.

**Gish Lab** Warren Gish was extremely helpful in providing physical space and computing resources for allowing me to continue my studies. I greatly appreciate all of his helpful feedback and support. I would also like to thank the members of the Gish lab who adopted me as a member of the lab. Raymond Yeh and Jarret Glasscock have been very kind to share their GoldenPath to RefSeq blast results, allowing me to calculate potential gene-pseudogene pairs. They have answered many questions dealing with Perl. Thanks to John Kloss for maintaining a stable computing environment and for thoughtful algorithmic discussions. Thanks also to German Leparc, Miao Zhang, Ting Wang and Magnus Bjursell for helpful discussions.

**Department of Genetics** I would like to thank the Genetics Department for all of their help and support. Specifically, I would like to thank Carol Jones and Mary Pichler for their help in dealing with administrative issues. Richard Mazzarella provided examples of regions of interest in studying CpG islands. A special thanks to Sean Eddy who has served on my committee and has provided much appreciated critical discussion regarding research direction.

**Computer Science Department** A special thanks goes to the people in the Computer Science Department who have been extremely helpful and flexible in letting me pursue

research at the medical school campus. In particular, I would like to thank Jean Grothe for providing an open line of communication and for all her patience in helping me figure out all of the administrative side of things. Thanks to Ron Cytron for his candidness and clarity in helping to make sure I was on the right track with the Computer Science Department and for his role in serving on my committee. I would also like to thank Subhash Suri who previously served on my committee, and Weixong Zhang and Michael Brent who are current members.

**Miscellaneous** We wish to thank Marco Marra and Bernard Brownstein for making experimental fingerprint data available for our use in the course of this study as well as Dr. Pui Kwok and the postdoc in his lab for their work in SNP validation. We wish to thank Jeffrey Bailey and Evan Eichler from the Department of Genetics and Center for Human Genetics at Case Western Reserve School of Medicine and University Hospitals of Cleveland for their help in obtaining and interpreting tables generated from their segmental duplications project. Ray Hookway and Compaq's Enterprise system lab generously provided CPU cycles through the use of their 100 CPU BioCluster.

**Financial Support** Financial support was provided in part by grants from the National Science Foundation, Department of Energy (ER61910 000261; DE-FG02-94ER61910, David States, PI), the National Human Genome Research Institute (R01-HG01391; R01-HG00201, David States, PI), the Merck Foundation (#225), and the National Institutes of Health (HG-R01-01391; 5-T32-HG00045).

# Chapter 1

# Introduction

## 1.1 Background of the Human Genome Project

The United States Human Genome Project, coordinated by the United States Department of Energy (DOE) and the National Institutes of Health (NIH), began in 1990 as a 15 year venture with a primary goal of sequencing the approximately three billion bases making up the human genome (Vaughan, 1996) using a clone-based sequencing approach. In May of 1998, The Perkin-Elmer Corporation, Dr. Craig Venter, and The Institute for Genomic Research (TIGR) announced plans to form the genomics company Celera with a strategy based on completing the sequencing of the human genome in three years using a shotgun based approach (Perkin-Elmer, 1998). At the same time, the United States Human Genome Project announced revised goals to continue the exponential growth of sequencing data and provide a complete human genome by 2003 (Collins *et al*., 1998) in conjunction with the 5oth anniversary of the discovery of the double helix structure of DNA (Watson and Crick, 1953).

In February, 2001, both the public and private efforts announced completion of a rough draft of the human genome (International Human Genome Sequencing Consortium, 2001; Venter *et al*., 2001). As of July 30, 2001, the public sequencing efforts have finished 1.04 billion bases representing 47.1% of the human genome to a

Bermuda quality level (Smith and Carrano, 1996). Plans to assemble and orient the remaining 53% of the human genome from a rough draft state into a finished product by 2003 are still in effect (Collins *et al*., 1998).

Human sequence data is available in more refined forms than raw genomic sequence. In particular, it is also available as the sequence of gene products expressed in the cells known as Expressed Sequence Tags (ESTs) (Adams *et al*., 1991) and sequences of experimentally known and predicted mRNAs (Pruitt and Maglott, 2001). The sequence data available from each of these projects continues to grow. NCBI's dbEST (Boguski, Lowe and Tolstoshev, 1993) release 030802 contains 4.17 million entries of human ESTs (`http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html`). A total of 14,823 known and predicted mRNAs are available through NCBI's REFSEQ (Pruitt and Maglott, 2001) as of March 15, 2002.

Due to the large-scale availability of differing types of sequence data, a focus has been placed on mining and modeling sequence information in order to understand biological systems. Tools to handle and analyze large amounts of sequence data are needed.

## 1.2 Computational Biology

Computational biology is a multidisciplinary field, bringing together biologists, computer scientists, chemists, physicists, mathematicians and others together with a common goal of modeling and extracting information concerning biological systems. The NIH's Biomedical Information Science and Technical Initiative Consortium defines

computational biology as "The development and application of data-analytical and theoretical methods, mathematical modeling and computation simulation techniques to the study of biological, behavioral, and social systems." (http://grants.nih.gov/grants/bistic/CompBioDef.pdf) One aspect of computational biology that has come to the forefront in recent years is genome sequence analysis. Due to the efforts of both large-scale sequencing centers and individual scientists throughout the world, abundant resources of sequence data are now available. The methods described are rooted in the field of computational biology and are presented as techniques to aid in the discovery of biologically significant data.

## 1.3 Specific Aims

The specific questions we set out to answer concern human sequence assembly and organization. In particular, a sequence-based assembly approach is analyzed. In the process, overlapping assembled regions can be mined for single nucleotide polymorphisms (SNPs). Additionally, once assembled regions are available, they can be compared to restriction fragment digests to examine sequence assembly validation and the presence of large-scale polymorphisms. Compositional analysis is performed and methods for maintenance of high and low G+C regions of the human genome are studied.

The overview of the research chapters 2 through 6 follows. Each of these short sections introduces the problems that are set up in more detail in the appropriate chapters. Each chapter flows in an Introduction, Methods, Results and Discussion manner whether or not the sections are implicitly stated as such.

### 1.3.1 Overview of Chapter 2: Assembly of Genomic Contigs

The size of human chromosomes range from the 50 megabase (Mb) chromosome 21 to the 263 Mb chromosome 1 (Morton, 1991). The International Human Genome Consortium has employed clone-based sequencing strategies in order to sequence the euchromatic regions of the human genome chromosome by chromosome. Due to limitations of current clone-based sequencing techniques, the genome must be broken down into smaller portions in the range of 20 kilobases (kb) for cosmid clones (Collins and Bruning, 1978) to 200-300 kb for bacterial artificial chromosomes (BACs) (Shizuya *et al*., 1992) and yeast artificial chromosomes (YACs) (Burke, Carle and Olson, 1987). We attempt to collate a definitive set of non-redundant extended segments of finished human genomic sequence by taking individual human entries in GenBank greater than 10 kilobases (kb) and extending them on either end. As the sequencing of the rough draft data nears a close (Macilwain, 2000) and finished data comes to the front, we report on our experiences in dealing with the difficulties that arise when attempting to assemble contigs using a sequence-based approach.

In addition to our set of finished human genomic contigs, groups at NCBI and UCSC have undertaken the task of assembling the whole human genome through the incorporation of both finished and draft sequence data. A comparison of our assembly to these two assemblies is made. A detailed comparison of both of these public assemblies is performed at both the clone order and orientation level as well as at the sequence level. The discrepancies found indicate the degree of uncertainty that must be understood when incorporating unfinished sequence data.

### 1.3.2 Overview of Chapter 3: Single Nucleotide Polymorphisms

Single nucleotide polymorphisms (SNPs) occur when two or more different nucleotides are found at the same position within the population, i.e. a nucleotide substitution occurs. SNPs can be used as stable genetic markers within a population. SNPs occurring within coding regions can be used to analyze the relationship between genotype and phenotype (Picoult-Newberg *et al*., 1999). They are used as markers for a specific trait since they add genetic variation to a population.

The overlapping regions between two clones can lead to insight concerning possible SNPs. As a result, the construction of human genomic contigs is important in being able to detect specific locations of variation within the human population. I will present a method for determining possible SNPs sites when looking at the overlapping regions.

### 1.3.3 Overview of Chapter 4: Sequence Assembly Validation

Genomic sequence analysis depends on the accurate assembly of short (400 to 1000 base pair) sequence reads into contigs that cover extended regions as a necessary step in deriving finished sequence. Errors at the fragment layout assembly stage may be difficult or impossible to detect later in the editing process, and fragment assembly errors may have a serious impact on the biological interpretation of the data. Since assembly errors are difficult to detect and can impact the utility of the finished sequence, experimental validation of the fragment assembly is highly desirable. We propose a dynamic programming algorithm to match up experimental restriction fragments with

expected restriction fragments based on a reference sequence taken from the genomic contigs assembled previously.

### 1.3.4 Overview of Chapter 5: Breakpoint Segmentation

Once genomic sequence is available in either a rough draft (Kent and Haussler, 2001) or finished (Rouchka and States, 1999) state, we can begin to study how the human genome is constructed. One particular characteristic of interest is CpG islands, which are regions rich in the dinucleotide CG. These regions are interesting due to their association with upstream regions of genes. A method to detect and visualize CpG islands using log-likelihood and changepoint methods is given. Generalizations of this method can be applied to other compositional analysis as well.

### 1.3.5 Overview of Chapter 6: Compositional Analysis of Homogeneous Regions in Human Genomic DNA

The bulk of the genomic analysis lies within chapters 6 and 7. In chapter 6, we use the available human genome assemblies to study how the human genome is constructed into regions of homogeneous G+C content. We examine the previous isochore definitions of Bernardi (1993) that are based on density gradient centrifugation techniques. We show that a 5-class isochore definition is no longer applicable when sequence data is examined.

### 1.3.6 Overview of Chapter 7: Accounting for Regions of High and Low G+C Content Found in Human Genomic DNA

While sequence analysis indicates that a 5-class isochore system is too broad when human genomic sequence data is brought into play, there is still significant evidence in the presence of regions of high and low G+C composition within the human genome. In chapter 7, we examine two hypotheses for the maintenance of these regions by studying the G+C content of repetitive elements and by looking at the substitution rates between copies of repetitive elements and between genes and pseudogenes. Our results rule out the possibility of the G+C content of repetitive elements determining regions of high and low G+C regions in the human genome. We determine that there is a compositional bias for mutation rates. However, these biases are not responsible for the maintenance of high G+C regions. In addition, we show that regions of the human under less selective pressure will mutate towards a higher A+T composition, regardless of the surrounding G+C composition.

# Chapter 2

# Assembly of Genomic Contigs

## 2.1 Motivation

Since the beginning of the Human Genome Project (HGP) in 1990, the International Human Genome Sequencing Consortium has been using a clone-based strategy to sequence the human genome. Finished data is deposited into databases such as the DNA Data Bank of Japan (DDBJ) (Tateno *et al*., 2000), the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database (Stoesser *et al*., 2001), and GenBank (Benson *et al*., 2000). As more data has become available, the presence of overlapping clones, whether sequenced at the same center or different centers, have become more prevalent. We attempt to collate a definitive set of non-redundant extended segments of finished human genomic sequence by taking individual human entries in `GenBank` greater than 10 kilobases (kb) and extending them on either end. As the sequencing of the rough draft data nears a close (Macilwain, 2000) and finished data comes to the front, we report on our experiences in dealing with the difficulties that arise when attempting to assemble contigs using a sequence-based approach.

As of February 26, 2001, our largely automated process has resulted in 4,360 contigs covering a total of nearly 1081 megabases (MB) of non-redundant finished human genomic sequence. This figure represents 34% of the complete human genome

and includes nearly complete euchromatic data for chromosomes 21 and 22. Our sequence-based method was able to correctly piece together 92.73% of all fragments using a simulation study while at the same time avoiding any incorrect merging of two non-adjacent segments.

## 2.2 Introduction

The U.S. Human Genome Project, coordinated by the United States Department of Energy (DOE) and the National Institutes of Health (NIH), began in 1990 as a 15-year public venture to sequence the approximately three billion bases making up the human genome using clone-based techniques (Vaughan, 1996). As of February 26, 2001, 1081 million bases (34%) of the human genome has been sequenced to a Bermuda-quality (Smith and Carrano, 1996) finished state. In addition, a rough draft of the human genome has been announced as complete (Macilwain, 2000).

The International Human Genome Consortium has employed clone-based sequencing strategies in order to sequence the human genome. Due to limitations of current clone-based sequencing techniques, the genome must be broken down into smaller portions in the range of 20 kilobases (kb) for cosmid clones (Collins and Bruning, 1978) to 200-300 kb for bacterial artificial chromosomes (BACs) (Shizuya *et al*., 1992) and yeast artificial chromosomes (YACs) (Burke *et al*., 1987). Since a complete sequence of each human chromosome is desired, a method to assemble these smaller sequences into larger contiguous regions (contigs) is produced.

Physical maps of the human genome have been constructed using restriction fragment fingerprint data (The International Human Genome Consortium, 2001; Cheung *et al*., 2001; Stewart *et al*., 1997). Because of the large number of clones and limited information available from restriction fingerprints, this is a challenging task. In addition, clone tracking errors and microbiological contamination can lead to errors in the labeling of clones. Extended sequence overlaps are highly informative and provide a final arbiter as to how clones relate to one another. However, because the human genome contains regions of very recently duplicated sequence, even near identity sequence overlaps may be ambiguous.

An additional source of error is the presence of chimeric clones in the BAC collection. While chimeric clones are far less common in BACs than in YACs, they cannot be completely excluded. Chimeric clones can lead to false joins in assembly, potentially even placing sequence data on the wrong chromosome. Correlation of the sequence assembly with other map data is therefore a valuable source of confirmation.

Since December, 1998, we have been concerned with automating a process to assemble clones into contigs maintained at Washington University's Institute for Biomedical Computing, now known as The Center for Computational Biology. We report on the status of our work, as well as the limitations and difficulties we have faced in the last two years.

## 2.3 System and Methods

### 2.3.1 Contig Construction

GenBank is used as the reference database for the human genomic DNA used in building the contigs. The results are based upon release 122.0, which includes sequences submitted to GenBank up until February 15, 2001. In addition, we have downloaded all of the finished human genomic sequence data submitted between February 15, 2001 and February 26, 2001 to be included in our studies. The GenBank primate division is used in order to create stable human contigs based on finished data. In release 122.0, this is divided into gbpri1, gbpri2, gbpri3, gbpri4, gbpri5, gbpri6, gbpri7, gbpri8, and gbpri9. Table 2-1 shows a breakdown of the sequences in the primates division by sequence size.

**Table 2-1: Size of Primate GenBank Entries.** This table indicates the number of sequences in the primate divisions (gbpri1, gbpri2, gbpri3, gbpri4, gbpri5, gbpri6, gbpri7, gbpri8 and gbpri9) of GenBank release 122.0 as well as the human entries between February 15, 2001 and February 26, 2001.

| Sequence Size (in nucleotides) | Number of GenBank entries |
|---|---|
| > 200,000 | 490 |
| 150,000-199,999 | 2836 |
| 100,000-149,999 | 2939 |
| 75,000-99,999 | 1370 |
| 50,000-74,999 | 763 |
| 25,000 -49,999 | 1894 |
| 10,000-24,999 | 1093 |
| **TOTAL > 10,000** | 11,385 |

We create most of the contigs using an automated procedure highlighted in Figure 2-1. The first step is to retrieve human sequences from GenBank greater than 10 kb in length. After these sequences are retrieved their ends are searched against the primate division of GenBank for overlapping regions at least 70 base pairs (bp) long, and at least

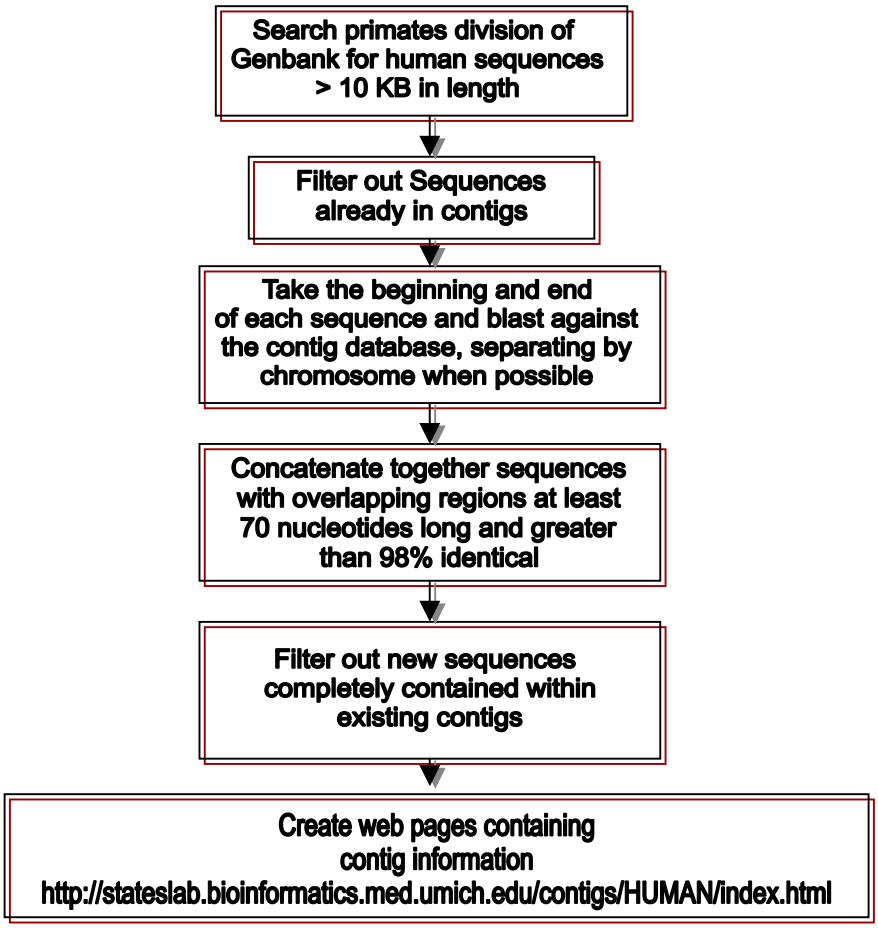



**Figure 2-1: Contig Creation Flowchart.** This figure indicates the steps that are followed in creating the human genomic contigs from `GenBank` entries.

98% identical. These searches are performed using `wublastn` version 2.0 (Gish, 1996-2001) with the user options `-gapw=256` and `-W=99`.

Contigs can be extended by looking for blast hits to their ends. When overlapping clones are found, they are merged together into a contig based on the alignment. The character N in the contig marks discrepancies in the alignment resulting from gaps and mismatches. In some cases, the restrictions need to be relaxed for automatic assembly to occur. Other contigs need to be assembled by hand in order to create the overlapping

region. Since the volume of sequencing data is growing exponentially, these steps are largely automated using Perl scripts.

Each assembled contig, including singletons, is noted in a contig description file. For each clone entry, the clone locus name, clone size, beginning and ending position in the current contig, strandedness (+ denotes the strandedness found in GenBank; - is its reverse complement), and the center at which sequencing took place. Table 2-2 indicates an example of a contig entry. The current list of contig descriptions can be downloaded at http://stateslab.bioinformatics.med.umich.edu/contigs/HUMAN/contigList.dat.

**Table 2-2: Sample Contig Entry.** Shown in this table is the entry for contig IBC_chr7-ctg51 dated 1/08/01. The first line lists the generated contig name, its size, and its cytogenetic position, if available. The second and third lines are historical and have no meaning at the current time. The NOTES line can contain various information about the clones, as entered by hand. Under the column headings "LOCUS LENGTH OVERLAP START END STRAND SOURCE" is a list of the individual GenBank entries used to create the contig. The first column list the locus name of the individual GenBank entry. The second column lists the length of the entry. The third column lists the overlap between two adjacent clones. If the overlap is a 100% identity, only a single value is given; otherwise, both the number of matching nucleotides and total number of nucleotides in the overlap are given. The fourth and fifth columns list the position of the given entry within the current contig. The sixth column lists the strandedness of the GenBank entry relative to the current contig, and the final column lists the sequencing center, if it can be automatically ascertained.

```
**************************************************************************
IBC_chr7-ctg51  (504,868)       7q22
GENOME CHANNEL:  ????
NCBI:            7ct113
NOTES:

 LOCUS         LENGTH    OVERLAP       START          END   STRAND  SOURCE
AC005072        69367       --            1         69367     +      WUGSC
AC005103       146394      200         69168        215661     +      ????
AC005086       129586      200        215462        345047     +      WUGSC
AF024533        84912  39828/39843    305207        390119     -      JENA
AF030453       125108    10359        379761        504868     +      JENA
**************************************************************************
```

## 2.3.2 Contig Validation

In order to test the validity of our sequence-based contig assembly algorithm, we attempted to assemble twelve different contigs extracted from our set of assembled

contigs dated 01/10/01 ranging in size from 2.0 MB to 5.5 MB (see results; Table 2-9). Rather than break the contigs up at the clone level, we randomly fragmented them using a uniform distribution into pieces ranging in size from 50 kb to 200 kb. A uniformly distributed overlap between segments of the size 100 bp to 20 kb was imposed. Once the fragments were created, sequencing errors and single nucleotide polymorphisms (SNPs) were introduced at a rate of 1/10000 bp and 1/2000 bp, respectively. All of these are followed in order to simulate the observed conditions between overlapping clones. Once all of the simulated fragments were created, they were piped through the contig assembly process and the resulting contigs were analyzed.

## 2.4 Results

### 2.4.1 Genomic Contig Database

GenBank release 122.0 contains 11,385 human genomic sequences greater than 10 kb in length. Table 2-1 indicates the breakdown of these clones. As of February 26, 2001, we have assembled a total of 4,360 contigs. These contigs cover a total of 1,080,908,685 bases. Note that there are more clones in the assembled contigs than entries in GenBank greater than 10 kb due to the fact that several contigs contain clones shorter than 10 kb. Most of these shorter clones were sequenced in order to close gaps between neighboring clones.

Table 2-3 indicates the breakdown of the contigs by their size. Most of the contigs are comprised of either one or two clones. There are sixteen examples that contain 20 or more clones, including a 33,626,454 base contig composed of 105 clones

**Table 2-3: Size of Generated Contigs.** The left-hand portion of this table indicates the number of contigs falling within a size range where the size is the number of GenBank entries that are concatenated together to produce them. The right-hand portion of the table reports the number of contigs falling within a certain size range, where the size is based on the number of nucleotides in the contig.

| Contig Size (in clones) | Number of contigs |
|---|---|
| 1 | 2691 |
| 2 | 791 |
| 3 | 325 |
| 4 | 192 |
| 5 | 101 |
| 6 | 61 |
| 7 | 45 |
| 8 | 37 |
| 9 | 22 |
| 10 | 16 |
| 11-20 | 63 |
| 20+ | 16 |

| Contig Size (in kilobases) | Number of contigs |
|---|---|
| 0-50 | 528 |
| 50-100 | 374 |
| 100-150 | 739 |
| 150-200 | 1251 |
| 200-300 | 608 |
| 300-400 | 368 |
| 400-500 | 139 |
| 500-1000 | 266 |
| 1000+ | 87 |

on chromosome 21 and a 23,109,284 contig composed of 334 clones on chromosome 22. Both of these chromosomes have been announced as complete (Hattori *et al*., 2000; Dunham *et al*., 1999) and contain only a few minor gaps.

Shown in Figure 2-2 is a plot of the number of contigs found of various size ranges. Also indicated is the total percentage of all finished human genomic sequence covered by contigs of various lengths. Since the majority of contigs (2691 out of 4360; 67.9%) are single clone contigs (singletons), the vast majority of contigs (2598 out of 4360; 59.6%) lie in the 100-300 kb range. Upon further examination, we see that although only 353 out of 4,360 (8.1%) contigs are greater than 500 kb, these contigs account for 36.5% of the total finished sequence available through the contig database.

The breakdown by chromosome is presented in Table 2-4. According to this data, the contigs cover about 34% of the human genome through February 26, 2001. Table 2-4

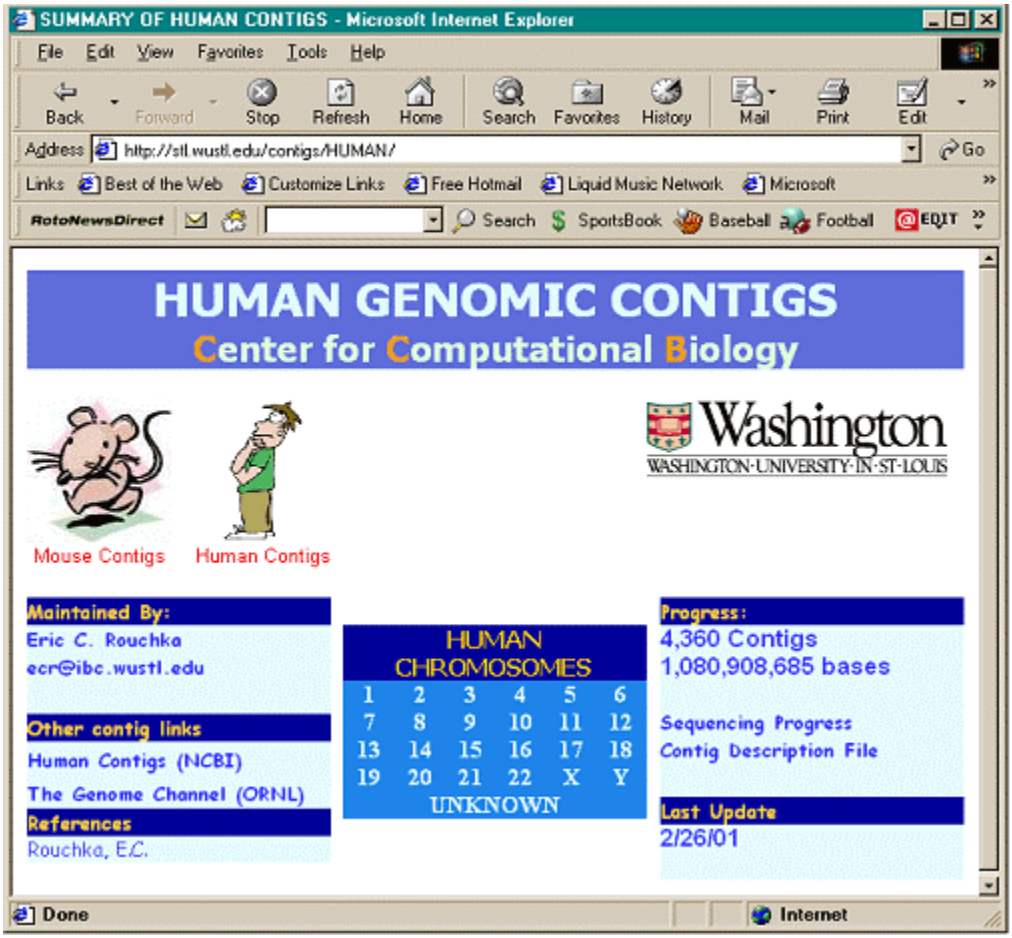



**Figure 2-2: Human Genomic Contigs Web Page.** Shown is a screen shot of the Human Genomic Contigs web page which can be found at the URL: `http://stateslab.bioinformatics.med.umich.edu/contigs/HUMAN/index.html`.

indicates that chromosomes 21 and 22 are complete, while chromosomes 6 and 7 have produced the largest amounts of sequence data.

In addition to the data presented in Table 2-4, there are 2417 additional sequenced clones that overlap contigs already assembled. Several of these refer to multiple entries under different accession numbers within `GenBank`. This data will be compared with the assembled clones. These extra sequences are useful in detecting SNPs. In addition,

they lend some information into the distribution of single nucleotide polymorphisms and mutational hotspots (Blackwell, Rouchka and States, 1999).

The growth of our sequence-based contigs since their inception has been linear (Figure 2-3). Using a projected linear growth based on the current finishing rates of 44.3 MB per month (the rate of growth from 2/29/00 to 2/26/01), the human genomic sequence will be prepared to a finished state in February 2005.

**Table 2-4: Current Sequencing Progress.** These figures are non-redundant finished sequence data taken from the Human Genomic Contigs Database (http://stateslab.bioinformatics.med.umich.edu/contigs/HUMAN/index.html) dated 2/26/2001. Note that the second column for the total euchromatic chromosome size is taken from NCBI. (http://www.ncbi.nlm.nih.gov/genome/seq/page.cgi?F=HsProgress.shtml&&ORG=Hs).

| Chromosome Number | Total Size (MB) | Aggregate Contig Length (MB) | Percent Completed |
|---|---|---|---|
| 1 | 263 | 64.59 | 24.5 |
| 2 | 255 | 61.14 | 23.9 |
| 3 | 214 | 35.22 | 16.4 |
| 4 | 203 | 18.69 | 9.2 |
| 5 | 194 | 57.13 | 29.4 |
| 6 | 183 | 112.04 | 61.2 |
| 7 | 171 | 114.55 | 66.9 |
| 8 | 155 | 15.77 | 10.1 |
| 9 | 145 | 36.88 | 25.4 |
| 10 | 144 | 25.53 | 17.7 |
| 11 | 144 | 22.52 | 15.6 |
| 12 | 143 | 45.15 | 31.5 |
| 13 | 98 | 44.19 | 45.0 |
| 14 | 93 | 60.87 | 65.4 |
| 15 | 89 | 9.40 | 10.5 |
| 16 | 98 | 27.58 | 28.1 |
| 17 | 92 | 32.87 | 35.7 |
| 18 | 85 | 6.30 | 7.4 |
| 19 | 67 | 38.53 | 57.5 |
| 20 | 72 | 58.23 | 80.8 |
| 21 | 34 | 35.05 | 103.1 |
| 22 | 34.5 | 35.27 | 102.2 |
| X | 164 | 89.54 | 54.6 |
| Y | 35 | 20.94 | 59.8 |
| UNKNOWN | N/A | 12.79 | N/A |
| **TOTALS** | **3175** | **1080.90** | **34.0** |

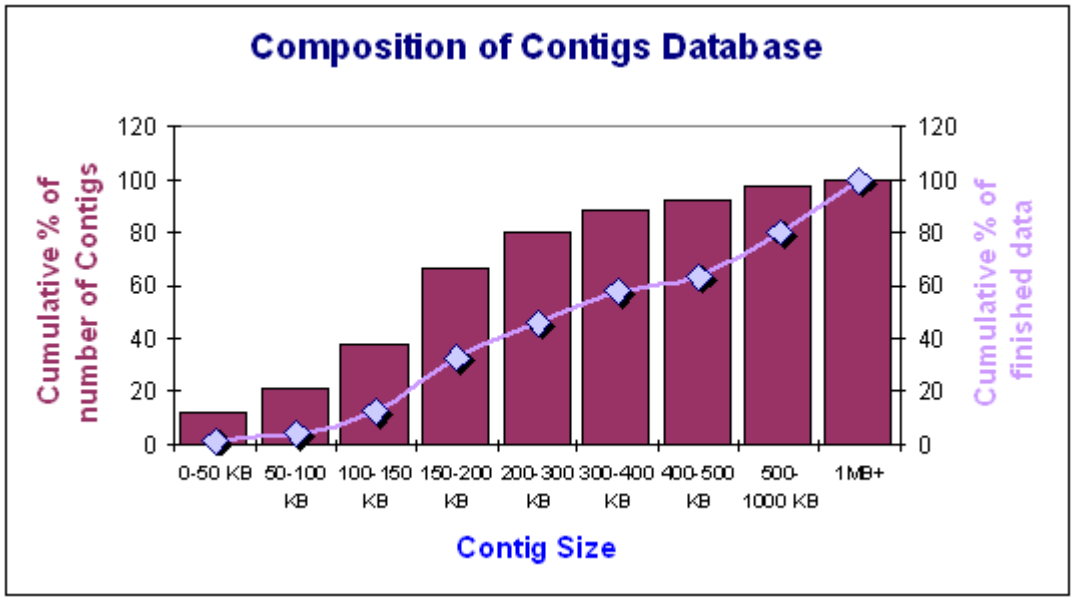



**Figure 2-3: Composition of Contigs Database.** Shown in the x-axis is the various size groupings of contig length. There are two y-axes: the one to the left represents the cumulative percentage of the total number of contigs falling into a particular size range (corresponding to the bar data). The y-axis to the right indicates the cumulative percentage of finished data falling into the various size ranges (corresponding to the diamond data). This graph shows that the majority of contigs lie in the 100-300kb range, which is the expected range for single clone contigs. It is also shown that while there are few contigs > 400 kb in length (8.1%), they still account for a large percentage (36.5%) of all of the finished data.

### 2.4.2 Difficulties in Contig Assembly

**Overlapping Clone Information.** Some genome sequencing centers incorporate neighboring clone information into their `GenBank` entries. Table 2-5 shows some examples of how this data is entered into the comments section. Use of this information could help in the creation of genome contigs. However, as Table 2-5 indicates, this data is not standardized among the sequencing centers. The data is entered by hand in a manner that is easy for a human to read, but not easily parsed by a computer. The

overlap between two clones, if given, is present only in a positional manner. An alignment between two overlapping clones is not given.

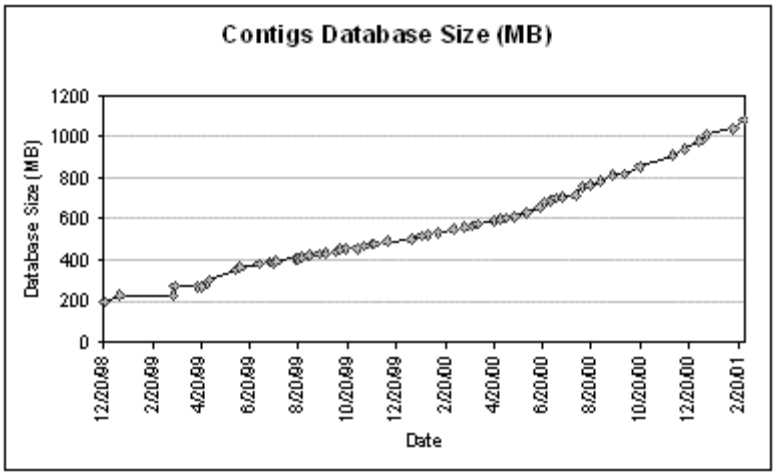



**Figure 2-4: Growth of Contigs Database.** Shown in this figure is the growth of the non-redundant contigs database from 12/20/98 to 2/26/01. Each data point represents an update to the contigs database.

Assembly of adjacent clones into larger contigs is not always a straightforward process. For instance, the orientation of two adjacent clones might be different. Our routines handle both the forward and reverse complement of each `GenBank` entry when assembling contigs.

The length of the overlap between two adjacent clones varies greatly. Some sequencing centers such as Washington University Genome Sequencing Center (WUGSC) and Sanger Centre have a relatively constant sequence overlap length for known overlapping sequences. (In the case for WUGSC it is 200 bp; for Sanger Centre it is 100 bp.) For the assembled contigs, the size ranges from 0 base pair overlaps from the Japan Science and Technology Corporation efforts on chromosome 21 to a 155,954 base

**Table 2-5: Overlapping Clone Information.** The second column contains examples of overlapping clone information contained within the COMMENT section of the GenBank reports for the GenBank entries located in the second column. The overlapping clone information is typical for the sequencing centers shown in the first column.

| GenBank Accession | Overlapping Information In COMMENT section |
|---|---|
| Z99715 | The true right end of clone 1114G22 is at 104.<br>The true left end of clone 262D12 is at 51983. |
| AC004398 | Overlapping Sequences:<br>5': UWGC: g1248a010 (Accession: AC004107)<br>3': UWGC: g1248a139 |
| AC005303 | Only 90.0 kilobases from the middle of this clone are being submitted. The remainder overlaps either accession AC003664 (WICGR project L281) or accession AC005277 (WICGR project L351). |
| AC002378 | NEIGHBORING SEQUENCE INFORMATION:<br>The clone being sequenced to the left is BK085E05; the clone being sequenced to the right is DJ102K02. Actual start of this clone is at base position 1 of DJ438O4. |
| AC002523 | Begining of sequence overlaps with AF007262, end of sequence overlaps with AF011889.<br><br>(*Note that Beginning is misspelled here*) |

overlap between GenBank accession AC012634 and AC004782 from Lawrence Berkley National Labs on chromosome 5. Note that those sequences with less than a 70 base pair overlap are hand assembled. The GenBank entries for these sequences have been used to aid in the detection and assembly of these contigs. For the shorter overlapping segments, running wublastn to find the alignment between two sequences takes a matter of seconds, but for larger regions, the time spent to find the alignment can take hours.

**Repetitive Elements.** Repetitive elements pose a serious problem in assembling contigs. The composition of the human genome is at least 35% repetitive elements (Jurka, 1998). These can come in the form of interspersed repeats (Smit, 1999) as well as large regions of chromosome specific (Shakh *et al*., 2000) and human specific (Choo *et al*., 1988)

repetitive elements. Table 2-6 indicates a partial list of clones that cannot be extended, due to the fact that their ends contain interspersed repeats (SINES such as ALUs or LINES). Table 2-7 indicates a list of clones on various chromosomes whose overlaps cannot be resolved due to the occurrence of large-scale repeats occurring only on that particular chromosome, or uniquely within the human genome. As a result, the end of the clones listed in Tables 2-6 and 2-7 match multiple clones and the true neighboring clone cannot be determined.

