Microarray Approaches to Experimental Genome Annotation

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

By
Paul Nicolas Bertone

Dissertation Directors: Mark Gerstein and Michael Snyder

May 2005

Microarray Approaches to Experimental Genome Annotation

Paul Bertone

2005

Abstract

This work describes the development and application of genomic DNA tiling arrays: microarrays designed to represent all of the DNA comprising a chromosome or other genomic locus, regardless of the genes that may be annotated in the region of interest. Because tiling arrays are intended for the unbiased interrogation of genomic sequence, they enable the discovery of novel functional elements beyond those described by existing gene annotation. This is of particular importance in mapping the gene structures of higher eukaryotes, where combinatorial exon usage produces rare splice variants or isoforms expressed in low abundance that may otherwise elude detection. Issues related to the design of both oligonucleotide- and amplicon-based tiling arrays are discussed; the latter technology presents distinct challenges related to the selection of suitable amplification targets from genomic DNA. Given the widespread fragmentation of mammalian genomes by repetitive elements, obtaining maximal coverage of the non-repetitive sequence with a set of fragments amenable to high-throughput polymerase chain reaction (PCR) amplification represents a non-trivial optimization problem. To address this issue, several algorithms are described for the efficient computation of optimal tile paths for the design of amplicon tiling arrays. Using these methods, it is possible to recover an optimal tile path that maximizes the coverage of non-repetitive DNA while minimizing the number of repetitive elements included in the resulting sequence fragments. Tiling arrays were constructed and used for the chromosome- and genome-wide assessment of human transcriptional activity, via hybridization to complementary DNA derived from polyadenylated RNA expressed in normal complex tissues. The approach is first demonstrated with amplicon arrays representing all of the non-repetitive DNA of human chromosome 22, then extended to the entire genome using maskless photolithographic DNA synthesis technology. A large-scale tiling array survey revealed the presence of over 10,000 novel transcribed regions and verified the expression of nearly 13,000 predicted genes, providing the first global transcription map of the human genome. In addition to those likely to encode protein sequences on the basis of evolutionary sequence conservation, many of the novel transcripts constitute a previously uncharacterized population of non-coding RNAs implicated in myriad structural, catalytic and regulatory functions.

Microarray Approaches to Experimental Genome Annotation

Paul Bertone

PhD 2005

Yale University

Preface and Acknowledgments

This dissertation has been a long time coming. I originally went to college for music and received my first degree in composition. A growing interest in technology, along with a sense of commonality between art and mathematics, led me to computer science. I studied computation in graduate school for five years, specializing in artificial intelligence, programming language design and computability theory. Many aspects of computing are of great interest to me, but I also wanted to find an application for some of the techniques I had specialized in, particularly machine learning. Over time I became interested in molecular biology for the wealth of interesting and complex research problems in the field.

Despite the current popularity of bioinformatics, working on real biological problems was not straightforward. Without the large-scale data sets we deal with today, biologists had little need for computer scientists. When computational techniques were used, I saw a disappointing trend in the way computer scientists approached the situation. Given limited knowledge of biology, one is tempted to simply apply familiar techniques (e.g., string algorithms) to nucleic acid and protein sequences with little understanding of the underlying purpose or implications. I therefore decided that if I wanted to make a meaningful contribution to the field, it would be necessary for me to pursue molecular biology and become an experimentalist in addition to working in computation. This was one of the best decisions I ever made, due in large part to the many people who gave me the opportunity to develop my career in this direction.

As each year passes, I am always surprised at how much my life has improved and how more knowledgeable I am than I was the previous year. This continued realization is a product of my interactions with friends, family and coworkers. These are the people who have influenced my life and shaped my goals, bringing me closer to the person I hope to become.

I first wish to thank my research advisors, Mark Gerstein and Michael Snyder. The advantages of working in both a computational and experimental environment cannot be overstated, particularly given the range of projects I have been able to work on through their sponsorship. It was immensely satisfying to be able to work in Mike's lab to generate experimental data, then perform computational analyses in Mark's group. This joint arrangement has given me a unique outlook on science, and for that I will always be grateful.

I also want to acknowledge the input of my research committee: Sherman Weissman, Perry Miller, Doug Kankel, Frank Ruddle and Frank Slack. I am particularly thankful for the many excellent conversations I have had with Sherm and Perry over the years about genetics, computers and life. I would also like to thank Joe Ecker (Salk Institute) for reading my dissertation.

I want to thank my labmates and other colleagues with whom I have had many inspiring discussions: Bob Adams, Vadim Alexandrov, Sid Altman, Wilhelm Ansorge, Peter Bertone, Suganthi Balasubramanian, Scott Bidlingmaier, Metin Bilgin, Anthony Borneman, Ron Breaker, Nick Carriero, Antonio Casamayor, Alexandra Charos, Paolo Coelho, Rajdeep Das, Enrique De La Cruz, Xing-Wang Deng, Geeta Devgan, Savithramma Dinesh-Kumar, Shawn Douglas, Nat Echols, Olof Emanuelsson, Don Engelman, Ghia Euskirchen, George Farr, Zareen Gauhar, Dan Gelperin, Tara Gianoulis, Yoav Gilad, Tom Gingeras, Bob Green, Dov Greenbaum, Jim Griffith, Roderic Guigó, Paul Harrison, Steve Hartman, Hedi Hegyi, Christine Horak, Art Horwich, Christine Jacobs, Ron Jansen, Ted Johnson, Ghil Jona, Jochen Junker, John Karro, Yuval Kluger, Anuj Kumar, Ian Laurenzi, Bob Leamnson, Sue Leclair, Peter Little, Nick Luscombe, Becky Martone, Chris Mason, John Mattick, Greg Michaud, Duncan Milburn, Janine Mok, Ken Nelson, Sara Nichols, Svante Pääbo, Alberto Paccanaro, Shae Padrick, Jason Ptacek, John Quackenbush, Bing Ren, John Rinn, Tom Royce, Joel Rozowsky, Bill Segraves, Mike Seringhaus, Tom Scheuermann, Falk Schubert, Chern Sing-Goh, Mike Smith, Dieter Söll, Ned Staples, Valery Trifonov, Richard Upchurch, Alex Urban, Marc Vidal, Eric White, Kevin Wise, Joe Wolenski, Yu Xia, Haiyuan Yu, Zhaolei Zhang, Deyou Zheng, Heng Zhu and Xiaowei Zhu.

I am lucky to have assembled an outstanding group of friends outside of Yale who have offered strong and consistent encouragement, including Alec Antobenedetto and family, James and Maggie Burke, John Choquette, Craig Dahlberg, Clif and Joyce Fischlein, Jack and Kim Foehl, Chris and Amy Goffredo, Arno Grbac and family, Adrian and Alison Griffiths, Osha Gula, Jon and Francie Harris, the Josti family, Todd Lander, Matthew Mandelbaum, Jamie and Karen Quinn, Bruce Rader, Jay and Suzy Rouhana, Craig and Trish Rousseau, Bob Sellers and family, Eric Staffier, Matt and Kelly Stevens, Scott Tarulli and family, Petie Thorndike, Steve and Diane Troppoli, Theresa Willett and Tim Zimmerman.

My family has been a source of unwavering support over the years; my parents Felice and Joan, brother Mark, sister Diana and niece Aislyn constitute the emotional core of my life. I also wish to thank my extended family, as well as the Liponis family, for all of their love and kindness. I am particularly grateful to Steve, Marilyn and Tom Liponis for making me feel like a member of their family from the outset.

Most of all I want to thank my wife, Ruth. She has been a source of infinite love and encouragement, providing friendship and support through all my experiences and endeavors. Throughout this section I have emphasized how fortunate I am to have known and worked with such amazing people. While that is certainly true, most of all I am lucky that Ruthie married me.

ORGANIZATION AND SYNOPSIS

Chapter 1 contains an introduction to some of the problems inherent in eukaryotic genome annotation, with respect to characterizing expressed mRNAs and mapping transcribed sequences. Here I outline a brief history of this area and describe why the microarray approach is well suited to this problem.

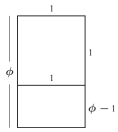
Chapter 2 describes computational approaches for designing microarrays for a variety of applications. In particular, the chapter addresses the problem of recovering a population of non-repetitive sequence fragments amenable to PCR amplification. To facilitate the design of amplicon (PCR-based) tiling arrays, I developed new algorithms to derive optimal tile paths through highly repetitive genomic DNA. This approach maximizes the coverage of non-repetitive DNA while minimizing the number of repetitive elements included in the amplified sequence fragments. This work was instrumental in the construction of amplicon arrays spanning human chromosome 22 (Chapter 3). Work on similar problems has been an ongoing collaboration with several people, including Ming-Yang Kao, Bhaskar DasGupta, Falk Schubert, Valery Trifonov and Joel Rozowsky. The theoretical analysis of various tiling problems is expanded in Appendices B–D.

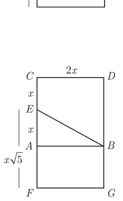
Chapter 3 introduces a DNA microarray of human chromosome 22, developed by our group over a period of about three years. This was a large project carried out by a relatively small group of individuals. The work was initiated by Ghia Euskirchen in early 2000, shortly after the sequence of chromosome 22 was first published. Ghia began painstakingly acquiring and cataloging BAC clones from various labs in the sequencing consortium, while I began work on computational methods that would allow us to select thousands of sequences across the chromosome for PCR amplification. It eventually became clear that the BAC approach would be infeasible, so it was decided to amplify the DNA from genomic template (HeLa cells) instead. Ghia undertook this effort and PCR-amplified 50% of the chromosome herself (over 10,000 reactions). During this time I began building a database and analysis system to support the project, integrating existing gene annotation from numerous sources and developing software to coordinate the capture of experimental data. The microarray construction effort was later shared between Ghia, Becky Martone and John Rinn, who joined the project in 2001. In addition to his participation in constructing the arrays, John would go on to probe them with a variety of complex tissue RNAs. Nick Luscombe and I used this data to map transcriptional activity across the chromosome. Nick was instrumental in developing a statistical scoring method to determine positive hybridizing fragments, and I performed the remaining analyses in consultation with Paul Harrison. Paul and I used a number of gene prediction methods to guide the design of a secondary array, which was used to determine strand-specific transcription. This oligonucleotide-based approach was essentially a precursor to the genome-wide transcript mapping experiment described in Chapter 4. Subsequent to this study, several members of our group have used this microarray platform to identify the chromosome-wide binding sites of several transcription factors via hybridization to chromatin immunoprecipitated DNA, as reviewed in Chapter 5.

Chapter 4 describes the application of a new DNA microarray technology called maskless photolithography, and extends the experiment described in Chapter 3. Using high-resolution oligonucleotide arrays, our group conducted the first complete survey of transcriptional activity across the genome. This work is based on a multi-year collaboration I led with the nanotechnology group at the NASA Ames Research Center. John Rinn was a great help as we carried out pilot experiments together to develop new hybridization protocols. I would later work with Alex Urban and Xiaowei Zhu to carry out these techniques for a variety of different experiments. Using this technology I eventually developed a series of high-resolution microarrays to represent all of the non-repetitive DNA in the human genome. Alex, Xiaowei and I used these arrays to map transcriptional activity in polyadenylated liver tissue RNA; these experiments were performed over a year-long period, and their enthusiastic dedication to the project did not go unnoticed. I was also fortunate to be able to work closely with Tom Royce and Joel Rozowsky in the analysis of the microarray data. It was largely through their efforts that we were able to formulate an accurate picture of the results, and I greatly enjoyed working with them. This study produced the first mammalian genome-wide transcription map, simultaneously identifying over 10,000 novel transcribed sequences and verifying the expression of nearly 13,000 predicted genes.