**Table 2-6: `Genbank` Clones with Repetitive Elements at the Ends.** Shown in this table is a list of finished human genomic clones which have a previously defined repetitive element sequence at one or both end(s). Such clones cannot be extended, due to the inability to determine which overlapping clone is its true neighbor.

| GenBank Accession | Repeat Family |
|---|---|
| AC006525 | LINE1 |
| HSJ433F14 | ALU |
| HS503N11 | LINE1 |
| HS1043E3 | LINE1 |
| HS179P9 | LINE1 |
| HS271G9 | LINE1 |
| AC004935 | ALU-Sb; ALU-Sc |
| AC002461 | LINE1 |
| AC007459 | ALU-Sb |
| HUM7501 | LINE1 |
| AC000100 | LINE1 |
| HSU161B10 | LINE1 |
| HS296K21 | LINE1 |
| HS884M20 | LINE1 |
| HSV602D8 | LINE1 |
| HSV618H1 | LINE1 |
| HSAF002997 | LINE1 |
| HSU86H4 | LINE1 |
| HSU19F10 | LINE1 |
| HS1168A5 | LINE1 |
| AF068624 | ALU |
| AF036876 | LINE1 |
| AC004389 | LINE1 |

**Table 2-7: `GenBank` Clones with Human-Specific Repeats at the Ends.** This table indicates those finished human genomic clones with a previously unidentified human specific repeat occurring at one or both end(s). These clones cannot be extended due to the occurrence of multiple clones that could be the adjacent clone.

| GenBank Accession | Repeat Classification |
|---|---|
| AF186194 | HUMAN SPECIFIC |
| AC002402 | HUMAN SPECIFIC |
| U73649 | HUMAN SPECIFIC |
| AC010196 | HUMAN SPECIFIC |
| HUAC002544 | CHR16 SPECIFIC |
| HUAC002045 | CHR16 SPECIFIC |
| HUAC002425 | CHR16 SPECIFIC |
| AC015853 | HUMAN SPECIFIC |
| HS138B7 | HUMAN SPECIFIC |
| AC012398 | CHR22 SPECIFIC (BOTH ENDS) |
| AC007981 | HUMAN SPECIFIC |
| AC023490 | HUMAN SPECIFIC |
| AC007324 | CHR22 SPECIFIC |
| HSA191C22 | HUMAN SPECIFIC |
| HS179D3A | HUMAN SPECIFIC |
| HS411B6 | HUMAN SPECIFIC |
| AC006314 | HUMAN SPECIFIC |
| HS884M20 | HUMAN SPECIFIC |

Less frequently observed are recent duplications between two chromosomes. We have observed and studied one such region involving two overlapping clones originating from two separate chromosomes in detail. The first entry is `GenBank` accession AL021921 and the second entry is `GenBank` accession U95738. The 135 kb AL021921 is sequenced by Sanger Centre and is annotated as 1p36.13. The 171 kb entry U95738 is sequenced by The Institute for Genome Research (TIGR) and is annotated as 16p13.11. According to the blast hits, AL021921 lies completely within U95738 with 100 mismatches, 74 of which are transitions (A⇔G; C⇔T) and 26 are transversions (A⇔T, G⇔C, A⇔C, G⇔T). There are also 22 gaps composed of 123 indel events. At random, it is expected to have twice as many transversions as transitions. However, in this case, there are almost three times as many transitions as transversions. In addition, the 105 kb

GenBank accession AL161638, annotated as 1p34.2-35.3 and sequenced by the Sanger Centre, overlaps AL021921 with a 100% 100 bp overlap. The beginning of AL161638 overlaps the end of U95738 with 25405 matches, 57 mismatches (26 transitions and 31 transversions) and 20 gaps composed of 431 indel events. The higher number of transitions in both of these cases indicates a possible evolutionary relationship (Kimura, 1980).

**Polymorphisms.** One of the major challenges in assembling contigs is the occurrence of polymorphisms in the human population. These can range from single nucleotide polymorphisms (SNPs) to large-scale polymorphisms. In most cases, large-scale polymorphisms occurring between two adjacent clones result from differences in repeat copy numbers. However, there are also large insertion and deletion events occurring between adjacent clones. A dramatic example occurs on chromosome 22 between GenBank accessions AP000351 and AP000352, both sequenced at Keio University in Tokyo, Japan. The GenBank record for AP000351 indicates a 94,726 base pair overlap with AP000352, while the GenBank record for AP000352 indicates a 40,455 base pair overlap with AP000351. Blast analysis on these two sequences indicates the end of AP000351 overlaps with the beginning of AP000352 with a 55,248 base insertion in AP000351. Table 2-8 indicates the beginning and ending positions of the overlap.

Since clones may not overlap with 100% identity due to sequencing errors and polymorphisms, we have crafted our scripts to allow for overlapping sequences greater than 98% identical. This is an empirical cutoff, which reduces spurious matches, while

**Table 2-8: Overlapping Clones with 50kb Insertion.** Shown in this table are the corresponding beginning and ending nucleotide positions of two overlapping clones on chromosome 22 sequenced at Keio University in Tokyo, Japan. Note that the beginning of AP000352 matches the end of AP000351, with an additional 50kb insertion in AP000351. The GenBank entries for these two clones list them as being adjacent to one another.

| CLONE ACCESSION | CLONE LENGTH | BEGIN OF OVERLAP | END OF OVERLAP | BEGIN OF OVERLAP | END OF OVERLAP |
|---|---|---|---|---|---|
| AP000351 | 118,999 | 24,274 | 53,787 | 108,035 | 118,999 |
| AP000352 | 152,244 | 1 | 29,527 | 29,528 | 40,455 |

allowing for naturally occurring single nucleotide polymorphisms at a rate of 7/1000 (Taillon-Miller, *et al*., 1998) and acceptable sequencing error rates of 1/10000 (Collins, *et al*., 1998). Some overlaps such as the example on chromosome 22 can still be missed through this automated process, but most overlapping segments should be detected.

**Mislabeled GenBank Entries.** One of the difficulties in relying on physical map data in the annotation sections of GenBank entries is that these data are not completely reliable. We have uncovered at least two instances where it appears that GenBank entries have been mislabeled. These clones were discovered while looking for overlapping clones from different chromosomes forming chimeric contigs. In one case, the sequencing center involved acknowledged the missanotation and has since updated the GenBank entry. The second case appears to have arisen from a data-tracking problem where clones from two different chromosomes with similar names were confused.

## 2.4.3 Contig Assembly Validation

Table 2-9 summarizes the results of contig assembly validation. For the contig assembly, the 12 original contigs to reassemble were broken down into 356 fragments.

As a result, there are 344 total expected merges between fragments. A total of 319 true merges were calculated, leaving a total of 25 false negatives. In addition, there were no false merges (false positives) calculated. The resulting sensitivity, calculated as the number of correctly calculated merged segments divided by the number of known merged segments, is 319/344, or 92.73%. The specificity, calculated as the number of correctly calculated merged segments divided by the total number of predicted merged segments, is 319/319, or 100%. These findings suggest that our model is highly specific, while producing an acceptable level of sensitivity. This supports our methods as a valid approach to assemble individual GenBank entries into larger contiguous regions.

**Table 2-9: Contig Assembly Validation.** Shown is the list of contigs used for contig assembly validation and their respective sizes, in nucleotides. The contigs are taken from the set of IBC contigs dated 1/10/01. The fourth column indicates the number of expected merge events. Column five and six indicate the number of merges found and the number of merges missed, respectively. The eighth column indicate the true merge rate (sensitivity), calculated as the number of true merges found divided by the number of merges expected. Since there are no false merges found, the specificity is 100%.

| CONTIG NAME | CONTIG SIZE | Total Frags | Merges Found | Merges Missed | False Merges | True Merge Rate | False Merge Rate |
|---|---|---|---|---|---|---|---|
| IBC_chr14-ctg5 | 2,444,856 | 23 | 20 | 2 | 0 | **90.9%** | 0% |
| IBC_chr14-ctg50 | 2,087,975 | 15 | 14 | 0 | 0 | **100%** | 0% |
| IBC_chr17-ctg2 | 2,834,939 | 27 | 25 | 1 | 0 | **96.2%** | 0% |
| IBC_chr20-ctg12 | 5,549,661 | 52 | 46 | 5 | 0 | **90.2%** | 0% |
| IBC_chr20-ctg20 | 5,530,385 | 45 | 40 | 4 | 0 | **90.9%** | 0% |
| IBC_chr22-ctg11 | 2,488,705 | 24 | 23 | 0 | 0 | **100%** | 0% |
| IBC_chr6-ctg1 | 4,562,704 | 44 | 40 | 3 | 0 | **93.02%** | 0% |
| IBC_chr7-ctg1 | 2,044,635 | 20 | 15 | 4 | 0 | **78.95%** | 0% |
| IBC_chr7-ctg34 | 2,880,961 | 23 | 20 | 2 | 0 | **90.91%** | 0% |
| IBC_chr7-ctg49 | 2,204,146 | 20 | 17 | 2 | 0 | **89.47%** | 0% |
| IBC_chrY-ctg10 | 4,210,264 | 39 | 38 | 0 | 0 | **100%** | 0% |
| IBC_chrY-ctg3 | 3,063,814 | 24 | 21 | 2 | 0 | **91.3%** | 0% |
| **TOTAL** | **39,903,045** | **356** | **319** | **25** | **0** | **92.73%** | **0%** |

## 2.5 Discussion

### 2.5.1 Whole Genome Assemblies

Since the inception of the IBC Finished Genomic Contig Data set in 1998, other groups including the National Center for Biotechnology Information (Jang *et al*., 1999), Oak Ridge National Labs (Mural *et al*., 1999), The University of California-Santa Cruz (Kent and Haussler, 2001) and Celera Genomics (Venter *et al*, 2001) have entered the arena of assembling human genomic contigs. In the case of the UCSC's GoldenPath Working Draft data and the more recent NCBI assemblies, high throughput genomic sequence (HTGS) is incorporated to create a whole genome assembly, even though over 50% of human genomic data is available only in a rough-draft form.

Both GigAssembler, which is the algorithm used to construct the GoldenPath contigs, and NCBI incorporate additional information besides sequence similarity in ordering and orienting genomic sequences relative to one another. The information used by GigAssembler includes the alignments of mRNA, paired plasmid ends, ESTs and BAC end pairs as well as additional information (Kent and Haussler, 2001). NCBI takes advantage of clone-overlap information provided by the genome centers in their clone annotation as well as looking for STS markers and BAC end pairs in their assembly (Jang, *et al*., 1999). Additional information may be incorporated into the current NCBI assembly.

### 2.5.2 Comparison to Whole Genome Assemblies

Since our assemblies do not incorporate any mapping information, they cannot be ordered and oriented relative to one another. Hence, when comparing our contigs to the NCBI and UCSC contigs respectively, we order them in the following manner: for each of our contigs, we take the first clone listed in the contig. All of these first clones are then packed together and ordered according to where they are placed in the NCBI or UCSC contig relative to one another. This position then denotes the ordering of all of the IBC contigs. Since the NCBI and UCSC assemblies are not in complete agreement, this is done two times: once when comparisons are made to the NCBI data set, and once when comparisons are made to the UCSC data set.

**Clone Ordering Comparison**. Clone ordering comparisons are graphically shown by drawing a polygon between the absolute positioning of a clone on the IBC data set to the absolute positioning within the reference set. Figure 2-5 shows the results of such a clone ordering comparison for chromosomes 7, 20 and 21. A complete set of clone comparison graphs is available at `http://sapiens.wustl.edu/~ecr/COMARE/`. It can be seen from these results that clone ordering within finished contigs is consistent. However, when rough-draft data is incorporated into the genomic assemblies, inconsistencies start to arise, even when these assemblies are aided by mapping information.

**Sequence Level Comparison.** Whole genome sequence comparisons are made using a tool called multi (States, unpublished). multi creates a deterministic finite automaton

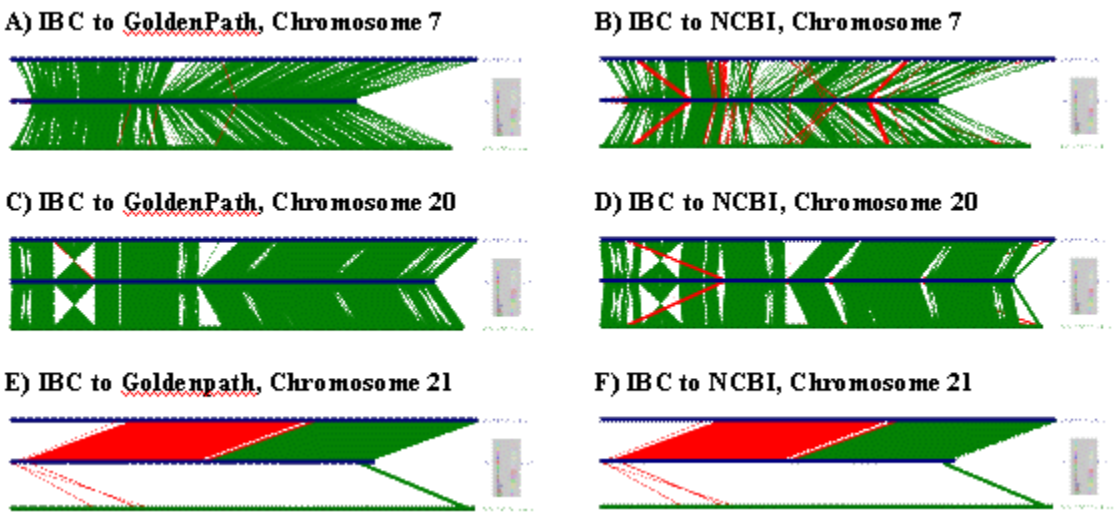



**Figure 2-5: Clone Ordering Comparison.** Shown in each of the six images is a comparison between the UCSC Goldenpath clone position (top) to the IBC clone position (middle) and the IBC clone position to the NCBI clone position (bottom). For images A, C and E, the IBC contig ordering used is adjusted according to the Goldenpath clone ordering. In images B, D and F, the IBC contig ordering is based on the NCBI clone ordering. In image A, there is only one disagreement between the UCSC and IBC clone orderings, and several disagreements between the IBC and NCBI clone orderings. In image B, there are several disagreements between the Goldenpath and IBC, and fewer between the IBC and NCBI clone orderings. These discrepancies are a result in disagreements between the Goldenpath and NCBI clone orderings, which is shown in Figure 2-8. Figures C and D show indicate that some of the IBC clones are in opposite orientation with respect to the IBC and NCBI orientation. The large gaps in figures E and F are the result of different clone names used in the NCBI assembly.

(DFA) which is searched for exact matches of a specified length. Since two assemblies of the human genome are being compared, the match length is set to 1000 and the window size is set to 500. Graphical results of the multi output for chromosomes 7, 20 and 21 are shown in figure 2-6.

### 2.5.3 Comparison of NCBI and GoldenPath Assemblies

In an ideal situation, there would be only one way for the clone pieces of the genomic puzzle to fit together. However, due to events such as repetitive elements (Smit, 1999), gene duplication (Lynch and Conery, 2000) and segmental duplications (Bailey *et al*.,

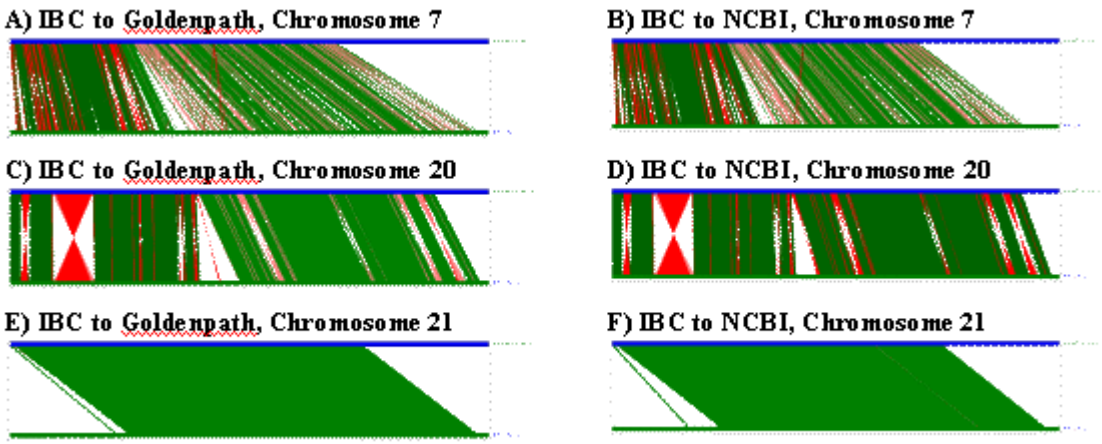



**Figure 2-6: Sequence Level Comparison.** Shown in images A, C and E are comparisons between the Goldenpath sequence (top) and the IBC sequence (bottom). Images B, D and F show a comparison between the NCBI sequence (top) and the IBC sequence (bottom). Each line indicates a perfect match between the two assemblies of at least 1000 nucleotides. The graphs were constructed from data resulting from multi. These graphs show that the biggest discrepancies are the result of individual clone orientation. Green represents matches in the same orientation and red represents matches in opposite orientations.

2001) there is nonrandomness associated with human genomic data. This makes it difficult to verify whether or not two clones do indeed belong in a contig or they just happen to have some similarities in their ends. This will become a more prevalent problem as more and more finished data becomes available through the Human Genome Project.

In order to illustrate the difficulty involved with whole genome assembly, a comparison was made between UCSC Goldenpath's April, 2001 release and NCBI's MapViewer build 22 (April 1, 2001) assemblies at a clone ordering and sequence similarity level. As figure 2-7 indicates, there are widespread inconsistencies in clone ordering. This is especially evident with chromosomes X and Y. Other chromosomes at or near completion as of the April releases indicate a greater level of consistency in clone ordering, such as 20, 21 and 22. Even so, there are still areas where inversions of clones

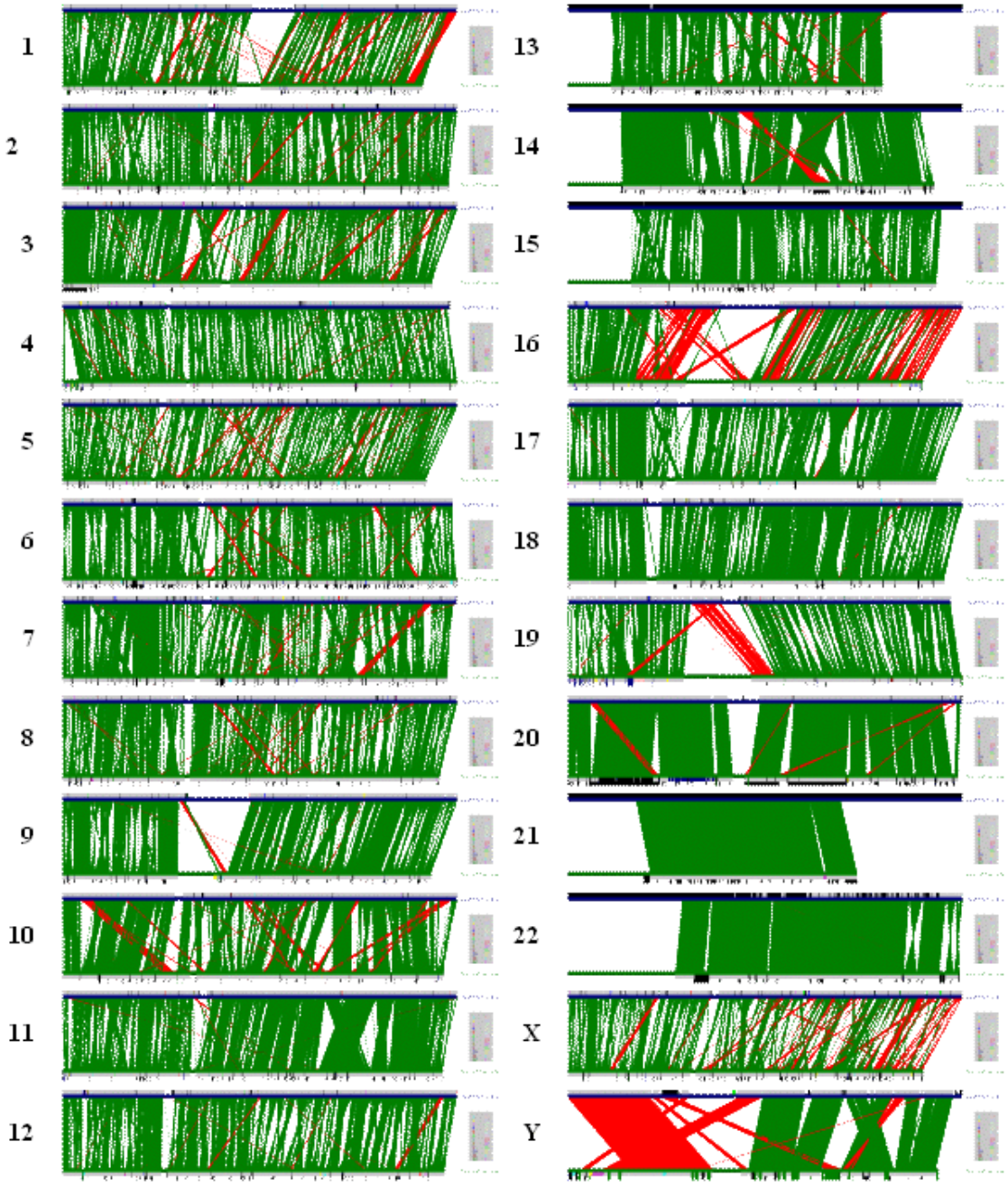



**Figure 2-7: NCBI Build 22 vs. GoldenPath April 2001 Clone Ordering Comparisons.** Shown in each one of these images is a graph relating the location of clones in the GoldenPath assembly (top) to their location in the NCBI assembly (bottom). If the clone position on both assemblies is within 10%, then the polygon is drawn in green. If the clone position between assemblies differs greater than 10%, the polygon is drawn in red.

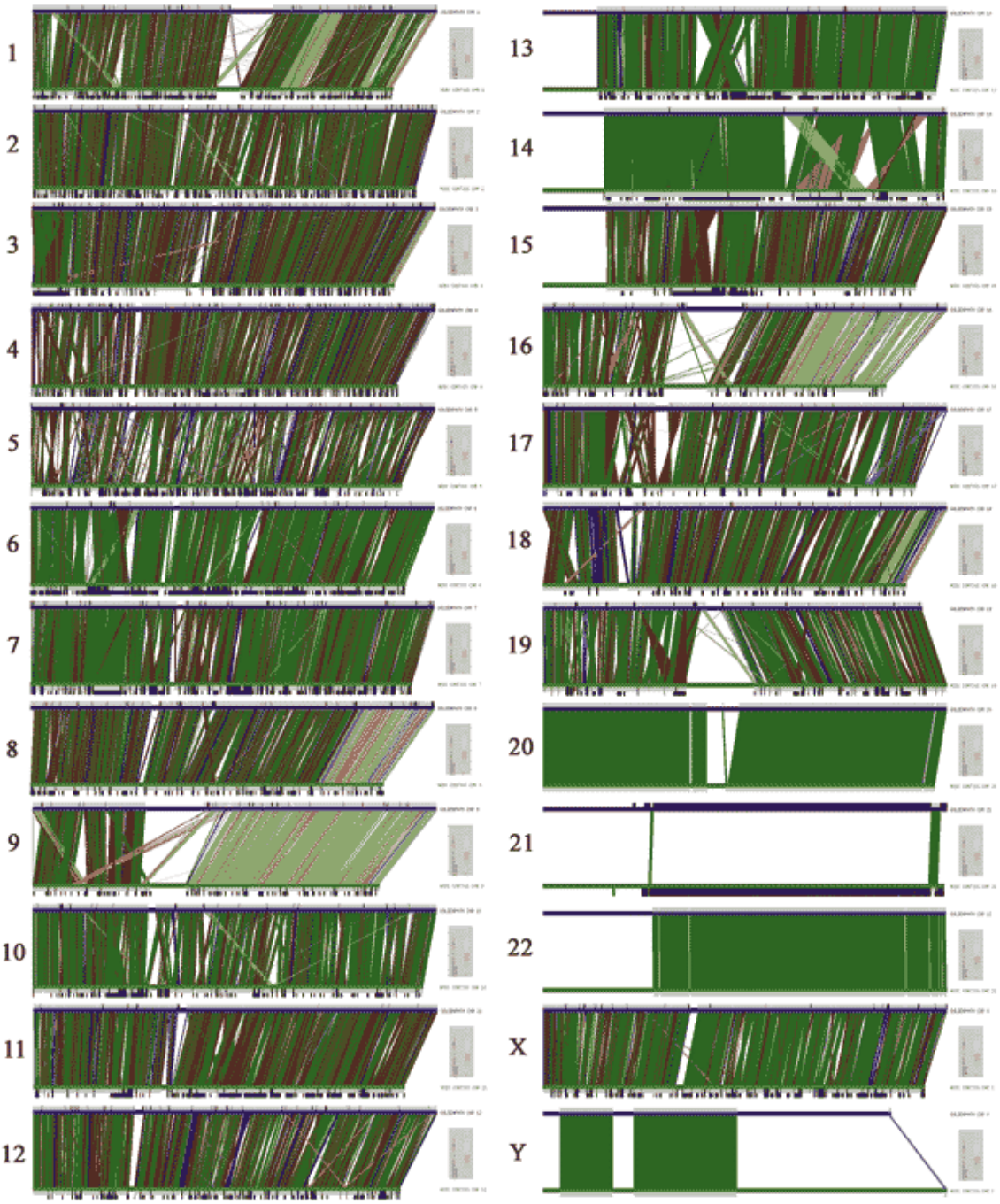



**Figure 2-8: NCBI Build 26 vs. Goldenpath August 2001 Clone Ordering Comparisons.** Shown in each of these images is a graph relating the location of clones in the Goldenpath assembly (top) to their location in the NCBI assembly (bottom). In this figure, clone orientation data is included as well. If the orientation of both clones is the same, they are colored green. If they are different, they are colored red. If the orientation is unknown, it is drawn in blue. If the difference between the clone locations on the two assemblies differs by more than 10%, it is drawn in a lighter color.

**Table 2-10: Summary of Accessions Used in the August 6, 2001 Goldenpath Assembly.** The second column indicates the total number of GenBank sequences used by UCSC to construct the chromosome labeled in the first column. The third column indicates the number of sequences UCSC uses that remain unordered in the NCBI assembly. The fourth column indicates the number of sequences assigned to a chromosome by UCSC that NCBI labels as unknown. The fifth column lists those sequences used by both UCSC and NCBI that are identical, but different accession ids are given. The sixth column lists those sequences that UCSC assigns to one chromosome and NCBI assigns to another. The seventh column indicates the total number of sequences used in the UCSC chromosome assembly that are not found anywhere in the NCBI assembly. The final column lists the total number of sequences that are used in both assemblies.

| Chromosome | Total used | Unordered | Unknown | Different Accession | Different Chromosome | Unmatched Accessions | Matched Accessions |
|---|---|---|---|---|---|---|---|
| **1** | 2704 | 20 | 0 | 11 | 20 | 71 | 2593 |
| **2** | 1965 | 6 | 0 | 50 | 13 | 82 | 1864 |
| **3** | 2004 | 22 | 5 | 28 | 19 | 121 | 1837 |
| **4** | 1723 | 21 | 5 | 150 | 26 | 69 | 1602 |
| **5** | 2084 | 37 | 0 | 0 | 7 | 42 | 1998 |
| **6** | 1932 | 11 | 3 | 0 | 13 | 37 | 1868 |
| **7** | 1561 | 10 | 2 | 10 | 15 | 46 | 1488 |
| **8** | 1444 | 17 | 17 | 0 | 19 | 37 | 1354 |
| **9** | 1117 | 13 | 0 | 0 | 12 | 46 | 1046 |
| **10** | 1300 | 15 | 0 | 1 | 5 | 21 | 1259 |
| **11** | 1666 | 12 | 3 | 0 | 8 | 17 | 1626 |
| **12** | 1323 | 10 | 1 | 0 | 13 | 44 | 1255 |
| **13** | 893 | 2 | 0 | 0 | 8 | 12 | 871 |
| **14** | 678 | 1 | 0 | 0 | 3 | 9 | 665 |
| **15** | 826 | 2 | 0 | 0 | 12 | 18 | 794 |
| **16** | 856 | 9 | 0 | 2 | 10 | 33 | 804 |
| **17** | 763 | 5 | 1 | 0 | 8 | 5 | 744 |
| **18** | 968 | 10 | 0 | 0 | 9 | 8 | 941 |
| **19** | 819 | 5 | 1 | 0 | 1 | 14 | 798 |
| **20** | 629 | 0 | 0 | 0 | 0 | 0 | 629 |
| **21** | 103 | 0 | 0 | 0 | 0 | 99 | 4 |
| **22** | 527 | 0 | 0 | 0 | 0 | 0 | 527 |
| **X** | 1465 | 9 | 1 | 0 | 6 | 16 | 1433 |
| **Y** | 200 | 0 | 0 | 0 | 0 | 0 | 200 |
| **TOTALS** | **29,550** | **237** | **39** | **252** | **227** | **847** | **28,200** |

seem to be occurring. Figure 2-8 shows a clone ordering comparison of NCBI build 26 to the Goldenpath August 2001 release. When figures 2-7 and 2-8 are compared, it can be seen that as sequences reach a finished state, the assemblies merge to agreement. This is particularly evident when looking at the assemblies of chromosomes 20 and X. Note

that in figure 2-8, that the NCBI and Goldenpath assemblies use different clones in order to create chromosome 21, thus leading to only a few matches in the clone ordering comparison. A summary of the GenBank entries used in the August 2001 Goldenpath and NCBI build 26 assemblies are given in Tables 2-10 and 2-11, respectively. Table 2-12 summarizes the orientation agreements when comparing the August 6, 2001 Goldenpath assembly to the NCBI build 26.

**Table 2-11: Summary of Accessions Used in the NCBI Build 26.** The second column indicates the total number of GenBank sequences used by NCBI to construct the chromosome labeled in the first column. The third column indicates the number of sequences NCBI uses that remain unordered in the UCSC assembly. The fourth column indicates the number of sequences assigned to a chromosome by NCBI that UCSC labels as unknown. The fifth column lists those sequences used by both UCSC and NCBI that are identical, but different accession ids are given. The sixth column lists those sequences that NCBI assigns to one chromosome and UCSC assigns to another. The seventh column indicates the total number of sequences used in the NCBI chromosome assembly that are not found anywhere in the UCSC assembly. The final column lists the total number of sequences that are used in both assemblies.

| Chromosome | Total used | Unordered | Unknown | Different Accession | Different Chromosome | Unmatched Accessions | Matched Accessions |
|---|---|---|---|---|---|---|---|
| **1** | 3088 | 9 | 1 | 11 | 30 | 455 | 2593 |
| **2** | 2134 | 3 | 0 | 50 | 15 | 252 | 1864 |
| **3** | 2078 | 2 | 0 | 28 | 7 | 232 | 1837 |
| **4** | 1765 | 11 | 5 | 150 | 17 | 130 | 1602 |
| **5** | 2348 | 13 | 1 | 0 | 14 | 322 | 1998 |
| **6** | 2337 | 7 | 0 | 0 | 12 | 450 | 1868 |
| **7** | 1716 | 1 | 0 | 10 | 9 | 218 | 1488 |
| **8** | 1556 | 3 | 0 | 0 | 10 | 189 | 1354 |
| **9** | 1193 | 1 | 1 | 0 | 9 | 136 | 1046 |
| **10** | 1463 | 4 | 2 | 1 | 11 | 187 | 1259 |
| **11** | 1970 | 20 | 0 | 0 | 19 | 305 | 1626 |
| **12** | 1438 | 2 | 0 | 0 | 14 | 167 | 1255 |
| **13** | 1078 | 0 | 0 | 0 | 8 | 199 | 871 |
| **14** | 817 | 3 | 0 | 0 | 5 | 144 | 665 |
| **15** | 891 | 0 | 0 | 0 | 3 | 94 | 794 |
| **16** | 894 | 4 | 0 | 2 | 5 | 71 | 804 |
| **17** | 816 | 14 | 0 | 0 | 7 | 51 | 744 |
| **18** | 1055 | 2 | 0 | 0 | 2 | 110 | 941 |
| **19** | 901 | 14 | 0 | 0 | 12 | 77 | 798 |
| **20** | 629 | 0 | 0 | 0 | 0 | 0 | 629 |
| **21** | 475 | 78 | 4 | 0 | 5 | 384 | 4 |
| **22** | 527 | 0 | 0 | 0 | 0 | 0 | 527 |
| **X** | 1661 | 7 | 1 | 0 | 13 | 207 | 1433 |
| **Y** | 200 | 0 | 0 | 0 | 0 | 0 | 200 |
| **TOTALS** | **33,020** | **198** | **15** | **252** | **227** | **4,380** | **28,200** |

**Table 2-12: GenBank Entry Orientations.** The orientation of a GenBank entry is considered consistent if the entry occurs in the same orientation in both the NCBI and Goldenpath assemblies. An inconsistent orientation occurs when the orientation of the entry is different in both assemblies. In the case that the orientation is marked as unknown in at least one of the assemblies, the entry is marked with an unknown orientation. The distance threshold used means that the GenBank entry positions must agree within 10% in both assemblies.

| | Consistent Orientation | | Inconsistent Orientation | | Unknown Orientation | |
|---|---|---|---|---|---|---|
| **Chromosome** | **within Threshold** | **outside Threshold** | **within Threshold** | **outside Threshold** | **within Threshold** | **outside Threshold** |
| **1** | 1327 | 269 | 802 | 143 | 37 | 4 |
| **2** | 1125 | 37 | 558 | 13 | 78 | 3 |
| **3** | 906 | 29 | 701 | 18 | 151 | 4 |
| **4** | 690 | 8 | 588 | 2 | 160 | 4 |
| **5** | 1008 | 60 | 676 | 40 | 176 | 38 |
| **6** | 1562 | 44 | 233 | 12 | 17 | 0 |
| **7** | 1229 | 1 | 209 | 0 | 39 | 0 |
| **8** | 623 | 123 | 395 | 100 | 81 | 32 |
| **9** | 224 | 496 | 111 | 196 | 8 | 11 |
| **10** | 856 | 43 | 298 | 14 | 46 | 1 |
| **11** | 924 | 2 | 642 | 1 | 55 | 2 |
| **12** | 728 | 8 | 385 | 11 | 122 | 1 |
| **13** | 728 | 1 | 122 | 0 | 20 | 0 |
| **14** | 582 | 36 | 18 | 25 | 4 | 0 |
| **15** | 453 | 0 | 313 | 3 | 25 | 0 |
| **16** | 351 | 195 | 155 | 81 | 7 | 13 |
| **17** | 414 | 5 | 271 | 3 | 39 | 12 |
| **18** | 481 | 36 | 324 | 22 | 71 | 7 |
| **19** | 589 | 17 | 154 | 7 | 27 | 4 |
| **20** | 628 | 0 | 0 | 0 | 1 | 0 |
| **21** | 4 | 0 | 0 | 0 | 0 | 0 |
| **22** | 527 | 0 | 0 | 0 | 0 | 0 |
| **X** | 1160 | 2 | 211 | 2 | 56 | 2 |
| **Y** | 198 | 0 | 0 | 0 | 0 | 2 |
| **TOTALS** | **17317** | **1412** | **7166** | **693** | **1220** | **140** |

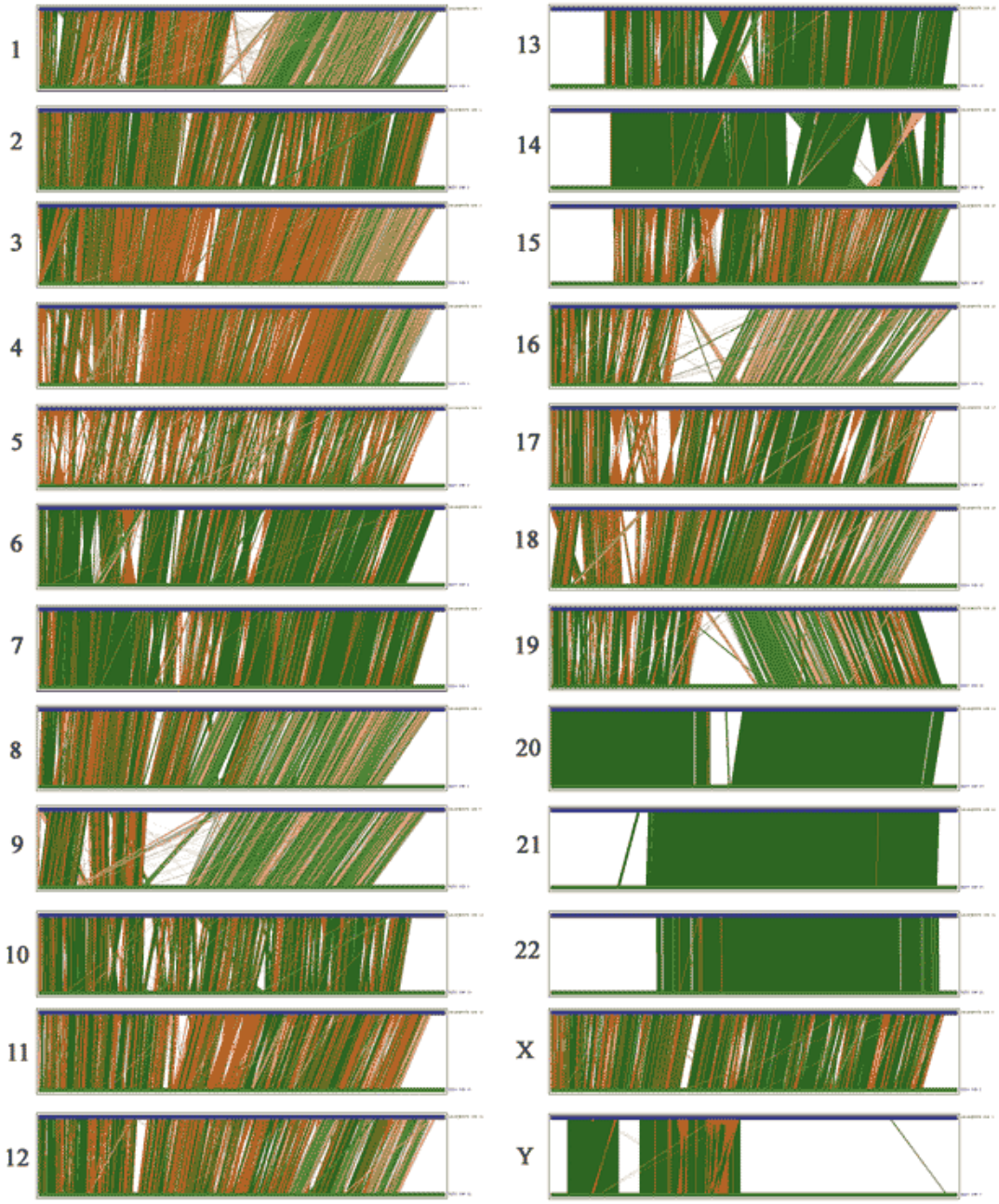



**Figure 2-9: Sequence Level Comparison of NCBI Build 26 vs. Goldenpath August 2001.** Shown in each of these graphs are the results from aligning the NCBI build 26 data to the Goldenpath August 2001 data using multi (States, unpublished). In each of the graphs, if a sequence similarity is found in the same orientation, it is drawn in green. If the orientation is in opposite directions, it is drawn in red. For those sequence similarities falling in close proximity on both assemblies, the color used is a darker color. A lighter color is used if they fall outside of a distance threshold.