Chapter 5 concludes with a view of alternate applications of genomic DNA tiling arrays, and how these methods relate to the problem of transcript mapping discussed earlier. At the end of the chapter I take the opportunity to highlight the excellent work of two labmates, Nick Luscombe and Haiyuan Yu, in the analysis of transcription factor-target relationships and the derivation of complex regulatory networks from these data. Due to the availability of large-scale data sets, this work has focused on the budding yeast *Saccharomyces cerevisiae* as a model organism. However, it seems apparent that the global analysis of transcriptional regulation will be of primary interest in human genetics as more comprehensive experiments are undertaken, so I felt it appropriate to underscore this work as an area where I think the field is headed.

ABOUT THE DOCUMENT





The general page design follows a grid system based on the de Stijl form, inspired by the work of Dutch Constructivist painter Piet Mondrian. The layout is defined by a number of intersecting golden rectangles whose dimensions are defined by the relation $\phi = \frac{1}{2}(1 + \sqrt{5}) = 1.618$. When the square $\square ABDC$ is removed from the rectangle $\square CFGD$, a new rectangle is produced with identical proportions. The golden rectangle is a fractal whose infinite recursion is derived from the Fibonacci sequence and follows the harmonic series $\sum_{k=1}^{\infty} \frac{1}{k}$.

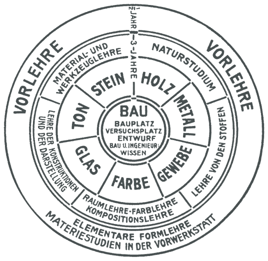
The principal typeface is Perpetua, designed by Eric Gill in 1928 for Samuel Morris of the Monotype Corporation. The font is an adaptation of the letterforms etched on the column of Trajan in Rome, and was first used in a limited edition book of *The Passion of Perpetua and Felicity*, after which the Roman and italic typefaces are named. The type is intended to have a chiseled quality that recalls Gill's work as a stonecutter. Mathematical symbols and equations are set in Euclid and Computer Modern fonts used in the TEX system developed by Donald Knuth.

Chapter headings and figure labels appear in a clean and direct sans-serif typeface designed by Adrian Frutiger in 1968, originally commissioned for a new signage system to complement the modern architecture of the aéroport de Roissy (later named after Charles de Gaulle) outside Paris.

Coming from a creative background, I appreciate a range of different art forms (trivia: I actually share a birthday with Vasarely, whose work I enjoy). I have a particular interest in Modernist design that was prevalent in work of the early twentieth century, when the introduction of machines began to change the landscape of society. Among the most prominent centers of modern design was the Bauhaus school, founded by architect Walter Gropius in 1919. Although the Wiemar and Dessau facilities were in operation only until 1933 when the institute was shut down by the Nazis, the innovations developed at the Bauhaus would have a profound impact on industrial design, architecture, furniture construction, photography, painting, theatre sets and typography. In addition to the obvious cultural influence affected by the products developed there, I think that much credit for the success of the institute can be attributed to the methodology promoted by its founders. The faculty of the Bauhaus consisted of a panel of artisans with diverse backgrounds: architecture, painting, graphic design, etc. However, the staff also included a number of master craftsmen and engineers. The reason for this deliberate

STAATLICHES
B
AUHAUS
WEIMAR
1919
1923
BAUHAUSVERLAG
MÜNCHEN
MAXIMILIANSTR.

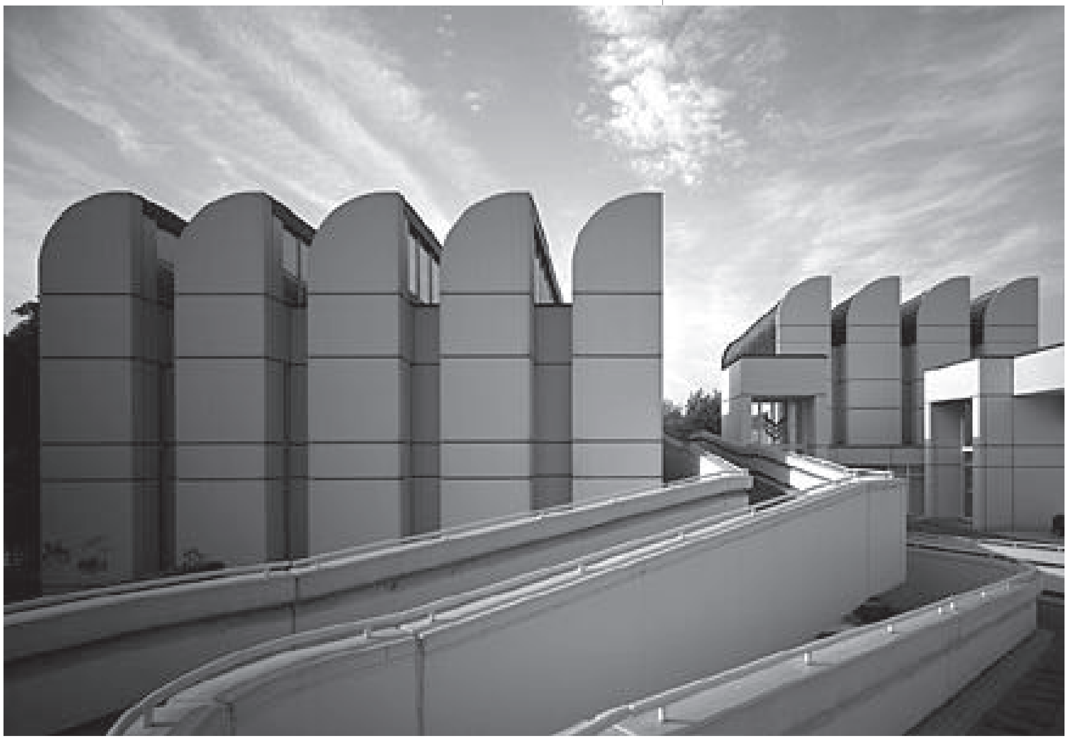
L. MOHOLY-NAGY



integration of various experts was to address the multifaceted demands of new design projects. Students often devised unique plans where the machines and fabrication techniques required to realize their ideas might not exist. They would therefore work not only with a faculty artisan, but also in consultation with craftsmen to develop the necessary instrumentation. Engineering staff did not need a comprehensive background in design theory to be effective; likewise, the artists involved in directing a given project may not have been intimately familiar with the details of its implementation. Through collaborative effort, people with different areas of expertise worked to combine their skills to make a designer's vision a reality.

I see many similarities in the way we approach modern scientific investigation, particularly where the application of emerging technologies is involved. Through the multi-disciplinary collaboration of experts, the Bauhaus school promoted a balance between technology development and ideas. In the same way, scientific research often demands the invention of new experimentation and analysis methods in order to realize large-scale projects. During my time at Yale I participated in many such projects, including the development of a silicone elastomer-based microwell system for biochemical analysis of proteins, a contact-printed microarray of the yeast proteome, DNA microarrays representing an entire human chromosome, and the first high-resolution oligonucleotide microarrays to span the human genome. With these and other innovations, our group has been able to carry out experiments that would have been nearly impossible only a few years ago. The people I have worked with shared their insight and expertise from a variety of disciplines, to create something from nothing in the pursuit of knowledge. Science offers the opportunity to draw on the diverse talents of many individuals to unite the concrete and abstract, producing the ideas and machines that will enable us to ask increasingly complex questions about the nature of life.