**Table 2-13: Aligned Bases Using multi.** The second column indicates the number of matching bases for each chromosome where the matches occur in the same orientation. The third column indicates the number of matching bases for each chromosome where the matches occur in a different orientation.

| Chromosome | Matching Bases Same Orientation | Matching Bases Different Orientation |
|---|---|---|
| **1** | 61.746.063 | 20.411.421 |
| **2** | 68.880.069 | 16.725.895 |
| **3** | 46.402.958 | 18.531.575 |
| **4** | 39.549.971 | 20.747.599 |
| **5** | 40.449.965 | 21.510.030 |
| **6** | 62.936.552 | 8.228.667 |
| **7** | 53.201.681 | 5.840.056 |
| **8** | 37.439.999 | 11.180.941 |
| **9** | 34.975.746 | 9.624.503 |
| **10** | 42.060.960 | 8.006.401 |
| **11** | 37.438.424 | 13.694.342 |
| **12** | 36.328.682 | 11.486.143 |
| **13** | 35.695.803 | 4.281.521 |
| **14** | 32.428.345 | 2.117.332 |
| **15** | 19.193.109 | 8.687.356 |
| **16** | 17.702.596 | 6.568.383 |
| **17** | 15.667.567 | 6.813.709 |
| **18** | 20.677.969 | 7.429.676 |
| **19** | 12.368.846 | 3.453.858 |
| **20** | 23.484.031 | 33.978 |
| **21** | 12.994.871 | 2.498 |
| **22** | 12.412.454 | 25.480 |
| **X** | 49.177.058 | 9.246.770 |
| **Y** | 10.897.336 | 2.787.913 |
| **TOTALS** | **824.111.055** | **217.436.047** |

Comparisons at the sequence level produce results consistent with clone ordering. Figure 2-9 indicates pairings of identical 1000 base matches between the two assemblies. As can be seen in this figure, there are large regions of sequence matches where the matches seem to be inverted. This is most evident in the red portions of chromosomes 3

and 4. When all of the sequence comparisons are taken into account nearly 30% of all of the matches occur in opposite orientations. Table 2-13 summarizes the aligned bases between the two assemblies. Bailey *et al*. (2001) show that 10.6% of the January 2001 Goldenpath assembly shows regions of greater than 1 kb in length and greater than 98% identity. Even if all of these segmental duplications occurred on the same chromosome and in a different orientation, they could not account for 30% of all matched regions. Thus, there is a large amount of inconsistently oriented data between the NCBI and Goldenpath assemblies.

In order to help determine the confidence in the assembly of any particular chromosome, we calculated a metric to determine the expected nucleotide length to the next major mismatch between the NCBI and UCSC assemblies. Each matching multi block includes begin and end positions within the NCBI and UCSC assemblies. Consecutive matching blocks were compared to determine whether or not they should be merged together. The nucleotide distance between two consecutive blocks was calculated for both assemblies. The distance for each of these was compared. If the difference was less than 1 kb, then the two blocks were merged together. Otherwise, they were kept separate. Once merging of consecutive blocks was finished, the length of each block was stored. For each chromosome, the percentage of nucleotides with at least 1, 10, 100, 1000, 10000, $10^5$, $10^6$, $10^7$ and $10^8$ bases before the end of a block was calculated. This gave a measure of the agreement between the UCSC and NCBI assemblies. The results for the chromosomes with the longest length to next mismatch (chromosome 20), shortest length to next mismatch (chromosome 4) and all

chromosomes are given in figure 2-10. These results suggest that the agreement between the assemblies falls off between 10 kb and 100 kb, or approximately the size of an individual clone.

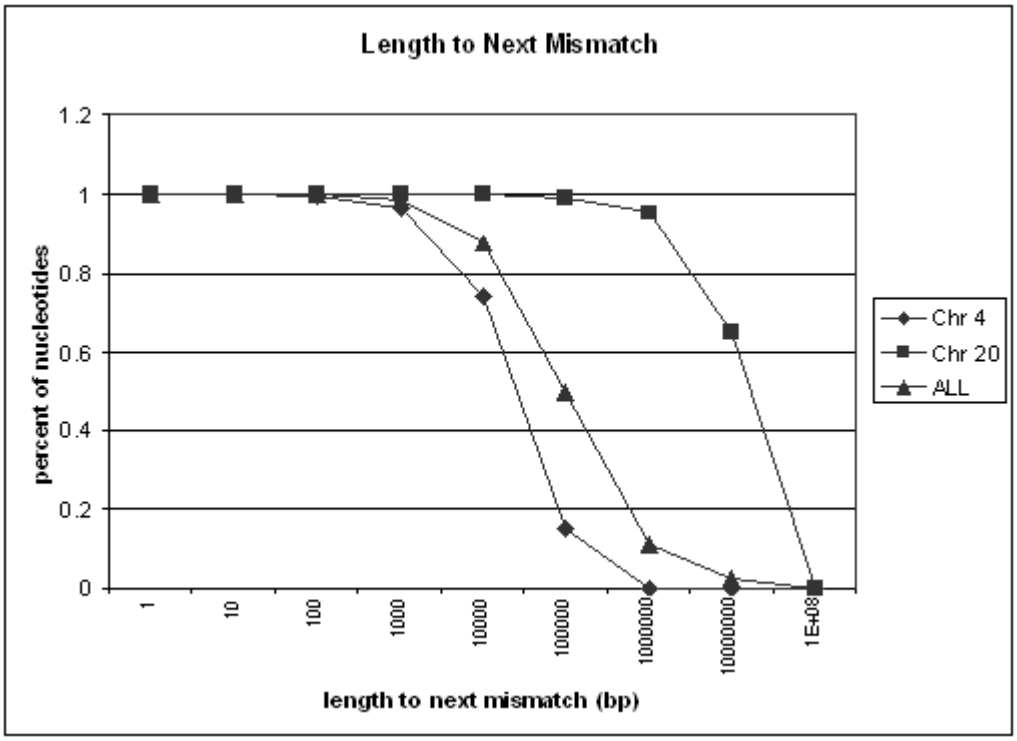



**Figure 2-10: Length to Next Major Mismatch.** Shown are the percentage of nucleotides which have a length to the next major mismatch at least as many nucleotides as specified in the x-axis. The results are shown for all chromosomes as well as chromosomes 4 and 20.

Major mismatches could result due to differences in gap lengths, repetitive regions and assembly errors or discrepancies. In order to illustrate these differences, dot plots of chromosomes 5 and Y are shown in figure 2-11. With chromosome 5, the major mismatches are due to sequencing differences, while chromosome Y agrees to a greater degree. The dot plot for chromosome Y also illustrates the presence of large scale repeats within the chromosome.

As more finished human sequence data becomes available, assembled human genomic contigs become a powerful resource. In addition to compositional analysis, genomic contigs can be mined for SNP detection and analysis (Blackwell, Rouchka and

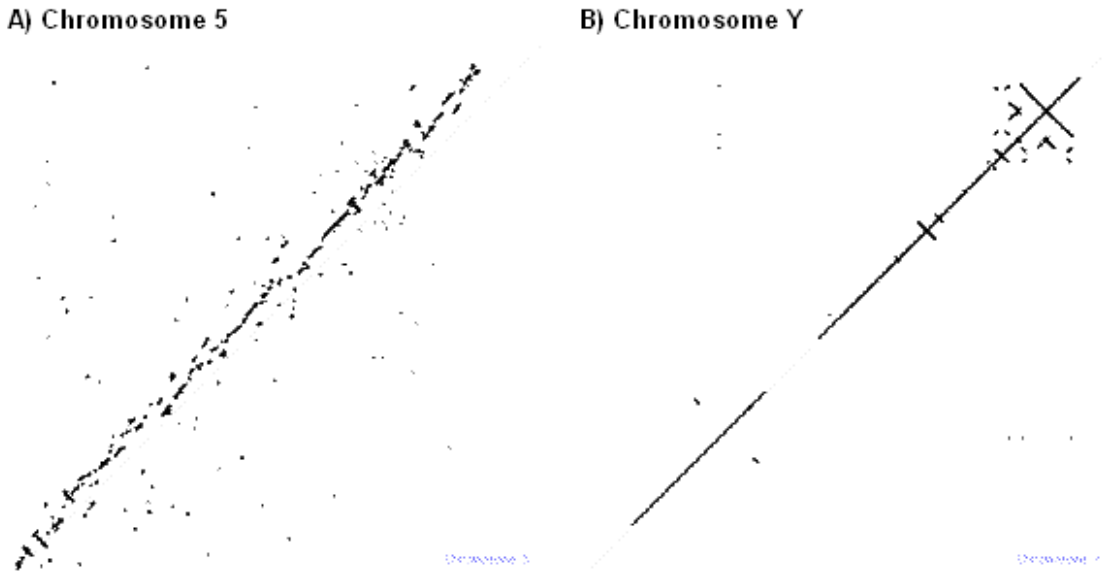



**Figure 2-11: Chromosome Dot Plots.** Both of these figures show dot plots resulting from a multi alignment of the NCBI assembly (x-axis) to the UCSC Goldenpath assembly (y-axis). Shown are the results for chromosome 5 (figure A) and chromosome Y (figure B).

States, 1999), transcriptional analysis (Kan *et al*., 2000), sequence assembly validation (Rouchka *et al*., 1998), and many other interesting problems. In applications such as these, confidence in the assembled sequences is paramount. An assembly incorporating only finished data provides a consistent starting point from which to base analyses.

## 2.6 Summary

Automated assembly of finished clone sequences into contiguous regions is a useful endeavor. Simulation results suggest that a sequence-based approach can piece

together nearly 93% of all fragments without adding false joins. While whole genome assembly incorporating draft sequences is useful, it leads to a large number of errors in order and orientation of clones and/or their trace fragments. As more clone sequence data reaches a finished state and physical maps are refined, the number of errors declines. This is observed in the agreement with the NCBI release 26 and Goldenpath August 2001 assemblies of chromosomes 20, 21, 22 and Y that are either at or near a finished state.

Due to the expected exponential growth of finished data available in the genomic databases, it is becoming imperative that procedures become automated to create and annotate these large sequences. It is equally important to determine which sequences are redundant and which offer novel information.

Once contigs are assembled, analysis can proceed into understanding different aspects of the human genome. In the subsequent chapters, assembled contigs are used as the basis for single nucleotide polymorphism (SNP) detection, sequence assembly validation, large scale polymorphism detection, CpG island segmentation, and an analysis of homogeneous regions of G+C content throughout the human genome.

# Chapter 3

# Single Nucleotide Polymorphisms

Single nucleotide polymorphisms (SNPs) occur when two or more different nucleotides appear at the same position within a population leading to genetic variation. In order to understand the relationship between the genetic makeup of an individual (the genotype) and the resulting observed properties, whether it be structural or functional (the phenotype), it is necessary to study genetic differences. For instance, a single nucleotide change accounts for the difference between a healthy individual and one with sickle-cell anemia (Lodish, *et al*., 1995). In addition, a single base mutation in the *APOE* gene is associated with Alzheimer's Disease (Chakravarti, 2001) and a one base deletion in the chemokine-receptor gene *CCR5* leads to resistance of HIV (Chakravarti, 2001). While the majority of genes and diseases within the human genome are more complex, detection of SNPs within the population can give a better understanding into the intricate interactions. As a result, methods for the detection of SNPs are necessary. An understanding of how SNPs cluster within the human genome is an important aspect that will be considered.

## 3.1 SNP Detection

Overlapping regions between clones within the assembled contigs are useful in detecting SNPs. Since the clones used to create the contigs are finished human sequence data, mismatches in these regions are less likely to be the result of sequencing errors which occur at a rate of 1 per 10000 (Smith and Carrano, 1996) and more likely to be actual SNPs which occur at a rate of 7 per 1000 (Taillon-Miller, *et al*., 1998). When the contigs are constructed, the length of the overlapping region between two clones is reported along with the percent identity between the clones in this region. If the identity is less than 100%, then there exists at least one gap or mismatch in one of the sequences. While gaps in the sequences can indicate SNPs in the sense of single nucleotide insertion or deletion events such as the resistance to HIV discussed earlier, we concentrate on the detection of single nucleotide substitution SNPs.

Once the contigs are created, the overlapping regions less than 100% identical are extracted and the alignment is reconstructed using wu2blastn (Gish, 1994-2001). The resulting alignment is then scanned, and all mismatches are treated as potential SNPs. Since at least 35% of the human genome is made up of repetitive elements (Jurka, 1998), it is possible that SNPs can occur in these regions. For analytical purposes, this information can be incorporated. However, for experimental validation of SNPs, it is important that the sites occur in unique regions. Thus, we are only interested in those regions where an SNP has at least 75 bases before and after it that do not occur within repeat regions. When possible, we report up to 500 bases to each side of the SNP that do

not occur within repeats. RepeatMasker (Smit and Green, unpublished) is used to locate repeat regions within the overlapping segments. This data can then be used for the purpose of creating PCR primers for amplification of SNP regions.

In a study to determine the effectiveness of such an approach to SNP detection, a collaboration was formed with Pui Kwok in the Dermatology Department at Washington University. Preliminary results detected and verified 10 novel SNPs. The SNPs were then deposited into NCBI's dbSNP (Sherry, *et al*., 2001). The accessions are G54158, G54159, G54160, G54161, G54162, G54163, G54164, G54165, G54166 and G54167.

Besides the non-redundant contig data that we have assembled, an additional 8 MB of redundant data completely lying within assembled contigs has been found. This data can also be used to screen for candidate SNPs using the same techniques outlined here.

## 3.2 SNP Clustering

In addition to the detection of SNPs, there are many other interesting questions to ask. Once of these questions concerns the evolution and clustering of SNPs. The information I gathered through the SNP detection was used in collaboration with Tom Blackwell at Washington University's Institute for Biomedical Computing for the purpose of testing a probabilistic population genetic theory for the expected distribution of SNPs (Blackwell, Rouchka and States, 1999). A visual inspection of a typical clustering of possible SNPs, such as that seen in figure 3-1, shows that mismatches tend to be clustered. Since SNP clustering is purely mathematical and therefore does not

require the same restrictions as experimental validation, all candidate SNPs are considered including those occurring within repetitive regions. One aspect of finding SNPs that are identical by descent is that these regions can now be studied as linkage events and how inheritance of several different combinations of SNPs can lead to disease.

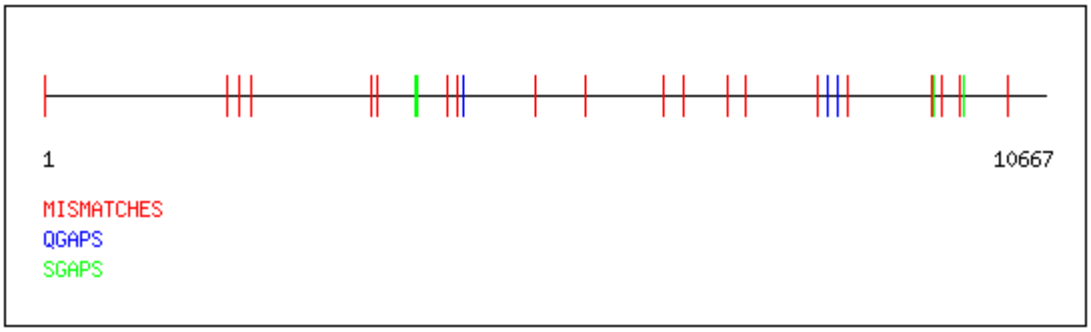



**Figure 3-1: Distribution of Candidate SNPs.** This figure illustrates the typical pattern of possible SNPs occurring within two overlapping regions. The two sequences illustrated here are GenBank accession AF003625 and AF035396. The red tick marks indicate a mismatch occurring between the two sequences, while a blue or green tick mark indicates a single nucleotide insertion or deletion event.

# Chapter 4

# Sequence Assembly Validation

One particular application which requires the use of extended regions of genomic data involves the validation of assembled sequence. Genomic sequence analysis depends on the accurate assembly of short (400 to 1,000 base pair) sequence reads into contigs that cover extended regions as a necessary step in deriving finished sequence. Errors at the fragment layout assembly stage may be difficult or impossible to detect later in the editing process, and fragment assembly errors may have a serious impact on the biological interpretation of the data. For example, entire regions of the genome could be inverted or swapped as a result of assembly errors. Such errors could impact the biological interpretation of the sequence data, potentially leaving groups of exons out, swapping exons or control elements onto the anti-sense strand, breaking genes into pieces, or dissociating genes from their control elements. Since assembly errors are difficult to detect and can impact the utility of the finished sequence, experimental validation of the fragment assembly is highly desirable.

Comparison of predicted and experimental restriction digests has been proposed as a means for validating fragment assembly. The pattern of fragment masses resulting from a restriction digest of the source DNA can be readily determined with a precision of $\pm$1%. This pattern of restriction fragment masses is commonly referred to as a restriction fingerprint. The cleavage sites for restriction enzymes are specific so it is easy to

electronically generate a set of predicted fragment masses from the finished sequence. Similarly, the location of each of the predicted fragments on the finished sequence is known. Errors in sequence assembly will either change fragment masses directly or rearrange the position of restriction sites resulting in new fragments with altered masses. Figure 4-1 shows the general flow of the concepts used in comparing predicted and experimental restriction digests.

Restriction fragment matching has been extensively used as the basis for physical map assembly (Riles *et al*. 1993; Waterston *et al*. 1993). Similarities in fingerprint are used to infer clone overlap. Since most clones overlap over only a fraction of their length and because restriction digest sites may be polymorphic, software has been developed to recognize common features of fingerprint patterns while ignoring the disparities. Most of the information in a fingerprint is accessible even if several bands in the digestion pattern are missed or a number of false positives are scored.

In this section, we examine the use of multiple restriction digest fingerprints for assembly validation. Both simulated and experimental results will be discussed as well as a specific application to clone mapping. We also compare the requirements for fingerprint mapping with the requirements for assembly validation.

## 4.1 Methods

Dynamic programming algorithms were first used in the context of computational biology for the purpose of finding the best alignment between two DNA or protein sequences (Needleman and Wunsch 1970; Sellers 1974; Smith and Waterman 1981). We

have developed a similar dynamic programming algorithm to determine the maximum alignment of error prone electrophoretic mobility data to predicted fragment mobilities. The expected fragment mobility information can be calculated when the sequence to validate and the restriction enzyme patterns used in creating the experimental data are known. String matching functions are used to find the exact location of a particular cutting site in the sequence.

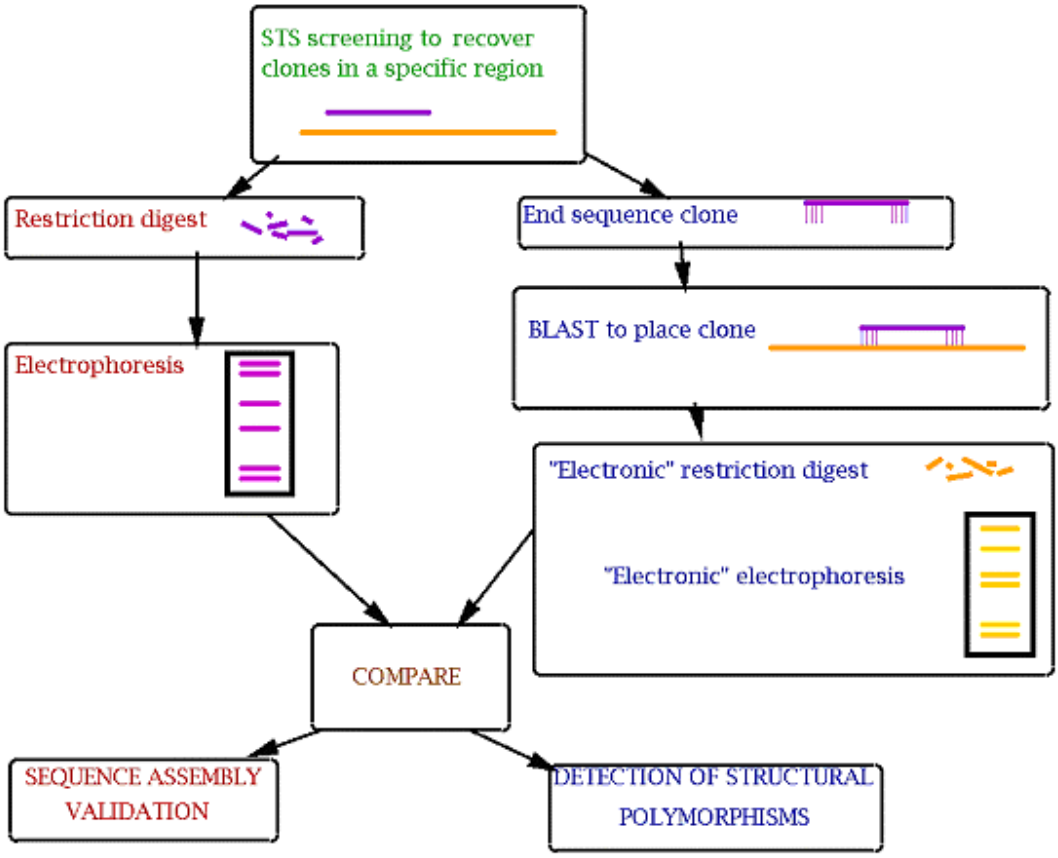



**Figure 4-1: Sequence Assembly Validation Flow Diagram.** This figure indicates the steps used in order to compare experimental restriction fragments to expected restriction fragments.

Predicted fragments are generated according to these locations. The mobility, *m*, for each of these expected fragments is calculated using the same formula from which the experimental data is derived according to equation 4-1.

$$m_{fragment} = 2Log\left(\frac{L_{tot}}{L_{fragment}}\right)$$

**Equation 4-1: Predicted Fragment Mobility.**

Here, $L_{tot}$ is the total length of the sequencing project. The factor of 2 is applied to give mobilities in the range typical of current experimental protocols, 0 to 20 cm. In these units, a standard deviation in determination of band position of 0.1 cm corresponds to a relative accuracy of mass determination of 0.5%.

Within the dynamic programming algorithm, fingerprint pattern alignments were scored using a log odds system based on the likelihood of deriving the observed fragment mobilities from the predicted digest mobilities relative to the odds of observing the fingerprint pattern at random. Table 4-1 indicates these scores.

**Table 4-1: Scores for Fingerprint Pattern Alignments.**

| **Relationship** | **Score** |
|---|---|
| Band match | $\text{Log}(P_{match}/P_{random})$ |
| False positive | $\text{Log}(P_{false\ positive})$ |
| False negative | $\text{Log}(P_{false\ negative})$ |

The probability, $P_{match}$, of a fragment having an observed mobility, $m_{obs}$, given a true mobility, $m$, and normally distributed errors in mobility determination (Drury *et al*. 1990, 1992), is given in equation 4-2.

$$P_{match}(m_{obs} \mid m) = \frac{1}{\sqrt{2\pi\sigma}} e^{(-\frac{(m_{obs}-m)^2}{2\sigma^2})}$$

**Equation 4-2: Observed Mobility Probability.**

Assuming that the fragment mobilities scale as the log of the molecular weight of the fragment (Maniatis, Jeffrey and van deSande, 1975), this formulation results in a constant fractional error in mass determination and agrees with empirical observations based on current data (M. Marra, personal communication, 1998).

Equation 4-3 gives the probability, $P_{random}$, of matching a band at random given a maximum mobility of $X$ and $N$ bands.

$$P_{random} = \frac{N}{X}$$

**Equation 4-3: Random Probability of Matching a Band.**

The values of $P_{false\ positive}$ (false positive "added" band probability), $P_{false\ negative}$ (false negative "missing" band probability), and $\sigma$ (standard deviation from true mobility) are calculated based on the precision with which the experimental data can be extracted.

This scoring system penalizes either matching a band with an error in the mobility or failing to match a band altogether. The false positive score represents the case where a band in the experimental data does not match up with a band in the expected data. The

false negative score represents the case where a band in the expected data does not match up with any experimental bands. The maximum score is the log likelihood that the query fingerprint was derived from the target pattern under the assumptions of our model relative to the likelihood of assuming the same match at random. Scores are reported in units of the natural logarithm of the likelihood ratio (nats). They may be converted to bits by dividing ln(2).

### 4.1.1 Coverage

Since the sequence to be validated is known, a map of the restriction enzyme cut sites can be created for each of the restriction enzymes used in the experiments. As a result, the location of each of the expected fragments within the sequence is known. Figure 4-2 shows an example of the known cutting sites for the restriction enzymes BamHI, EcoRI, HindIII, and KpnI within an example sequence.

For each of the four restriction enzymes, an experimental digest has been performed independent of the other three enzymes. The experimental fragments are compared to the expected fragments using the previously described dynamic programming algorithm. The purpose of the algorithm is to tell which of the expected fragments are matched with an experimental fragment. A region between two restriction sites in the sequence to be validated is said to be *covered* when it is matched with an experimental fragment. The results of the coverage analysis for each individual restriction enzyme can be combined to produce a total coverage map where the coverage for any particular fragment can range from 0% to 100%. When four enzymes are used,

the coverage for any fragment between two restriction sites can be 0% (not covered by any individual restriction enzyme coverage map), 25% (covered by one), 50% (covered by two), 75% (covered by three), or 100% (covered by all four restriction enzyme coverage maps).

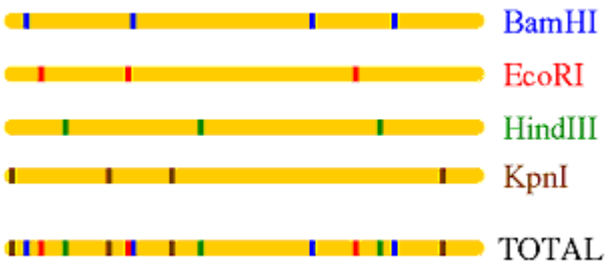



**Figure 4-2: Enzyme Fragment Coverage.** The sequences labeled BamHI, EcoRI, HindIII and KpnI show the location of the respective restriction enzyme recognition sites within an example sequence. The sequence labeled TOTAL indicates the location of all of the enzyme restriction sites within the sequence.

Analysis of coverage maps can indicate possible sequence assembly errors. For instance, suppose that one segment within the clone has been reversed in the sequence assembly. In such a case, we would expect two predicted restriction fragments from each digest not to be matched, resulting in a low coverage for the regions containing these fragments. The regions of low coverage contain within them the endpoints of the reversed segment.

### 4.1.2 Setting up the Simulations

Simulated restriction digest patterns were created by adding random perturbations to the computationally predicted mobilities. The predicted mobilities were created using a subset of the palindromic six base restriction sites EcoRI (GAATTC), BamHI

(GGATCC), HindIII (AAGCTT), BalI (TGGCCA), HpaI (GTTAAC), PstI (CTGCAG), SalI (GTCGAC), KpnI (GGTACC), NaeI (GCCGGC), and NarI (GGCGCC). The test fingerprints were compared with reference fingerprint patterns derived from sequences rearranged by introducing a segmental inversion between two randomly chosen points in the sequence. For each of the patterns, we find which target bands get matched up with an experimental band. Using this information, a coverage plot can be generated for the target sequence. By comparing the digest patterns of more than one restriction enzyme and overlapping their coverage results, it is proposed that errors in sequence assembly can be differentiated from false positive and false negative experimental bands. We ran simulations to test the effects of false positive and false negative band rates (ranging from .5% - 2%), band mobility resolution (ranging from .1% - 1%; 0.02mm - 0.2mm), and the number of restriction enzymes used. We looked at false negative rates (the percentage of time that one of the ends in the inversion is not detected by coverage analysis) and false positive rates (the percentage of time that an incorrect inversion location is detected by coverage analysis). The data presented is based on the simulations using a 219.4 kb interval derived from the human X chromosome (GenBank accession no. L44140) (Chen *et al*. 1996a). We will focus on the results using four restriction enzymes for a more detailed discussion.

Experimental results have also been achieved using a HindIII digest on the bWXD718 sequencing project at the Washington University Center for Genetics in Medicine. These results are discussed as well.

## 4.2 Results

The Washington University Center for Genetics in Medicine and Genome Sequencing Center have been collaborating in construction of sequence ready maps and reagents for the human X chromosome, and over 1,000 clones have now been fingerprinted. The precision of fragment mass determination was 1% (M. Marra personal communication, 1998). In the early phases of this work 30 clones were sent for repeat analysis making it possible to estimate the reliability of the fingerprint data. In this preliminary data set, one discrepancy in 25 bands was observed between identical clones implying a combined false positive and false negative rate of roughly 4%. As the lab has become more experienced with fingerprint analysis, performance has improved substantially.

### 4.2.1 Increasing the Number of Restriction Enzymes

Figure 4-3 illustrates the use of a single restriction enzyme. Fingerprint analysis is sensitive to false positive and false negative bands. As a result, it can be impossible to differentiate between false negative bands and regions of incorrect sequence assembly. A restriction site is expected every $4^6 = 4096$ bases in random sequence since six base restriction enzymes are used. It is well known that genomes are not randomly distributed. Thus, some restriction sites might be rare in a particular region. Two problems can result. The first is that an inversion can be missed because it has a greater likelihood of occurring between two sites where it cannot be detected. The second is that even though

a region of low coverage might be detectable, a greater area might have to be considered as a possible location for the inversion.

A second enzyme can help alleviate the problem of differentiating false negatives and areas of concern. However, if the restriction enzymes are not chosen carefully, relatively long stretches where there is not a restriction site for either enzyme can still exist. Figure 4-4 illustrates the results using a second restriction enzyme.

Coverage analysis of our simulations suggests that the use of four or more enzymes should produce the desired results (compare Figures 4-3, 4-4, and 4-5). Two enzymes still present the difficulty of an inversion occurring in between two restriction sites. Experimental errors will also have some effect when only two enzymes are used. We have analyzed the results using an even number of enzymes. This is done to balance the number of A+T restriction patterns with the number of G+C restriction patterns, so as to avoid compositional biases. Figure 4-5 illustrates the results using four restriction enzymes. If the restriction digests are repeated when a potential region of difficulty is observed, experimental gel errors can be filtered out and differentiated from sequence assembly errors. Figure 4-6 illustrates this point. Note that if a single enzyme is used (as in Figure 4-3), the digests would have to be repeated quite often due to false negative bands.

Table 4-2 and Figure 4-7 examine the effects on the percentage of time that a region of faithful sequence is found to have low coverage by restriction digest fragment mapping. Figure 4-8 shows the percentage of time that a region that is involved in a

segmental inversion is found to have high coverage. This corresponds to the fraction of the time that the rearrangement would be missed by our analysis.

### 4.2.2 Analysis of Experimental Data

One of the sequencing projects that the Washington University Center for Genetics in Medicine and Genome Sequencing Center is working on involves a region of the human X chromosome labeled bWXD718. In a preliminary assembly, the sequence appears to be 79,612 nucleotides long. The experimental HindIII digest of this clone indicates a total fragment size of 169,699 nucleotides, indicating the preliminary assembly contains errors.

All but two of the expected fragments match up with experimental fragments. The two fragments that do not match up are 558 and 145 nucleotides long. It is possible that some of the smaller fragments travel through the gel more rapidly, and thus there are greater errors, so the 558 nucleotide segment might actually map to an expected segment that is 520 nucleotides long. Also, the 145 nucleotide segment might have gone undetected in the gels. Thus, the validation program cannot discern where the problem is located, but rather alerts the biologists that there is an existing assembly problem or a molecular biological rearrangement that occurred between the fingerprint and sequence analysis stages.

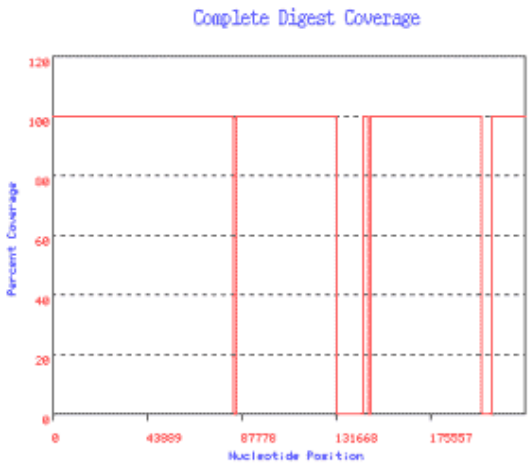



**Figure 4-3: Coverage Graph Using a Single Enzyme.**

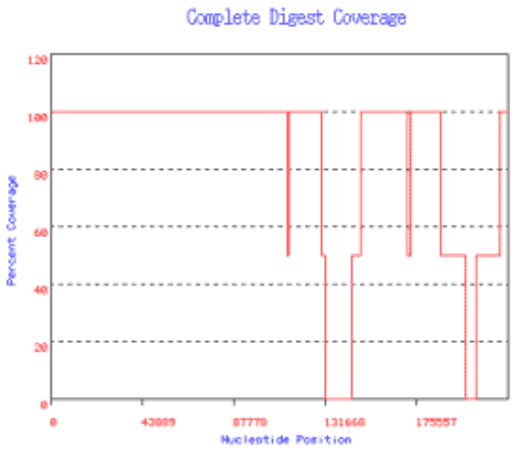



**Figure 4-4: Coverage Graph Using 2 Enzymes.**

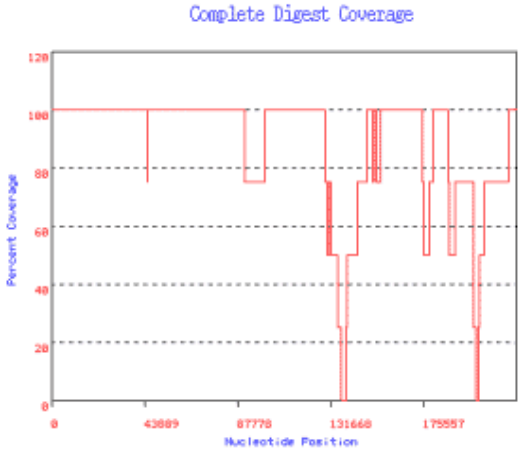



**Figure 4-5: Coverage Graph Using 4 Enzymes.**

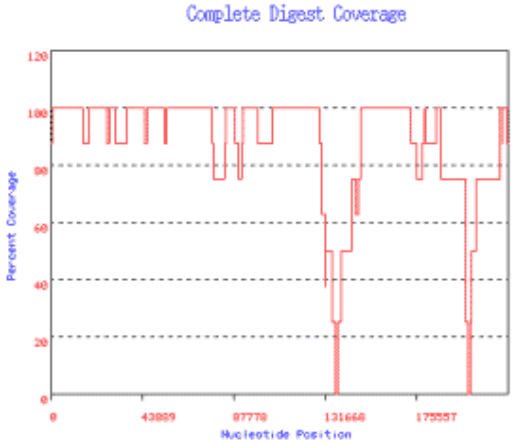



**Figure 4-6: Coverage Graph Using 4 Enzymes and Repeating the Digest Analysis.**

**Figures 4-3 to 4-6: Coverage Graphs.** Indicated in all four figures is the coverage for the 219.4 kb region with a segmental inversion between nucleotides 136,796 and 201,014. A single restriction enzyme is used in figure 4-3, resulting in four regions of zero coverage. Two of these are due to experimental false negative rates, suggesting that a single enzyme is not sufficient for sequence assembly validation. When two restriction enzymes are used as in figure 4-4, only the two regions where the inversion occurs have zero coverage, indicating that using a second restriction enzyme improves the analysis. Figures 4-5 and 4-6 show the results using four enzymes. In figure 4-5, the band around the segmental inversion endpoints has shrunk to 2175 nucleotides for the left end and 1161 nucleotides for the right end. Figure 4-6 repeats the restriction digest. Some bands begin to have better coverage and the area surrounding the left end has shrunk from 2175 to 1286 nucleotides.

**Table 4-2: Empirical Error Rates for Band Assignment.** The table presents the error rates for the assignment of segmental inversions to their corresponding segment of genomic sequence. The column on the far left represents experimental gel resolution values. False positives are the percentage of time that a region not involved in a segmental inversion is found to have low coverage. False negatives are the percentage of time that a region that is involved in a segmental inversion is not found. Within each section results are presented for simulations conducted with false negative and false positive band calling rates of 0.5%, 1% and 2%, and these results are presented separately. These results are based on four enzyme digests, each performed once, and a coverage cutoff of 50%.

| Gel Resolution | False Positive Result | | | False Negative Result | | |
|---|---|---|---|---|---|---|
| | .5% | 1% | 2% | .5% | 1% | 2% |
| 0.001 | 4.2% | 6.8% | 9.9% | 6.2% | 3.8% | 3.9% |
| 0.0025 | 5.5% | 7.5% | 11.9% | 6.1% | 4.2% | 5.8% |
| 0.004 | 5.9% | 7.2% | 11.2% | 2.8% | 3.8% | 6.3% |
| 0.0055 | 4.9% | 8.2% | 12.6% | 3% | 4.9% | 3% |
| 0.007 | 7.5% | 7.7% | 13.2% | 3.9% | 4.6% | 3.3% |
| 0.0085 | 5.5% | 7.2% | 13.5% | 5% | 3.5% | 5% |
| 0.01 | 5.2% | 8.5% | 11.4% | 4.3% | 3.6% | 6% |

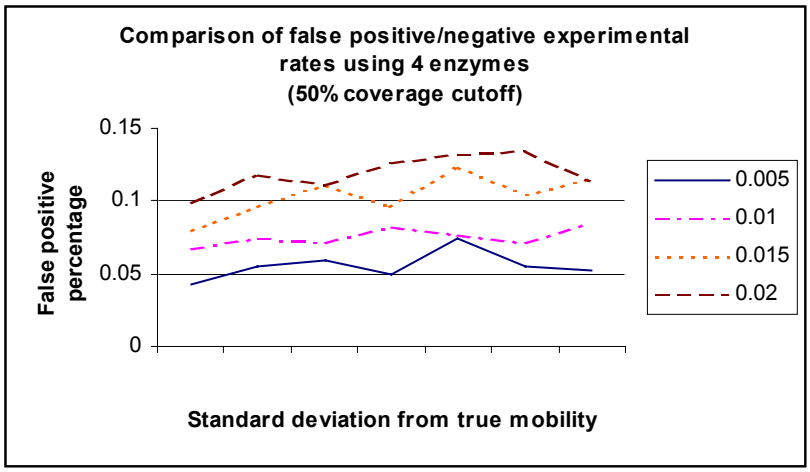



**Figure 4-7: False Positive Rates.** This figure corresponds to the data from Table 4-2. The x-axis represents the standard deviation from true mobility and the y-axis represents the false positive rates. By examining this graph, we can see that the experimental false positive and false negative rates have an effect on false positives. In particular, as the experimental rates increase, so does the percentage of time that a region that is not involved in a segmental inversion is found to have low coverage. At the same time, the standard deviation from true mobility does not seem to affect the false positive percentage.