Paul Bertone | New Haven, CT | March 2005 for my dad



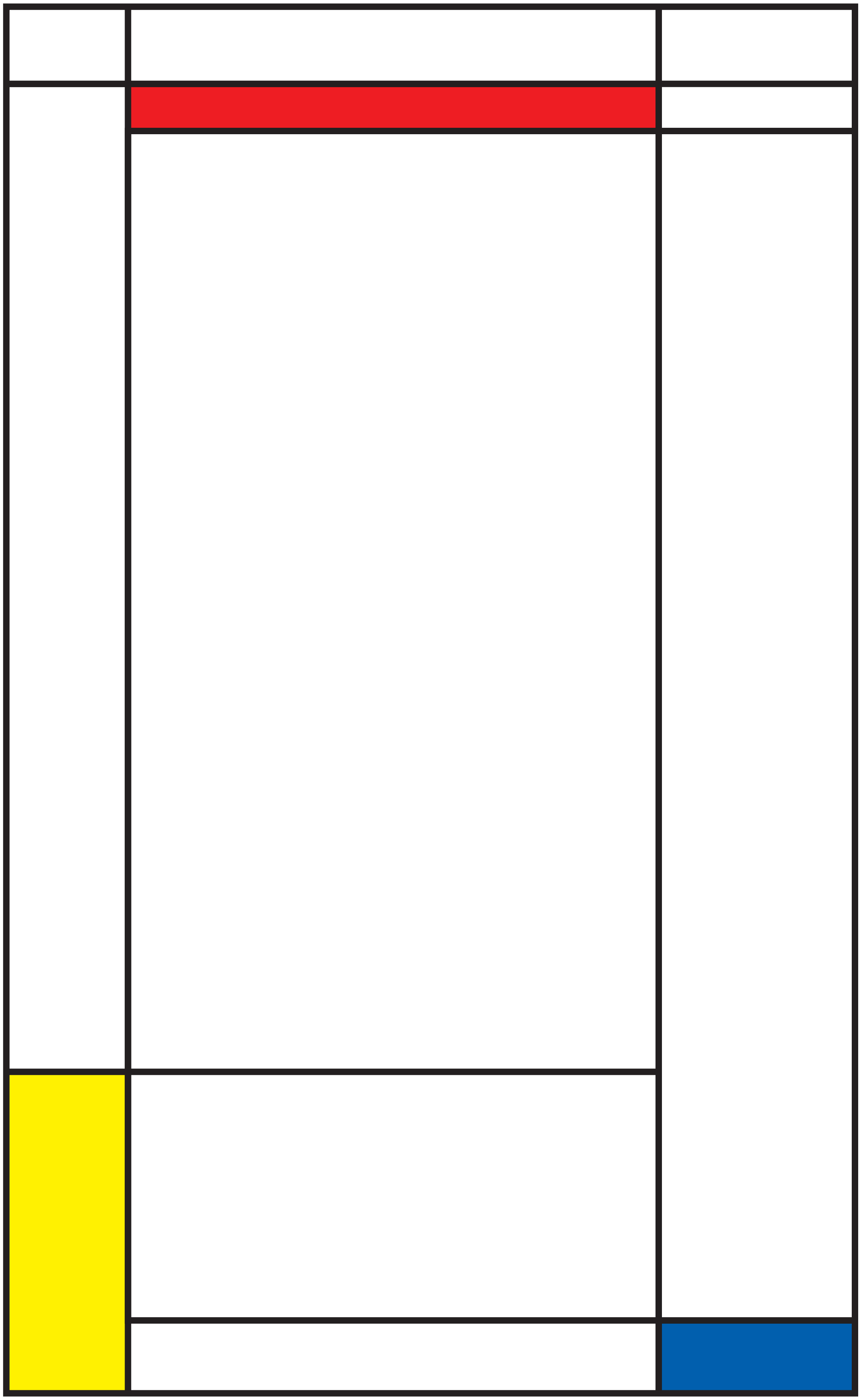
Contents

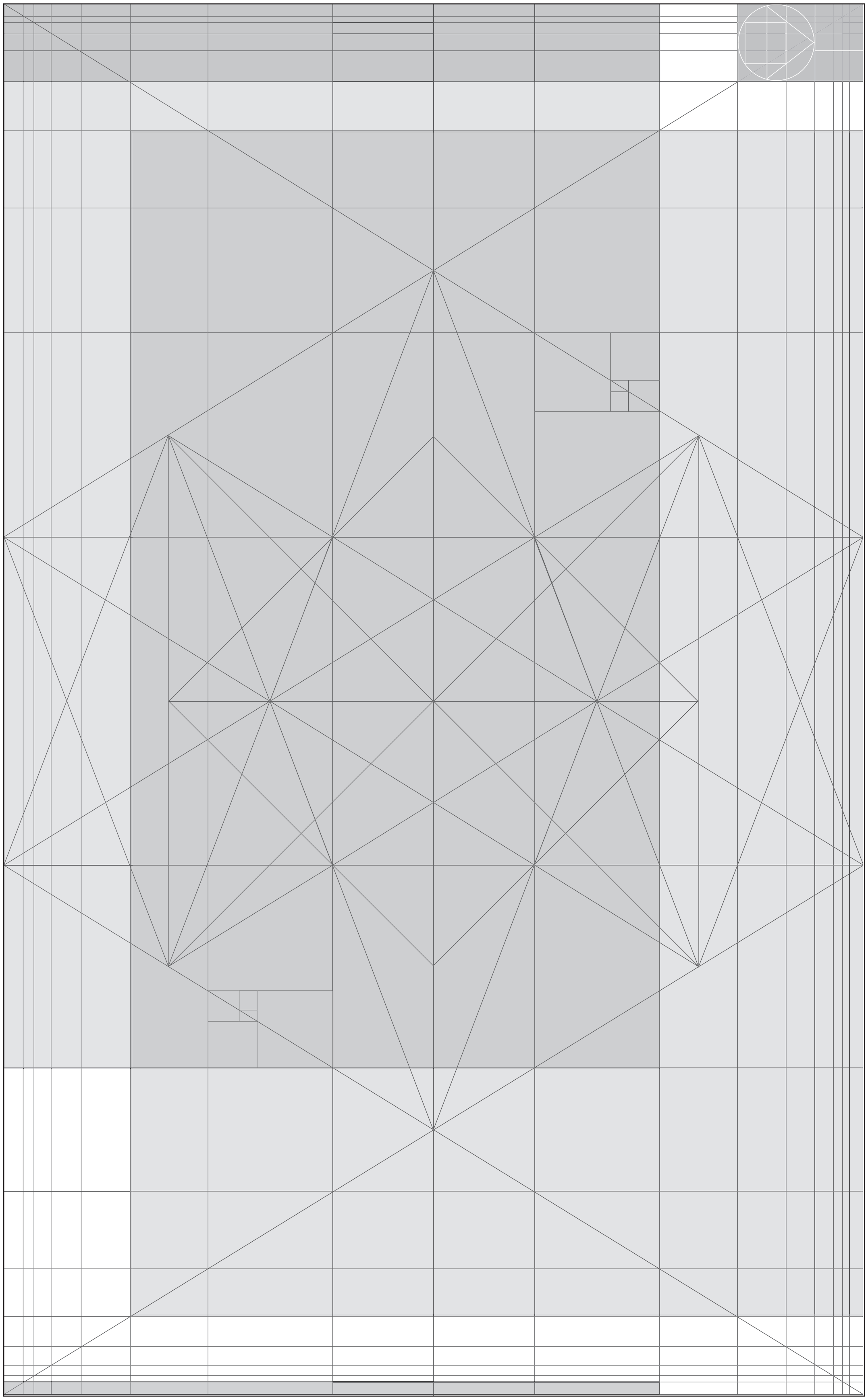
Tables and Illustrations

Nomenclature and Abbreviations

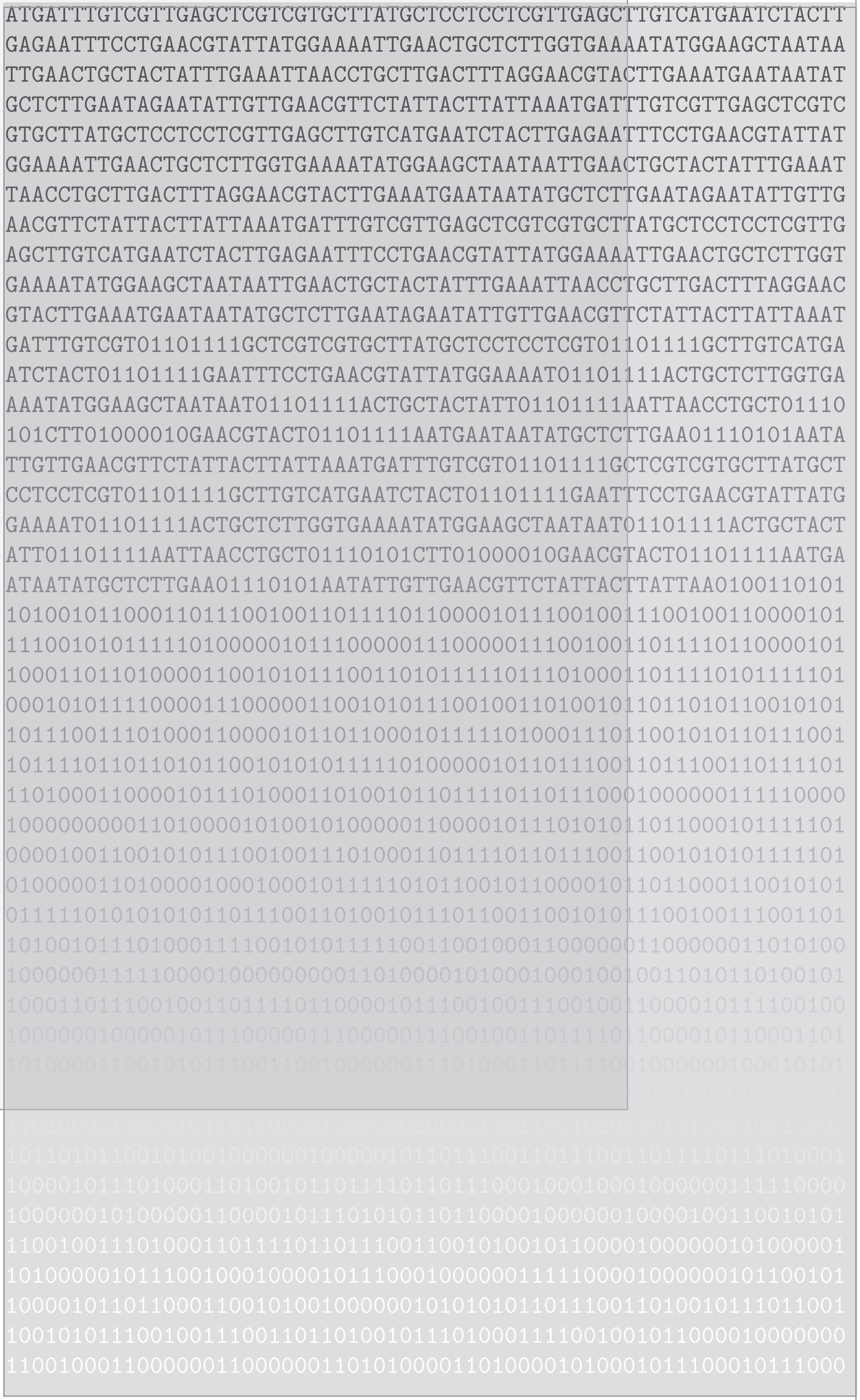
AA-dUTP	5-(3-aminoallyl)-2′deoxyuridine-5′-triphosphate
BLAST	Basic local alignment search tool
bp	Base pair
cDNA	Complementary DNA
ChIP	Chromatin immunoprecipitation
Cot	Concentration (M) × time (s)
CpG	Cytosine-phosphate-guanine
Cy[3,5]	Cyanine fluorochromes
Dam	DNA adenine methyltransferase
DHM	Differential hybridization mapping
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleoside triphosphate
EDTA	Ethylenediaminetetraacetic acid
ENCODE	Encyclopedia of DNA elements
EST	Expressed sequence tag
FBL	Feedback loop
FFL	Feed-forward loop
Gb	Gigabase
hnRNA	Heterogeneous nuclear RNA
kb	Kilobase
M	Molar concentration
MAS	Maskless array synthesis
Mb	Megabase
MES	2-(N-morpholino)-ethanesulphonic acid
MIM	Multiple-input motif
M-MLV	Moloney murine leukemia virus
mol	Molar quantity
mRNA	Messenger RNA
NAT	Natural antisense transcript
NCBI	National Center for Biotechnology Information
NHS	N-hydroxysuccinimide
n-mer	*n*-base polymer

NPPOC	2-nitrophenyl propoxycarbonyl
nrDNA	Non-repetitive DNA
nt	Nucleotide
ORESTES	Open reading frame expressed sequences tags
ORF	Open reading frame
PCR	Polymerase chain reaction
Poly(A)$^+$	Polyadenylation
RACE	Rapid amplification of cDNA ends
RNA	Ribonucleic acid
rpDNA	Repetitive DNA
RT	Reverse transcriptase
RT-PCR	Reverse-transcription PCR
SAGE	Serial analysis of gene expression
SIM	Single-input motif
siRNA	Small intervening RNA
snoRNA	Small nucleolar RNA
SNP	Single-nucleotide polymorphism
SQL	Structured query language
SSC	Sodium chloride sodium citrate
SSPE	Sodium chloride sodium phosphate + EDTA
SVM	Support vector machine
Taq	*Thermus aquaticus* DNA polymerase
TAR	Transcriptionally-active region
UTR	Untranslated region





Microarray Approaches to Experimental Genome Annotation



Introduction: Genome Annotation via Large-scale Transcript Mapping

1

Microarrays have become an important new technology for surveying global patterns in gene expression and regulation. A number of innovative experiments have extended microarray applications beyond the measurement of mRNA expression levels, in order to uncover aspects of large-scale chromosome function and dynamics. This has been made possible due to the recent development of tiling arrays, where all non-repetitive DNA comprising a chromosome or locus is represented at various sequence resolutions. Since tiling arrays are designed to contain the entire DNA sequence without prior consultation of existing gene annotation, they enable the discovery of novel transcribed sequences and regulatory elements through the unbiased interrogation of genomic loci. The implementation of such methods for the analysis of large eukaryotic genomes presents significant technical challenges. Nonetheless, tiling arrays are expected to become instrumental for the global identification and characterization of functional elements in a wide range of organisms. Combined with other experimental and computational methods to relate these data and map the complex interactions of transcriptional regulators, tiling array experiments can provide insight toward a more comprehensive understanding of fundamental molecular and cellular processes.

global measurement of gene expression

It is widely recognized that the availability of a complete genome sequence can significantly enhance our ability to analyze biological phenomena and elucidate molecular and cellular function. Beyond the initial determination of the DNA sequence, the most valuable resource produced by genome mapping efforts entails a comprehensive catalog of functional elements that encompass the genetic repertoire of an organism. Experimental methods for the global analysis of gene expression include subtractive hybridization (Hedrick *et al.* 1984), differential display (Liang & Pardee 1992), and representational difference analysis (Hubank & Schatz 1994). While these techniques are useful for characterizing differences in mRNA transcript populations, they are unable to generate comprehensive gene expression profiles.