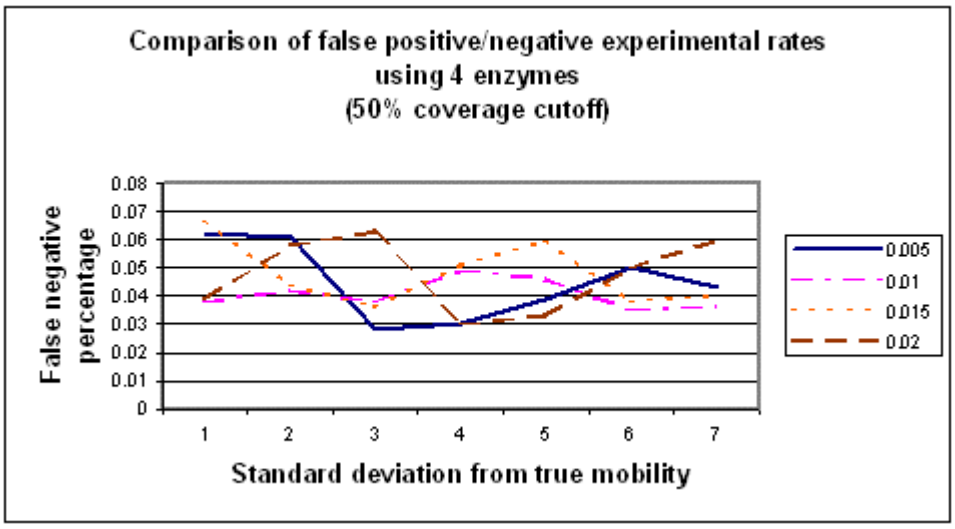



**Figure 4-8: False Negative Rates**. This figure corresponds to the data from Table 4-2. The x-axis represents the standard deviation from true mobility and the y-axis represents the false negative rates. By examining this graph, we can see that the experimental false positive and false negative rates do not have much of an effect on the rate of missing a rearrangement.

## 4.3 Discussion of Sequence Assembly Validation

The results presented here demonstrate that it is possible to detect most sequence fragment assembly errors using a set of four restriction digests and without reference to an overlying physical map. The confidence of sequence validation can be further improved by independently repeating the digests or by using additional enzymes (data not shown). The confidence of sequence validation improves with both the resolution of the electrophoretic fragment sizing and the accuracy of band calling.

### 4.3.1 False Negatives

There are four reasons why the simulated segment inversion sites may not be determined correctly. One reason is that the inversion could occur in a segment such that it does not overlap any restriction sites. Another explanation is that the inversion occurs

in such a way that the restriction sites are located near the middle of the inverted segment, resulting in similar fragment mobilities. Thirdly, an inversion occurs in such a way that the modified segments are similar to other existing segments, so coverage is preserved, albeit at a lower percentage than normal. Finally, the inversion could occur within a long repeat segment, resulting in no change with an inversion.

### 4.3.2 Application to Clone Mapping

We have been in collaboration with the Washington University Center for Genetics in Medicine and Genome Sequencing Center to use these assembly validation techniques to map locations of BAC and YAC clones within the human genome. For the purposes of our analysis, we are given both the end sequences of the clones and a set of restriction digest fragments for the enzymes BamHI, EcoRI, HindIII, and KpnI. Once we have the experimental data, the process begins by searching our assembled genomic contigs for homologies with the end sequences using a local sequence alignment technique. We find which, if any, of the contigs we have assembled have stretches of matching nucleotides longer than 30 nucleotides. If such a contiguous sequence exists, we can compare an expected digest covering this region with the experimental digests. A coverage graph of the results can then be analyzed. Such a study can be helpful because it places the clones within existing contigs, helping to determine whether or not the whole clone should be sequenced. This might help to bridge the gap between two contigs.

To test our methods, we began by analyzing three clones on chromosome X (bX759, bX691, and bX171) where complete sequence determination has been performed

by Dr. Ellison Chen. Bands with molecular weights between 1500 bp and 12,000 bp were scored. A match was scored as a positive if a band in the observed digest was identified within 2% of the molecular weight predicted from the electronic digest. 293 of the 302 (97%) of the bands were scored as matches. Of the nine bands that failed to match a band in the electronic digest, eight were within 2.5% of the predicted molecular weight, and one deviated by 3.3%. These are entirely within the expected experimental error. Four complete enzyme digests (HindIII, EcoRI, BamHI, and KpnI) were analyzed for each clone. In no case did a fragment that failed to match overlap with a second fragment in a different enzyme digest. These data verify the integrity and accuracy of the sequence data obtained from the Chen laboratory and validate our fingerprint analysis methods.

### 4.3.3 Detecting Structural Polymorphisms

When restriction digests for multiple clones within the same region are available, the results of the sequence assembly validation can be expanded upon to look for large scale structural polymorphisms. Restriction digest data has been made available to us for the breast cancer susceptibility region BRCA2 on chromosome 13; the T-cell receptor region on chromosome 7; and for the color vision region located on chromosome X. The contigs created for these regions are described in tables 4-3 (BRCA2), 4-4 (T-cell), and 4-5 (color vision).

Since we have available to us clones within these three regions, the first step is to take the end sequences and find out where they should be placed within the genomic contigs. This search is performed using Smith-Waterman dynamic programming

**Table 4-3: BRCA2 Contig (IBC_chr13-ctg1).** This table describes the clones making up the BRCA2 contig.

| BRCA2 Region (13q12) IBC_chr13-ctg1 | | |
|---|---|---|
| Clone | Length | Overlap |
| AC002525 | 140,942 | -- |
| HUM85D2 | 34,931 | 200 |
| HUM2G3A | 110,858 | 200 |
| AC002483 | 102,846 | 200 |
| HS214K23 | 127,079 | 200 |
| HS234I22 | 3,158 | 79 |
| HS92M18 | 68,903 | 68 |
| HS130N4 | 84,170 | 104 |
| HS26H23 | 91,835 | 104 |
| HS267P19 | 113,704 | 104 |
| HS49J10 | 137,246 | 99 |
| HS179I15A | 146,810 | 104 |
| HS46H23 | 129,098 | 104 |
| HS65O19 | 95,274 | 110 |
| TOTAL LENGTH = 1,385,178 bases | | |

**Table 4-4: T-cell Receptor Contig (IBC_chr7-ctg23).** This table describes the clones making up the T-cell receptor contig.

| T Cell Receptor Beta Chain (7q35) IBC_chr7-ctg23 | | |
|---|---|---|
| Clone | Length | Overlap |
| U66059 | 267,156 | -- |
| U66060 | 215,422 | 9,638 |
| U66061 | 232,650 | 20,617 |
| TOTAL LENGTH = 684,973 bases | | |

**Table 4-5: Color Vision Contig (IBC_chrX-ctg56).** This table describes the clones making up the color vision contig. note that the overlaps indicated with a * are not 100% identical.

| Color Vision Region (Xq28) IBC_chrX-ctg56 | | |
|---|---|---|
| Clone | Length | Overlap |
| HSU52112 | 174,424 | -- |
| AF030876 | 112,756 | 12965 |

| | | |
|---|---|---|
| HSQLL2C9 | 15,250 | 10986* |
| HSQC14G3 | 13,546 | 251 |
| HSQC8B6 | 21,480 | 120 |
| HSCG1160 | 28,230 | 6092* |
| HS14B7 | 36,429 | 241 |
| HUMFLNG6PD | 219,447 | 305 |
| TOTAL LENGTH = 590,587 bases | | |

methods. After the clones have been placed, the expected fragment sizes can be calculated. After the sequence assembly validation has taken place, an optimal alignment of the experimental and expected fragments is determined. Figures 4-9, 4-10 and 4-11 show the results for BRCA2, color vision, and T cell receptor, respectively. For each of these regions, restriction digest information was available for four different enzymes: BamHI, HindIII, KpnI, and EcoRI. Fragments lying within the range 1,500 to 12,000 base pairs were scored. Those bands not scored are colored in gray. When a predicted fragment which should be scored fails to match an experimental fragment, it is colored red. The patterns of red can then be examined as possible locations of structural polymorphisms.

In the preliminary work screening these three regions, at least 15 examples of structural polymorphisms have been detected. These polymorphisms can range in length from hundreds of base pairs to kilobases of sequence. Tables 4-6, 4-7, and 4-8 show candidate polymorphisms for the BRCA2, color vision, and T cell receptor regions.

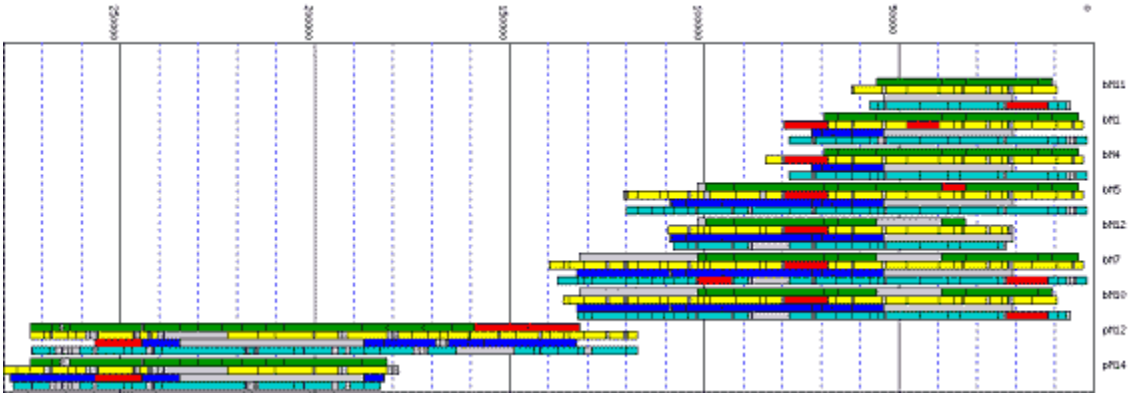


**Figure 4-9: BRCA2 Region Clone Alignment.** Shown in the figure is a graphical summary of the matching of restriction fragments to the electronic digest of the human genomic contig. Clones were positioned by end-sequence alignments. Matching fragments are shown in green - BamHI, yellow – HindIII, blue – KpnI, cyan – EcoRI. Indeterminate fragments are shown in gray, and red indicates regions where a predicted fragment is unambiguously missing from the observed digest.

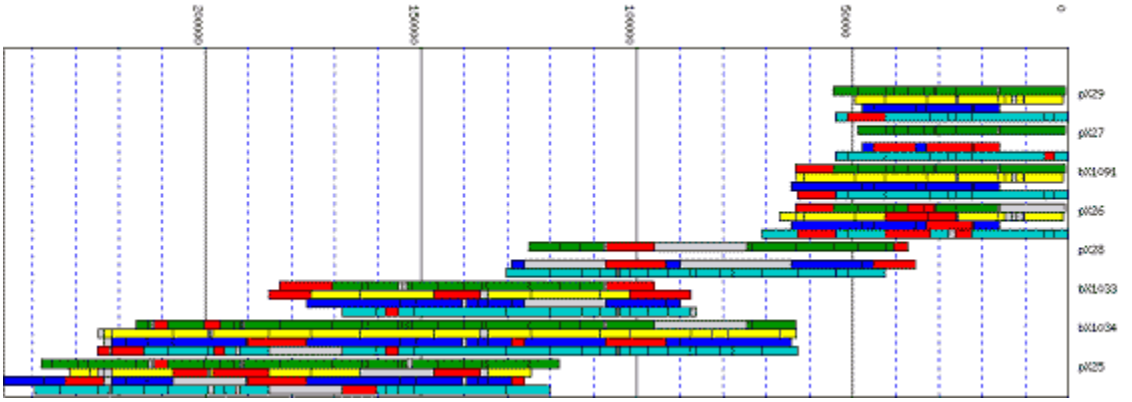


**Figure 4-10: Color Vision Clone Alignment.** Shown in the figure is a graphical summary of the matching restriction fragments for clones in the color vision region to the electronic digest of the human genomic contig. Clones were positioned by end-sequence alignments. Matching fragments are shown in green – BamHI, yellow – HindIII, blue – KpnI, cyan – EcoRI. Indeterminate fragments are shown in gray, and red indicates where a predicted fragment is unambiguously missing from the observed digest.

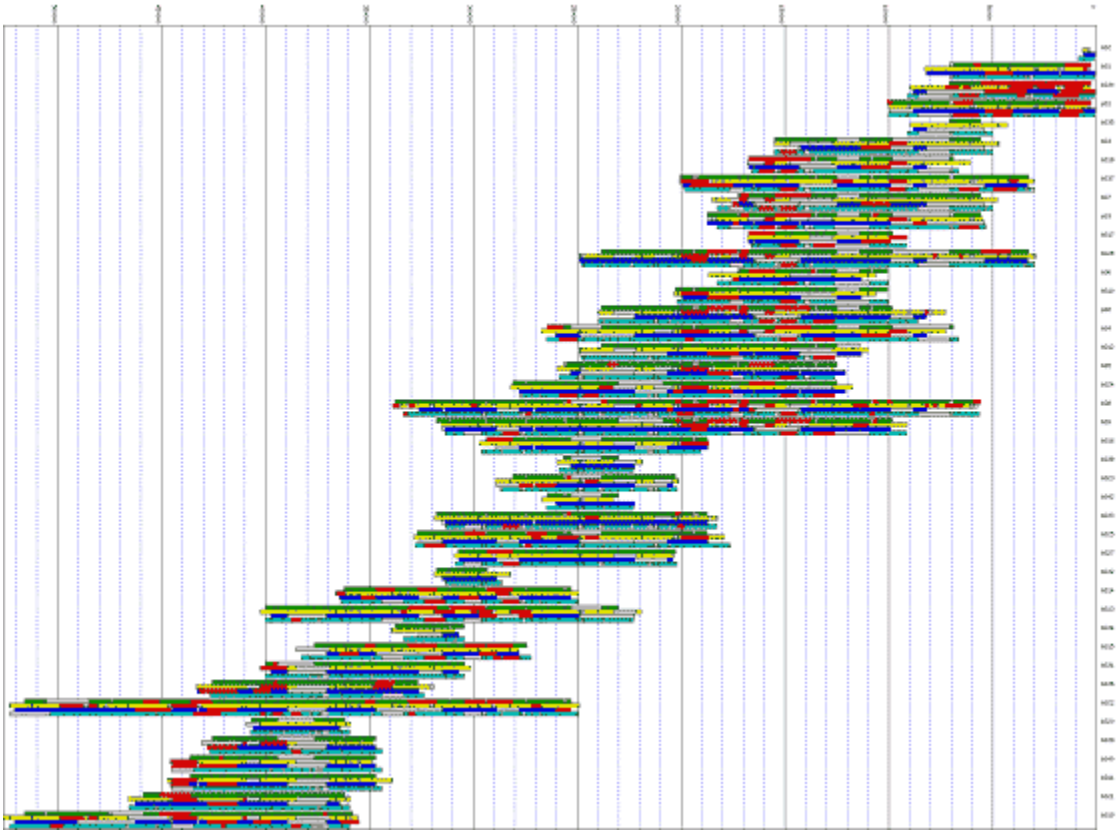


**Figure 4-11: T-cell Receptor Clone Alignment.** Shown in the figure is a graphical summary of the matching restriction fragments for clones in the color vision region to the electronic digest of the human genomic contig. Clones were positioned by end-sequence alignments. Matching fragments are shown in green – BamHI, yellow – HindIII, blue – KpnI, cyan – EcoRI. Indeterminate fragments are shown in gray, and red indicates where a predicted fragment is unambiguously missing from the observed digest.

**Table 4-6: Selected Polymorphic Sites from the BRCA2 Contig.** The left column on this table indicates the restriction enzyme and location of the fragment associated with a possible polymorphic site within the human genomic contig IBC_chr13-ctg1. The second column indicates those clones with matching fragments while the third column indicates those clones which do not have matching fragments.

| Fragment | Matching Clones | Fail to match |
|---|---|---|
| **EcoRI fragment**<br>From 11816 to 22552 (size 10,3760) | **bM1 bM4 bM5**<br>Matching bands:<br>Low 10,491 high 10,976<br>Mean 10,814+-228 | **bM10 bM11 bM7**<br>Nearest bands:<br>Low 11,237 high 11,407<br>Mean 11,294+-80 |
| **HindIII fragment**<br>From 68,458 to 79,572 (size 11,114) | **None** | **bM1 bM10 bM12 bM4 bM5 bM7**<br>Nearest bands:<br>Low 11,607 high 12,266<br>Mean 11,916+-237 |
| **EcoRI fragment**<br>From 78,467 to 87,817 (size 9,350) | **bM5**<br>matching band: 9,416 | **bM10 bM12 bM7**<br>Nearest bands:<br>Low 9,074 high 9,138<br>Mean 9,108 +-26 |
| **KpnI fragment**<br>From 244,546 to 256,644 (size 12,098) | **None** | **pM12 pM14**<br>Nearest bands:<br>Low 12,894 high 13,117<br>Mean 13,006+-111 |

**Table 4-7: Selected Polymorphic Sites from the Color Vision Contig.** The left column on this table indicates the restriction enzyme and location of the fragment associated with a possible polymorphic site within the human genomic contig IBC_chr13-ctg56. The second column indicates those clones with matching fragments while the third column indicates those clones which do not have matching fragments.

| Fragment | Matching Clones | Fail to Match |
|---|---|---|
| **KpnI fragment**<br>From 126,086 to 129,196 (size 3,110) | **bX1033 and pX28**<br>Matching bands:<br>Low 3,111 high 3,157<br>Mean 3,134+-23 | **bX1034 pX25**<br>Nearest bands:<br>Low 3,192 high 3,194<br>Mean 3,193 |
| **EcoRI fragment**<br>From 155,269 to 158,311 (size 3,042) | **pX25**<br>Matching band 3,095 | **BX1033 bX1034**<br>Nearest bands:<br>Low 3,119 high 3,122<br>Mean 3,120 |
| **BamHI fragment**<br>From 208,800 to 211,844 (size 3,044) | **None** | **bX1034 pX25**<br>Nearest bands:<br>Low 3,118 high 3,133<br>Mean 3,126 |

**Table 4-8: Selected Polymorphic Sites from the T-cell Receptor Contig.** The left column on this table indicates the restriction enzyme and location of the fragment associated with a possible polymorphic site within the human genomic contig IBC_chr7-ctg23. The second column indicates those clones with matching fragments while the third column indicates those clones which do not have matching fragments.

| Fragment | Matching Clones | Fail to Match |
|---|---|---|
| **HindIII fragment**<br><br>From 76,727 to 79,129 (size 2,402) | **bG1 bG18 bG3 bG30 bG35 bG37 bG4 bG7 pG1 pG3 pG6**<br><br>matching bands:<br>Low 2,382 high 2,442<br>Mean 240+-17 | **bG28 bG8**<br><br>Nearest bands:<br>Low 2,499 high 2,529<br>Mean 2,514 +-15 |
| **BamHI fragment**<br><br>From 98,690 to 100,771 (size 2,081) | **None** | **bG8 pG1**<br><br>Nearest bands:<br>Low 2,419 high 2,807<br>Mean 2,613+-194 |
| **BamHI fragment**<br><br>From 135,639 to 137,734 (size 2,095) | **bG10 bG12**<br><br>Matching bands:<br>Low 2,080 high 2,086<br>Mean 2,083+-3 | **bG6 bG8**<br><br>Nearest bands:<br>Low 2,332 high 2,807<br>Mean 2,570 +-237 |
| **BamHI fragment**<br><br>From 167,634 to 171,578 (size 3,944) | **bG10 bG12 bG24 bG37 bG4 bG5 bG6**<br><br>Matching bands:<br>Low 3,902 high 4,011<br>Mean 3952+-38 | **bG28 bG7 bG9 pG3 pG6**<br><br>Nearest bands:<br>Low 3,497 high 4,685<br>Mean 4,073+-43 |
| **BamHI fragment**<br><br>From 200,722 to 202,951 (size 2,229) | **bG14 bG16 bg24 bg25 bG28 bG4 bG5 pG6**<br><br>Matching bands:<br>Low 2,189 high 2,244<br>Mean 2,223+-17 | **bG33 bG8 bG9**<br><br>Nearest bands:<br>Low 2,102 high 2,807<br>Mean 2,563 +-326 |
| **BamHI fragment**<br><br>From 253,216 to 255,322 (size 2,106) | **bG14 bG16 bG25 bG27 bG39 bG42**<br><br>Matching bands:<br>Low 2,083 high 2,132<br>Mean 2,109+-18 | **bG23 bG33 bG8**<br><br>Nearest bands:<br>Low 2,781 high 2,940<br>Mean 2,843+-69 |
| **HindIII fragment**<br><br>From 298,045 to 301,664 (size 3,619) | **bG14 bG15 bG22 bG25 bG27 bG32 bG33 bG9**<br><br>Matching bands:<br>Low 3,605 high 3,670<br>Mean 3,634+-20 | **bG13 bG8**<br><br>Nearest bands:<br>Low 3,500 high 3,521<br>Mean 3,510+-10 |
| **HindIII fragment**<br><br>From 480,807 to 485,169 (size 4,362) | **None** | **bG19 bG22**<br><br>Nearest bands:<br>Low 4,257 high 4,265<br>Mean 4,261+-4 |

### 4.3.4 Differences Between Physical Mapping and Assembly Validation

Restriction digest fingerprinting has been an effective and useful tool in physical map assembly (Riles *et al*. 1993; Waterston *et al*. 1993), but there are several critical differences between genome mapping and sequence assembly validation. In physical mapping, the problem is to identify overlapping clones by similarity in their digest patterns. The presence of one or more discrepant bands in comparing fingerprints in overlapping clones is expected. Clones are rarely the same length, rarely overlap over their full extent, and may be derived from different haplotypes in a heterogeneous population. Fingerprint matching algorithms have been developed that recognize the common features of an overlapping pair and ignore the discrepancies. False positives and false negatives in scoring the bands on a gel are readily tolerated. In physical mapping, all comparisons are made between experimental data so the precision of electrophoretic analysis is important but the absolute accuracy is not. Fragments exhibiting anomalous migration behavior in gel electrophoresis (Chastain *et al*. 1995) match reliably as long as their anomalous behavior is reproducible.

The goal in sequence assembly validation is to recognize the possible presence of a small number of disparities between the experimentally observed fingerprint and the pattern inferred from the sequence. Many rearrangements, such as a segmental inversion, will alter only two or three of the fragments in a digest that may contain 50 or more bands. Comparisons must be made between experimental data and theoretically derived predicted patterns so the absolute accuracy as well as the precision of mass determination are important. False positive and false negative band calls are potentially confounding

and could be mistaken for fingerprint disparities resulting from an incorrect sequence assembly.

The difficulty of sequence assembly validation by fingerprint comparison increases with the size of the project being analyzed. There are several reasons for this dependence. As the size of the clone increases, the number of bands in the restriction pattern will also increase. This makes it more likely that matches will occur at random, decreasing the information content of a match. As the number of bands in the pattern increases, the number that are expected to deviate from their predicted migration behavior also increases. In a digest with 50 bands, 2 or 3 are expected to deviate from the predicted position by $P<0.05$. The number of disparities arising from a sequence rearrangement is constant while the number of uninformative bands increases. For all of these reasons, the task of assembly validation by fingerprint matching becomes more difficult as the size of the project increases. Trends in high-throughput sequencing are moving toward the use of very large insert clones (200kb BACs and YACs). It is important to be aware that experience in assembly validation based on previous generations of small (10 kb lambda) to moderate (35 kb cosmid) insert vector systems may not be applicable to the case of current BAC or YAC scale projects.

### 4.3.5 Alternative Sequence Assembly Validation Techniques

**High Coverage Clone Maps.** To address the problem of experimental sequence assembly validation, several methods appear worth exploring. The first is the use of high coverage clone maps assembled from restriction fingerprint data to bin the fingerprint

markers by clone content. For a map with a 5X mean clone coverage, there will, on average, be 5 clone ends and 5 clone beginnings in the interval spanned by the sequencing project of interest. These endpoints will define 10 intervals. By comparing the fingerprint content of the overlapping clones, it should be possible to assign most fragments to a unique interval. Comparing this binned set of fingerprint markers to the digest predicted from the assembled sequence will provide a more powerful test of sequence integrity. This strategy is particularly attractive because the necessary data are likely to be available as a result of clone retrieval and mapping work done prior to the initiation of sequence analysis. The strategy needs to be tested in a production setting. Phenomena such as restriction site polymorphisms in the clone libraries, errors in fingerprint band calling, and uncertainty in the physical map may confound analysis.

**Multiple Complete Digest (MCD) Mapping.** Multiple complete digest (MCD) mapping (Gillett 1992; Gillett *et al*. 1996) is a more demanding physical map assembly process that utilizes multiple restriction enzyme digests and complete fragment accounting in the physical map assembly. MCD data should provide a powerful test of sequence assembly. Compared to single digest analysis with complete fragment accounting, MCD offers two advantages. Even if it is not possible to uniquely assign all fragments of each enzyme digest to unique intervals in an MCD map, a uniquely assigned fragment will likely cover every base in the assembled sequence for at least one enzyme digest (as we show above). A single restriction fragment map may be insensitive to some rearrangements if the fragment mass pattern for the rearranged sequence fortuitously matches the original pattern, but it is very unlikely that this will be the case for all of the

enzymes in an MCD data set. MCD mapping requires the analysis of multiple enzyme digests for each clone increasing the necessary experimental work by several fold. Experimental and analytical studies are needed to determine if the additional work of multiple complete digest analysis is warranted.

**Optical Restriction Mapping.** Optical restriction mapping determines both fragment mass and order through the use of advanced microscopy technology to visualize the digest patterns for individual DNA molecules. In principle, the technique is ideally suited to the problem of assembly validation. Optical mapping is capable of determining accurate fragment masses and orders even for large insert clones (Cai *et al*. 1995) and requires very little input DNA, but production scale throughput remains to be demonstrated. A second alternative is the use of 2-dimensional gels (Peacock *et al*. 1985) in which the first dimension is a rare cutting enzyme and the second dimension is a frequent cutting (4-cutter) digest. The resulting data set is a two-dimensional fingerprint for the clone in which each column represents 4-cutter fragments derived from a rare-cutter fragment. Comparing the experimental fingerprint with a pattern predicted from the sequence would provide a powerful test of assembly validity. While only the sequenced clones need be analyzed, 2-D gel analysis is labor intensive, difficult to standardize, and difficult to run reproducibly.

**Ordered Shotgun Sequencing (OSS).** Finally, some sequencing strategies, notably Ordered Shotgun Sequencing (OSS) (Chen *et al*. 1993), incorporate high coverage intermediate length clone end sequences into the sequence assembly. The map built from

these end pair overlaps serves as an intrinsic verification of assembly fidelity and can be used for assembly validation as long as this information has not already been used in assembling the project. Given the high clone coverage (typically 10X) used in OSS framework map generation, it should be possible to choose an initial tiling set of lambda clones from the framework map and to reserve the remaining lambda end pair relationships for assembly validation. Bootstrap procedures could be used to independently verify the validation.

## 4.4 Summary of Sequence Assembly Validation

In summary, comparison of experimental restriction digest fingerprints with inferred patterns derived from finished sequence data may identify some errors in sequence assembly, but high-resolution electrophoretic analysis and accurate scoring of bands are necessary. The problem of assembly validation by fingerprint comparison becomes more difficult as the size of the sequencing project increases. Even with state-of-the-art experimental technology, it is difficult to exclude the possibility of an undetected assembly error such as a large segmental inversion in a BAC-scale sequencing project. In the work presented here, we demonstrate that reliable validation of assembly integrity is possible using multiple restriction digests without the necessity of constructing a full MCD physical map.

# Chapter 5

# Breakpoint Segmentation

Contained within the DNA of the human genome are many different signals that give rise to the genetic blueprint of life. Once extended regions of human genomic DNA are available, compositional analysis on a larger-scale basis can be explored. One approach is to partition a contig according to the frequency of a particular pattern. Among the patterns that could be looked for include tandem repeats, single nucleotides, dinucleotides, higher order oligonucleotides and isochore regions.

One pattern of particular interest is the dinucleotide CG, which is can also be written as CpG (a cytosine linked to a guanine through a phosphate bond). Regions of DNA rich in CpG dinucleotides, also known as CpG islands, are often located upstream of the transcription start site in both tissue specific and housekeeping genes. By identifying the CpG islands, it is thought that regions of DNA coding for housekeeping or tissue-specific genes can be located (Antequera and Bird, 1993) even in the absence of transcriptional activity.

A method we have developed to detect different signals including CpG islands involves a heuristic algorithm employing classic changepoint methods and log-likelihood statistics. A comparison to score-based methods (Karlin and Altschul 1990; Karlin 1994) is provided. A Java applet has been created to allow for user interaction and visualization of the segmentation resulting from the changepoint analysis. The model is tested using

several sequences obtainable from GenBank (Benson, *et al*., 2000), including a 220 Kb fragment of human X chromosome from the filanin (FLN) gene to the glucose-6-phosphate dehydrogenase (G6PD) gene which has been experimentally studied (Rivella, *et al*., 1995; Chen, *et al*., 1996a). Also examined are sequences from two regions of the human X chromosome where subtle CpG islands previously undetected are found. The GenBank accession numbers and clone names are L44140 (HUMFLNG6PD), AF0033528 (bWXD3), and AF003530 (bWXD42).

## 5.1 Introduction

Deoxyribonucleic acid, also known as DNA, is the genetic blueprint for life. DNA is composed of a linear chain of four nucleotide bases: adenine (A), cytosine (C), guanine (G), and thymine (T). Information is encoded in the genome in independently heritable units known as genes. A gene typically includes control signals that determine when it will be active, a promoter which signals where the sequence should be copied into DNA, and a protein-coding region. There are two basic types of genes: housekeeping and tissue specific. Housekeeping genes are genes that are transcriptionally active (i.e. produce proteins), in cells throughout the body. Tissue specific genes, on the other hand, are transcriptionally active only in certain cells. Experimental results suggest that all housekeeping genes and 40% of the tissue specific genes in humans have an associated CpG island (Bird, 1993). It is proposed that by locating CpG islands in sequences of vertebrate DNA gene positions can be postulated. This section will present characteristics of CpG islands in vertebrates and how they can

be distinguished in a statistical fashion. The methods used can later be extended to incorporate segmentation according to other compositions, including tandem repeats and higher order oligonucleotides.

## 5.2 CpG Island Characteristics

Chemically, DNA is composed of nucleoside monomers ("bases") linked by a phosphate from the 3' hydroxyl of one sugar to the 5' hydroxyl of the next. CpG islands are regions of DNA high in the dinucleotide composition CG; that is, where a cytosine residue (C) is immediately followed by a guanine residue (G). The existence of CpG islands in vertebrates, particularly humans and mice, has been studied (Antequera and Bird, 1993; Aissani and Bernardi, 1991; Cross and Bird, 1995; Gardiner, 1996; Macleod, *et al*., 1994). Aissani and Bernardi (1991) and Bernardi (1993) have studied the location of genes in the DNA of vertebrates and have grouped regions of chromosomes into isochores based on the nucleotide composition. It has been determined that both the majority of genes (Antequera and Bird, 1993; Gardiner, 1996) and CpG islands (Bernardi, 1993; Cross and Bird, 1995) are found on the Giemsa light or reverse bands of chromosomes, which are rich in the nucleotides C and G.

The CpG islands studied so far are mainly located upstream (5') of the gene that they are associated with, even though a few are located downstream (3') (Cross and Bird, 1995). Chen *et al*. (1996) discuss this association by examining candidate genes occurring within a region of high G + C DNA. It is possible that CpG islands can be

found in a region where no genes have previously been mapped. This information could help in setting up experiments to determine gene location.

## 5.3 Why CpG Islands can be Statistically Determined

If successive nucleotides in a DNA sequence were independent and identically distributed and residues occurred with equal frequency, it would be expected that by chance a nucleotide G or C would be observed at any given location 50% of the time. However, in genomic DNA, G + C occurs only 40% of the time. One simple method to find interesting regions of DNA would be to look for regions where the observed number of G's and the observed number of C's together exceeds 40%.

Since there are 4 different choices of nucleotides, it is expected that CpG dinucleotides will occur once in every 16 positions or 6.25% of the time by chance alone. As a result of methylation, CpG occurs at 25% the expected frequency (Bird, 1993). Over evolutionary time, this 5' methylcytosine decay has mutated the dinucleotide CpG into TpG (CpA on the complementary strand) so that both TpG and CpA are both over represented (Bird, 1980). A technique that Antequera and Bird (1993) use to locate possible CpG islands is to look at regions of DNA, at least 200 nucleotides in length, where the G + C content is at least 50% and an observed: expected CpG ratio is above 0.6. This criterion has also been used with the software package *CpG Isle* (Larsen *et al*., 1992; Lopez, 1995) which characterizes CpG islands from sequences in the EMBL database. (*CpG Isle* can be obtained from the Internet at the URL ftp://ftp.ebi.ac.uk/pub/databases/cpgisle.)

CpG islands are also known as HTF islands (*HpaII* tiny fragments) since they are cut by the restriction enzyme *HpaII* (Cross and Bird, 1995). Other methods to experimentally determine the location of these islands include looking for rare-cutter sites and G/C boxes within DNA (Aissani and Bernardi, 1991). While these locations can be found experimentally in a wet lab, they can also be located using string-matching algorithms due to their specificity.

## 5.4 Algorithm

### 5.4.1 Segmentation Algorithm

As previously described, determining CpG island location by using the criterion that the G + C content is at least 50% and an observed:expected CpG ratio is above 0.6 will provide some clues as to where CpG islands will occur. However, such an approach can leave undetected CpG islands. It is also very specific to human nucleotide composition. A more sequence and organism independent approach is proposed that will help to detect even subtle CpG islands. Our aim is to implement this approach to search for other regions of compositional bias.

The problem can be approached as a classic changepoint problem (Carlin, Gelfand, and Smith 1992). Lawrence and Reilly (1985) have proposed changepoint methods to determine subsequence conservation within amino acid sequences using maximum likelihood estimation. Similar techniques can be used to determine the location of the breakpoints according to dinucleotide composition. The idea is to segment the DNA sequence into regions adjacent to one another with different CpG distribution.

**First Phase: Breakpoint Segmentation.** A heuristic approach involving greedy optimization through random sampling has been applied to the changepoint problem in order to determine breakpoint locations. The general idea is to iterate a number of times, randomly choosing whether a new breakpoint should be tested, an existing one should be moved, or two adjacent regions should be merged. Each segment is assigned a score according to the formula in Equation 5-1.

$$S = \overline{CpG} * \ln\frac{\overline{CpG}}{N} + CpG * \ln\frac{CpG}{N}$$

**Equation 5-1: Segment Log-Probability Score.**

Here, $S$ is the log probability score, $\overline{CpG}$ is the number of dinucleotides in the segment that are not CpG, $CpG$ is the number of CpG dinucleotides in the segment, and $N$ is the total number of dinucleotides in the segment. Note that $N = L - 1$ where $L$ is the length of the segment in nucleotides. Table 5-1 indicates the dinucleotide counts for an example segment.

**Table 5-1: Dinucleotide Counts for the Sequence ACGGTACGCGCGA.**

| Dinucleotide | Counts | Dinucleotide | Counts |
|---|---|---|---|
| AA | 0 | GA | 1 |
| AC | 2 | GC | 2 |
| AG | 0 | GG | 1 |
| AT | 0 | GT | 1 |
| CA | 0 | TA | 1 |
| CC | 0 | TC | 0 |
| CG | 4 | TG | 0 |
| CT | 0 | TT | 0 |

DNA is typically found in a double stranded conformation where one strand is complementary to the other and running in the opposite direction. Since it may not be known which strand the gene is transcribed from, both strands should be searched for dinucleotides. A nice property of the CpG dinucleotide is that its complement is the dinucleotide GpC. For the sequence ACGGTACGCGCGA, its complement is TGCCATGCGCGCT. The location of CpG islands in the complement should be looked for in the reverse direction due to the orientation. The CpG islands are as follows:

```
5′ ---- ACGGTACGCGCGA ---- 3′
3′ ---- TGCCATGCGCGCT ---- 5′
```

Note that the locations of CpG islands in both strands are identical. Thus, it is only necessary to search one strand for the location of CpG islands.

In order to determine whether or not a given breakpoint is significant, consider the diagram in Figure 5-1. The threshold needs to be chosen in such a way as to ensure that all possible breakpoints are found, yet that no false breakpoints will result. It has been empirically determined that threshold values between 15 and 20 work best. It is also possible that the segmentation can over segment a CpG island. To overcome this problem, a post-processing step is invoked.

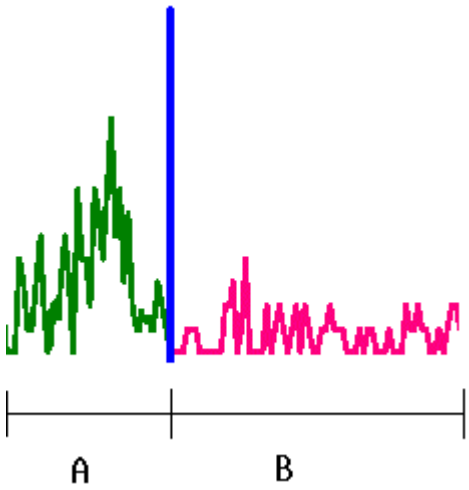



**Figure 5-1. Breakpoint Segment Example.** Let $A_{score} = S$ for segment *A*, $B_{score} = S$ for segment *B, and* $C_{score} = S$ for segment $A+B$ where $S$ can be calculated using Equation 5-1. If $A_{score} + B_{score} > C_{score} + Threshold$, then it is significant and a new breakpoint should be inserted at the location separating segments A and B.

**Second Phase: Post Processing.** The purpose of the post-processing step is to further refine the boundaries of the segments found in the breakpoint segmentation phase. This can be accomplished in one of two ways. The first method is to merge segments together using a lower threshold value. The second method is to determine if two adjacent segments should be merged by determining if they are both above or both below the expected dinucleotide content based on the composition of the DNA sequence being studied. This in effect reverts back to the previous method of testing an observed:expected CpG ratio. Since this is done as a post-processing step, subtle islands will not be missed. By processing the breakpoints in this manner, false positives and fractionation of segments can be eliminated without loss of the true positives.

### 5.4.2 Generalization of the CpG Detection Algorithm

Location of CpG islands is only one application of the segmentation algorithm. Equation 5-1 can be easily changed to allow the user to determine breakpoints in other biologically significant locations. The user is given the option of finding breakpoints according to the C + G content (for the purpose of isolating isochores), mononucleotide content, purine/pyrimidine content (for structural purposes), and dinucleotide content. Equation 5-2 shows a generalization of Equation 5-1.

$$S = \sum_{i=1}^{K} C_i * \ln F_i$$

**Equation 5-2: Generalized Log-Probability Score.**

Here, $K$ is the number of different compositions to segment by, $C_i$ is the count of items in the segment of composition $i$ and $F_i$ is the frequency of items in the segment of composition $i$.