The genome-wide identification of transcribed sequences was made possible with the development of the SAGE (serial analysis of gene expression) technique (Velculescu *et al*. 1995). SAGE enables the quantitative estimation of mRNA expression levels by sampling short (10nt–14nt) subsequences of transcribed messages, and using these to deduce the identity of the specific transcripts from which they are derived. The advantages of this approach are twofold: first, it is not necessary to use a unique hybridization probe to detect each individual transcript; second, multiple SAGE tags may be concatenated and sequenced together, providing several measurements simultaneously. A caveat inherent in the SAGE technique is that the use of relatively short sequence tags can result in ambiguous transcript identification. This deficiency can be overcome by using 200nt–600nt expressed sequence tags (ESTs) (Adams *et al*. 1991). Although EST methods predate SAGE technology, they afford a higher degree of specificity and can produce long stretches of transcribed sequence.

DNA microarrays are by far the most widely adopted platform for the high-throughput analysis of gene expression. The advent of cDNA (Schena *et al*. 1995, DeRisi *et al*. 1997), inkjet (Shoemaker *et al*. 2001, Hughes *et al*. 2001) and oligonucleotide (Fodor *et al*. 1993, Pease *et al*. 1994, Lockhart *et al*. 1996, Lipshutz *et al*. 1999) arrays has allowed researchers to simultaneously monitor the expression levels of thousands of genes in a single experiment. The cDNA format consists of mechanically-deposited DNA sequences, amplified via the polymerase chain reaction (PCR) and representing the entire coding sequence of annotated genes. Oligonucleotide arrays (e.g. Affymetrix GeneChips) typically contain one or more complementary oligomer sequences internal to spliced mRNA transcripts, generally positioned near the 3′ end to ensure hybridization to incomplete cDNAs.

While all of these approaches provide the ability to measure genome-wide expression levels of annotated genes, only when a complete corpus of transcribed sequences has been defined can they be exploited to their full potential. Once an organism's complement of transcribed sequences is known, high-throughput analysis methods can be used to comprehensively investigate the dynamics of gene expression over the entire transcriptome.

CHALLENGES IN GENOME ANNOTATION

The early characterization of genes from prokaryotes and model eukaryotes revealed simple gene structures consisting almost entirely of protein-coding sequences. For these organisms, there usually exists a one-to-one relationship between the open reading frames (ORFS) that delineate transcribed sequences and the proteins they encode. In contrast, the genome sequences of higher eukaryotes tell a far different story. Here the predominant gene structures are often fragmented, largely due to the widespread integration of repetitive elements. The transcribed regions of larger, more complex genomes typically embody many short exons interspersed with long intron sequences (Figure 1.1A). The separation of coding sequences into discrete units provides the opportunity for additional genetic variation through the mechanism of alternate splicing. Through selective exon usage, many different protein isoforms may arise from a single gene, greatly amplifying the potential coding complexity of the genome (Figure 1.1B,C). This is particularly true in mammals, where a typical gene may consist of dozens of exons and various combinations of these might be included in spliced messages expressed in different cell types or under different environmental conditions. Thus, a single gene may give rise to a family of protein products that confer a wide range of functional roles. This mechanism is believed to account for the disproportionate increase in organismal complexity in relation to the number of genes it encodes.

Given the fragmented nature of mammalian genes, predicting coding regions from genomic DNA has proven a difficult computational challenge. Some introns may exceed tens of kilobases (kb) in length, making it difficult to aggregate the much shorter coding sequences they divide into plausible gene structures. As a result, many genomes are annotated through homology to characterized protein sequences from evolutionarily-related organisms. However, this approach is inherently biased in that the putative genes identified through sequence similarity must, by definition, be related to genes that are already known. The discovery of unique or highly divergent transcribed sequences is therefore precluded by this approach. Further, the problem of identifying non-coding RNA transcripts is largely neglected by current homology-based prediction methods.

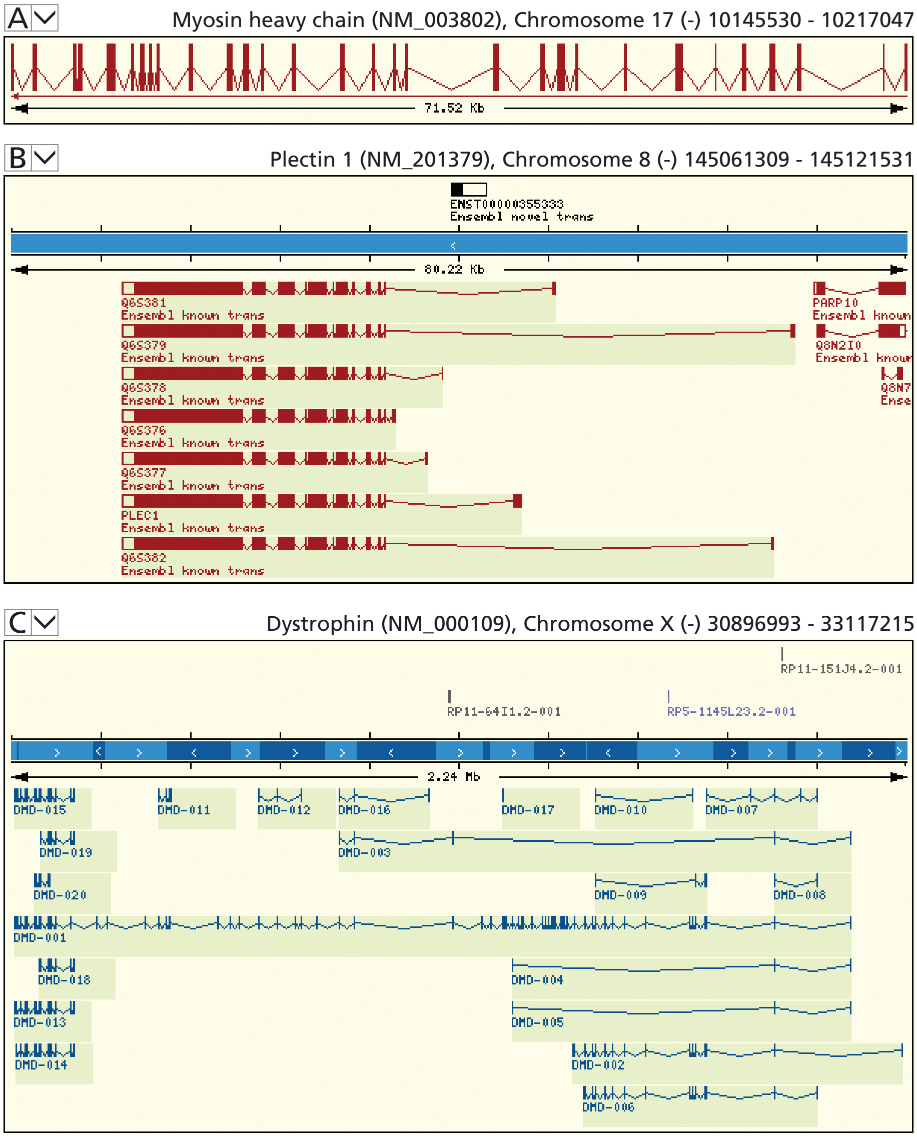


Figure 1.1. Differential exon usage and splice variation in human transcripts. **A**) Complex gene structure is evident the myosin heavy chain. **B**,**C**) Splice variation accounts for much of the complexity observed in higher eukaryotes, where a single gene may encode multiple functionally-distinct isoforms. Seven related splice variants arise from the plectin gene (**B**), while twenty isoforms can be mapped to the dystrophin locus implicated in Duchenne and Becker muscular dystrophy (**C**). In each figure, vertical bars indicate exons (protein-coding segments) separated by intron sequences.

Experimental methods of determining full-length mRNA sequences usually involve the cloning and sequencing of cDNA collections (Adams *et al.* 1991, Strausberg *et al.* 1999, Kawai *et al.* 2001, Ota *et al.* 2004). Once identified, cDNAs can be mapped onto the genome based on sequence similarity to yield a preliminary annotation of expressed gene structures. Although this approach captures a wealth of information about genes transcribed under specific cellular conditions, it often fails to identify rare splice variants or messages expressed in low abundance. Additionally, 5′ ends of genes may be under-represented due to the low fidelity of the viral polymerases used to reverse-transcribe polyadenylated RNA.

Various techniques such as primer extension and 5′ RACE can be used to more precisely map transcriptional start sites; however, these methods are difficult to implement in a high-throughput manner. To address this problem, Marayuma and Sagano (1994) developed a protocol for ligating a primer to the modified 5′ ends of RNA transcripts, thereby providing a template sequence from which to amplify the message for more accurate sequencing. The group went on to generate full-length cDNAs for the entire RefSeq collection (Pruitt *et al.* 2003), revising over one-third of the existing sequences (Suzuki *et al.* 1997).

EMPIRICAL DISCOVERY OF TRANSCRIBED SEQUENCES

The first microarray experiments designed to address the problem of gene annotation were performed with the *E. coli* genome. Selinger *et al.* (2000) developed an oligonucleotide tiling array to represent the genome sequence at 30-bp resolution, using the array for both transcript mapping and differential expression analysis. Nearly all of the annotated sense-strand ORFs were detected as well as 3,000–4,000 antisense ORFs. Even though the genome of *E. coli* is among the best studied, subsequent microarray analysis by Tjaden *et al.* (2002) revealed a 25% increase in the number of transcriptional units detected beyond those previously annotated.

The level of transcriptional activity detected within unannotated regions of genomic DNA appears to increase with the size and complexity of the genome in question. Recently, the entire genome of the flowering plant *Arabidopsis thaliana* was surveyed using oligonucleotide array technology. Yamada *et al*. (2003) developed a series of 12 tiling arrays to characterize transcriptional activity in four complex tissue RNAs, producing the first comprehensive expression map of a eukaryotic genome. Many transcribed sequences were detected within intergenic regions devoid of existing gene annotation, and approximately 30% of antisense transcription was found to be coincident to sense-strand coding regions.

Tiling arrays have also been used for global expression analysis of the fruit fly *Drosophila melanogaster*. Stolc *et al*. (2004) used maskless photolithographic DNA synthesis (Nuwaysir *et al*. 2002, Albert *et al*. 2003) to fabricate oligonucleotide arrays representing all of the predicted exons and exon splice junctions, as well as intergenic and intronic regions throughout the genome. RNA transcript levels were measured at six developmental stages in the organism's life cycle, profiling the expression levels and splice variation of known genes but also revealing the presence of novel transcribed sequences. Comparison with the *Drosophila pseudoobscura* genome indicated that transcriptionally-active sequences within unannotated regions exhibit a greater degree of sequence conservation than those from which transcription was not observed.