## 5.5 Implementation

### 5.5.1 Java Applet Interface

A Java applet interface has been developed using Sun's JDK 1.1.1. It can be run using any Java-enabled browser at the URL `http://www.ibc.wustl.edu/~ecr/CPG/segment.html`. The purpose of the interface is to allow the user to input a nucleotide sequence in fasta format and then segment it into significant pieces based on the various compositions, the default of which is CpG islands.

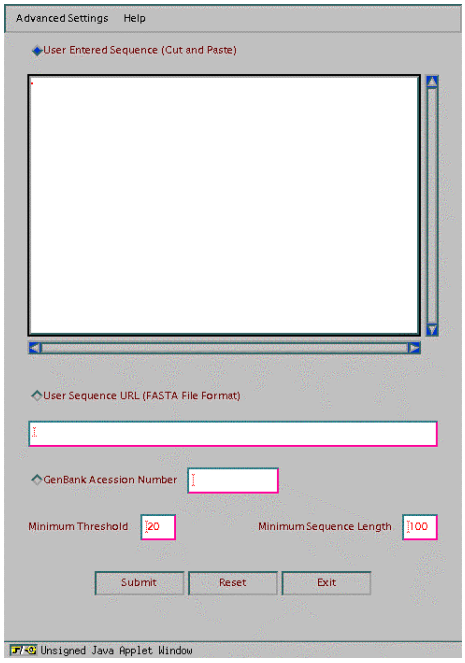



**Figure 5-2: Sequence Fragmentation Interface.** Sequences can be entered either through cut/paste methods, a URL pointing to a valid FASTA file, or by Genbank identification number.

The results will be returned graphically to the user who can then analyze them interactively.

Two frames should initially appear when the applet is run. The first frame is the Sequence Fragmentation Interface frame (see Figure 5-2) which is the main user interface. The second frame is the Status and Message Frame where error messages will be displayed as they occur. Other messages will also appear in this frame in order to inform the user of the status of the breakpoint segmentation.

**Segment Composition.** Clicking on an "Advanced Settings" menu, going to the "Segmentation Criteria" submenu, and clicking on the desired composition can change the criterion used for segmentation. There are currently five different compositions that

can be used for segmentation criteria: mononucleotide, dinucleotide, CpG dinucleotide, purine/pyrimidine, and isochore (C+G) content.

**Fasta Sequence File.** The interface allows the user to input a DNA sequence in fasta format in one of three ways. One method is to input the sequence in a cut and paste fashion. A second method is to enter in a URL that points to a valid fasta file. The third method is to enter in the GenBank id number of a sequence.

Regardless of which method is chosen, a valid fasta file must be present. Fasta file format specifies that the first line begins with a '>' followed by the GSDB sequence accession number, the International Collaboration accession number, and a sequence description. The sequence follows the one line header. For the purposes of this segmentation program, it is only required that the first line begins with a '>'. Valid nucleotide characters of the sequence should follow the standard IUB/IUPAC nucleic acid codes as seen in Table 5-2 (Corhish-Bowden, 1984). Note that the case of the characters can be mixed. In addition, spaces, tabs, and carriage returns are valid characters that will be stripped out prior to segmentation.

Note for the segmentation program, U will be converted to T, and anything besides A, C, G, or T will get set randomly according to the codes in Table 5-2. If an invalid FASTA file is present, an error message will be displayed.

**Minimum Threshold.** The minimum threshold parameter allows the segmentation program to tell when segmentation should occur due to two segments being significantly different. If not enough breakpoints are appearing, lowering the threshold should introduce more. If too many breakpoints are appearing, then raise the threshold. A default

value of 20 generally produces acceptable results. The user can change this value by changing the text box located to the right of the "Minimum Threshold" label. Note that this value is a real number.

**Table 5-2: IUB/IUPAC Nucleic Acid Codes.**

| Symbol | Representation | Symbol | Representation |
|---|---|---|---|
| A | Adenine | M | A C (amino) |
| C | Cytosine | S | G C (strong) |
| G | Guanine | W | A T (weak) |
| T | Thymine | B | G T C |
| U | Uridine | D | G A T |
| R | G A (purine) | H | A C T |
| Y | C T (pyrimidine) | V | G C A |
| K | G T (keto) | N | A G C T (any) |

**Minimum Sequence Length.** The minimum sequence length parameter refers to the minimum number of nucleotides that must be present in a segment. This parameter has been introduced, because without it, over segmentation becomes a problem. A default value of 100 is set. Updating the text box located to the right of the “Minimum Sequence Length” label can change this.

**Post-processing.** There is an additional post-processing parameter that can be set under the "Advanced Settings" menu. By checking the post-processing parameter, the segmentation program will attempt to merge breakpoints back together to form the most optimal results. This option is turned off by default.

### 5.5.2 Interpretation of the Results

Once the breakpoint segmentation has occurred, two windows will pop up. One window indicates “Breakpoint Statistics” (Figure 5-3) while the other is a “Choices”

frame (Figure 5-4). The window in Figure 5-3 contains information on the nucleotide and possibly dinucleotide composition for each of the segments, as well as their beginning and ending points, their length, and the logP value. Included as well are C + G composition, purine composition, and pyrimidine composition statistics for each of the segments. The title of the frame indicates the composition criteria for segmentation. In this case, segmentation is based on CpG content. When looking at the table of dinucleotide composition, the column labels refer to the first nucleotide and the row labels refer to the second nucleotide.

By clicking on any of the buttons in the Choices Frame as shown in Figure 5-4, a graph will appear showing the content of the nucleotide(s) or dinucleotide(s) indicated on the button labels. Color-codes for the graphs are defined in Table 5-3. Note that when

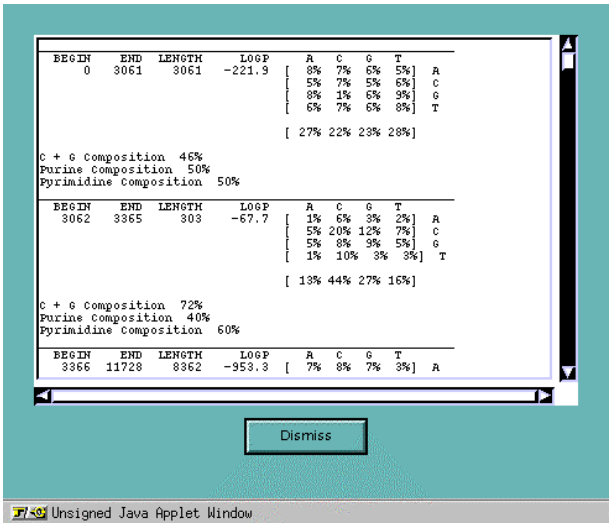



**Figure 5-3: Breakpoint Statistics Frame**. Statistics include beginning and ending points for the segment, the segment length, the logP score for the segment, and various frequencies. For the dinucleotide frequencies, the column label refers to the first nucleotide and the row label refers to the second nucleotide.

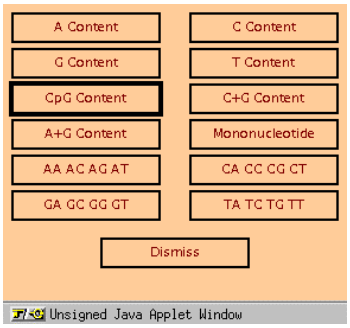



**Figure 5-4: Choices Frame**. Clicking on any of these buttons will cause a graph to appear showing the content of the indicated nucleotide(s) or dinucleotide(s).

**Table 5-3: Nucleotide Color Codes.**

| | |
|---|---|
| Adenine | **Green** |
| Cytosine | **Blue** |
| Guanine | **Black** |
| Thymine | **Red** |
| All Others | **Purple** |

multiple dinucleotides are shown together, the color corresponds to the second nucleotide.

Figure 5-5 shows all of the breakpoints, which are indicated by the vertical dark blue lines. In this case, the breakpoints were determined according to CpG content. Note that the graphs are based on a running average over a specified window size. Editing the text to the right of the “Widow Size” label can change the window size. The graph will change according to the new window size once the "Redraw" button is pressed. The breakpoints might shift slightly to follow this window. If the graphs appear too cluttered, it would be best to increment the window size.

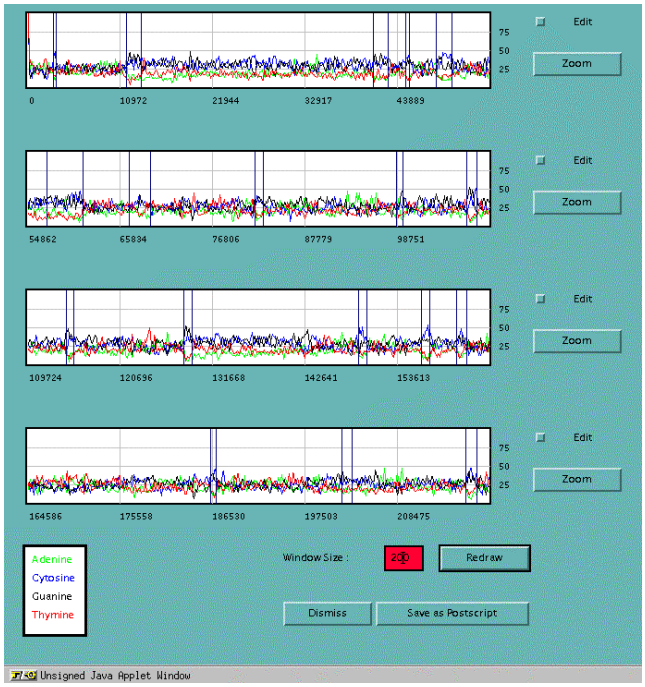



**Figure 5-5: Mononucleotide Content using CpG Segmentation.** Breakpoints are indicated by the dark blue vertical lines. The four lines in the graphs represent a running average of the frequency of the four nucleotides.

Located directly to the right of each of the graphs is an "Edit" choice button. By clicking on this button, a blue background will appear on the associated graph. The user can then select a specific portion of the graph by either clicking or dragging the mouse to the desired location. Once the desired area is covered, the user can press the associated zoom button to zoom in on this region of the graph.

Figure 5-6 shows an example of a zoomed in portion of a graph. The resulting zoom graph is very similar to the previous graphs. There are two main differences. The first is that when only a single composition is to be displayed, there will be blue tick marks underneath the graph indicating where it occurs within the sequence. The second difference is that there is a "View Sequence" button. By pressing this button, the nucleotide sequence will be displayed in a frame as shown in Figure 5-7.

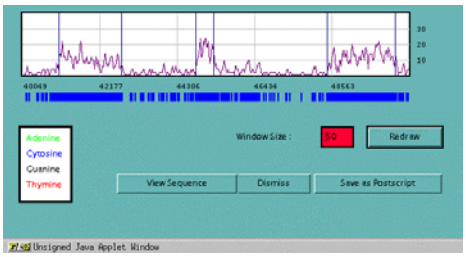



**Figure 5-6: Zoom Graph of CpG Content**. The blue tick marks underneath the graph indicates the occurrence of an item of a particular composition (in this case, CG dinucleotide.)


**Figure 5-7: Nucleotide Sequence Frame**.

## 5.5.3 Implementation Issues

Due to the limitations of Java security, a client/server application is used in order to retrieve sequences from remote locations and to run the segmentation algorithm. Once the user has entered in the desired parameters in the client side applet, the parameters are sent to the server side executable. The server is responsible for taking in the parameters, retrieving the DNA sequence, and segmenting the sequence according to the provided parameters. Once the segmentation is finished, the server sends the results back to the

client where they can be viewed graphically. All of the communication takes place through the use of sockets.

### 5.5.4 Code Statistics

For the client side Java applet, there are currently twenty-one different classes containing a total of 2104 lines of code. The server side consists of a single Java class that is 48 lines long and a C program that has 435 lines of code. The segmentation routines, written in C, take up six files containing a total of 943 lines of code.

Performing the actual segmentation in Java has been attempted, but is not feasible due to the nature of Java as an interpreted language. The bottleneck in the process is in I/O. Table 5-4 shows the runtime comparisons of the Java segmentation program versus the standalone executable created from compilation of C code. Testing was performed using a 55 MHz HyperSparc as the web server. The client side was run on a 200 MHz Pentium Pro machine. This data indicates that the Java interface slows down processing by a factor of 10.

**Table 5-4: Average Runtime Comparisons on a 55 MHz HyperSparc Web Server and 200 MHz Pentium Pro Client.**

| Length (in Nucleotides) | JAVA Sequence Retrieval Time | JAVA Segmentation Time | Total JAVA Time | Standalone Segmentation Time |
|---|---|---|---|---|
| 5828 | 17.3 Seconds | 4.3 Seconds | 21.6 Seconds | 1.06 Seconds |
| 93964 | 19.4 Seconds | 9.3 Seconds | 28.8 Seconds | 2.33 Seconds |
| 219446 | 36.0 Seconds | 30.5 Seconds | 66.5 Seconds | 3.66 Seconds |

## 5.6 Results

### 5.6.1 Human Xq28 Region

Figure 5-8 shows the breakpoint locations calculated within a sequence in the human Xq28 chromosomal region. The default parameters are used with the exception of the post-processing step being allowed. The location of breakpoints found is consistent with the results found by Chen, *et al*. (1996). Our segmentation routine finds all of the CpG islands postulated. An additional CpG island is found between bases 201,861 and 203,041. The implications of this additional CpG island are discussed in Figure 5-8.

### 5.6.2 Human bWXD3 Region

A subtle CpG island that cannot be picked out by the more traditional methods is shown in Figure 5-9 for the bWXD3 region of the X chromosome. The minimum nucleotide length required for a segment is increased to 150. All other parameters take on their default values. Two CpG islands are postulated using these parameters. There is an exon located between bases 68,432 and 68,633 associated with the 3' end of the EDA gene. The first postulated CpG island is located between bases 85,472 and 85,727. This indicates that it is a good candidate located upstream of an exon associated with a gene. The second detected CpG island may indicate that there is another exon within this region. A discussion of predicted exons is included in Figure 5-9.

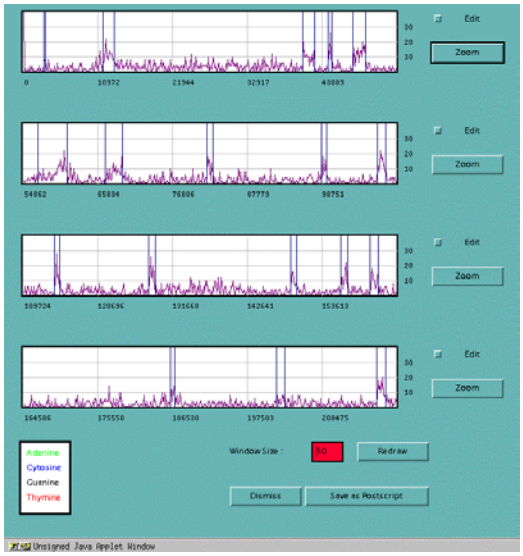



**Figure 5-8: CpG Segmentation for Human Xq28 Chromosomal Region.** Default parameters are used. All seventeen of the CpG islands postulated by Chen, *et al*. (1996) are located. An additional CpG island is found as well between bases 201, 861 and 203, 041. *GenScan* (Burge and Karlin, 1997) and *Grail* (Guan, *et al*., 1992) do not predict any exons in the '+' strand in the region of the additional CpG island, while *GeneID* (Guigo, *et al*., 1992) predicts one. The additional CpG island partially covers exons in the 3' end of the glucose-6-phosphate dehydrogenase gene (E.Y. Chen, *et al*., 1991) on the '-' strand, which has exons spanning from bases 201,336 to 217,196.

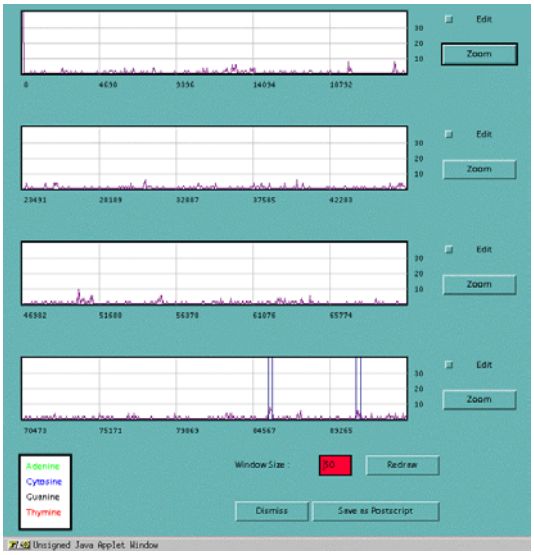



**Figure 5-9: CpG Segmentation Results for bWXD3.** A minimum nucleotide length is set to 150. All other parameters are set according to their default values. Two CpG islands are postulated. An exon associated with the 3' end of the EDA gene is located between bases 68,432 and 68,633, indicating that the 5' end of this gene might be closer to one of the two postulated CpG islands. The two CpG Islands are located between positions 85,501-85,733 and 90,856-91,133. *Grail* and *GenScan* both predict and exon from locations 84,012-84,045. Such an exon could be associated with the first CpG island. *GenScan* predicts an additional exon between locations 93,433-93,592 that may be associated with the second CpG island.

### 5.6.3 Human bWXD42 Region

Figure 5-10 shows the results for the bWXD42 region of the X chromosome that has a hint of a subtle CpG island. There is a cdx4 gene in this region with exons extending between bases 43,025 (3' end) and 50,304(5' end). Using the breakpoint segmentation program with a minimum threshold of 24 and default values for all of the other parameters, a single CpG island is located between bases 48,716 and 50,710. This indicates that that CpG island is actually located in the 5' end of the gene. More research will be pursued to determine the association between CpG island location and

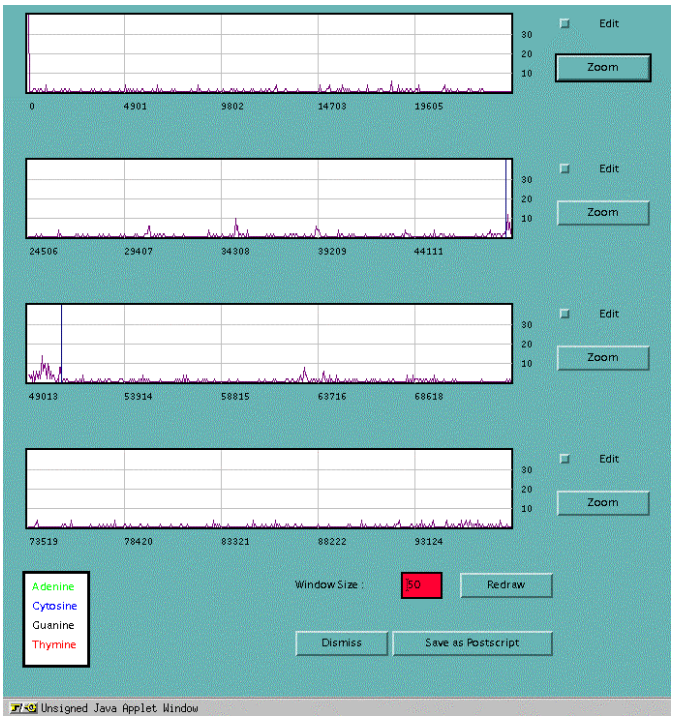



**Figure 5-10: CpG Segmentation Results for bWXD42.** There is a cdx4 gene in this region on the '-' strand with exons extending between bases 43025 (3' end) and 50304 (5' end). A minimum threshold of 24 and default values for all other parameters is used. The postulated CpG island is located in the 5' end of the cdx4 gene.

the 5' end of genes. The sequences for the bWXD3 and bWXD42 regions have been shared by the Washington University Medical School Center for Genetics in Medicine (CGM, 1997).

## 5.7 Comparison to Score-based Methods

As a testing measure, a program incorporating a score-based method (Karlin and Altschul, 1990; Karlin, 1994) to detect CpG islands was created. Traditional scoring criterion was used. In order for a region to be considered a CpG island, the C+G content must be at least 50% and an observed:expected CpG ratio for that region must be above 0.6.

For the Xq28 region (as discussed in Figure 5-8), the score-based method finds all eighteen candidate CpG islands that our algorithm finds. There are also five additional candidate CpG islands postulated by the score-based method. Upon further examination, one of the additional CpG islands results from the splitting of one of our CpG islands. The other four additional candidates actually lie within the coding region of genes. Three of them are contained within the FLN gene, and one within the 2-19 gene, as discussed by E.Y. Chen, *et al*, (1996). Since we are only interested in the CpG islands that signal genes, the results indicate that score-based methods using the traditional criterion are actually over-sensitive to high C+G regions while our algorithm produces the expected results.

The score-based method picks out a CpG island between bases 49,645 and 50,564 for the bWXD42 sequence (as discussed in Figure 5-9). This is consistent with the results of our algorithm.

There are no possible CpG islands for the bWXD3 region (as discussed in Figure 5-10) according to the results of the score-based method. This deviates from the results of our algorithm that proposes two subtle CpG islands.

In order to give the score-based approach a more fair evaluation, we took the resulting CpG frequencies found in the CpG islands using the traditional criterion as the expected value of CpG frequency within CpG islands. We also disregarded the condition that a CpG island must have a C+G content at least 50%. The results for both bWXD42 and bWXD3 are consistent with the previous score-based results. For the Xq28 region, more postulated CpG islands are found. As with the previous score-based method, all of these CpG islands either result from the splitting of previously postulated CpG islands, or they are located within the coding regions of genes.

These comparisons indicate that our approach can produce more useful results than score-based approaches. The score-based approaches modeled here indicate a decrease in specificity without increasing sensitivity when compared to our approach. As a result, score-based methods do not have the ability to detect subtle CpG islands given the traditional segmentation criterion.

## 5.8 Discussion

In addition to the detection of CpG islands, it would be interesting to determine if there are any conserved sequence signals in either the beginning or end of the CpG islands that could lead to the conservation of CpG islands over the course of evolution. Gibbs Sampling (Lawrence, *et al*., 1993) and other similar motif identification programs could be used in this analysis. Other analyses could be performed in order to determine other conserved characteristics of CpG islands, including length, total CpG content and locations relative to the 5' start exons of genes.

There is room to improve the segmentation process. One area is to make a more accurate post-processing procedure to merge breakpoints without losing minor islands. Hopefully this would reduce false positives. Analytical methods to determine segments taking segment length into account could be explored. Perhaps such a method will eliminate the need for a post-processing step.

The goal with the segmentation algorithm is to be able to develop an automatic method to annotate databases with added CpG island information. Hopefully this will add insight into the location of genes. While testing out the capabilities of this algorithm, it will be possible to assimilate a database of CpG islands more extensive than anything else currently available by looking at the human genomic contigs I have assembled.

Discussion of CpG islands has traditionally been limited to vertebrates. A comparison of homologous regions of DNA in mice and humans is possible. Through such a study, it can be determined which islands are conserved and which are lost. Future studies could also include analysis of other organisms including *S. cerevisiae* and *C.*

*elegans* to determine if they have subtle CpG islands. Traditional methods suggest they do not.

The analysis performed so far suggests that there are at least 3 classifications of CpG islands: those having gradual signals, those having sharp left-handed signals, and those having sharp right-handed signals. A method using Kolmogorov-Smirnov testing (Lilliefors, 1967) could be explored in an attempt to classify CpG islands.

Segmentation can be applied to other sequence problems in addition to CpG island detection. The segmentation algorithm could be improved by allowing for the detection of other forms of compositional bias, introduction of higher-order oligomers, repeat sequences, and searching through amino acid sequences in addition to nucleic acid sequences.

Isochores are relatively large regions of DNA which are compositionally homogeneous in their C+G content. Isochores have been studied and classified extensively before high throughput human genomic sequence was available (Bernardi, 1993.) There are four main classes of isochores that have been classified based on density gradient centrifugation. Now that large amounts of sequence data is available, segmenting the genome into isochore regions according to sequence can be accomplished and compared to the earlier results.

The current Java 1.1 implementation could be updated to Java 1.2. Hopefully a newer implementation would lead to faster speed and greater flexibility while maintaining a high degree of available user interaction. A newer version could include enhanced features, such as reading in complete GenBank records. This would allow the

user to view additional features such as predicted and experimental gene locations as well as EST homologies.

# Chapter 6

# Compositional Analysis of Homogeneous Regions in Human Genomic DNA

Due to the increased production of human DNA sequence, it is now possible to explore and understand human genomic organization at the sequence level. In particular, we have studied one of the major organizational components of vertebrate genome organization previously described as isochores (Bernardi, 1993), which are compositionally homogeneous DNA segments based on G+C content. We have examined sequence data for the existence of compositionally differing regions and report that while compositionally homogeneous regions are present in the human genome, current isochore classification schemes are too broad for sequence-level data.

## 6.1 Introduction

It has been proposed that vertebrate genomes, including human, are made up of compositionally homogeneous DNA segments based on G+C content (Bernardi, 1993). These regions, known as isochores, have been studied experimentally using density gradient centrifugation on mechanically sheared DNA in the range of 50-100 kb (Bernardi, 1993) since their discovery in the late '70s (Macaya, Thiery and Bernardi, 1976). Isochores are biologically interesting due to the association between increasing

G+C content and high gene density (Mouchiroud, *et al*., 1991; Gardiner, 1996; Zoubiak, Clay and Bernardi, 1996).

According to Bernardi's theories, there are five families of isochores, each having a different level of cytosine and guanine (C and G, respectively) as described in Table 6-1. There are two G+C-poor isochore families L1 and L2 that make up approximately 60% of the human genome. The isochore family L1 is defined to be regions corresponding to less than 37% G+C content; L2 is defined to be regions containing between 37% and 41% G+C. The isochore family H1 forms 24% of the human genome and corresponds to regions between 41% and 46% G+C. The other G+C rich isochore family H2 forms 7.5% of the human genome and corresponds to those regions containing between 46% and 53% G+C. The final isochore family, H3, forms almost 5% of the genome and corresponds to those very G+C rich regions which are greater than 53% G+C. Since the overall composition of the human genome is approximately 60% AT and 40% G+C, the L1 and L2 families correspond to isochore regions containing less than average G+C content while the H1, H2, and H3 families correspond to isochore regions containing higher than average G+C content. The availability of human genomic sequence makes it possible to explore and understand human DNA composition at a sequence level. We attempted to correlate Bernardi's isochore family definition to sequence data.

**Table 6-1: Isochore Classifications.** This table indicates the GC ranges for the five isochore family classification as defined by Bernardi (2000). The remaining 3.8% of human genomic DNA corresponds to satellite repeats and ribosomal sequences (Bernardi , 2000). [A]Note that the L1 and L2 isochore classes together represent 60 percent of the human genome.

| Isochore Class | Range | Percent of Genome |
|---|---|---|
| L1 | 0-37% GC | 60[A] |
| L2 | 37-41% GC | |
| H1 | 41-46% GC | 24 |
| H2 | 46-53% GC | 7.5 |
| H3 | 53-100% GC | 4.7 |

# 6.2 Methods

## 6.2.1 Analyzing Homogeneous Segments

In order to study the validity of Bernardi's definitions on a sequence level and to examine more properties of the homogeneous regions found in human sequence data, we took the contig sequences for each chromosome available in the April 2001 release of UCSC's Goldenpath (Kent and Haussler, 2001). For each of these chromosomes, we examined the effect of varying the fragment size. This was accomplished by segmenting each chromosome into all possible fragments of 1 kb, 5 kb, 10 kb, 20 kb, 50 kb, 75 kb and 100 kb. For each fragment size, there are 101 possible bins into which each fragment could be placed. Each bin represents a G+C percentage, from 0 to 100. We calculated the G+C percentage for each fragment, and then increased the total counts for the appropriate bin. The histograms were compared to determine the effect of variable fragment size and compositional variation from one chromosome to another. Chi-squared analysis was applied in order to compare the G+C distributions among the chromosomes. In addition, we calculated the frequency of the dinucleotide CG within

each bin in order to test whether or not a correlation exists between G+C content and the occurrence of CpG dinucleotides.

An attempt to validate Bernardi's classifications was made by calculating where isochore boundaries should be based on the percentage of the genome that belongs to each of his classifications. This was accomplished by calculating which histogram bin represents the first 60% of the genome, the next 24%, the next 7.5%, and the next 4.7%.

### 6.2.2 Sequence Homogeneity

The term "isochore" implies a level of high sequence homogeneity. In order to test the validity of this point, we examined 80 different contigs greater than 10 MB in length available through the August 2001 Goldenpath human genome assembly (Kent and Haussler, 2001). The total sequence length of these contigs is over 2 GB in length, representing nearly 2/3 of the human genome. At 1 KB intervals, we calculated the G+C percentage for a surrounding 1 KB, 10 KB, 50 KB, 100 KB, 500 KB, 1 MB and 3 MB window. The variation in the G+C content was calculated and reported. In addition, random sequences were generated corresponding to the lengths of each of the contigs with the following frequencies: A = 0.30, C = 0.20, G = 0.20 and T = 0.30. The same tests in variation were tested for the randomized sequences.

# 6.3 Results

## 6.3.1 Isochore Classifications

Chi-squared analysis was performed on the seven different window sizes (1 kb, 5 kb, 10 kb, 20 kb, 50 kb, 75 kb and 100 kb) for each chromosome in a pair-wise fashion. In each case, the null hypothesis that the distributions of G+C fragments are independent of the window size can be rejected (results not shown). Thus, the isochore classification schemes are highly dependent on the fragment sizes being studied. In the case of the five-class system, the results were skewed towards fragments in the range of 50 kb to 100 kb due to the use of density gradient centrifugation. Figure 6-1 graphically illustrates a

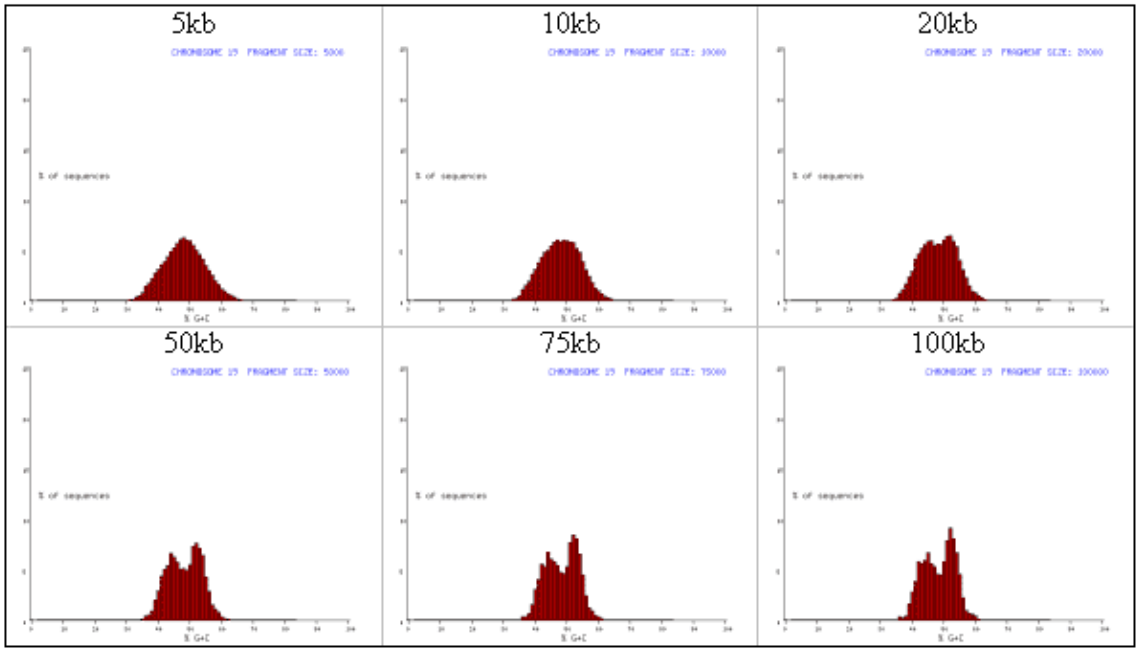



**Figure 6-1: Chromosome 19 G+C Histograms.** Shown in this figure from top left to bottom right are the resulting C+G histograms for chromosome 19 (extracted from the Goldenpath April 2001 release) using 5kb, 10kb, 20kb, 50kb, 75kb, and 100kb fragments. This graph illustrates that the distribution of C+G within a particular chromosome is dependent on the fragment sizes that are used.

dependence on window size with chromosome 19. By looking at this figure, it can be seen that when a smaller fragment size (5 kb) was used when studying chromosome 19, a unimodal distribution of G+C fragments is observed. When the window size was increased (50 kb - 100 kb), a bimodal distribution of G+C fragments can be seen.

In order to determine whether or not G+C content distribution is chromosome specific, Chi-squared analysis was performed (results not shown). The distributions of G+C fragments using 75 kb windows was compared for each pair of chromosomes. The null hypothesis that the G+C content distribution of any two given chromosomes is similar was rejected, no matter which two chromosomes were compared. Displayed in

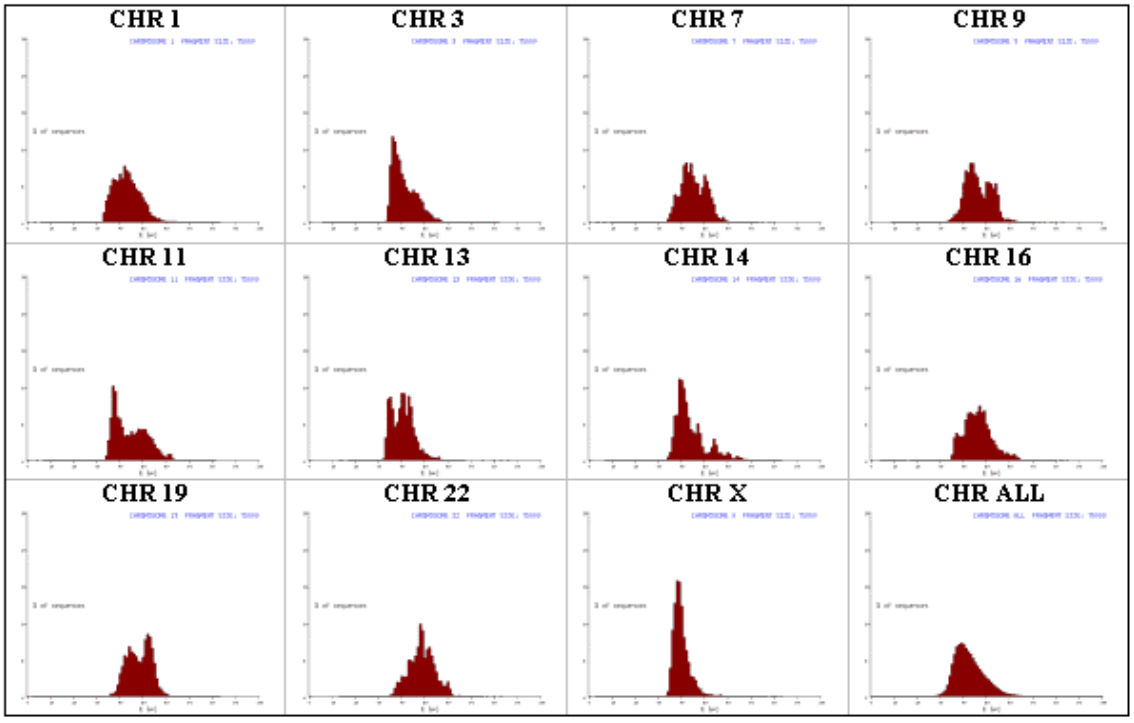



**Figure 6-2: Chromosomal Histograms for 75 kb Fragments.** Shown in this figure are the resulting G+C histograms for the following chromosomes: Row 1: (left to right): 1, 3, 7, 9. Row 2: 11, 13, 14, 16. Row 3: 19, 22, X, ALL. The X-axis represents the G+C content, and the Y-axis represents the percentage of fragments falling within a given G+C content. These histograms were created using the April 2001 Goldenpath release (http://genome.ucsc.edu).

figure 6-2 is the distribution of G+C fragments using a 75 kb window for eleven different chromosomes and the genome as a whole. As this figure shows, there are vast differences in the G+C fragment distribution among chromosomes. Some chromosomes, such as 1 and X, appear to have a distinct unimodal distribution of fragments at the 75 kb window level. Other chromosomes, such as 9, 11 and 19 seem to have distinct bimodal distributions in the G+C fragments. However, in none of the cases were there more than two distinct peaks in the distribution of G+C fragments. Our results show the difficulty of defining isochore boundaries based on sequence data alone. We do see, however, that there does appear to be two distinct isochores that were observable: the majority that are in low G+C, and those that are high in G+C. Further division of these two major groups based on sequence data appears to be a difficult, if not impossible, task.

According to the density gradient centrifugation experiments performed by Bernardi, 60% of the human genome falls into an L1+L2 isochore classification, 24% is H1, 7.5% is H2, and 4.7% is H3. Table 6-2 was created using these guidelines to split the histograms for 75 kb fragments for the various chromosomes into densities of 60%, 84%, and 91.5%, which would theoretically find the isochore boundaries. Not surprisingly, we see that when all of the chromosomal data was inspected, 60% of the histograms lie at 43% G+C or less, which is just above the cutoff for the L2-H1 isochore boundaries. 84% of the histograms lie at 48% G+C or less, which is just above the cutoff for H1-H2 isochores. 91.5% of the histograms lie at 51% G+C, or slightly less than the H2-H3 isochore cutoff of 53% G+C. However, Table 6-2 also shows that these cutoffs do not

**Table 6-2: Boundary Locations Based on Total Percent of all Fragments.** Shown in column 1 is the chromosome label. Column 2 indicates the breakpoint where 60% of all 75 kb fragments for the given chromosome lie. Column 3 indicates the breakpoint under which 84% of all 75kb fragments lie. Column 4 indicates the breakpoint under which 91.5% of all 75 kb fragments lie. Note that the breakpoints of 60%, 84%, and 91.5% indicate breakpoints for the defined isochore classes L2-H1, H1-H2, and H2-H3 (Bernardi, 2000).