The use of tiling arrays for human genome annotation has met considerable technical challenges, mainly due to the large size of mammalian genomes. As part of a study involving inkjet oligonucleotide arrays to survey annotated exon usage in the human transcriptome, Shoemaker *et al*. (2001) developed a tiling approach to accurately map the coding sequence of a novel transcript located within a 113-kb locus of chromosome 22. Although this analysis was carried out on a limited scale, the results clearly illustrated the value of using tiling arrays to delineate transcript boundaries, exon content and splice junctions.

The first tiling array developed to cover the sequence of an entire human chromosome was described by Kapranov *et al*. (2002). In this study, a series of oligonucleotide arrays representing all non-repetitive DNA on chromosomes 21 and 22 was interrogated with cytosolic polyadenylated RNA from 11 cell lines. Surprisingly, a roughly two-fold increase in transcribed DNA was measured over that predicted by existing gene annotation. This finding was reproduced by Rinn *et al*. (2003) using a microarray representing all non-repetitive DNA of chromosome 22 with approximately 21,000 PCR products. Transcriptional activity was

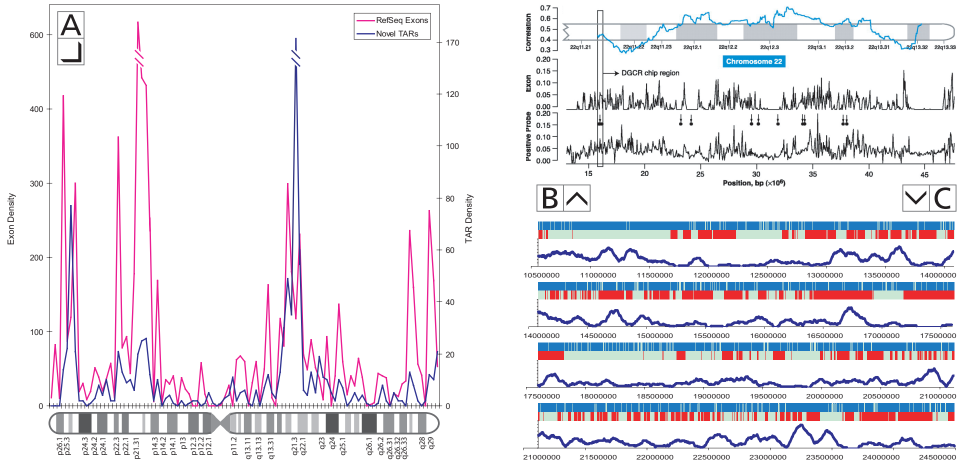


Figure 1.2. Transcriptional profiling of the human genome using three different tiling array platforms. **A**) The distribution of transcriptionally-active regions (TARs) is coincident with annotated exon density on a global scale, as illustrated on human chromosome 3 (Bertone *et al.* 2004). **B**, **C**) Unbiased surveys of human chromosome 22 using oligonucleotide (**B**, Kapranov *et al.* 2002) and PCR-based (**C**, Rinn *et al.* 2003) tiling arrays reveal evidence of RNA transcription originating from previously unannotated regions as well as known genes.

measured across the chromosome in normal placental tissue RNA, followed by strand-specific hybridization of novel transcribed sequences to a contact-printed oligonucleotide array. This work is presented in detail in Chapter 3.

Chapter 4 describes the construction of 134 high-resolution oligonucleotide arrays representing both sense and antisense strands of the entire human genome (Bertone *et al.* 2004). The arrays comprise nearly 52 million 36-nt probe sequences, synthesized via maskless photolithography. Hybridization to polyadenylated liver tissue RNA revealed over 10,000 new transcribed sequences throughout the genome, and verified the transcription of nearly 13,000 predicted genes. A large fraction of novel transcripts exhibited a high degree of similarity to the mouse genome and other mammalian protein sequences, suggesting they may be functional on the basis of evolutionary conservation. Approximately 11% were found to intersect retroprocessed pseudogenic sequences identified in previous studies (Harrison *et al.* 2002, Zhang *et al.* 2003). A small number of these were not determined to be homologous to other annotated genes, decreasing the likelihood of cross-hybridization and indicating that some of the detected pseudogenes may be transcribed. Many other novel transcribed sequences are presumed to correspond to exons retained in rare splice variants, under-represented untranslated regions (UTRs) of annotated genes, protein-coding transcripts expressed in low abundance, and non-coding RNAs. All three human transcript mapping studies identified previously unannotated transcription units located distal to known genes, indicating they originate from distinct messages (Figure 1.2).

The studies described in Chapters 3 and 4 measure differential hybridization of RNA to sense and antisense strands of transcriptionally-active regions and the entire genome, respectively. In both experiments, strand-specific transcription was detected antisense to annotated gene components, notably introns. The initial transcriptome analysis of chromosomes 21 and 22 by Kapranov *et al*. interrogated one strand of genomic DNA with double-stranded cDNA, and therefore could not discern the strand from which transcription originated. However, a subsequent study by Kampa *et al*. (2004) used end-labeled RNAs to obtain strand-specific information, finding 11% of novel transcription to occur antisense to annotated coding sequences and 50% of transcription within intron regions to originate from the antisense strand, consistent with previous studies.

Using a computational approach to select regions where antisense transcription may occur, Yelin *et al*. (2003) conducted a microarray survey of 2,667 sense-antisense sequence pairs to assay for strand-specific transcription. Hybridization to RNA from 19 cell lines and four normal complex tissues confirmed transcription on both strands for 60% (1,600) of the sequences interrogated. A subset of these were confirmed by Northern blot hybridization to strand-specific RNAs, confirming the detection of endogenous natural antisense transcripts (NATS).

Together, these findings reinforce an emerging view of widespread antisense and non-coding RNA transcription throughout the human genome (Cawley *et al*. 2004, Mattick 2003, Mattick 2004, Johnson *et al*. 2005). The repeated identification of novel transcribed sequences by several independent research studies provides compelling evidence of a complex transcriptome encompassing novel protein-coding, regulatory and structural RNAs that have previously eluded detection by conventional genetic approaches. ⊞

Design Optimization Methods for Genomic DNA Tiling Arrays

2

A RECENT DEVELOPMENT IN MICROARRAY DESIGN ENTAILS THE UNBIASED coverage, or tiling, of non-repetitive genomic DNA for the purpose of discovering unannotated transcribed sequences and regulatory elements. Tiling arrays represent intergenic sequences as well as the exons, introns and untranslated regions comprising annotated genes. Although many of the issues in tiling array design are straightforward in principle, the solutions can become computationally intensive when applied to real genomes. This chapter describes the general problem of designing tiling arrays for either oligonucleotide or amplicon (PCR-based) implementations and the issues particular to each platform. In particular, an efficient method for rapidly determining the uniqueness of oligonucleotide sequences is presented, followed by two algorithms for finding an optimal tile path over genomic sequences for the design of amplicon tiling arrays. The first algorithm, a dynamic programming approach, finds an optimal partitioning in linear time and linear space; the second applies a heuristic search to reduce the space complexity to a constant requirement. These methods are applied to several eukaryotic genomes to illustrate the degree to which optimal tiling differs from a trivial partitioning of the sequence. The improvement in non-repetitive sequence coverage is most pronounced in complex mammalian genomes, which exhibit a much higher degree of sequence fragmentation due to increased repeat content. These approaches enable the construction of amplicon arrays that maximize the amount of non-repetitive DNA for the discovery of novel functional elements in eukaryotic genomes.

INTRODUCTION

DNA microarrays have become ubiquitous in genomics research as tools for the large-scale analysis of gene expression. Microarrays allow the simultaneous interrogation of thousands of nucleotide sequences, providing a genome-wide snapshot of transcriptional activity. Until recently, the design of DNA microarrays has focused entirely on annotated genes, constructed either from PCR products comprising entire cDNA sequences (Schena *et al*. 1995), or from short oligonucleotides representing internal regions of spliced mRNA transcripts (Lipshutz *et al*. 1999). Since its introduction, microarray technology has advanced to a degree that currently accommodates enough individual array features to represent all the annotated genes in a mammalian genome.

Commensurate with these technological improvements has been the development of tiling arrays: microarrays that represent a complete non-repetitive tile path over a chromosome or locus, irrespective of any genes that may be annotated in that region (Figure 2.1). This unbiased approach has enabled the discovery of many novel transcribed sequences (Kapranov *et al*. 2002, Rinn *et al*. 2002, Yamada *et al*. 2003, Kampa *et al*. 2004, Bertone *et al*. 2004), as well as the global identification of transcription factor binding sites (Ren *et al*. 2000, Iyer *et al*. 2001, Lee *et al*. 2002, Horak *et al*. 2002b, Martone *et al*. 2003, Cawley *et al*. 2004, Euskirchen *et al*. 2004, Odom *et al*. 2004).

In addition to genes and regulatory elements, eukaryotic genomes contain thousands of repetitive DNA sequences that have been introduced in high copy number over evolutionary time. The frequency and diversity of repetitive elements increases with the size and complexity of higher eukaryotic genomes, accounting for approximately 45% of the total nucleotide content of mammalian genomes (see Appendix A for more information). In selecting sequence fragments for inclusion on a microarray, it is important to exclude as many repetitive elements as possible. The reasons for this are twofold: first, microarray features whose sequences contain repeats present highly redundant hybridization targets; these generate non-specific background signals that mask the fluorescence resulting from specific probe hybridization. Second, the inclusion of repeats would significantly increase the number of DNA sequences assigned to array elements. Many of these would contain superfluous information, thereby wasting a large number of available features on a given array platform.

When tiling genomic DNA with oligonucleotides, near-optimal coverage of the non-repetitive sequence can be achieved in a relatively straightforward manner, although a number of important factors should be considered for probe selection. A more problematic situation arises when selecting non-repetitive sequence fragments intended for pcr. For this application, it is necessary to derive a tile path of larger sequences whose sizes are amenable to high-throughput amplification. Small pcr products can be difficult to resolve in a high-throughput setting, while fragments of several kilobases (kb) in length can restrict the