## Isochore Boundary locations based on total percent of all fragments

| Chromosome | 60% of all fragments<br>L2-H1 Boundary | 84% of all fragments<br>H1-H2 Boundary | 91.5% of all fragments<br>H2-H3 Boundary |
|---|---|---|---|
| BERNARDI | 42% G+C | 47% G+C | 53% G+C |
| 1 | 44% G+C | 49% G+C | 51% G+C |
| 2 | 44% G+C | 47% G+C | 49% G+C |
| 3 | 41% G+C | 47% G+C | 49%G+C |
| 4 | 40% G+C | 43% G+C | 45% G+C |
| 5 | 41% G+C | 44% G+C | 46% G+C |
| 6 | 39% G+C | 43% G+C | 45% G+C |
| 7 | 46% G+C | 51%G+C | 52% G+C |
| 8 | 42% G+C | 45% G+C | 49% G+C |
| 9 | 47% G+C | 53% G+C | 54% G+C |
| 10 | 44% G+C | 48% G+C | 49% G+C |
| 11 | 46% G+C | 52% G+C | 55% G+C |
| 12 | 44% G+C | 48% G+C | 50% G+C |
| 13 | 41% G+C | 44% G+C | 47% G+C |
| 14 | 43% G+C | 51% G+C | 55% G+C |
| 15 | 43% G+C | 46% G+C | 47% G+C |
| 16 | 47% G+C | 51% G+C | 55% G+C |
| 17 | 49% G+C | 52% G+C | 54% G+C |
| 18 | 41% G+C | 44% G+C | 46% G+C |
| 19 | 51% G+C | 54% G+C | 55% G+C |
| 20 | 47% G+C | 50% G+C | 53% G+C |
| 21 | 50% G+C | 55% G+C | 56% G+C |
| 22 | 50% G+C | 54% G+C | 56% G+C |
| X | 40% G+C | 43% G+C | 45% G+C |
| Y | 39% G+C | 42% G+C | 43% G+C |
| ALL | 43% G+C | 48% G+C | 51% G+C |

correlate with isochore boundaries for all chromosomes. Some chromosomes, such as chromosomes 9, 11, 14, 16, 17, 19, 21 and 22 have more fragments that are G+C rich, while other chromosomes such as 4, 5, 6, 13, 18, X and Y have more fragments that are G+C poor. These results suggest that calculating the isochore boundaries based on the fragment density is not valid when applied to individual chromosomes.

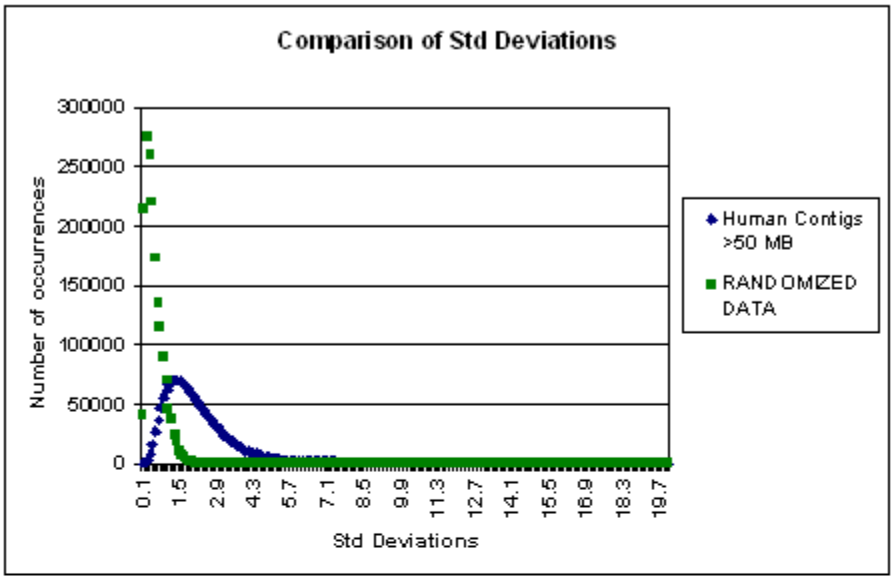
### 6.3.2 Sequence Homogeneity

Figure 6-3(A) illustrates the distribution of standard deviations in G+C content for every 1000$^{th}$ base in both the randomly generated contigs and Goldenpath contigs greater than 10 MB in length. The mean was computed by calculating the G+C content for windows of 1 KB, 10 KB, 50 KB, 100 KB, 500 KB, 1 MB and 3 MB. As figure 6-3(A) shows, the distribution of standard deviations for the random sequence is much tighter and closer to zero than the distribution of standard deviations for the actual human sequence. Figure 6-3(B) shows the calculated cumulative percentage of standard deviations. Examination of this data indicates that in random sequence data, 50% of the points examined have a standard deviation in G+C content of ± 0.4%, while for the real sequence data this number is ± 1.8%. 75% of all random points have a standard deviation of ± 0.7% or less, while this number grows to ± 2.6% in the real sequence data. 95% of all random fragments have a standard deviation of ± 1.2%. This number grows to ± 4.5% in the real sequence. In fact, only 24% of all real sequence data had a standard deviation of ± 1.2% or less. These results indicate that the human genome is much more heterogeneous than the theories of Bernardi (1993) lead one to believe.

## 6.4 Discussion

In order to understand the concept of a 5-class isochore system as proposed by Bernardi, it is important to revisit the experimental procedures performed over 25 years ago. In the article where isochores were first described (Cuny *et al*., 1981), human genomic DNA was found to be fractionated into five major components using CsCl

A) Distribution of Standard Deviations from a Mean G+C Content




B) Cumulative Percentage of Standard Deviations from a Mean G+C Content

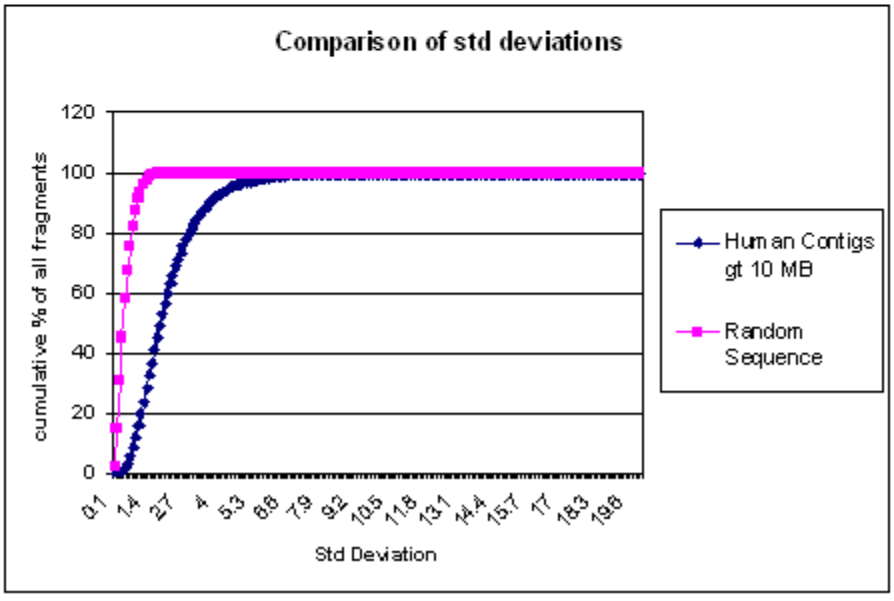



**Figure 6-3: Distribution of Standard Deviations from a Mean G+C Content.** Shown in A) is the count of each standard deviation calculated for every 1000$^{th}$ base in human and randomized contigs using window sizes of 1 KB, 10 KB, 50 KB, 100 KB, 500 KB, 1 MB and 3 MB. B) shows the cumulative percentage of standard deviations from figure 3 falling under a certain percentage.

density profiles. Each component represents a set of DNA segments that sediments differently based on different buoyant densities. The results presented are based on earlier analyses of the composition of eukaryotic genomes (Thiery, Macaya and Bernardi, 1976). Thiery *et al*. (1976) looked at the separation of human DNA using thirteen different density gradients. What results are thirteen different Gaussian distributions of absorbance, each representing a different distribution of genomic DNA based on G+C content. Three main observations of the experimental work are discussed.

First of all, the decision to choose five major components (later given the label "isochores" by Cuny, *et al*., 1981) seems somewhat arbitrary. In fact, examination of Figure 1 of Thiery, *et al*. (1976) indicates that any of the thirteen different results could have been chosen as major components. In addition, if more than thirteen different density gradients were examined, a different distribution of major components could potentially result.

The second critique is that the Gaussian distributions resulting for each of the labeled major components are overlapping. This means, for instance, that a fragment of human genomic DNA containing an average G+C content of 47% could potentially wind up belonging to multiple major components, or isochore families. This is a major problem when looking at a sequence level comparison. It is a necessary requirement that each individual sequence fragment be assigned to a single classification, or at most, belong to an unknown area between two breakpoints.

The final critique is that density gradient centrifugation experiments can only allow for the fractionation of DNA based on the overall G+C content of any segment. It

does not seem to be in any way possible to determine the homogeneity. In fact, the only means by which homogeneity can be discerned is by looking at finished sequence data.

The density gradient centrifugation experiments are important in that they indicate that there are larger regions of the human genome with a conserved low or high G+C content. However, the previous school of thought of a five-class isochore system for the human genome with strict boundaries appears to be out-of-date in light of the availability of sequence data.

Our results have shown the difficulty of defining isochore boundaries based solely on sequence data. This is supported by failed attempts of window-based sequence segmentation resulting in arguments against strict definitions of isochore classes (IHGSC, 2001; Nekrutenko and Li, 2000; Häring and Kypr, 2001). We do see, however, that there does appear to be two different classes of isochores that can be observed: the majority that are low in G+C, and those that are high in G+C. Further breakdown of these two major groups based on the sequence data appears to be a difficult task.

# Chapter 7

# Accounting for Regions of High and Low G+C Content Found in Human Genomic DNA

The increased availability of finished human genomic sequence data has made it possible to analyze human genomic organization at the sequence level. Examination of sequence data indicated regions of high and low G+C content exist within the human genome. Different hypotheses were presented examining why these regions are present in the human genome, including the widely studied hypotheses that these regions are maintained in the human genome via various mechanisms. Preliminary tests of one of these hypotheses strongly suggested high and low G+C regions have *not* been maintained by the presence of repetitive elements with a high or low G+C content within them. Examination of a mutational hypothesis supports the conclusion that compositional mutation biases influenced the evolution of the human genome. However, the observed mutation biases did not seem to have maintained the regions of high G+C content. Rather, preliminary results indicated different substitution rates were in effect in different regions of the genome. This led to a detailed examination of a separate hypothesis that the human genome began as a G+C rich ancestral genome that mutated towards the present-day A+T rich genome. Different regions of the genome may have mutated at different rates, presenting the current mosaic view of the human genome. The preliminary study of composition specific substitution rates in repetitive elements and

pseudogenes suggested that features inserted into the human genome under less selective pressure appear to be mutating towards a higher A+T composition with a rate dependent upon the local G+C context at the insertion site.

## 7.1 Introduction

It has been proposed that vertebrate genomes, including human, are made up of compositionally homogeneous DNA segments based on G+C content (Macaya *et al*., 1976; Cuny *et al*., 1981). These regions, known as isochores, have been studied for nearly 30 years using experimental density gradient centrifugation techniques (Bernardi, 1993; Macaya *et al*., 1976). The theories of Bernardi *et al*. (reviewed in Bernardi, 1993) suggest five separate classes of isochores are found within the human genome. These five classes are defined and separated from each other by different levels of G+C composition. The availability of bulk human genomic DNA sequence has made it possible to study these regions in more detail.

Recent sequence level studies argue the human genome is not nearly as homogeneous as Bernardi's 5-class system of isochore classification might lead one to believe. Rouchka and States (2002) show isochore classifications are specific to the fragment size and chromosome being studied. Nekrutenko and Li (2000) show the human genome is highly heterogeneous both within and between chromosomes and suggest the previous isochore definitions of Bernardi should be relaxed to allow for the high heterogeneity index within human genomic DNA. The International Human Genome Sequencing Consortium (2001) tested the variance of G+C content within

windows of the human genome and concluded that, within a given window, the variance is far too large to be in agreement with a definition implying regions with strong homogeneities. While these recent studies have brought to light that a strict five-class system based on G+C content may not be the best approach for sequence segmentation in human DNA, all of the authors seem to agree with Bernardi that large regions of long-range variation in high and low G+C content are present in the human genome (IHGSC, 2001).

At least two categories of theories have emerged to account for these regions. The first category, the maintenance hypotheses, states regions of high and low G+C content are present in the human genome due to various poorly specified mechanisms that promote compositional maintenance. The second category hypothesizes regions of high and low G+C content are observed within the human genome due to regional variations in mutational rates across the genome.

### 7.1.1 Overview of Maintenance Hypotheses

Several theories have recently been proposed arguing in support of maintenance mechanisms (see Eyre-Walker and Hurst, 2001, and Bernardi, 2000, for reviews). The two main arguments stem from a selectionist hypothesis that a selective process was at work to promote G+C compositional regions and a neutralist hypothesis that states no selection has occurred. The neutralist theories can be broken down into two camps, those subscribing to biased gene conversion theories and those who believe some sort of mutational mechanism was at work.

**Selectionist Hypothesis**. The selectionist argument for the presence of high and low G+C regions suggests these regions arose due to selective advantages. In particular, G:C base pairs contain three hydrogen bonds while A:T base pairs contain two, and thus G:C base pairs should provide greater stability at higher temperature levels (Wada and Suyama, 1986). The argument for the presence of high and low G+C regions in warm-blooded vertebrates due to selection stems from the apparent observation of an "isochore" structure in mammals and birds, while genomes of cold-blooded vertebrates including fish and amphibians are devoid of such structure (Bernardi, 1993). One explanation is an increase in G+C content could provide thermodynamic stability against degradation by heat (Bernardi, 2000). Ohama *et al*. (1987) show that in some bacterial genomes, the overall G+C content is related to different selective pressures in the environment, including thermostability. However, a conflicting study by Galtier and Lobry (1997) shows genomic G+C content is not correlated with optimal growth temperature when 224 different prokaryotes were examined. Bernardi (2000) suggests this lack of correlation could be due to other selective factors such as DNA-binding proteins (Robinson *et al*., 1998) and thermostabile chaperonins (Taguchi *et al*., 1991) that act to stabilize genomic DNA. This hypothesis of high/low G+C structure as a selective advantage to homeothermy has additionally been questioned due to the apparent presence of an "isochore" structure in the genomes of the cold-blooded Nile crocodile and red-eared slider turtle when 16 different genic regions are studied (Hughes *et al*., 1999). This result indicates the strong possibility that "isochore" evolution predated homeotherm evolution.

Conflicting studies by Galiter and Lobry (1997) and Hughes *et al*. (1999) appear to argue against the selectionist hypothesis presented by Bernardi by giving counterexamples. The most unfortunate property of the selectionist hypothesis as presented is that it cannot be easily tested using scientific rigor. While the argument may have some merit, it appears to be grounded more at a philosophical rather than factual level. Therefore, in the remainder of the discussion, the selectionist hypothesis as proposed by Bernardi is not considered and tested.

**Biased Gene Conversion.** During meiotic recombination, two homologous genomic fragments originating from sister chromosomes form a DNA heteroduplex. Since these fragments originate from sister chromosomes, heterozygous sites are possible. Gene conversion is the molecular process in which one allele of a gene is converted into the other at these heterozygous sites. The biased gene conversion (BGC) hypothesis states regions of the human genome have been maintained at a higher (lower) G+C composition due to a bias in A|T→G|C (G|C→A|T) gene conversion events (Galtier *et al*., 2001). Biased gene conversion has been shown to play a potential role in the maintenance of high G+C regions, due to the high G+C content of regions in recombination hotspots such as regions encoding ribosomal operons, tRNAs and histones (Galtier *et al*., 2001). Galtier *et al*. (2001) suggest the BGC hypothesis could account for the bias in G|C→A|T vs. A|T→G|C mutations found within single nucleotide polymorphisms (Eyre-Walker, 1999).

**Mutational Bias**. A third hypothesis for the presence and maintenance of high and low G+C regions is the mutational bias hypothesis. This hypothesis states these regions were maintained by biases in mutational mechanisms favoring A|T→G|C mutations in G+C rich regions and G|C→A|T mutations in G+C poor regions. Thus, if a G+C poor segment of DNA inserted into a region that was G+C rich, over time the G+C poor segment would mutate and evolve to the surrounding G+C composition.

Filipski (1987) study the correlation between coding regions and their surrounding G+C content and codon usage, a phenomenon now well studied (Knight, Freeland and Landweber, 2001; D'Oniofro and Bernardi, 1992). Filipski suggests differences in composition arise from mutational biases contributed by the fidelity of α and β polymerases. The α polymerase, the main replicating enzyme, maintains higher sequence fidelity. The β polymerase, a DNA repair enzyme, is much more error prone. The β polymerase mostly acts on relaxed G+C rich chromatin regions. Thus, Filipski argues, regions of differing G+C content have been maintained due to mutational biases caused by the actions of the β polymerase.

Wolfe, Sharp and Li (1989; also see Wolfe, 1991) study mutation rates in silent sites in thirteen genes and two pseudogenes found in humans and Old World monkeys and 88 genes found in mouse and rat. Their results provide evidence for a significant difference in mutation rates in different regions of G+C content in mammals. They suggest compositional biases could be due to differences in replication conditions. They note high G+C regions replicate early in the S-phase of the cell cycle when dGTP and dCTP is high in the dNTP pools. As the S-phase progresses, the dGTP and dCTP

concentrations decrease, and low G+C regions replicate. As a result, A|T→G|C mutations are more likely to occur early in S-phase replication (or in high G+C regions) and C|G→A|T mutations are more likely to occur later in low G+C regions.

Casane *et al*. (1997) perform similar experiments on three argininosuccinate-synthetase-processed pseudogenes and the surrounding non-coding regions in human, orangutan, baboon and colobus. Their results show the ratio of the G|C →A|T mutation rate to the A|T→G|C mutation rate varied according to G+C content of the genomic position. This indicates a mutational bias was at work.

Francino and Ochman (1999) suggest high and low G+C regions result from mutation events in their study of α and β globin clusters of genes and pseudogenes in humans and Old World monkeys. Their results from this limited data set indicate the ratio of G|C→A|T to A|T→G|C mutations produces strikingly different results when the composition of the genes and pseudogenes is considered. They conclude a compositional bias in mutation rates existed which in turn promoted the formation of high and low G+C content regions.

Mutational biases have also been observed within bacterial genomes. Ohama *et al*. (1987) examine the G+C composition of the streptomycin operon in two separate bacterial organisms with different overall G+C content. The *Escherichia coli* genome is approximately 45% G+C while the *Micrococcus luteus* genome is approximately 74% G+C. The high G+C content of the *M. luteus* genome affects the G+C composition of the *str* operon which has a mean G+C content of 67%, much higher than found in *E. coli*

(51%). In addition, 95% of all wobble bases in the *M. luteus str* operon are either G or C compared to only 52% in *E. coli*.

Fryxell and Zuckerkandl (2001) suggest context dependent mutational biases is possibly due to cytosine deamination, which causes C→T and G→A transitions within mammals. It decreases in rate two-fold for each 10% increase in G+C content. This implies the higher the G+C content, the lower the rates of C→T and G→A mutations will be, and similarly, lower G+C content will produce a higher rate of C→T and G→A mutations through cytosine deamination. This bias could be due to a higher concentration of methylation/deamination enzymes in regions of lower A+T composition. Cytosine deamination would then function as a positive feedback loop, promoting maintenance of both high and low G+C regions.

### 7.1.2 Overview of Regional Variation in Mutation Hypothesis

In 1972, before the notion of isochores in vertebrates was introduced, Cox argued (albeit with little hard scientific evidence) that the spontaneous mutation rate within mammalian DNA varies over the entire genome. This conflicts with the previous assumption that mutation rates were uniform throughout genomes (Sueoka, 1962). More recent studies have begun to illustrate that variation in mutation rates across a genome appears to be present. Wolfe, Sharp and Li (1989) discuss significant variation they observed in silent site mutation rates along the human genome in their discussion of mutation within pseudogenes found in humans and old world monkeys. Casane *et al.* (1997) looked at pseudogenes within four closely related species. Among the results of

their work is the suggestion that a regional variation in the mutation rate exists. This stems from their observation that pseudogenes appear as mutational "hot" spots located within mutationally "cold" regions. Castresana (2002) studied the rates of evolution in a set of mouse and human genes, comparing the rates within the exonic and alignable intronic regions. Castresana concludes the most likely explanation for the observed correlation in evolutionary rates in exonic and intronic regions was the existence of local nonrandom fluctuations in mutation rates of a nonrandom nature.

Regions of high and low G+C could potentially arise due to regional variations in mutation rates. A hypothesis studied herein is that the human genome evolved over time from a G+C rich ancestral genome. As discussed in the results, substitution rates within the human genome appear to have moved the genome towards A+T richness. This rate would appear to have been slower, but nonetheless present, in regions of high G+C. The variability in the mutation rate hypothesis suggests regions of high G+C are seen in the present view of the human genome due to their location in regions of low mutation while regions of low G+C tend to be located in mutation hot spots.

### 7.1.3 Understanding Large-scale G+C Variation

The interest in understanding large-scale G+C variation within the human genome led to the exploration of experiments designed to test the maintenance hypothesis. However, preliminary results appear to support instead the regional variation in mutation rate theory. Two hypotheses for the maintenance of high and low G+C regions were explored. The first hypothesis states high and low G+C regions were maintained by the

presence of repetitive elements with a high or low G+C content within them. The second hypothesis tested was that a compositional bias for mutation rates existed which promoted the maintenance of such regions. The results ruled out the possibility of the G+C content of repetitive elements determining regions of high and low G+C composition. Based on the study of compositional specific mutation rates in repetitive elements and pseudogenes, it is believed that compositional biases in mutation rates did occur within the human genome. However, these biases do not seem responsible for the maintenance of high G+C regions. In addition, features likely to be under less selective pressure inserted into the human genome appear to have mutated towards a higher A+T composition, regardless of the G+C context in which they were placed.

## 7.2 Exploration of Two Maintenance Hypotheses

One of the shortcomings of previous studies into the mechanisms suggesting maintenance of regions of high and low G+C content is they are largely based on looking at genic regions within the genome. While an underlying association between genes and G+C content does exist (Zoubak, Clay and Bernardi, 1996), genes only account for 3-5 percent of the human genome (Gardiner, 1996). In order to understand regions of high and low G+C composition more completely, potential maintenance of these regions was studied by looking at two features in the human genome less likely to be under selective pressure. Such an approach may rule out other evolutionarily advantageous mechanisms that were at work. The first feature is repetitive elements, which make up at least 35

percent of the human genome (Jurka, 1998). The second feature is processed pseudogenes.

Two separate hypotheses for the maintenance of regions of high and low G+C content were studied. The first hypothesizes regions of high and low G+C content were determined by the G+C content of the repetitive elements contained within them. The second suggests regions of high and low G+C content were evolutionarily maintained by mechanisms promoting compositional mutational bias. The second hypothesis is an expansion of the mutational bias hypothesis previously discussed. The methods were based on analysis of the University of California-Santa Cruz's Goldenpath rough draft assembly of the human genome (Kent and Haussler, 2001; `http://genome.ucsc.edu/`).

## 7.3 Maintenance Hypothesis 1: Regions of High/Low G+C Result from Repetitive Element Composition

Previous studies show the densities of certain types of repetitive elements such as ALU, L1, and MIR are not uniform throughout the human genome (Belle and Eyre-Walker, 2002; IHGSC, 2001; Pavlíček *et al*., 2001; Matasi, Labuela and Bernardi, 1998; Jabbari and Bernardi, 1998). The pattern of distribution of G+C rich SINE elements (the mean G+C content of the representative ALUs is 52%) and G+C poor LINE elements (L1 elements are 37% G+C) is particularly intriguing (Belle and Eyre-Walker, 2002; IHGSC, 2001; Eyre-Walker and Hurst, 2001). SINEs and LINEs both incorporate the LINE transcription mechanism (Jurka, 1997). In both cases, the LINE endonuclease selectively chooses the cleavage site TTTT/A to prime reverse transcription (Feng *et al*., 1996). It would be thought that such an insertion mechanism promotes SINEs and LINEs

both within A+T rich regions due to an increased likelihood of finding a cleavage site. However, it has been shown LINEs tend to be found in A+T rich regions, while SINEs are found in more G+C rich regions (IHGSC, 2001; Eyre-Walker and Hurst, 2001), although more recent ALUs are more evenly distributed in the genome (Eyre-Walker and Hurst, 2001).

One potential explanation leading to the appearance of high and low G+C regions in the human genome is regions of G+C variation are caused by the presence of repetitive elements within them. Under this hypothesis, regions of high G+C will exist in the human genome due to a high density of G+C rich SINEs within them. Similarly, regions of low G+C should be observed due to the high density of G+C poor LINEs in these areas. If repeats alone were responsible for regional variation, there should be no correlation between regional G+C content and the G+C content of the unique sequence contained within.

### 7.3.1 Calculating Repetitive and Non-repetitive G+C Composition

One method used to determine whether or not the G+C content of repetitive elements biases regions towards a given G+C distribution was to compare the G+C composition of the region as a whole to the G+C composition of the repetitive and potentially unique (non-repetitive) regions. If the repetitive elements were the driving force behind the overall G+C composition, then there should be a higher correlation between the G+C content of the repetitive elements and the G+C content of the overall

region. At the same time, the G+C content of the unique regions should remain neutral and randomly vary based on the G+C content of the repetitive elements.

This hypothesis was explored by examining the Goldenpath December 2001, assembly of the human genome, which breaks apart the human genome sequence into 2,992 contigs comprising 2.8 billion bases. Only contigs mapped to a particular chromosome were considered. Known repeats from the `Repbase` database version 6.10 (Jurka, 2000) were masked out using `RepeatMasker` (Smit and Green, unpublished; `http://repeatmasker.genome.washington.edu/`). Each of the contig sequences was run through `RepeatMasker` twice. One run was performed in the slower, native settings for the detection of low complexity and simple repeats (using the `-int` option). The second run took advantage of the `-w` option, which incorporates `wublastn` as the underlying alignment algorithm (Bedell, Korf and Gish, 2000) resulting in a significant speed up in the detection of interspersed repetitive elements.

### 7.3.2 Repetitive Element Composition Results

A total of 51.6% of the bases were masked out, indicating they contained some form of repetitive sequence structure. For each of the 2,992 contigs, the G+C composition of the overall, masked, and unmasked regions was recorded. The G+C composition of each overall contig was compared to the G+C composition of the masked regions and unmasked regions looking for correlations. Figure 7-1 shows the resulting plot for those contigs greater than 250 KB in length. As the graph clearly shows, there was a positive correlation between both the unmasked (potentially non-repetitive) G+C

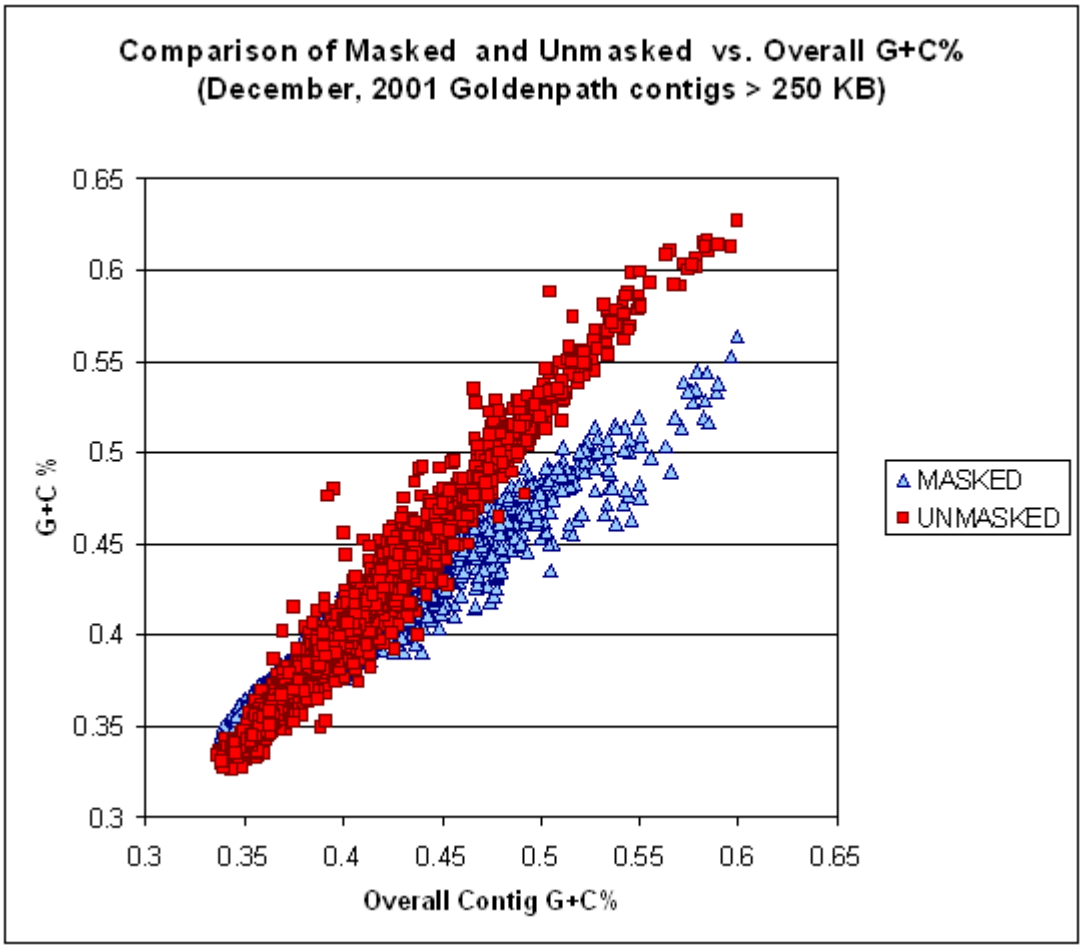



**Figure 7-1: Comparison of G+C Content.** Shown in this figure is the comparison of the G+C content of masked (repetitive) and unmasked (potentially non-repetitive) regions compared to the overall G+C content for each of the 1,927 Goldenpath contigs greater than 250 KB in length from the December, 2001 build. The x-axis represents the overall G+C content of each contig. Regions were masked using `RepeatMasker` (Smit and Green, unpublished; `http://repeatmasker.genome.washington.edu`).

content and the overall G+C content, as well as between the masked portion G+C content and the overall G+C content. Correlation coefficients and *t*-scores were calculated for each of these comparisons. In the case of the masked/overall comparison, the correlation coefficient of 0.9620 yielded a *t*-score of 192.55. For the unmasked/overall comparison, the correlation coefficient is 0.9532, corresponding to a *t*-score of 172.45. In each of

these cases, the *t*-score was much greater than the critical value of 2.58 (using a *p*-value of 0.995; $\alpha = 0.005$). Thus, these correlations were highly statistically significant.

A positive correlation between the G+C content of the masked regions and the overall contigs was expected. This is due to the previously reported positive correlation between increasing genomic G+C content and G+C rich SINE elements and the negative correlation between increasing genomic G+C content and the density of A+T rich LINE elements (IHGSC, 2001; Eyre-Walker and Hurst, 2001). However, such a strong positive correlation between the overall G+C content and the G+C content of the unmasked regions was not expected. Since the unique regions were highly correlated with the overall G+C content, it cannot be concluded that the G+C content of repetitive regions was responsible for the variable G+C content within the human genome.

It could be postulated there was some sort of mechanism for preferential insertion of low G+C repetitive elements into genomic regions of low G+C, while high G+C repetitive elements were inserted into genomic regions high in G+C content. However, as previously discussed (Feng *et al*., 1996), SINEs and LINEs use the same mechanism of insertion. This indicates both SINEs and LINEs would be preferentially located in regions of low G+C. Eyre-Walker and Hurst (2001) show this is the case when only recently inserted SINE elements are considered. So why do older SINE insertions tend to be found in higher G+C regions? Pavlíček *et al*. (2001) propose this may occur if the excision of ALUs was fast enough to remove new copies before they had a chance to fixate in the population. They discuss the possibility of positive selection of the CpG rich ALUs in G+C rich regions due to hypomethylation in germline cells. In addition, it is

suggested there are different recombination rates that could be affected by the short length of SINE elements (on the order of 300 bases) when compared to LINE elements (several KB long).

The parameters in `RepeatMasker` have been designed so potentially interesting, unique regions are not falsely masked as repetitive. This is based on a cutoff alignment score. As a result, repetitive elements that have sufficiently diverged from the consensus for their repeat family will not be detected. This does not pose a problem in the analysis, since those repetitive elements closest in identity to the `Repbase` consensus are detected. These result from more recently active transposable elements within the human genome. Since these recent transposable events do not lead to the creation and maintenance of regions of high and low G+C content within the human genome, it is unlikely ancient copies of the same repetitive elements would have any different effect. In fact, these ancient copies should behave in the same manner due to the same mechanisms of insertion. In addition, `Repbase` consensus sequences have been carefully constructed to address the problem of detecting diverse repeats by representing the best available approximation of the elements that generated the repeats (Jurka, 1998).

The variance of the G+C content in unmasked regions was small. Ancient copies of repeats currently undetected are expected to have properties similar to the detected repeats. If methods to detect these repeats were available, then a migration of the data points in figure 7-1 from the unmasked fraction to the masked fraction would result. This migration should have little effect on the correlation between the unique region and overall contig G+C% due to the low variance. Therefore, even if all ancient copies of

repeats were detected, a positive correlation between unique and overall contig G+C% would be expected to exist.

Low copy number repeats and repeats that have not been characterized in the human genome will not be detected when using `RepeatMasker`. Undetected low copy repeats are not likely to contribute much to the maintenance of regions of high and low G+C content. This is due to the definition that each family of a low copy repeat is found only in a small portion of the genome due to the small copy number. The human genome has been available at least to a rough draft level since February of 2001 (IHGSC, 2001). Since `Repbase` has been carefully examining and collating information on repetitive elements within the human genome, it is highly unlikely there are any high copy number repeats that remain uncharacterized. Any remaining uncharacterized repeat families or subfamilies will likely have a relatively low copy number, and constitute a low percentage of the human genome. Thus, currently uncharacterized repeats should contribute little information into the origin and maintenance of high and low G+C regions within the human genome.

Based on the information gathered, the first hypothesis should be rejected. Regions of G+C content within the human genome do not appear to result from the presence of repetitive elements; rather it appears as though the presence of regions of high and low G+C concentration determines the density of certain repetitive elements within the human genome.

## 7.4 Hypothesis 2: Mutational Biases Revisited

As previously discussed, one of the hypotheses for high and low G+C region maintenance is it was due to biological mechanisms favoring compositional bias in mutation rates. Previous studies in favor of the mutational bias theory have focused on a limited set of genes and pseudogenes within human and primate populations (Filipski, 1987; Wolfe, Sharpe and Li, 1989; Casane *et al*., 1997).

The shortcoming of these approaches is two-fold. While an association between genes and G+C content can be demonstrated (Zoubak, Clay and Bernardi, 1996), genes only account for 3-5 percent of the genome (Gardiner, 1996). Secondly, these studies are closely tied to genic regions and, as such, selective pressure is a factor. Thus, it is not easy to separate the conclusions of results suggesting a mutational bias mechanism for the maintenance of high and low G+C regions from the biased gene conversion hypothesis.

In order to work around selection mechanisms that may play a role, two elements likely to be under less selective pressure were studied: processed pseudogenes and repetitive elements. In an ideal case, the rate of A|T→G|C and G|C→A|T mutations would be compared when elements deriving from the same ancestor were placed in differing neighborhoods of G+C concentration. However, it is not always possible to determine whether a mutation has occurred within the ancestor or the descendant sequence (see section 7.6). Therefore, the rate of A|T→G|C and G|C→A|T substitutions were studied as to how they related to the surrounding G+C composition.

### 7.4.1 Studying Compositional Bias in Processed Pseudogenes

Processed pseudogenes are non-functional copies of processed mRNAs from functional genes that have been retrotransposed (reverse-copied) into a region of the genome. Processed pseudogenes are characterized by the presence of direct repeats on both the 5' and 3' ends, which result due to target site duplication with the retrotransposable insertion mechanisms employed. Depending on the processed pseudogene, this mechanism for insertion is borrowed from either the human endogenous retrovirus (HERV) or LINE retrotransposition machinery (Pavlíček *et al*., 2002). Since processed pseudogenes are derived from processed mRNA, intronic regions are spliced out and poly-A tails are present at the 3' end (Lodish *et al*., 1995). In addition, insertion mechanisms incorporated by processed pseudogenes can cause truncation at the 5' end (Pavlíček *et al*., 2002). Multiple mutations may occur that disrupt the reading frame or introduce stop codons. This is particularly important in pseudogenes where the 5' end has not been truncated (Lodish *et al*., 1995). Figure 7-2 illustrates the steps in which a gene and processed pseudogene pair are generated.