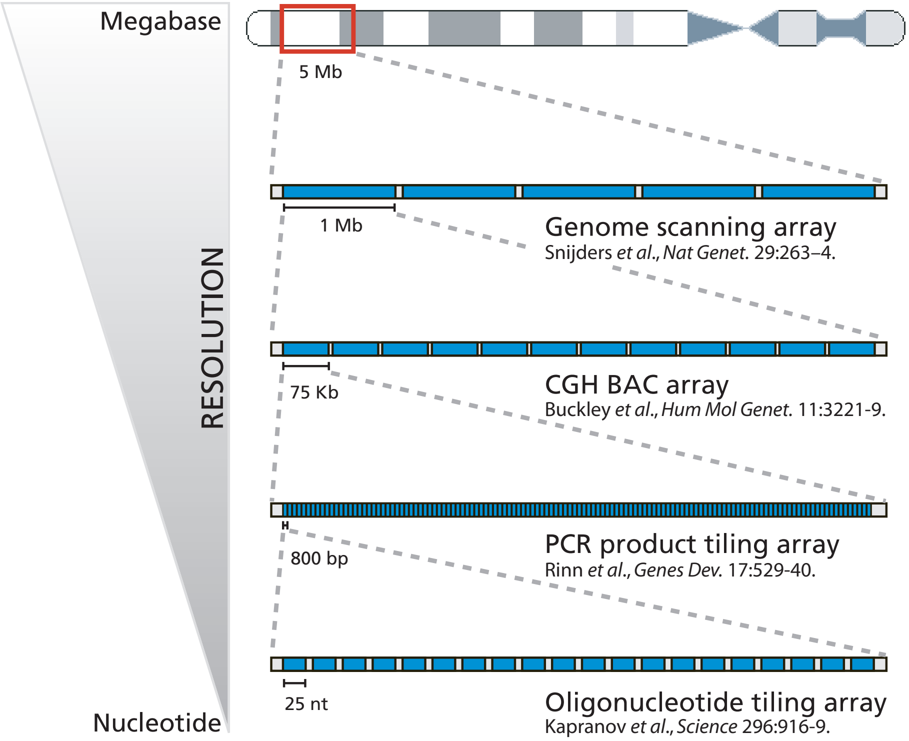

Megabase
5 Mb
RESOLUTION
1 Mb
Genome scanning array
Snijders *et al.*, *Nat Genet.* 29:263–4.
75 Kb
CGH BAC array
Buckley *et al.*, *Hum Mol Genet.* 11:3221-9.
800 bp
PCR product tiling array
Rinn *et al.*, *Genes Dev.* 17:529-40.
25 nt
Oligonucleotide tiling array
Kapranov *et al.*, *Science* 296:916-9.
Nucleotide

Figure 2.1. Evolution of genomic tiling arrays. Representing large spans of genomic DNA with bacterial artificial chromosome (BAC) clones facilitates global experimentation using relatively few array features, at the expense of low tiling resolution. Higher-resolution designs using PCR products or oligonucleotides allow precise mapping of transcripts and regulatory elements, but require labor-intensive or technologically sophisticated approaches to implement.

ability to identify the precise locations of hybridizing sequences by reducing the tiling resolution of the array. Balancing these criteria to select appropriate target sequences presents a unique optimization problem.

TILING DISCONTIGUOUS GENOME SEQUENCES

Genomic tiling arrays are intended to maximally cover a span of non-repetitive DNA with representative sequence fragments, or tiles, whose sizes fall within a prescribed range. The number of repetitive elements included in the tile path should be minimized, while partitioning the sequence into the fewest number of tiles that can maximally cover the non-repetitive DNA. The sequences included on the array are either PCR-amplified and deposited mechanically onto glass slides via contact printing (Schena *et al.* 1995), or partitioned further and represented as oligonucleotides that may be printed mechanically or synthesized in situ using photolithographic (Lipshutz *et al.* 1999, Nuwaysir *et al.* 2002) or piezoelectric (Hughes *et al.* 2001) technologies (Table 2.1). Since the number of available features on a given microarray platform is often dependent on development costs, an optimal tiling solution should comprise the fewest number of non-repetitive sequence fragments whose lengths approach a pre-determined upper bound.

Repeat identification and low-complexity filtering. For the purpose of designing microarrays, it is necessary to locate repetitive elements in genomic DNA with local sequence alignment methods (Smith & Waterman 1981, Altschul *et al.* 1990). This is most easily accomplished through the use of publicly available software such as RepeatMasker (Smit & Green, unpublished), CENSOR

Table 2.1. Comparison of DNA microarray formats. Although contact-printed arrays allow for unlimited customization, initial production costs can be prohibitive compared to the relatively affordable but fixed array designs produced commercially. The recent development of maskless photolithographic synthesis technology represents a trade-off between these platforms, allowing customized design while maintaining very high feature density.

	Contact printing	Inkjet synthesis	Affymetrix	NimbleGen
Arraying method	Mechanical deposition	Phosphoramitide in situ DNA synthesis, piezoelectric printing	3′→5′ phosphoramitide in situ DNA synthesis, chromium masked photolithography	5′→3′ phosphoramitide in situ DNA synthesis, laser-based maskless photolithography
DNA size limit	None	~60nt	25nt	~100nt
Feature type	Various	Oligonucleotides	Oligonucleotides	Oligonucleotides
Features/ slide	≤ 20K	Typically < 5K; 25,575 for Agilent synthesizers	60K – 1.6M commercial; 6.2M in development	200K–800K
Array design	Flexible	Flexible	Fixed	Flexible
Fabrication cost	High (DNA preparation)	Moderate	High	Low
Array cost	Low	Moderate	Moderate	Moderate

(Jurka *et al*. 1996), Tandem Repeats Finder (Benson 1999) and RECON (Bao & Eddy 2002). Of these, RepeatMasker is widely used and is capable of identifying repeats in a variety of genomes using a database of characterized repetitive elements (Jurka 2000).

In addition to identifying instances of canonical repeats, it is often desirable to screen genomic DNA for low-complexity sequences: stretches of polypurine/polypyrimidine bases, or regions of extremely high A/T or G/C content. RepeatMasker is able to filter some low-complexity DNA by default; more extensive filtering is often performed using programs such as DUST (Tatusov & Lipman, unpublished) and NSEG (Wooton & Federhen 1993). DUST is included as a component of the NCBI BLAST distribution; NSEG is a member of the SEG family of programs and affords more flexible control over low-complexity filtering by using an information entropy-based model of sequence analysis.

tiling sequences with oligonucleotides

Designing oligonucleotide tiling arrays constitutes a different problem than selecting oligonucleotides for gene-based arrays, primarily because end-to-end or overlapping tile layouts present fewer options with regard to sequence selection. However, several strategies can be used to improve both the annealing specificity and thermodynamic properties of oligonucleotides selected for tiling arrays. A number of factors should be considered when tiling genomic DNA with oligonucleotides, including tiling resolution, uniqueness of oligonucleotide sequences, and hybridization affinity.

Tiling resolution. An important factor in microarray design entails determining how the remaining non-repetitive DNA should be subdivided and how densely it should be represented by oligonucleotide probes. The serial placement of oligonucleotides along non-repetitive genomic DNA can either be contiguous, covering all of the available sequence, or discontiguous, where gaps of a predetermined size range are allowed between adjacent probes (Figure 2.2). This determination should be made according to the type of experiment for which the microarray is intended, and what kind of biological information the

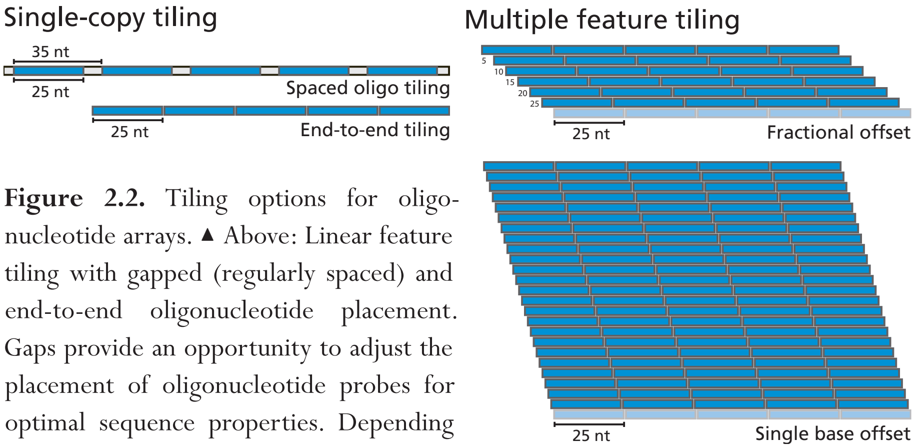


Figure 2.2. Tiling options for oligonucleotide arrays. ▲ Above: Linear feature tiling with gapped (regularly spaced) and end-to-end oligonucleotide placement. Gaps provide an opportunity to adjust the placement of oligonucleotide probes for optimal sequence properties. Depending on the experimental design, it is seldom necessary to interrogate every nucleotide of the genome sequence to obtain the desired measurement. ► Right: Overlapping tiles placed with a fractional offset (e.g., one 25mer probe placed every 5nt) and single-base offset. The latter strategy provides a finer-resolution tiling of the genomic sequence and can give a more precise indication of where hybridization is occurring along the chromosome.

array is capable of measuring given a particular experimental sample. In the case of ChIP-chip experiments, chromatin-immunoprecipitated DNA is hybridized to an intergenic microarray to locate transcription factor binding sites (Horak & Snyder 2002, Cawley *et al.* 2004). The immunoprecipitated DNA is sonicated prior to hybridization to shear the molecules into smaller fragments; even so, fragments smaller than approximately 500bp will be largely unaffected by sonication. Since the sample DNA comprises a population of molecules whose sizes will generally exceed 500bp, it is reasonable to represent the genomic sequence with oligonucleotide probes spaced under 500bp apart. Although closer probe spacing will yield more precise hybridization data, larger gaps are still appropriate for ChIP-chip experiments because this layout will ensure adequate hybridization to the sample DNA.

For the fine-resolution mapping of transcribed sequences, much closer probe spacing is required. Because many exons span only tens of nucleotides, most of these would elude detection if the genomic sequence is tiled with large gaps. Further, if the experiment is intended to measure exon-intron boundaries, it may be desirable to cover the genomic DNA with multiple oligonucleotides such that the starting position of each probe is shifted by several nucleotides in order to overlap the previous oligonucleotide's coordinates (Figure 2.2, right). Although this strategy increases the tiling resolution, the number of probes required will eventually occupy many more features on the array. It is therefore important to select the desired tiling resolution in a manner that considers the intended microarray platform and optimizes the use of the available array elements.

Oligonucleotide probes selected for microarray applications are typically short (25nt–80nt) and uniform in length. These assumptions allow the non-repetitive regions to be tiled by adopting a straightforward approach in which the sequences are subdivided into fixed-size partitions. There will naturally be many cases where the oligomer length does not divide evenly into the size of a non-repetitive sequence fragment and the remainder is therefore omitted from the tile path. However, the resulting loss in sequence coverage is inconsequential given the typically short length of the oligonucleotides.

Uniqueness of oligonucleotide sequences. When developing gene-based microarrays, one or more oligonucleotides are typically selected to represent each gene. These are designed to be highly specific to the target gene, to anneal within a suitable affinity range, to occur within annotated exons so that they will hybridize to the mature spliced transcript, and are typically positioned proximal to the 3′ end of the gene to increase the likelihood of detecting partially reverse-transcribed messages. Subdividing contiguous genomic DNA in a naïve, end-to-end fashion offers little opportunity to select optimal probe sequences because the aim is to cover the non-repetitive regions using predetermined spacing constraints. However, the non-repetitive regions will inevitably include many contiguous sequences that are not equally divisible by the length of the oligonucleotide probes. In these situations, it is desirable to adjust the placement of oligonucleotides in order to bias the sequence selection toward the optimal criteria, thereby reducing the potential for cross-hybridization to sequences elsewhere in the genome.

To implement this approach the degree of uniqueness is computed for any given oligonucleotide sequence in a large region (i.e., an entire chromosome).

This problem can be stated as follows: given a large genomic sequence and an oligomer of length n, find all oligomers in the sequence differing from the input in no more than m places. In theory, we need only create a direct hash of each sequence to a list of all subsequence occurrences. However, the space required to implement the hash quickly becomes impractical. With $4n$ possible oligomer sequences, we find that a hash of size 14 requires 1 gigabyte of storage, in addition to the space needed to store each of the possible index coordinates of the input sequence – another gigabyte for large chromosomes. In practice, these requirements impose a limitation on the size of hash tables such that $n \leq 14$. This is insufficient for most microarray applications where oligonucleotide sizes are typically $\geq$ 25nt.

To work around these memory constraints and deal with possible mismatches, a blast-like scheme similar to the approach described in Wang and Seed (2003) is adopted. A hash table is first created based on oligomers of size $k < n$. When considering a given oligonucleotide sequence, we look up each of the oligonucleotide's $n-k$ substrings of length k, extending each hit to full length as dictated by the substring's position in the oligonucleotide and comparing it to the input sequence. In doing so we can also allow for mismatches, knowing that we will detect all oligonucleotides with no more than two mismatches to the input so long as $k \geq (n-m) / (m+1)$. Given a random model of a chromosome of length c, a substring of length k will have an expected $c / 4k$ matches, each of which can be processed in constant time. In such a model the algorithm runs in an expected time of $O((n-k) / 4k)$.

Thermodynamic properties of oligonucleotide probes. A third factor concerns the selection of oligonucleotide sequences for tiling arrays based on their predicted hybridization affinities (SantaLucia 1998). When representing individual genes with one or more probes, careful consideration is made to select sequences unique to each gene, having thermodynamic characteristics that are optimal for hybridization. For sequences longer than 13nt, hybridization affinity can be approximated by calculating the melting temperature (Tm) of each oligomer using the following standard formula:

$$\mathrm{Tm} = 64.9 + 41(n\mathrm{G} + n\mathrm{C} - 16.4) \;/\; (n\mathrm{A} + n\mathrm{T} + n\mathrm{G} + n\mathrm{C})$$

where n[A,C,G,T] indicates the number of instances of each nucleotide present in the DNA sequence. For more precise calculations, a base-stacking approach can be used that takes the exact sequence into account rather than the overall nucleotide composition:

$$\mathrm{Tm} = \left[\Delta H(\mathrm{kcal} / {}^{\circ}\mathrm{C}{*}\mathrm{Mol}) / \Delta S + R \ln([\mathtt{oligo}] / 2)\right] - 273.15^{\circ}\mathrm{C}$$

where ΔH is the enthalpy of base stacking interactions, ΔS is the entropy of base stacking, [oligo] indicates the oligonucleotide concentration, and R is the universal gas constant 1.987 Cal / °C*Mol (Rychlik & Rhoads 1989).

Considering these criteria, it is useful to shift the placement of oligonucleotides within each region of non-repetitive DNA in order to reduce the variability of the melting temperatures associated with each probe sequence. In the case of spaced oligo tiling an individual probe is selected from within each available region such that the calculated Tm is closest to the optimal temperature. For overlapping tiling designs either the entire set of oligos can be shifted together such that their aggregate Tm is optimized, or the previous approach can be taken and the available regions for oligo placement simply overlap with adjacent regions instead of considering gaps between them.

TILING APPROACHES FOR AMPLICON MICROARRAYS

Designing PCR-based microarrays involves a number of challenges beyond oligonucleotide selection. Each of the non-repetitive sequence fragments must necessarily be larger in size to facilitate high-throughput amplification. Typically, the size distribution of sequences amenable to both PCR amplification and microarray analysis falls between 300bp and 1.5kb. Although it is certainly feasible to amplify sequence fragments far exceeding this upper limit, it becomes difficult to determine the locations of hybridizing sequences within larger fragments. Conversely, amplifying thousands of small sequence fragments complicates the production of large-scale projects. This problem is examined in detail and several methods are introduced to derive an optimal tile path that maximizes the coverage of non-repetitive DNA while minimizing the number of repetitive elements included in the resulting sequences.

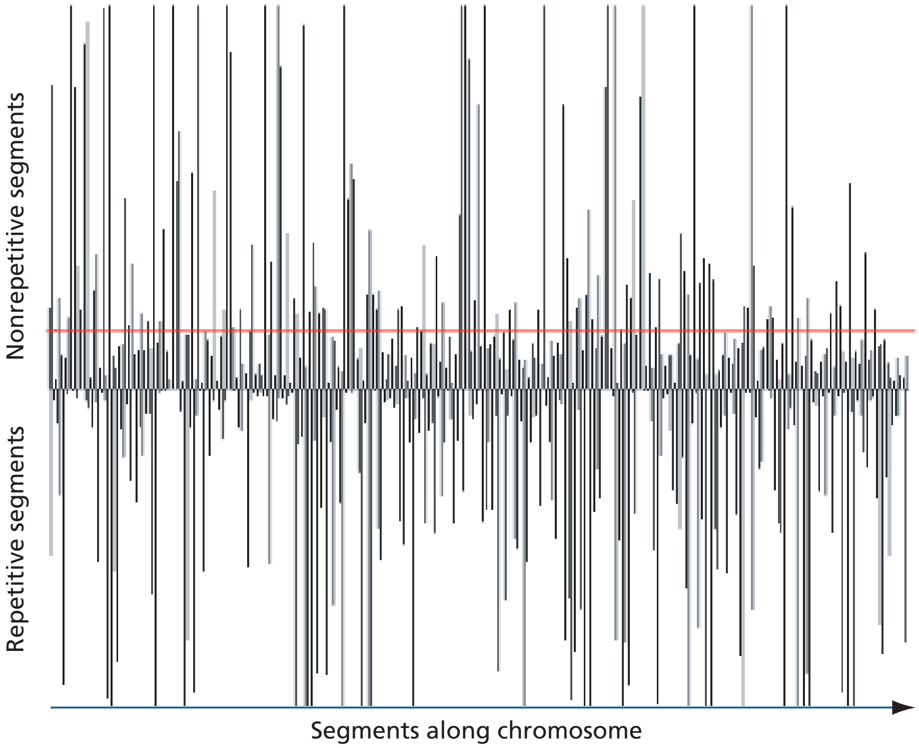


Figure 2.3. Repeat-masked region of human chromosome 10 showing alternating repetitive and non-repetitive contiguous segments (plotted vertically). The sizes of these subsequences are reflected in the length of the vertical bars. The high level of fragmentation is clear, as is the wide range of sizes in both repetitive and non-repetitive sequences. The red bar indicates a size of 300bp; a large number of non-repetitive sequences below this threshold are omitted when using naïve tiling methods that simply avoid repeats.

The repetitive elements present in eukaryotic genomes introduce a high degree of fragmentation of the non-repetitive DNA. Avoiding repeats and targeting only the remaining sequence fragments 300bp and larger results in suboptimal coverage of the non-repetitive DNA (Figure 2.3). In order to improve the sequence coverage, a scheme must be devised to recover some of the non-repetitive fragments that are too small to be efficiently amplified. This can be accomplished by strategically incorporating short repeat elements that lie between these non-repetitive sequences, effectively joining the adjacent fragments into larger contiguous tiles (Figure 2.4).

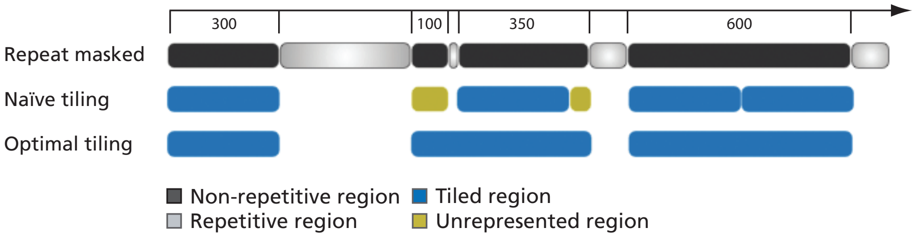


Figure 2.4. Graphical representation of repetitive and non-repetitive segments in repeat-masked DNA. In the naïve tiling case, the sequence is subdivided into segments of equal length; consequently, many small non-repetitive regions are lost as indicated in yellow. These can be recovered by using partitioning methods that generate an optimal tile path over the sequence, as illustrated in the bottom example.

With regard to sequence tiling, a repeat-masked genome sequence can be viewed as containing two categories of nucleotide information: 1) coding, regulatory and intergenic sequences located in euchromatic regions, together viewed as non-repetitive DNA (nrDNA), and 2) nucleotides belonging to repetitive elements and low-complexity regions (rpDNA). Tiling of repeat-masked sequences can therefore be viewed as a two-class partitioning problem: Given a sequence with some subwords marked as repeat nucleotides and the remaining subwords composed of non-repetitive nucleotides, the sequence is partitioned into non-overlapping tiles of either type such that the total amount of non-repetitive sequence covered by tiles is maximized, while the number of repetitive nucleotides included in the resulting tile path is minimized.

OPTIMAL SEQUENCE TILING ALGORITHMS

Scoring potential tile paths. Given a sequence of nucleotides $S_{1..n}$, we would like to find an optimal tile path (possibly not unique) comprising a set of non-overlapping tiles, potentially separated by excluded regions, that maximizes a scoring function V over all possible tile paths, given by

Eq. 2.1
$$V\left[TilePath\,\{S_{1..n}\}\right] = \sum_{i=1}^{n} w_i - mC$$

where w_i is the weight associated with the ith nucleotide, m is the number of tiles and C is the cost for opening a tile (in this way, fewer longer tiles are favored over the creation of many smaller ones). For a given tile path each nucleotide in the sequence is either in a tile (which have weights w_{nr}^{T} and w_{rp}^{T} for non-repetitive and repetitive nucleotides, respectively) or in an excluded region (which have weights w_{nr}^{X} and w_{rp}^{X} for non-repetitive and repetitive nucleotides, respectively). Thus

Eq. 2.2
$$w_i^{T \text{ or } X} = \begin{cases} w_{nr}^{T \text{ or } X} & \text{if non-repetitive} \\ w_{rp}^{T \text{ or } X} & \text{if repetitive} \end{cases}$$

We can also use the scoring function V to evaluate the score of either an individual tile $T_{i..j}$ or an excluded region $X_{i..j}$,

Eq. 2.3
$$V\left[T_{i..j}\right] = \sum_{k=i}^{j} w_k^{T} - C\ , \quad V\left[X_{i..j}\right] = \sum_{k=i}^{j} w_k^{X}$$

Therefore the scoring function evaluated over an entire tile path is the sum of all scores for individual tiles and excluded regions,

Eq. 2.4
$$V\left[TilePath\,\{S_{1..n}\}\right] = \sum_{\{T_a\}} V\,[T_a] + \sum_{\{X_a\}} V\,[X_a]$$

where $\{T_a\}$ is a set of all tiles in the tile path and $\{X_a\}$ is analogously defined. A brute force algorithm would enumerate all tile paths to find an optimal solution; however, this approach would take exponential time to compute. We impose an additional constraint, that tiles are restricted to lengths between a lower bound l and an upper bound u. Given this constraint, the algorithm we present here solves the problem in linear time.

A dynamic programming solution. Dynamic programming solutions have been applied many times in bioinformatics. Examples include sequence alignment (Needleman & Wunsch 1970, Smith & Waterman 1981, Gotoh 1982), gene prediction (Gelfand & Roytberg 1993, Snyder & Stormo 1993) and RNA secondary structure prediction (Zuker & Sankoff 1984). The key idea behind dynamic programming is the reuse of intermediate results. This is usually accomplished by breaking down an exponential search space into subparts, which are evaluated and whose results are tabulated for reuse. The analysis of large search spaces can then be done in polynomial time.