For the purpose of the study, it was assumed that the gene locus existed first, and then at some point in the evolutionary history of humans, the pseudogene arose. Once the gene and pseudogene were in place, they could evolve and mutate independently of one another. However, genes are under selective pressure, so mutations within them were expected to be fewer than in neutrally mutating pseudogenes. When a nucleotide difference was observed between a gene and pseudogene, it would be more likely to have occurred within the pseudogene. An exception would be when a mutation occurred in

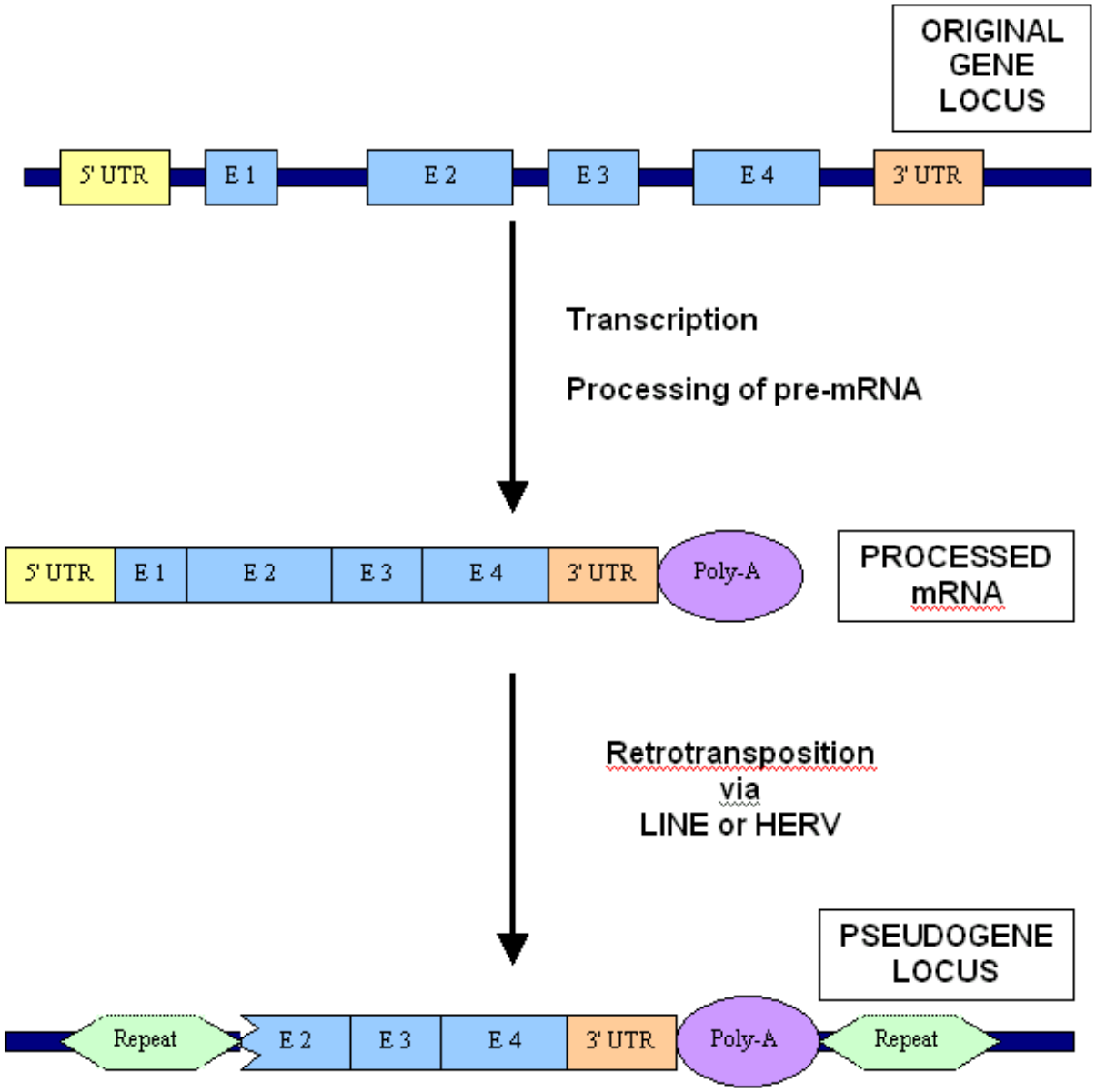



**Figure 7-2: Gene-to-Pseudogene Mechanism.** Shown is the native locus (top) and pseudogene locus (bottom). The processed mRNA is created by the transcription of the genic region into a pre-mRNA intermediary that is subsequently spliced to remove the introns and polyadenylated to add the poly-A tail. The pseudogene is created by the retrotransposition of the processed mRNA via either a LINE or HERV retrotransposition mechanism. The light blue boxes represent the exons, labeled E1, E2, E3, and E4. The dark blue boxes represent the genomic location.

the third codon position (also known as the wobble base). When mutations occurred in this location and were synonymous (did not change the encoded amino acid), they were not expected to alter the fitness of the genic region in a significant way.

The directionality of the change would be of interest as well. Details on how to incorporate directional information gathered using genomic comparisons is given in section 7.6. However, the directionality of the mutation is not nearly as important as whether or not it changed the overall G+C composition of the gene or pseudogene. Therefore, substitutions were reported as A|T→G|C and G|C→A|T where the nucleotide of the gene was listed first, and the nucleotide of the pseudogene second. If the original nucleotide was an A or T in the gene and the nucleotide in the pseudogene was a C or G, the effect will be the same as if the original gene nucleotide was a C or G that mutated to an A or T over time. Thus, the rates of A|T→G|C and G|C→A|T substitutions were compared when the gene was in one G+C composition and the pseudogene was in another. This allowed the examination to see if a compositional bias in substitution rates within genes and pseudogenes potentially exists. Limitations to this approach are discussed in section 7.6.1.

### 7.4.2 Obtaining Pseudogene Data

The first step in this analysis required gathering gene-pseudogene pairs.. Potential processed pseudogenes were obtained by searching individual mRNA entries of RefSeq (Pruitt and Maglott, 2001) against the University of California-Santa Cruz's Goldenpath assembly of the human genome (Kent and Haussler, 2001) using `wublastn`

(Gish, 1996-2001). For the data sets, `RefSeq` was downloaded on April 18, 2002, when 15,199 human mRNAs were available. The December 2001 `Goldenpath` assembly was used.

Only `RefSeq` entries hitting multiple loci in the `Goldenpath` assembly were considered. `RefSeq` entries mapping to more than one location were likely to contain both a native locus location as well as one or more other locations that were potential paralogs or pseudogenes. For the entries with multiple loci, a native locus scoring system was implemented in the following fashion. For each individual `BLAST` *HSP* (**H**igh-scoring **S**egment **P**air -- it can be thought of as a single local alignment), a score $S_{HSP}$ (Equation 7-1) was assigned a value equal to the fractional percentage identity multiplied by the fraction of the mRNA that the HSP covered. $L_{HSP}$ is the length of the HSP and $L_{REFSEQ}$ is the length of the RefSeq entry. Scores for all of the HSPs occurring within a single locus (a total of *n* HSPs) were summed into a single score, $S_{LOCUS}$ (Equation 7-2).

$$S_{HSP} = \%id * \frac{L_{HSP}}{L_{REFSEQ}}$$

**Equation 7-1: Individual HSP Score.**

$$S_{LOCUS} = \sum_{i=1}^{n} S_{HSP_i}$$

**Equation 7-2: Native Locus Score.**

The native locus should produce an $S_{HSP}$ score close to 1, which represents a locus that is 100% identical over 100% of the bases of the RefSeq mRNA. Therefore, the locus

with the highest (optimal) $S_{LOCUS}$ score was considered to be the native locus. All other suboptimal loci were treated as potential candidates for paralogs and pseudogenes, both processed and unprocessed.

Each HSP within an alignment should roughly correspond to an alignment of exonic regions. RefSeq hits were further filtered to only contain entries where the native locus contained at least three HSPs. Such a filter was applied to increase the likelihood that at least one intron (two exons) was in the native gene. This helped to reduce the problem of differentiating between paralogs, unprocessed pseudogenes and processed pseudogenes corresponding to single exon genes. Since processed pseudogenes have intronic regions spliced out, they should map continuously with the RefSeq mRNA. Thus, an additional restriction that the non-native loci contained only a single HSP was applied. A final restriction required non-native loci to align within 20 basepairs (bp) of the 3' end of the RefSeq sequence, since processed pseudogenes are often truncated at the 5' end. This helped to reduce spurious matches. While these restrictions would not allow detection of all of the processed pseudogenes within the human genome, the detected gene-pseudogene pairs had a greater likelihood of being true positives.

Gene and pseudogene pairs were separated into one of four categories based on their G+C content (Table 7-1). The four different categories are: (LOW, LOW), (LOW, HIGH), (HIGH, LOW), and (HIGH, HIGH). The first element in the ordered pair represents the regional G+C composition flanking the gene while the second element represents the regional G+C composition flanking the pseudogene. These neighboring

**Table 7-1: Number of Genes and Pseudogenes Found.** This table indicates the number of genes and corresponding pseudogenes found after processing the results of searching the April 18, 2002 version of `RefSeq` (Pruitt and Maglott, 2001) against the December, 2001 assembly of the `Goldenpath` (Kent and Haussler, 2001; `http://genome.ucsc.edu/`). The G+C content is listed as either LOW (less than 41% G+C) or HIGH (greater than 44% G+C).

| Gene G+C | Pseudogene G+C | Number of Genes | Number of Pseudogenes |
|---|---|---|---|
| HIGH | LOW | 242 | 564 |
| HIGH | HIGH | 233 | 464 |
| LOW | LOW | 173 | 250 |
| LOW | HIGH | 52 | 79 |
| **TOTALS** | | **700** | **1,357** |

compositions were calculated from the 25 kb flanking both sides of the gene or pseudogene. A region containing less than 41% G+C was categorized as LOW, while regions containing greater than 44% G+C were categorized as HIGH. The total neighborhood size of 50-kb (25-kb on two ends) was used to maintain consistency with Bernardi's earlier density gradient centrifugation experiments. In addition, the boundaries of 41% and 44% G+C were chosen due to their correspondence with major breakpoint divisions within Bernardi's isochore definitions (Bernardi, 1993).

## 7.4.3 Calculation of Gene -Pseudogene Substitution Rates

Once the genes and pseudogenes were separated into the appropriate category, they were aligned to one another using `Sim4` (Florea *et al*., 1998). `Sim4` is an algorithm for aligning cDNAs to genomic sequence. `Sim4` attempts to delineate intron/exon boundaries by looking for donor and acceptor sites, thus adding more information to the alignments. Whenever a mismatch appeared between the gene and pseudogene, it was treated as a substitution event.

Since substitutions occurring in different regions of genes have the potential to be under different selective pressure, the context of each substitution was recorded. The annotated coding sequence (CDS) was parsed out of each `RefSeq` entry. Since the frame of the CDS was known, the third base (wobble base) of each coding triplet was extracted. Substitutions in the CDS were recorded and separated into wobble base and non-wobble base positions. Anything outside of the CDS was labeled as a non-coding substitution. Non-coding substitutions were separated into 5' UTR mutations and 3' UTR mutations, depending on their relationship to the start and end of the CDS. The alignment of introns was not a problem since they would have been removed from the processed pseudogenes that made it into our test set.

Genes are likely to be under more selective pressure than processed pseudogenes. Thus, the direction of each substitution was more likely to be FROM the gene TO the pseudogene. However, mutational directionality is not nearly as important as how each substitution is reflected when compared to the overall G+C context of the gene or pseudogene. These limitations are discussed further in section 7.6. Once all of the alignments were made, the number of each of the 16 substitution events (gathered from the Cartesian product A x B where A,B = {A, C, G, T} and A represents the nucleotide in the gene and B represents the corresponding nucleotide in the processed pseudogene) were calculated for the following categories: coding regions, wobble bases, non-wobble coding bases, non-coding regions, 5' UTRs and 3' UTRs.

### 7.4.4 Approaches to Looking at Mutation and Substitution Events

In 1962, Noboru Sueoka introduced the concept of effective base conversion rates. These values, *u* and *v*, are described by Sueoka as the rates of conversion at any given point in the genome from A|T→G|C and G|C→A|T nucleotides, respectively. These rates are explained in terms of the observed inherited rates of nucleotide substitution within a single organism from generation to generation. These values are used in more recent studies to measure the mutation rates within different genomic regions (Piganeau *et al*., 2002; Smith and Eyre-Walker, 2001; Casane *et al*., 1997; Gu and Li, 1994).

Using these models as guidelines, the rate of A|T→G|C substitutions (*u*) was calculated as the probability that a G or C nucleotide was found at a given location in the pseudogene, conditioned on the nucleotide in the gene being an A or T. In addition, the rate of G|C→A|T substitutions (*v*) was calculated as the probability that an A or T nucleotide was found at a given location in the pseudogene, conditioned on the nucleotide in the gene being a C or G. The difficulty in determining the exact directionality of mutation within a single species is discussed in section 7.6.1.

The G+C bias (*f*) was calculated as $f=u/(u+v)$ (Piganeau *et al*., 2002). A measure of the A+T bias can be obtained as $1-f$. The G+C bias ranges from 0 to 1. A value of 0 means there were no A|T→G|C substitutions in a region for a given feature. A value of 1 indicates there were no G|C→A|T substitutions. If the A|T→G|C and G|C→A|T substitution rates were equal in any given region, then the G+C bias and A+T bias would

both be equal to 0.5. A G+C bias less than 0.5 indicates a region will drift to A+T richness over time, while a value greater than 0.5 indicates a drift towards G+C richness.

In order to test for compositional bias in substitution rates, a ratio of the G+C bias in high G+C regions ($f_{HIGH}$) to the G+C bias in low G+C regions ($f_{LOW}$) was computed. A ratio, *r*, consistently greater than 1 indicates a compositional bias in substitution rates was likely to exist, where high G+C regions acquired more G's and C's over time and low G+C regions were adding more A's and T's over time. A ratio less than 1 on a consistent basis indicates there was likely to be a negative correlation where G+C rich regions would be mutating towards A+T and A+T rich regions would be mutating towards G+C. If the ratios randomly fluctuate above and below 1, a compositional bias for substitution rates cannot be demonstrated for the feature being studied.

### 7.4.5 Gene-Pseudogene Mutational Bias Results

In order to test for a possible compositional bias for substitution rates in gene-pseudogene pairs, two different comparisons were made: one where the gene originated in a low G+C region, and one where the surrounding content of the gene was high G+C. In each comparison, two different cases were examined. The first case involved the pseudogene occurring in a low G+C region, and the second case was when the pseudogene was in a high G+C region.

If a compositional bias for substitution rates exists, the G+C bias, *f*, would be expected to increase as the G+C context of the pseudogene increases. This would indicate the ratio of A|T→G|C to G|C→A|T mutations increase as the surrounding G+C

context increases. In order to test this hypothesis, the G+C bias, *f*, was calculated for the four cases defined by the Cartesian product A x B where A,B={HIGH, LOW} and A=G+C context of the gene; B=G+C context of the pseudogene. The resulting G+C biases were labeled as follows: $f_1$ = {LOW,LOW}; $f_2$ = {LOW,HIGH}; $f_3$ = {HIGH,LOW}; $f_4$ = {HIGH,HIGH}. In order to test for potential compositional biases for substitution rates, the ratios $r_1 = f_2:f_1$ and $r_2 = f_4:f_3$ were calculated. If a compositional bias exists, the values of $r_1$ and $r_2$ would be expected to be greater than 1.

The results are listed in Table 7-2. Table 7-2 (A) lists the results for the first comparison of a gene in a low G+C region while Table 7-2 (B) lists the results when the gene was in a high G+C region. Table 7-2 (A) gives the value of $r_1$ calculated in 5' UTRs, coding sequences, wobble bases, non-wobble coding bases, and 3' UTRs. Table 7-2 (B) gives the value of $r_2$ calculated for each of these regions.

For each of the features studied, the values of $r_1$ and $r_2$ were greater than 1, with $r_1$ ranging from 1.173 to 1.362 and $r_2$ ranging from 1.134 to 1.175. This indicates A|T→G|C and G|C→A|T substitutions were 17-36% higher in the first case, and 13-17% higher in the second case. These increases indicate that, when pairs of genes and pseudogenes were examined, there appeared to be a compositional bias for substitutions. Table 7-2 yields an interesting result. When the G+C bias, *f*, was compared in the 5' UTRs, CDS, non-wobble CDS, and 3' UTRs, the values were always less than 0.5. This indicates these portions of the pseudogenes had higher rates of G|C→A|T substitutions than A|T→G|C substitutions no matter what the original gene and pseudogene G+C contexts were. As a result, as pseudogenes aged, these regions tended towards A+T

**Table 7-2: Comparison of G+C Bias in Gene and Pseudogene Pairs.** Table 7-2 A) lists the results when the gene was located in a region of low G+C content (<41% G+C). Table 7-2 B) lists the results when the gene was located in a region of high G+C content (>44% G+C). In each case, the second and third columns list the G+C bias when the pseudogene was located in a region of high and low G+C, respectively. The G+C bias was calculated as $f=u/(u+v)$ where $u$ was the rate of A|T→G|C substitutions and $v$ was the rate of G|C→A|T substitutions within a particular region. The fourth column lists the ratio of the HIGH:LOW G+C biases.

**A)**

| Gene in Low G+C | | | |
|---|---|---|---|
| | **Pseudogene HIGH G+C** | **Pseudogene LOW G+C** | **Ratio of HIGH:LOW** |
| **5' UTR** | 0.4632 | 0.3797 | 1.220 |
| **CDS** | 0.4288 | 0.3309 | 1.296 |
| **WOBBLE** | 0.4721 | 0.3467 | 1.362 |
| **NON-WOBBLE** | 0.3986 | 0.3145 | 1.268 |
| **3' UTR** | 0.3674 | 0.3132 | 1.173 |

**B)**

| Gene in High G+C | | | |
|---|---|---|---|
| | **Pseudogene HIGH G+C** | **Pseudogene LOW G+C** | **Ratio HIGH:LOW** |
| **5'UTR** | 0.4721 | 0.4032 | 1.171 |
| **CDS** | 0.4159 | 0.3600 | 1.155 |
| **WOBBLE** | 0.5710 | 0.5036 | 1.134 |
| **NON-WOBBLE** | 0.3331 | 0.2835 | 1.175 |
| **3' UTR** | 0.4376 | 0.3765 | 1.162 |

regardless of the surrounding G+C content. However, the rate of this substitution trend was slowed when the surrounding region was G+C rich.

Substitutions found within the non-wobble coding positions are likely to have occurred within the pseudogene since most mutations within the first two codon positions of a gene will cause a change to the amino acid encoded by that codon. Such a change can affect the fitness of the gene. Therefore, the results listed in Table 7-2 suggesting

that a compositional bias for substitution has occurred within non-wobble coding regions is likely to have a directionality associated with it.

The study of gene and pseudogene pairs indicates there was a strong possibility of a compositional bias for substitution rates. However, the rate of A|T→G|C substitutions was always less than the rate of G|C→A|T substitutions. This indicates pseudogenes within the human genome were likely to accumulate more A+T sequence over time regardless of the surrounding G+C context. However, as the G+C context of the pseudogene increased, the rate of this change slowed. As a result, a compositional bias in substitution rates was observed, but this rate cannot be the determining factor for maintaining regions of low and high G+C composition.

### 7.4.6 Studying Compositional Bias in Repetitive Elements

A large portion of human genomic DNA has been derived from the dispersion of transposable elements throughout the genome (Prak and Kazazinan, 2000; Smit, 1999). The International Human Genome Sequencing Consortium's analysis found that 45% of the human genome is made up of identifiable transposable elements (IHGSC, 2001). The two largest types of these are long interspersed elements (LINEs) and short interspersed elements (SINEs). There are approximately 868,000 copies of LINEs in the human genome, making up over 20% of the total genomic sequence. In addition, there are over 1.5 million copies of SINEs, accounting for over 13% of the genome (IHGSC, 2001). Due to the large abundance of repetitive elements in the human genome, substitution

rates within them were studied to determine if a compositional bias for substitution potentially existed in these segments.

### 7.4.7 Detecting Repetitive Elements

Instances of SINE and LINE repeats were located within the human genome using `RepeatMasker`. The repetitive regions were obtained by running `RepeatMasker` release 6/19/01 (Smit and Green, unpublished; `http://repeatmasker.genome.washington.edu/`) using the Repbase update 6.6 (Jurka, 2000) repeat definitions. `RepeatMasker` was run using the faster `-w` option, which employs `wublastn` as the alignment algorithm.

Once the contigs were masked, the generated `.out` files containing tables of repeat information were parsed. Files were generated to group together the Goldenpath contig name, contig location and orientation of the repeat instances for each type of repeat. The repeat regions were extracted from the contigs, and the G+C content of the surrounding 50-kb (25 kb on each side) window was noted. Each instance of a repeat was placed into one of two files for each repeat type based on whether the G+C content of the surrounding window was less than 41% or greater than 44%, labeled low and high G+C, respectively. Those repeat elements falling in the intermediate range of 41% to 44% G+C were discarded from the study.

Repetitive element families and subfamilies with the greatest number of instances currently detectible in the human genome were studied. The resulting data set analyzed

included 8 ALU families/subfamilies and 34 LINE families/subfamilies (see Table 7-3 for the family/subfamily names).

### 7.4.8 Calculating Repetitive Element Substitution Rates

With repetitive elements, it is difficult to assign directionality for each mutation since it cannot easily be determined which copy of a repeat was present first in a genome, and whether or not a second repeat was derived as a direct ancestor. In addition, once a copy is in place, it mutates and evolves independently of its parent copy. One possible scenario is that a C or G nucleotide is observed at one position in a copy of an element situated in a region of high G+C composition. At the same time, an A or T could be observed at the same position when a copy of the element was found in a low G+C region. The difficulty of determining directionality is discussed in detail in section 7.6.1.

The `Repbase`-defined consensus was taken as the ancestral repeat element. Such an approach is justified in the sense that the consensus sequence has been derived to be the best approximation of the original transposable element that generated a given repeat subfamily (Jurka, 1998). Such an approach assumes a master/slave model of repetitive element propagation (Shen, Batzer and Deringer, 1991; see 7.6.1 for a discussion). Substitution rates were measured as the difference from the `Repbase` sequence.

Each instance of a given repetitive element was compared against the `Repbase` consensus sequence using `wublastn` with the parameters `-S2=200 -S=250`. These parameters were chosen to eliminate smaller matching regions by requiring higher

scoring hits with a final score of at least `250`. Using the default `wublastn` scoring parameters of `+5,-4` for matches and mismatches, this corresponds to an ungapped alignment of at least 50 bp at 100% identity, or 78 bp at 80% identity.

The total number of substitution events FROM the `Repbase` consensus TO the instance of the repeat was noted. The total substitution events for repeat instances in low G+C (<41%) and high G+C (>44%) were calculated. The rate of A|T→G|C (*u*) and G|C→A|T (*v*) substitutions were computed as well as the G+C bias (*f*) for two categories: HIGH and LOW for each of the repetitive element families studied. HIGH represents those repetitive regions occurring in >44% G+C regions and LOW represents those repeats occurring in <41% G+C regions. A ratio of the HIGH:LOW G+C biases was calculated for each repeat family studied. A ratio greater than 1 indicates that the rate of A|T→G|C vs. G|C→A|T mutations is likely to be higher in high G+C regions.

### 7.4.9 Repeat Instance Substitution Bias Results and Discussion

Table 7-3 lists the resulting G+C biases calculated for each of the repeat families for the instances in low and high G+C. For the Alu repeat families studied, the ratio ranged from 0.937 to 1.080. Six of the eight Alu families had ratios greater than 1 (with the exception of the AluYa5 and AluYb8 families). This suggests for six of these families, a slight compositional bias for mutation rates exists. All 34 of the LINE families studied had ratios greater than 1. In fact, these ratios tended to be larger than the ratios for Alu families, ranging from 1.047 for the L1PA6 family, to 1.437 for the

L1MB3 family. These values show the LINE families have a potentially stronger compositional bias for mutation rates.

The G+C biases for nearly all of the repetitive families were much less than 0.5, yielding results similar to the gene-pseudogene substitution rates. This indicates no matter what the surrounding G+C content is for an instance of a repetitive element, the repeat copy will likely drift towards A+T richness over time. Since the ratios were greater than 1 (indicating there was a compositional bias for substitution rates), the rate of drift should be slower when the surrounding G+C content is higher. These results indicate there seems to be a compositional bias for substitution rates; however, this bias is unlikely be the cause for the maintenance of high G+C regions containing the features studied.

**Repeats on Chromosome Y.** As previously discussed, one potential problem is the mutational bias and biased gene conversion theories are not necessarily mutually exclusive. In order to address this concern, another study examining only instances of repetitive elements occurring on chromosome Y was performed. Chromosome Y contains a non-recombining region making up over 95% of the chromosome (Tilford *et al*., 2001). The non-recombining region of chromosome Y does not recombine with chromosome X or any other chromosome (Lahn, Pearson and Jegalian, 2001). Non-recombining regions will not allow for gene conversion, and biased gene conversion could not be the cause of any biases in G+C composition that are observed in such regions.

**Table 7-3: Comparison of G+C Bias in Instances of Repeat Families.** Listed in the first column is the repeat family studied. The second and third columns contain the G+C bias ***f=u/(u+v)*** (where ***u*** was the rate of A|T→G|C substitutions and ***v*** was the rate of G|C→A|T substitutions) calculated for instances of repeats occurring in HIGH and LOW G+C regions, respectively. The fourth column lists the ratio of HIGH:LOW G+C biases.

| Repeat Family | HIGH G+C | LOW G+C | RATIO HI:LOW |
|---|---|---|---|
| AluYa5 | 0.3821 | 0.4077 | 0.937 |
| AluYb8 | 0.5121 | 0.5440 | 0.941 |
| AluYc | 0.2486 | 0.2467 | 1.008 |
| AluY | 0.2479 | 0.2397 | 1.034 |
| AluSg1 | 0.2017 | 0.2091 | 1.036 |
| L1PA6 | 0.3018 | 0.2883 | 1.047 |
| L1PA3 | 0.3217 | 0.3057 | 1.053 |
| L1PA4 | 0.3418 | 0.3222 | 1.061 |
| L1 | 0.2888 | 0.2708 | 1.066 |
| L1PA2 | 0.4242 | 0.3955 | 1.073 |
| AluSq | 0.2332 | 0.2173 | 1.073 |
| AluSc | 0.2333 | 0.2160 | 1.080 |
| AluSp | 0.2109 | 0.1952 | 1.080 |
| L1PA8A | 0.3291 | 0.2978 | 1.105 |
| L1PA7 | 0.2989 | 0.2687 | 1.112 |
| L1PA5 | 0.3500 | 0.3134 | 1.117 |
| L1PB1 | 0.3428 | 0.3003 | 1.141 |
| L1PA10 | 0.3493 | 0.3020 | 1.157 |
| L1PA8 | 0.3367 | 0.2870 | 1.173 |
| L1PA11 | 0.3601 | 0.3049 | 1.181 |
| L1MA3 | 0.3373 | 0.2829 | 1.192 |

| Repeat Family | HIGH G+C | LOW G+C | RATIO HI:LOW |
|---|---|---|---|
| L1MA2 | 0.3439 | 0.2870 | 1.198 |
| L1PB2 | 0.3549 | 0.2930 | 1.211 |
| L1PA15 | 0.3227 | 0.2659 | 1.214 |
| L1PB3 | 0.3213 | 0.2621 | 1.226 |
| L1PA14 | 0.3469 | 0.2830 | 1.226 |
| LAMA4A | 0.3369 | 0.2743 | 1.228 |
| L1PA13 | 0.3646 | 0.2968 | 1.229 |
| L1MA4 | 0.3370 | 0.2741 | 1.229 |
| L1PA16 | 0.3376 | 0.2701 | 1.250 |
| L1MB4 | 0.3527 | 0.2810 | 1.255 |
| L1ME1 | 0.3489 | 0.2758 | 1.265 |
| L1PA17 | 0.3275 | 0.2579 | 1.270 |
| L1PB4 | 0.3511 | 0.2739 | 1.282 |
| L1MB8 | 0.3558 | 0.2770 | 1.284 |
| L1MA9 | 0.3616 | 0.2811 | 1.286 |
| L1MB7 | 0.3659 | 0.2756 | 1.327 |
| L1MA8 | 0.3691 | 0.2780 | 1.328 |
| L1MB2 | 0.3635 | 0.2732 | 1.331 |
| L1MC1 | 0.3836 | 0.2811 | 1.365 |
| L1MB5 | 0.3838 | 0.2726 | 1.408 |
| L1MB3 | 0.4035 | 0.2808 | 1.437 |

The analysis on chromosome Y was limited to only those Alu and LINE elements having at least five different instances in LOW G+C regions and five different instances in HIGH G+C regions. The G+C bias was calculated for instances occurring in HIGH and LOW G+C for those repetitive elements fitting this criterion. In addition, the ratio of the HIGH:LOW G+C biases was computed.

A total of five different Alu families and twelve LINE families were studied on chromosome Y. The results are listed in Table 7-4. The only repeat subfamily with a ratio less than 1 was the AluY subfamily, the youngest repeat studied with an age less than 1 million years old (IHGSC, 2001). The ratio of G+C biases for all of the other

repeat subfamilies was greater than 1. This indicates these 16 repetitive element families on chromosome Y likely have a compositional bias affecting substitution rates. The G+C biases were significantly less than 0.5, indicating instances of repetitive elements on chromosome Y are likely to tend toward A+T richness over time. Since 95% of chromosome Y is not subject to recombination, it is highly unlikely the compositional bias for substitution rates within repetitive elements on chromosome Y was due to biased gene conversion. Although it cannot be certain that biased gene conversion does not largely contribute on other chromosomes, the results observed for chromosome Y were consistent with the previous repeat study. As a result, biased gene conversion is thought to contribute little to the observed compositional bias.

**Table 7-4: Comparison of G+C Bias for Repeats Found on Chromosome Y.** Listed in the first column is the repeat family studied. The second and third columns contain the G+C bias ***f=u/(u+v)*** (where ***u*** was the rate of A|T→G|C substitutions and ***v*** was the rate of G|C→A|T substitutions) calculated for instances of repeats occurring in HIGH and LOW G+C regions, respectively. The fourth column lists the ratio of HIGH:LOW G+C biases. Only repetitive elements occurring at least five times in both HIGH and LOW G+C regions on chromosome Y were included.

| Repeat Family | HIGH G+C | LOW G+C | RATIO HIGH:LOW |
|---|---|---|---|
| AluY | 0.2194 | 0.2210 | 0.993 |
| L1PA2 | 0.3916 | 0.3898 | 1.005 |
| L1PB1 | 0.2876 | 0.2836 | 1.014 |
| AluSq | 0.2165 | 0.2099 | 1.031 |
| L1MA9 | 0.3008 | 0.2778 | 1.083 |
| L1PA4 | 0.3321 | 0.3046 | 1.090 |
| L1PA14 | 0.2740 | 0.2503 | 1.095 |
| L1PA3 | 0.3350 | 0.2957 | 1.133 |
| AluSp | 0.2185 | 0.1910 | 1.144 |
| AluSc | 0.2478 | 0.2150 | 1.153 |
| AluSx | 0.2614 | 0.2122 | 1.231 |
| L1 | 0.3238 | 0.2472 | 1.310 |
| L1MB7 | 0.4076 | 0.2866 | 1.422 |
| L1PA7 | 0.3784 | 0.2547 | 1.486 |
| L1MA8 | 0.4465 | 0.2984 | 1.497 |
| L1PB4 | 0.4125 | 0.2563 | 1.609 |
| L1PA15 | 0.4394 | 0.2570 | 1.709 |

## 7.5 Testing for Drift to an A+T Rich Genome Using Long Terminal Repeats (LTRs)

The results of looking at gene/pseudogene pairs and instances of repetitive elements suggest elements inserted into the human genome are likely to mutate towards a higher A+T composition over time. This phenomenon was observed when comparing the rate of A|T→G|C and G|C→A|T substitutions and was independent of the G+C content of the surrounding region. In order to test this hypothesis, elements inserted at different points in time were studied to determine whether or not older elements tend to be more A+T rich.

One class of repetitive elements of particular interest is those caused by LTR retroviral integration events. These elements are useful to study since the mechanism of LTR retroviral integration produces two identical long terminal repeats (LTRs) which flank the 5' and 3' end of the virus (Lodish *et al*., 1995). The divergence between the 5' and 3' LTRs can be used to calculate an approximate integration date for any particular instance (Tristem, 2000).

LTR retroviruses that have become integrated into the human germline cells are one such example. Human endogenous retroviruses (HERVs) have been studied in detail (Barulescu *et al*., 1999; Tristem, 2000; Griffiths, 2001). Approximately 1.3% of the human genome is composed of HERV elements, representing roughly half of the LTRs found in humans (Smit, 1996). One recent study looked at classification and integration age of the various HERV families (Tristem, 2000). The classification and naming convention for HERV families is based on the similarity of the HERV binding site to host

tRNAs. The study of Tristem (2000) estimates the HERV-H, HERV-K, and HERV-L families have the largest copy number in the human genome.

### 7.5.1 Detecting Copies of HERVs

Representative sequences for HERV-H, HERV-K, and HERV-L families as described by Tristem (2000) were obtained from `Genbank`. The accessions obtained were as follows: D11078 (HERV-H) (Hirose *et al*., 1993); M14123 (HERV-K) (Ono *et al*., 1986); and X89211 (Corodonnier, Casella and Heidmann, 1995). Each of these sequences was searched against the December 2001 release of the Goldenpath assembly of the human genome using `wublastn`. Score cutoff parameters of `-S=2000` and `-S2=2000` were used to filter spurious hits. A score of 2000 using the default `wublastn` scoring scheme of `+5,-4` requires a 400 bp ungapped alignment at 100% identity, or a 625 bp ungapped alignment at 80% identity. In addition, the parameter `-gapw=2000` was used to close longer alignment gaps.

The search matched 1001 HERV-H locations, 409 HERV-L locations, and 723 HERV-K locations. However, many of these instances were truncated, missing one or both of the LTR sequences due to recombination events leading to a solitary LTR (Prak and Kazazian, 2000). These matches were manually filtered to include only full-length copies. The resulting datasets included 14 HERV-H, 21 HERV-K, and 72 HERV-L copies.

### 7.5.2 Determining Insertion Age and G+C Composition

GenBank accessions for the HERV-H, HERV-K and HERV-L representative sequences contain various annotations including the 5' and 3' LTR sequences. The representative 5' and 3' LTR sequences were extracted and placed into separate files. Each of the full-length copies were searched against the appropriate 5' LTR using `wublastn` with the parameters `-S=300 -S2=300 -gapw=200`. Since the 5' and 3' LTRs should be identical at the time of insertion, searching full-length repeats for the presence of the 5' or 3' LTR should produce the same results. The 5' LTR was arbitrarily chosen, which in every instance located the 3' LTR as well. The resulting `wublastn` output was parsed to extract the 5' LTR sequence and 3' LTR sequence. These were aligned to each other using `wublastn` with the parameters `-S=200 -S2=200 -gapw=128`. The approximate edit distance for each instance was determined based on the number of mismatched bases in the alignment of the 5' and 3' LTRs. Gaps were ignored.

After the 5' and 3' LTRs were located in each full-length copy, the G+C content of the repeat copy was calculated. The edit distance for each instance was compared to the G+C content to see if more distant elements tend to be more A+T rich. Figure 7-3 shows a graph plotting the G+C composition against the percent divergence for the 72 full-length HERV-L copies. For this figure, the percent divergence was calculated as the percentage of mismatching bases when the 5' and 3' LTRs were aligned. The assumption is the higher the percent divergence, the older the insertion date will be.

### 7.5.3 LTR Results

A correlation coefficient was calculated to determine whether or not a correlation exists between the edit distance and the G+C content. An *r*-value of -0.3279 was calculated for the 72 HERV-L instances, indicating a slight negative correlation between the LTR divergence and the repeat G+C content. This suggests the older the date of insertion, the greater the accumulation of A's and T's will be. A *t*-score was calculated for the *r*-value of -0.3279 with 72 instances to determine the level of significance for this correlation coefficient. The resulting *t*-score was -2.946. Using 70 degrees of freedom and a two-tailed test, this *t*-score yields a *p*-value of 0.0087, indicating the observed correlation is likely to exist between the insertion date and G+C content.

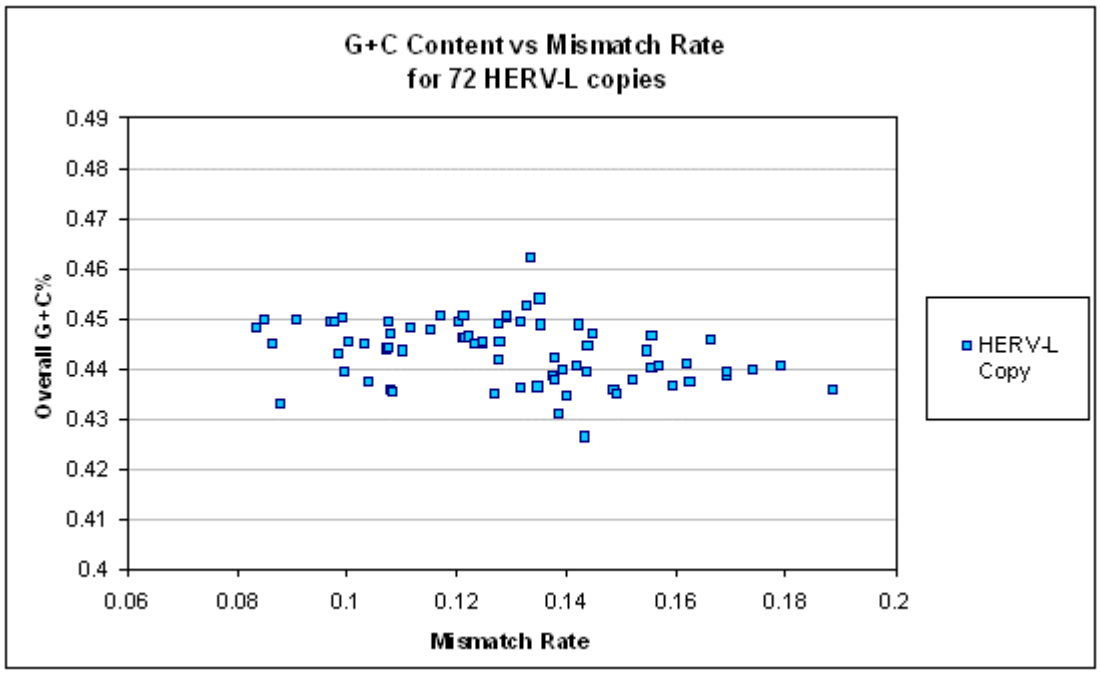



**Figure 7-3: Plot of Divergence Rate vs. G+C Composition in HERV-L Repeats.** Shown in this figure is a plot of the divergence rate versus the overall G+C percentage for each of the 72 full-length HERV-L copies found within the human genome. The divergence rate (x-axis) is calculated as the percentage of bases mismatched in an alignment between the 3' and 5' LTRs of the HERV-L copy. The overall G+C percentage (y-axis) is based on the G+C content of the complete HERV-L copy.

While a correlation between the insertion date and G+C content has been demonstrated with the HERV-L repeat family, it would be useful to locate more instances of high copy number elements in which the relative date of insertion can be determined. There are two main difficulties in obtaining such data for human LTR retrotransposons. The first problem is homologous recombination events often remove one or both of the LTRs (Prak and Kazazian, 2000). The second problem is the human genome contains relatively few LTR elements (Smit, 1996), many of which are solitary LTRs. Next to the HERV families of LTR retrotransposons, the mammalian apparent LTR-retrotransposon (MaLR) superfamily is the most interesting to study. However, most of the LTR copies from the MaLR superfamily are found as solitary LTRs in the genome (Smit, 1993), making it difficult to determine an insertion date.

It has been shown through examination of full-length copies of the HERV-L family of LTR retrotransposons that a correlation between the relative insertion date of an element and its G+C content likely exists. This upholds the previously described observations of mutation rates in gene/pseudogene pairs and instances of repetitive elements. Such a result was not expected, yet it leads to an interesting conclusion.

## 7.6 Discussion

### 7.6.1 Shortcomings in Determining Fixed Mutation Directionality

**Gene – Pseudogene Pairs.** One of the shortcomings of the approach of looking at mutation rates in the gene-pseudogene case is the direction in which a substitution has occurred cannot be inferred with a high degree of certainty. A fairly good idea of the

direction of mutation is obtained in the gene-pseudogene case, since genes are under high selective pressure, and therefore are likely to have fewer mutations than pseudogenes. However, there are regions such as synonymous wobble bases, where mutations can occur in genic regions with little consequence to fitness. One method of getting around this would involve constructing an evolutionary phylogeny of the genes in the data set using sequences from three or more related species. This would allow us to determine with greater confidence what the original nucleotide was in the human gene, and therefore directionality could be assigned more reliably, although still not with absolute certainty. Shown in Figure 7-4 is an example of how phylogentic inference could be used to determine the likely direction of mutation, given the nucleotide sequence of four present-day organisms and a phylogenetic relationship between them.

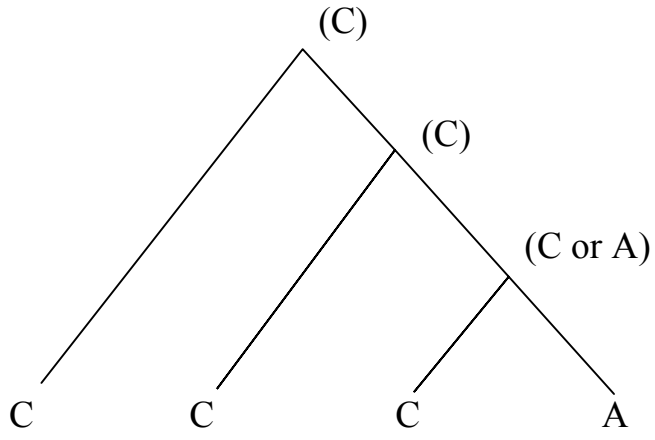



**Figure 7-4: Phylogenetic Inference.** Shown in this figure is a tree calculating the nucleotide at a specific location in a gene most likely to be present in the most recent common ancestor (parent nodes) given the currently observed nucleotide in four present-day species (leaf nodes). In the fourth species, if the pseudogene nucleotide is a C, that it is likely that there has been a C→A mutation within the gene.

Such an approach is taken by Wolfe, Sharp and Li (1989) and Casane *et al*. (1987) in their studies of small sets of genes and pseudogenes in the primate genome. While such a study may not currently be possible on a large set of genes due to the lack of large scale genomic sequence information for comparative species, it will shortly be possible in this era of genomics. Assemblies of the human genome (IHGSC, 2001) and the mouse genome (`http://genome.ucsc.edu/`) are already available and other complete genomes are likely to become available in the not too distant future.

**Repetitive elements.** Repeats within the human genome are thought to have evolved in one of two ways (Shedlock and Okada, 2000). The master gene model (Shen, Batzer and Deringer, 1991) suggests only a few Alu loci are capable of amplification, and all subsequent copies found within the genome are direct descendants from these loci. The multiple source gene model (Matera and Hellman, 1990) states offspring copies of repetitive elements may also be amplified.

Depending on which model actually holds for the human genome, the study of substitution rates in repetitive element instances has some potential pitfalls as well. Substitution rates were calculated from the `Repbase` defined sequence to the copies found in the human genome. If the master gene model was the actual mechanism, the assumptions made should be correct to the degree that the `Repbase` sequences were the actual master genes. However, if the multiple source gene model was the mechanism, some of the substitutions reported could actually be due to a single substitution occurring at some point in time in an intermediary copy, which subsequently proliferated throughout the genome.

Since the issue of which mechanism was involved is hard to resolve, we cannot be completely confident in assuming the master gene model was the only mechanism at work. At the same time, comparing substitutions to the Repbase defined consensus sequences is promising, since the Repbase repeats have been arduously studied. Therefore, while intermediary subfamilies may still exist, it seems likely a majority of substitutions observed between the Repbase sequence and a particular copy in the genome are due to accumulated substitution events in the genomic loci rather than a long line of mutational intermediaries.

### 7.6.2 Repeat Composition

The resulting studies of repetitive elements give insight into how regions of high and low G+C content are maintained within the human genome. Included is the first hypothesis accounting for the maintenance of high and low G+C regions within the human genome. This hypothesis states that the presence and G+C composition of repetitive elements was the cause of high and low G+C regions within the human genome. By looking at the occurrences of repetitive elements, however, it appears as though their G+C content was not the driving factor into the appearance of high and low G+C regions. Rather it appears as though the unique sequence DNA mirrors the G+C pattern of the surrounding sequence. Thus, the repeat composition and distribution hypothesis cannot be accepted as the cause for the maintenance of high and low G+C regions within the human genome.

### 7.6.3 Compositional Bias

The second hypothesis states that high and low G+C regions within the human genome were caused by biases in mutational mechanisms. The studies of G+C biases found in gene and pseudogene pairs as well as instances of repetitive elements indicate a high likelihood for compositional biases in substitution rates existing within the human genome. However, this compositional bias cannot be the cause for maintaining high and low G+C regions. This is due to the observed G+C biases suggesting the human genome is mutating towards A+T richness independently of the surrounding G+C content.

The ratio of G|C→A|T to A|T→G|C observed substitution rates is much higher in regions of high A+T. Such a finding suggests the human genome evolved from a G+C rich ancestral genome, and regions of high and low G+C arose as a result of the variance in mutation rates where some regions (high A+T regions) mutated faster than others (high G+C regions).

One of the difficulties with the selectionist, biased gene conversion, and mutational bias hypotheses is they are not mutually exclusive. For instance, it is possible a substitutional bias could be observed due to biased gene conversion. It is also possible substitutional biases are observed since they provide evolutionary advantages, and therefore fall under a selectionist hypothesis. Biased gene conversion could also provide changes that are advantageous and can fall under the selectionist theories.

Pseudogenes and repetitive elements are features likely to be under less selective pressure. In these regions, the bias observed is unlikely to have been caused by selection. The study of repetitive elements on the non-recombining chromosome Y yields similar

results. This indicates biased gene conversion is not likely to be the cause of the compositional biases in substitution we observe in these regions.

As described in the introduction, this context dependent substitution rate could be caused by mechanisms involved in DNA synthesis. The mechanism involved could possibly be related to the fidelity of α and β polymerases (Filipski, 1987), modification in the components of DNA synthesis (Muto and Osawa, 1987), or cytosine deamination (Fryxell and Zuckerkandl, 2000). Of course these mechanisms must be tied to germline cells in order for the mutations to become fixed in the population.

### 7.6.4 Shift Towards an A+T Rich Genome

**Shift Towards an A+T Rich Genome**. Perhaps the most intriguing result of the substitutional bias study was that the G+C biases for nearly all of the cases looked at were less than 0.5. This indicates no matter what the surrounding G+C context was, the rate of A|T→G|C substitutions seemed to be higher than the rate G|C→A|T substitutions. Such a result suggests over time, regions under less selective pressure within the human genome evolve into more A+T rich regions. The rate of this evolution appeared to be slower in high G+C regions, although it was still observed. The study of LTR retrotransposons within the human genome supports these results, since older copies tended to contain a higher A+T concentration.

**Maintenance of High G+C Regions**. The results suggest the human genome began from an ancestral genome higher in G+C composition that has evolved into a progressively lower G+C genome. However, the regions studied involved those features

(pseudogenes and repetitive elements) less likely to be involved in selection. Since there are regions of high G+C content observed within the human genome, there is likely to be some other mechanism at work to preserve these regions. One explanation for this might be that the presence of functionally and structurally important features in these regions makes the genome less tolerant of changes in their G+C composition. This would explain the high association between increasing G+C content and a higher gene density (Zoubak, Clay and Bernardi, 1996). If this is the case, the selectionist (and possibly biased gene conversion) hypotheses would hold true for these regions.

Comparing the G+C content of conserved and non-conserved regions in mouse and human could test this hypothesis. It is postulated conserved regions would have a higher G+C composition than non-conserved regions, if some sort of selection maintained high G+C regions. Otherwise, these regions would be subject to the compositional bias in substitution rates that are observed, and therefore the overall genome should mutate towards a higher A+T genome.

The main conclusions of the studies show repetitive element composition was not responsible for the maintenance of high and low G+C regions within the human genome. In addition, compositional biases in substitution rates were observed. However, the G+C biases for these substitution rates show this mechanism could not be responsible for maintenance of high G+C regions since they appear to move regions of the human genome towards a higher A+T composition over time. The study of LTR elements upholds these results, suggesting regions inserted into the human genome and under less selective pressure will mutate towards an A+T rich composition.

# Chapter 8

# Discussion

The Human Genome Project as well as the sequencing of other organisms provides the biological community with a wealth of genomic data waiting to be understood. The discipline of computational biology has provided a gateway between the biologist and the genomic data through the development of tools for mining important information.

Sequencing of the human genome at the finished quality level is proceeding at a rapidly increasing pace. As a result, assembly of finished human genomic sequences into larger contiguous regions (contigs) has proven to be a useful endeavor. Non-uniformity and redundancy in the human genome in the form of repetitive elements, pseudogenes, duplicated genes and other genomic duplications pose as obstacles that must be overcome.

We have provided a technique for conservative assembly of finished human genomic clones into larger contigs using a sequence-based method. Simulation studies indicate that approximately 93% of all overlapping fragments can be correctly assembled using this technique. The two most popular human genomic assemblies, NCBI and UCSC's Goldenpath, were examined. While both of these assemblies are based on the same input data, they contain inconsistencies in clone ordering and orientation which

leads to conflicting sequence data. Thus it is important that research using genomic assemblies as an underlying template be made aware of the inconsistencies that are present.

The availability of large contigs of human genomic data allows for the analysis of polymorphisms within the human genome. We have shown that overlapping regions between two clones originating from different haplotypes are excellent sources for mining single nucleotide polymorphisms. Mismatches in these regions allow for both the detection and clustering of potential SNPs that lead to informative genetic markers of disease.

A dynamic programming technique for aligning restriction fragment digests to contig regions has been discussed. Large-scale polymorphisms within a population can be detected using this approach. In addition, alignment of experimental and theoretical restriction digest fragments lends its hand to sequence assembly validation.

The availability of large contigs of human genomic data allows for compositional analysis of the human genome. Specifically, we have examined the organization of the human genome into CpG islands and homogeneous regions. A heuristic algorithm utilizing changepoint methods and log-likelihood statistics to detect and visualize different organizational components is discussed. Other knowledge can be mined as well, including information pertaining to gene structure, alternative splicing and paralagous sequences.

The human genome is made up of organizational components. We have shown that traditional approaches to isochore organization are not applicable when analyzing the

human genome at a sequence level. However, there are homogeneous regions that are maintained within the human genome. We have shown that repetitive element composition is not responsible for the maintenance of high and low G+C regions within the human genome. In addition, our analysis of gene to pseudogene mutations and repeat instances indicates there is an apparent compositional bias for mutation. G+C biases for these substitution rates show this mechanism cannot be responsible for maintenance of high G+C regions since they appear to move regions of the human genome towards a higher A+T composition over time. Our study of LTR elements upholds these results, suggesting regions inserted into the human genome and under less selective pressure will mutate towards an A+T rich composition.

Advancements in sequencing technology due to the human genome project have made it possible to sequence other organisms as well at a fraction of the cost in time and funds as previously was possible. In fact, over 800 genomes are represented in part or whole in the NCBI's Entrez nucleotide database (`http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome`). As more of these genomes become available in with greater genomic coverage, comparative genomics will become an important endeavor. The generality of the techniques outlined here will allow them to be applied to the genome of choice.

# References

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., *et al*. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, *252*, 1651-1656.

Aissani,B., Bernardi,G. (1991). CpG islands, genes and isochores in the genomes of vertebrates. *Gene*, *106*, 185-195.

Antequera, F., Bird, A. (1993). Number of CpG islands and genes in human and mouse. *Proceedings of the National Academy of Sciences USA*, *90*, 11995-19999.

Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J., Eichler, E.E. (2001). Segmental duplications: organization and impact within the current human genome project assembly. *Genome Research*, *11*, 1005-1017.

Barbulescu, M., Turner, G., Seaman, M.I., Deinard, A.S., Kidd, K.K., Lenz, J. (1999). Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. *Current Biology*, *9*:861-868.

Bedell, J.A., Korf, I., Gish, W. (2000). MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics*, *16*, 1040-1041.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., Wheeler, D.L (2000). GenBank. *Nucleic Acids Research*, *28*, 10-14.

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

Bernardi, G. (2000). Isochores and the evolutionary genomics of vertebrates. *Gene*, *241*, 3-17.

BioTech Resources. (1996). *BioTech Life Science Dictionary*. Retrieved from Indiana Institute for Molecular and Cellular Biology Web site:
`http://biotech.chem.indiana.edu/`

Bird, A.P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research*, *8*, 1499-1504.

Bird, A.P. (1993). Functions for DNA Methylation in Vertebrates. *Cold Spring Harbor Symposia on Quantitative Biology*, *LVIII*, 281-285.

Blackwell, T.W., Rouchka, E.C., States, D.J., (1999). Identity by Descent Genome Segmentation Based on Single Nucleotide Polymorphism Distributions. *ISMB*, *7*, 54-59.

Boguski, M.S., Lowe, T.M., Tolstoshev, C.M. (1993). dbEST -- database for 'expressed sequence tags'. *Nature Genetics*, *4*, 332-333.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology, 268*, 78-94.

Burke, D.T., Carle, G.F., Olson, M.V. (1987). Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science*, *236*, 806-182.

Cai, W., Aburatani, H., Stanton, V.P., Jr, Housman, D.E., Wang, Y.K., Schwartz, D.C. (1995). Ordered restriction endonuclease maps of yeast artificial chromosomes created by optical mapping on surfaces. *Proceedings of the National Academy of Sciences USA*, *92*, 5164-5168.

Carlin, B.P., Gelfand, A.E., and Smith, A.F.M. (1992). Hierarchical Bayesian Analysis of Changepoint Problems. *Applied Statistics*, *41*, 389-405.

Casane, D. (1997). Mutation Pattern Variation Among Regions of the Primate Genome. *Journal of Molecular Evolution*, *45*:216-226.

Castresana, J. (2002). Estimation of genetic distances from human and mouse introns. *Genome Biology*, *3*:0028.1-0028.7.

CGM. (1997). *Center for Genetics in Medicine*. Retrieved from Washington University School of Medicine Center for Genetics in Medicine Web site: `http://genome.wustl.edu/cgm/`

Chakravarti, A. (2001). Single nucleotide polymorphisms… to a future of genetic medicine. *Nature*, *409*, 822-823.

Chastain, P.D. II, Eichler, E.E., Kang, S., Nelson, D.L., Levene, S.D., Sinden, R.R. (1995). Anomalous rapid electrophoretic mobility of DNA containing triplet repeats associated with human disease genes. *Biochemistry*, *34*, 16125-16131.

Chen, C., Su, Y., Baybayan, P., Siruno, A., Nagaraja, R., Mazzarella, R. *et al*. (1996). Ordered shotgun sequencing of a 135 kb Xq25 YAC containing ANT2 and four possible genes, including three confirmed by EST matches. *Nucleic Acids Research*, *24*, 4034-4041.

Chen, E.Y., Cheng, A., Lee, A., Kuang, W.J., Hillier, L., Green, P., *et al*. (1991). Sequence of human glucose 6-phosphate dehydrogenase cloned in plasmids and a yeast artificial chromosome (YAC). *Genomics*, *10*, 792-800.

Chen, E.Y., Zolla, M., Mazzarella, R., Ciccodicola, A., Chen, C., Zuo, L., *et al*. (1996a). Long-range sequence analysis in Xq28: thirteen known and six candidate genes in 219.4 kb of high GC DNA between the RCP/GCP and G6PD loci. *Human Molecular Genetics*, *5*, 659-668.

Chen, E.Y., Schlessinger, D., Kere, J. (1993). Ordered shotgun sequencing, a strategy for integrated mapping and sequencing of YAC clones. *Genomics, 17*, 651-656.

Cheung, V.G., Nowak, N., Jang, W., Kirsch, I.R., Zhao, S., Chen, X.N., *et al*. (2001). Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature, 409*, 953-958.

Choo, K.H., Vissel, B., Brown, R., Filby, R.G., Earle, E. (1988). Homologous alpha satellite sequences on human acrocentric chromosomes with selectivity for chromosomes 13, 14 and 21: implicatons for recombination between nonhomologues and Robertsonian translocations. *Nucleic Acids Research, 16*, 1273-1284.

Collins, F.S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., Walters, L. (1998). New Goals for the U.S. Human Genome Project: 1998-2003. *Science, 282*, 682-689.

Collins, J., Bruning, H.J. (1978). Plasmids useable as gene-cloning vectors in an in vitro packaging by coliphage lambda: 'cosmids'. *Gene, 4*, 85-107.

Corhish-Bowden, A. (1984). Nomenclature for incompletely specified bases in nucleic acid sequence: Recommendations 1984. *Nucleic Acids Research, 13*, 3021-3030.

Corodonnier, A., Casella, J.F., Heidmann, T. (1995). Isolation of novel human endogenous retrovirus-like elements with foamy virus-related pol sequence. *Journal of Virology*, *69*:5890-5897.

Cox, E.C. (1972). On the Organization of Higher Chromosomes. *Nature New Biology*, *239*:133-134.

Cross, S.H. and Bird, A.P. (1995). CpG islands and genes. *Current Opinion in Genetics and Development, 5*, 309-314.

Cuny, G., Soriano, P., Macaya, G., Bernardi, G. (1981). The Major Components of the Mouse and Human Genomes. *European Journal of Biochemestry, 115*, 227-233.

D'Onofrio, G., Bernardi, G. (1992). A universal compositional correleation among coding positions. *Gene, 110*, 81-88.

Drury, H.A., Green, P., McCauley, B.K., Olson, M.V., Politte, D.G., and Thomas, Jr., L.J. (1990). Spatial Normalization of One-Dimensional Electrophoretic Gel Images. *Genomics, 8*, 119-126.

Drury, H.A., Clark, K.W., Hermes, R.E., Feser, J.M., Thomas, Jr., L.J., and Donis-Keller, H. (1992). A Graphical User Interface for Quantitative Imaging and Analysis of Electrophoretic Gels and Autoradiograms. *BioTechniques, 12*, 892-901.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., *et al*. (1999). The DNA sequence of human chromosome 22. *Nature, 402*, 489-495.

Eyre-Walker, A., (1993). Recombination and mammalian genome evolution. *Philosophical Transactions of the Royal Society of London. Series B, 252,* 237-243.

Eyre-Walker, A. (1999). Evidence of Selection on Silent Site Base Composition in Mammals: Potential Implications for the Evolution of Isochores and Junk DNA. *Genetics*, *152*:675-683.

Eyre-Walker, A., Hurst, L.D., (2001). The evolution of isochores. *Nature Reviews Genetics, 2*, 549-555.

Feng, Q., Moran, J.V., Kazazian, H.H.Jr., Boeke, J.D. (1996). Human L1 Retrotransposons encodes a conserved endonuclease required for retrotransposition. *Cell*, *87*:905-916.

Filipski, J. (1987). Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. *FEBS Letters, 217*, 184-186.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Research, 8,* 967-974.

Francino, M.P., Ochman, H. (1999). Isochores results from mutation not selection. *Nature, 400*, 30-31.

Fryxell, K.J., Zuckerkandl, E. (2000). Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Molecular Biology and Evolution, 17,* 1371-1383.

Fullerton, S.M., Bernardo Carvalho, A., Clark, A.G. (2001). Local rates of recombination are positively correlated with GC content in the human genome. *Molecular Biology and Evolution, 18*, 1139-1142.

Galtier, N., Lobry, J.R. (1997). Relationships Between Genomic G+C Content, RNA Secondary Structures, and Optimal Growth Temperature in Prokaryotes. *Journal of Molecular Evolution*, *44*:632-636.

Galtier, N., Piganeau, G., Mouchiroud, D., Duret, L. (2001). GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. *Genetics*, *159*:907-911.

Gardiner, K. (1996). Base composition and gene distribution: critical patterns in mammalian genome organization. *Trends in Genetics, 12*, 519-524.

Gillett, W. (1992). DNA Mapping Algorithms: Strategies for Single Restriction Enzyme and Multiple Restriction Enzyme Mapping. Technical Report, Washington University, Department of Computer Science, WUCS-92-29.

Gillett, W., Hanks, L., Wong, G.K.S., Yu, J., Lim, R., Olson, M.V., (1996). Assembly of high-resolution restriction maps based on multiple complete digests of a redundant set of overlapping clones. *Genomics, 33*, 389-408.

Gish, W. (1996-2001). unpublished.

Griffiths, D.J. (2001). Endogenous retroviruses in the human genome sequence. *Genome Biology*, *2*:1017.1-1017.5.

Gu, X., Li W.-H. (1994). A Model for the Correlation of Mutation Rate with GC Content and the Origin of *GC*-Rich Isochores. *Journal of Molecular Evolution*, *38*:468-475.

Guan, X., Mural, R.J., Einstein, J.R, Mann, R.C., Uberbacher, E.C. (1992). GRAIL: An Integrated Artificial Intelligence System for Gene Recognition and Interpretation. *Proceedings Of The Eighth IEEE Conference on AI Applications*, 9-13.

Guigo, R., Knudsen, S., Drake, N., Smith, T. (1992). Prediction of gene structure. *Journal of Molecular Biology, 226*, 141-157.

Häring, D., Kypr, J., (2001). No isochores in human chromosomes 21 and 22? *Biochem. Biophys. Res. Comm., 280,* 567-573.

Hattori, M., Fujiama, A., Taylor, T.D., Watanabe, H., Yada, T., Park, H.S., *et al*. (2000). The DNA sequence of human chromosome 21. *Nature, 405*, 311-319.

Hirose, Y., Takamatsu, M., Harada, F. (1993). Presence of env genes in members of the RTVL-H family of human endogenous retrovirus-like elements. *Virology*, *192*:52-61.

Hughes, S., Zelus, Z., Mouchiroud, D. (1999). Warm-Blooded Isochore Structure in Nile Crocodile and Turtle. *Molecular Biology and Evolution*, *16*:1521-1527.

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*, 860-921.

Jabbari, K., Bernardi, G. (1998). CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families. *Gene, 224*, 123-128.

Jang, W., Chen, H.-C., Sicotte, H., Schuler, G.D. (1999). Making effective use of genomic sequence data. *Trends in Genetics, 15*, 284-286.

Jurka, J. (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proceedings of the National Academy of Sciences USA*, *94*:1872-1877.

Jurka, J. (1998). Repeats in genomic DNA: mining and meaning. *Current Opinion in Structural Biology, 8*, 333-337.

Jurka, J. (2000). Repbase update: a database and an electronic journal of repetitive elements. *Trends in Genetics, 16,* 418-420.

Kan, Z., Gish, W., Rouchka, E., Glasscock, J., States, D. (2000). UTR Reconstruction and Analysis Using Genomically Aligned EST Sequences. *ISMB, 8*, 218-227.

Karlin, S. (1994). Statistical studies of biomolecular sequences: score-based methods. *Philosophical Transactions of the Royal Society of London. Series B, 344*, 391-402.

Karlin, S., Altschul, S.F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences USA, 87,* 2264-2268.

Kent, J.W., Haussler, D. (2001). Assembly of the Working Draft of the Human Genome with GigAssembler. *Genome Research, 11*, 1541-1548.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution, 16*, 111-120.

Klug, W.S., Cummings, M.R. (1995). *Concepts of Genetics*. New York: Macmillan Publishing Company.

Knight, R.D., Freeland, S.J., Landweber, L.F. (2001). A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biology, 2*, research0010.1-0010.13.

Lahn, B.T., Pearson, N.M., Jegalian, K. (2001). The human Y chromosome, in the light of evolution. *Nature Reviews Genetics*, *2*:207-216.

Larsen, F., Gundersen, G., and Lopez, L. (1992). CpG islands as Gene Markers in the Human Genome. *Genomics, 13*, 1095-1107.

Lawerence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C. (1993). Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment. *Science, 262*, 208-214.

Lawrence, C.E., and Reilly, A.A. (1985). Maximum Likelihood Estimation of Subsequence Conservation. *Journal of Theoretical Biology, 113*, 425-439.

Lilliefors, H.W. (1967). On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *American Statistical Association Journal, 62*, 399-402.

Lodish, H., Baltimore, D., Berk, A., Zipursky, S.L., Matsudaria, P., Darnell, J. (1995). *Molecular Cell Biology*. New York: Scientific American Books.

Lopez, R. (1995). *CpG Island Database*. Downloaded 1997 from embonet.news Web site:
`http://www.no.embnet.org/embnet.news/vol2_2/contents.html`

Lynch, M., Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. *Science, 290*, 1151-1155.

Macaya, G., Thiery, J.P., Bernardi, G. (1976). An approach to the organization of eukaryotic genomes at a macromolecular level. *Journal of Molecular Biology, 108*, 237-254.

Macilwain, C. (2000). World leaders heap praise on human genome landmark. *Nature, 405*, 983-986.

Macleod, D., Charlton, J., Mullins, J., Bird, A.P. (1994). Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. *Genes & Development, 8*, 2282-2292.

Maniatis, T., Jeffrey, A., van deSande, H. (1975). Chain length determination of small double- and single-stranded DNA molecules by polyacrylamide gel electrophoresis. *Biochemistry, 14*, 3787-3794.

Matassi, G., Labuda, D., Bernardi, G. (1998). Distribution of the mammalian-wide interspersed repeats (MIRs) in the isochores of the human genome. *FEBS Letters, 439*, 63-65.

Matera, A.G., Hellman, U., Schmid, C.W. (1990). A transpositionally and transcriptionally competent Alu subfamily. *Molecular Cell Biology*, *10*:5424-5432.

Morton, N.E. (1991). Parameters of the human genome. *Proceedings of the National Academy of Sciences USA, 88*, 7474-7476.

Mouchiroud, D., D'Onofrio, G., Aissani, B, Macaya, G., Gautier, C. Bernardi, G. (1991). The distribution of genes in the human genome. *Gene, 100*, 181-187.

Mural, R.J., Parang, M., Shah, M., Snoddy, J., Uberbacher, E.C. (1999). The Genome Channel: a browser to a uniform first-pass annotation of genomic DNA. *Trends in Genetics, 15*, 38-39.

Muto, A., Osawa, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. *Proceedings of the National Academy of Sciences USA, 84*, 166-169.

Needleman, S.B., Wunsch, C.D. (1970). A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology, 48*, 443-453.

Nekrutenko, A., Li, W.H. (2000). Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome Research, 10*, 1986-1995.

Ohama, T., Yamao, F., Muto, A., Osawa, S. (1987). Organization and Codon Usage of the Streptomycin Operon in *Micrococcus luteus*, a Baterium with a High Genomic G+C Content. *Journal of Bacteriology, 169*, 4770-4777.

Ono, M., Yasunaga, T., Miyata, T., Ushikubo, H. (1986). Nucleotide sequence of human endogenous retrovirus genome related to the mouse mammary tumor virus genome. *Journal of Virology*, *60*:589-598.

Pavlíček, A., Jabbari, K., Pačes, J., Pačes, V., Hejnar, J., Bernardi, G. (2001). Similar integration but different stability of Alus and LINEs in the human genome. *Gene*, *276*:39-45.

Peacock, A.C., Bunting, S.L., Cole, S.P., and Seidman, M. (1985). Two-dimensional electrophoretic display of restriction fragments from genomic DNA. *Analytical Biochemistry, 149*, 177-182.

Perkin-Elmer. (1998). *Perkin-Elmer, Dr. J. Craig Venter, and TIGR Announce Formation Of New Genomics Company*. Retrieved May 9, 1998, from Perkin-Elmer Press Releases Web site:
http://www.perkin-elmer.com/press/prc5447.html.

Picoult-Newberg, L., Ideker, T.E., Pohl, M.G., Taylor, S.L., Donaldson, M.A., Nickerson, D.A., Boyce-Jacino, M. (1999). Mining SNPs From EST Databases. *Genome Research, 9*, 167-174.

Piganeau, G., Mouchiroud, D., Duret, L., Gautier, C. (2002). Expected Relationship Between the Silent Substitution Rate and the GC Content: Implications for the Evolution of Isochores. *Journal of Molecular Evolution*, *54*:129-133.

Prak, E.T., Kazazian, Jr., H.H. (2000). Mobile elements and the human genome. *Nature Reviews Genetics*, *1*:134-144.

Pruitt, KD, Maglott, DR. (2001). RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research, 29,* 137-140.

Riles, L., Dutchik, J.E., Baktha, A., McCauley, B.K., Thayer, E.C., Leckie, M.P., *et al*. (1993). Physical maps of the six smallest chromosomes of Saccharomyces cerevisiae at a resolution of 2.6 kilobase pairs. *Genetics, 134*, 81-150.

Rivella, S., Tamanini, F., Bione, S., Mancini, M. Herman, G., Chatterjee, A., *et al*. (1995). A Comparative Transcriptional Map of a Region of 250 kb on the Human and Mouse X Chromosome between the G6PD and the FLN1 Genes. *Genomics, 28*, 377-382.

Robinson, H., Gao, Y., Mccray, B.S., Edmondson, S.P., Shriver, J.W., Wang, A.H.J. (1998). The hyperthermophile chromosomal protein Sac7d sharply kinks DNA. *Nature*, *392*:202-205.

Rouchka, E.C., States, D.J. (1998). Sequence Assembly Validation by Multiple Restriction Digest Fragment Coverage Analysis. *ISMB, 6*, 140-147.

Rouchka, E.C., States, D.J. (1999). Assembly and Analysis of Extended Human Genomic Contig Regions. Technical Report, Washington University Department of Computer Science, WUCS-99-10.

Rouchka, E.C., States, D.J. (2002). Compositional Analysis of Homogeneous Regions in Human Genomic DNA. Technical Report, Washington University Department of Computer Science, WUCS-2002-2.

Sellers, P.H. (1974). On the theory of computation of evolutionary distances. *SIAM Journal of Applied Mathematics, 26*, 787-793.

Shaikh, T.H., Kurahashi, H., Saitta, S.C., O'Hare, A.M., Hu, P., Roe, B.A., *et al*. (2000). Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genome organization and deletion endpoint analysis. *Human Molecular Genetics, 9*, 489-501.

Shedlock, A.M., Okada, N. (2000). SINE insertions: powerful tools for molecular systematics. *Bioessays*, *22*:148-160.

Shen, M.R., Batzer, M.A, Deninger, P.L. (1991). Evolution of the master Alu gene(s). *Journal of Molecular Evolution*, *33*:311-320.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research, 29*, 308-311.

Shizuya, H., Birren, B., Kim, U.J., Mancino, V., Slepak, T., Tachiiri, Y., Simon, M. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proceedings of the National Academy of Sciences USA, 89*, 8794-8797.

Smith, N.G.C., Eyre-Walker, A. (2001). Synonymous Codon Bias Is Not Caused by Mutation Bias in G+C Rich Genes in Human. *Molecular Biology and Evolution*, **18**:982-986.

Smit, A.F.A. (1993). Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucleic Acids Research*, *21*:1863-1872.

Smit, A.F.A. (1996). The origin of interspersed repeats in the human genome. *Current Opinion in Genetics & Development*, *6*:743-748.

Smit, A.F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current Opinion in Genetics And Development 9*, 657-663.

Smit, A.F.A., Green, P. Unpublished.

Smith, D., Carrano, A. (1996). International Large-Scale Sequencing Meeting. *Human Genome News, 7*, 19.

Smith, T.F., Waterman M.S. (1981). Identification of Common Molecular Subsequences. *Journal of Molecular Biology, 147,* 195-197.

States, D.J. Unpublished.

Stewart, E.A., McKusick, K.B., Aggarwal, A., Bajorek, E., Brady, S., Chu, A., *et al*. (1997). An STS-based radiation hybrid map of the human genome. *Genome Research, 7*, 422-423.

Stoesser, G., Baker, W., vad den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., *et al*. (2001). The EMBL Nucleotide Sequence Database. *Nucleic Acids Research, 29*, 17-21.

Sueoka, N. (1962). On the genetic basis of variation and heterogeneity of DNA base composition. *Proceedings of the National Academy of Sciences USA*., *48*:582-592.

Taguchi, H., Konishi, J., Ishii, N., Yoshida, M. (1991). A chaperonin from a thermophilic bacterium *Thermus thermophylus*, that controls refolding of several thermophilic enzymes. *The Journal of Biological Chemistry*, *266*:22411-22418.

Taillon-Miller, P., Gu, Z., Li, Q., Hillier, L., Kwok, P.Y. (1998). Overlapping genomic sequences: A treasure trove of single-nucleotide polymorphisms. *Genome Research, 8,* 748-754.

Tateno, Y., Miyazaki, S., Ota, M., Sugawara, H., and Gojobori, T. (2000). DNA Data Bank of Japan (DDBJ) in collaboration with mass sequencing teams. *Nucleic Acids Research, 28*, 24-26.

Thiery, J.-P., Macaya, G., Bernardi, G. (1976). An Analysis of Eukaryotic Genomes by Density Gradient Centrifugation. *Journal of Molecular Biology, 108*, 219-235.

Tilford, C.A., Kuroda-Kawaguchi, T., Skaletsky, H., Rozen, S., Brown, L.G., Rosenberg, M., et al. (2001). A physical map of the human Y chromosome. *Nature*, *409*:943-945.

Tristem, M. (2000). Identification and Characterization of Novel Human Endogenous Retrovirus Families by Phylogenetic Screening of the Human Genome Mapping Project Database. *Journal of Virology*, *74*:3715-3730.

Vaughan, D. (ed.). (1996). *To Know Ourselves*. Retrieved from Oak Ridge National Labs Web site:
`http://www.ornl.gov/TechResources/ Human_Genome/tko/`

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., *et al*. (2001). The sequence of the human genome. *Science, 291*, 1304-1351.

Wada, A., Suyama, A. (1986). Local stability of DNA and RNA secondary structure and its relation to biological functions. *Progress in Biophysics and Molecular Biology*, *47*:113-157.

Waterston, R.H., Ainscough, R., Anderson, K., Berks, M., Blair, D., Connell, M., *et al*. (1993). The genome of the nematode Caenorhabditis elegans. *Cold Spring Harbor Symposium on Quantitative Biology, 58*, 367-376.

Watson, J.D., Crick, F.H.C. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature, 171*, 737-738.

Wolfe, K.H. (1991). Mammalian DNA Replication: Mutation Biases and the Mutation Rate. *Journal of Theoretical Biology*, *149*:441-451.

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

Zoubak, S., Clay, O., Bernardi, G. (1996). The gene distribution of the human genome. *Gene, 174*, 95-102.

# Vita

Eric C. Rouchka

| | |
|---|---|
| **Date of Birth** | July 12, 1972 |
| **Place of Birth** | Sedalia, Missouri, USA |
| **Undergraduate Study** | Rockhurst College, Kansas City, Missouri<br>B.S. Computer Science, Mathematics, Biology, May 1994 |
| **Graduate Study** | Rennselaer Polytechnic Institute, Troy, New York<br>M.S. Computer Science, May 1996 |
| | Washington University, St. Louis, Missouri<br>D.Sc. Computer Science, August 2002 |
| **Publications and Technical Reports** | Rouchka, E.C., States, D.J. (2002) "Compositional Analysis of Homogeneous Regions in Human Genomic DNA." Technical Report, Washington University Department of Computer Science, WUCS-2002-2. |
| | Rouchka, E.C., Gish, W., States, D.J. "Maintenance of Compositional Variation in the Human Genome." Under Review. |
| | Rouchka, E.C., Gish, W., States, D.J. "Comparison of Whole Genome Assemblies of the Human Genome." In Preparation. |
| | Kan, Z., Rouchka, E.C., Gish, W.R., States, D.J. (2001) "Gene structure prediction and alternative splicing analysis using genomically aligned ESTs." *Genome Research* **11**(5):889-900. |
| | Kan, Z., Gish, W., Rouchka, E., Glasscock, J., States, D. (2000) "UTR Reconstruction and Analysis Using Genomically Aligned EST Sequences." *ISMB* **8**:218-227. |
| | Rouchka, E. C., States, D.J. (1999) "Assembly and Analysis of Extended Human Genomic Contig Regions." Technical Report, Washington University Department of Computer Science, WUCS-99-10. |

Rouchka, E.C. (1999) "Pattern Matching Techniques and Their Applications to Computational Molecular Biology-- A Review." Technical Report, Washington University Department of Computer Science, WUCS-99-09.

Blackwell, T.W., Rouchka, E. C., States, D.J. (1999) "Identity by Descent Genome Segmentation Based on Single Nucleotide Polymorphism Distributions." *ISMB*, **7**:54-59.

Rouchka, E.C., States, D.J. (1998) "Sequence Assembly Validation by Multiple Restriction Digest Fragment Coverage Analysis" *ISMB*, **6**:140-147.

Rouchka, E.C., Mazzarella, R., States, D.J. (1997) "Computational Detection of CpG Islands in DNA" Technical Report, Washington University, Department of Computer Science, WUCS-97-39.

Rouchka, E.C., States, D.J. (1997) "Sequence Assembly Validation by Restriction Digest Fingerprint Comparison." Technical Report, Washington University, Department of Computer Science, WUCS-97-40.

Rouchka, E.C. (1997) "A Brief Overview of Gibbs Sampling." IBC Statistics Study Group, Washington University, Institute for Biomedical Computing.

Rouchka, E.C. (1996) "An Algorithmic Approach to Gene Regulatory Sequence Analysis." Master's Project, Rensselaer Polytechnic Institute, Department of Computer Science.

**Posters Presented**

"Large Scale Genome Sequence Composition Analysis" ISMB2000, August 2000, San Diego, CA

"Assembly and Analysis of Extended Human Genomic Contig Regions" Cold Spring Harbor Genome Sequencing and Biology Meeting, 1999, Cold Spring Harbor, NY

"Assembly and Analysis of Extended Human Genomic Contig Regions" DOE Grantees and Contractors, 1999, Oakland, CA

"Sequence Assembly Validation by Multiple Restriction Digest Fragment Coverage Analysis" ISMB98, June 1998, Montreal, Canada

"Computational Detection of CpG Islands in DNA Sequence" DOE Grantees and Contractors, 1997, Santa Fe, NM
"Sequence Assembly Validation by Restriction Fragment Analysis" DOE Grantees and Contractors, 1997, Santa Fe, NM

**Invited Talks**

"Maintenance of Homogeneous Regions in Human Genomic DNA" University of Louisville Department of Computer Science, March 2002, Louisville, KY

"A Resource of Finished Human Genomic Sequence Contigs Using a Sequence-Based Approach", RECOMB Satellite Meeting on DNA Sequence Assembly, May 2001, Los Angeles, CA

"Two computational issues in genome sequence analysis: Fragment assembly validation and sequence segmentation", St. Louis University Department of Computer Science Colloquium, April 2000, St. Louis, MO

"Sequence Assembly Validation by Multiple Restriction Digest Fragment Coverage Analysis" ISMB98, July 1998, Montreal, Canada

August, 2002