The main iteration of the algorithm can be described as follows: at an intermediate step in the computation we have evaluated the optimal tile paths and their associated scores for all subsequences $S_{1..1}$ to $S_{1..(k-1)}$. In order to find an optimal tile path for the subsequence $S_{1..k}$, for each $i \in [\max(1,\ k-u),\ \max(1,\ k-l)]$ we compute the score for the tile path consisting of the optimal tile path from $1..i$ and the tile $T_{(i+1)..k}$ using the score of the optimal tile path from $1..i$ and $V[T_{(i+1)..k}]$. Similarly, we also evaluate the score of the tile path consisting of the optimal solution from $1..(k-1)$ and the excluded region $X_{k..k}$ (the kth nucleotide). The optimal tile path for $S_{1..k}$ is then one of the preceding tile paths having the maximal score. This tile path and its associated score are then stored and the algorithm proceeds to the next nucleotide in the sequence, $k+1$. A schematic of the algorithm appears below.

Given optimal tiles paths for all subsequences $S_{1..1}$ to $S_{1..(k-1)}$ and associated scores $V\left[OptimalTilePath\,\{S_{1..1}\}\right]$ to $V\left[OptimalTilePath\,\{S_{1..(k-1)}\}\right]$:

STEP 1: For each $i \in [\max(1,\ k-u),\ \max(1,\ k-1)]$ we construct the following tile path:

$$TilePath\,\{S_{1..k}\} = OptimalTilePath\,\{S_{1..i}\} \cup T_{(i+1)..k}$$

and compute its score:

$$V\left[TilePath\,\{S_{1..k}\}\right] = V\left[OptimalTilePath\,\{S_{1..i}\}\right] + V\left[T_{(i+1)..k}\right]$$

We also construct an additional tile path

$$TilePath\,\{S_{1..k}\} = OptimalTilePath\,\{S_{1..(k-1)}\} \cup X_{k..k}$$

and compute its score:

$$V\left[TilePath\,\{S_{1..k}\}\right] = V\left[OptimalTilePath\,\{S_{1..(k-1)}\}\right] + V\,[X_{k..k}]$$

STEP 2: From the preceding tile paths computed in Step 1, we select one having the maximal score and store it as $OptimalTilePath\,\{S_{1..k}\}$, along with its associated score.

STEP 3: Repeat for subsequence $S_{1..(k+1)}$.

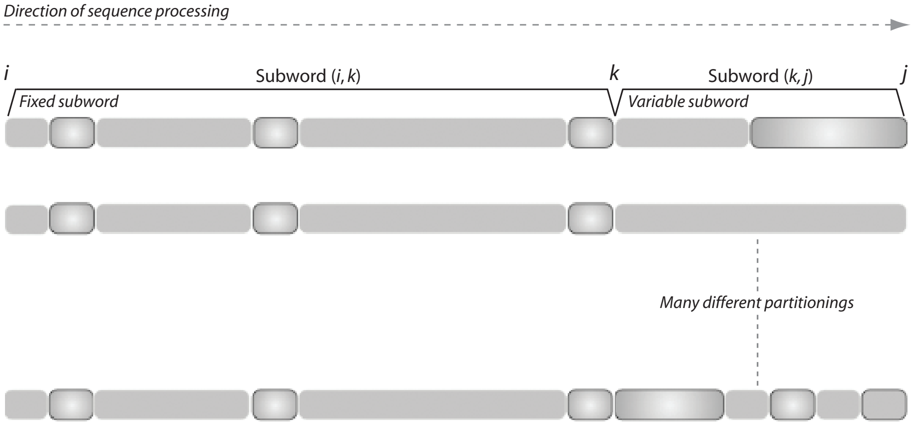

Direction of sequence processing
i
Subword (i, k)
k
Subword (k, j)
j
Fixed subword
Variable subword
Many different partitionings

Figure 2.5. Many different partitionings share common subparts. To compute any partitioning with a split at k, the best partitioning for (i, k) and for (k, j) must be known. Since there are many ways to partition the sequence with a split at k, we only need to recursively evaluate a subpartitioning for subword (i, j) and (k, j) once. In all cases where we need the optimal solution for these subwords again, we refer to the pre-computed result instead of considering all further possible partitionings of that subword.

The algebraic dynamic programming (adp) framework (Giegerich *et al.* 2000) was used to recursively construct all possible partitionings and apply the scoring scheme to each solution. Since many partitionings share common subpartitionings, we can tabulate their scores for reuse instead of recomputing them (Figure 2.5). Without the tile length constraints, the time and space complexity of this approach would be $O(n^2)$, which is inherent in the adp framework implementation. Given these constraints, the algorithm runs in linear time and space, specifically $O((u-l)n)$.

A linear time, constant space solution. The dynamic programming algorithm computes an optimal tiling solution over the target sequence. In practice, however, the time and space required to process real genomic DNA sequences preclude the use of this approach for large eukaryotic chromosomes (spanning up to ~250mb). Here we present an alternative method which traverses the sequence in a single pass, placing tiles according to local constraints instead of considering every possible tiling solution. In contrast to the dynamic programming algorithm, the result of this approach partitions the sequence into alternating included regions $I_{i..j}$ and excluded regions $X_{i..j}$. A post-processing step is then required to subdivide the included regions into individual tiles $T_{i..j}$ satisfying the length constraints.

The scores for included and excluded regions are given by

Eq. 2.5
$$V\left[I_{i..j}\right] = \sum_{k=1}^{j} w_k^I \ , \quad V\left[X_{i..j}\right] = \sum_{k=1}^{j} w_k^X$$

where the weights corresponding to included regions are the same as those for the tiles in the dynamic programming algorithm ($w_k^I = w_k^T$). Note that the score for included regions does not account for the tile cost C.

The algorithm partitions the sequence and outputs the region boundaries as processing continues. The sequence is scanned one nucleotide at a time, with the current position denoted by i. During the main iteration we keep track of an earlier position k, up to which an optimal partitioning has been determined. At each step, the algorithm attempts to determine if the window $S_{(k+1)..i}$ should be classified either as an extension of the last known region R (currently extending up to k), or as the prefix of a new region starting at $k+1$. Depending on the type of region R (included or excluded) and the difference $D = V[I_{(k+1)..i}] - V[X_{(k+1)..i}]$ between the values of the scoring function for the two potential classifications of the window $S_{(k+1)..i}$, the algorithm selects one of three possible options:

1) If R is an included region and D is positive, or if R is an excluded region and D is negative, then R is extended to include the nucleotides up to i (i.e., k is set to i);
2) If R is an included region and $D < -C$, or if R is an excluded region and $D >$ C, then R is terminated at k and a new region of the opposite type is initialized at $k+1$ and extended to position i;
3) Otherwise, neither action is taken.

Following this decision, the next nucleotide in the sequence is processed (i.e., i is incremented). The classification of the first and the last regions in the sequence is determined similarly, effectively assuming that the start of the sequence follows an excluded region, and only inspecting the sign of D if R is an included region at the end of the sequence (i.e., when $i = n-1$).

Since the number of times each nucleotide is examined is bounded by a constant, the overall time complexity is linear with respect to the size of the input sequence. The algorithm runs in constant space, as we need only keep a running value of D, the values of i and k, and the type of region R. A proof of optimality for this algorithm is presented in Appendix B.

This algorithm imposes no implicit upper bound on the size of nrDNA partitions, although C is effectively a lower bound on tile sizes. Therefore, included regions must be subdivided into smaller tiles whose sizes reflect the desired upper limit for PCR products. In terms of experimental preparation and subsequent microarray data analysis, it is preferable to create roughly equal-sized fragments whenever possible. Therefore the most straightforward tiling of long nrDNA partitions involves 1) taking the ceiling of the length of the partition divided by the maximum tile size, then 2) subdividing the partition into equal-sized fragments of this number.

Tiling statistics for eukaryotic genomes. A summary of tiling genomic DNA sequences of various sizes and repeat densities is presented in Table 2.2. Several model organisms were included in this analysis; some genomes have relatively few repeats, whereas the genomes of more recently sequenced rodents and primates contain large numbers of repetitive elements that constitute high percentages of nucleotide content relative to the size of each genome. The sequences were first tiled using a naïve approach, where the non-repetitive DNA was subdivided into tiles having lengths equal to the lower size bound (in this case 300bp). The linear time, constant-space tiling algorithm outlined above was then applied to the sequences to derive an optimal tile path for each.

In comparing these results, a number of observations become apparent. When the sequences are tiled in a naïve fashion, the coverage of non-repetitive DNA decreases dramatically as the target sequences progress from the relatively repeat-free *Arabidopsis* sequence to the larger mammalian genomes. This reflects the higher levels of genomic sequence fragmentation due to increased repeat content, a condition that clearly inhibits the optimal tiling of the sequence. The application of the tiling algorithm to the more complex eukaryotic genomes improves the non-repetitive sequence coverage significantly in the optimized tile path, while the percentage of included repeats remains very low.

SUMMARY

Tiling microarrays are becoming an important platform for empirical genome annotation, making available the maximum amount of non-repetitive DNA for interrogation. Numerous options exist for tiling genomic sequences with oligonucleotides, leading to microarray designs of various sequence resolutions and feature densities. Representing the non-repetitive DNA with spaced, serial, or overlapping oligonucleotides is straightforward in principle, but a number of probe selection optimizations can be made to address cross-hybridization issues. In particular, biasing the selection of oligonucleotides toward uniform thermal properties and uniqueness across the genome can improve the annealing characteristics and hybridization specificity of the probes.

Table 2.2. Comparison of optimal tiling and naïve tiling of various sequenced genomes for amplicon microarrays. The genome sequences vary in the degree of repeat density, ranging from mammalian genomes with nearly 50% repeat content to the relatively repeat-free *Arabidopsis* genome. Obtaining a high degree of non-repetitive sequence coverage for the genomes on the latter end of the spectrum is straightforward. However, as more complex genomes are considered it becomes impossible to optimally tile the highly repetitive sequences without further processing.

Organism	Genome size	Percent repeats	Linear sequence tiling			Naïve tiling	Comparison
			Percent non-repeat bp covered	Percent repeat bp included vs. all non-repeat bp	Efficiency	Efficiency	Percent improvement
Pan troglodytes	3,083,993,401	57.74	89.81	4.23	85.58	66.05	19.53
Homo sapiens	3,070,537,687	52.38	89.60	4.06	85.53	66.07	19.47
Mus musculus	2,638,213,512	45.62	91.09	5.51	85.58	66.18	19.41
Rattus norvegicus	2,795,745,218	48.75	91.43	5.54	85.89	66.86	19.03
Caenorhabditis elegans	100,277,879	11.26	98.54	3.10	95.44	84.29	11.16
Drosophila melanogaster	129,323,838	14.23	99.40	2.62	96.78	86.89	9.89
Fugu rubripes	349,519,338	15.06	99.07	2.13	96.94	87.97	8.97
Arabidopsis thaliana	119,186,497	0.16	100.00	0.00	100.00	99.97	0.02

A Transcriptional Survey of Human Chromosome 22

3

A microarray representing all of the non-repetitive sequence of human chromosome 22 was constructed and used to measure global transcriptional activity in polyadenylated [poly(A)$^+$] placental RNA. Hybridization to fluorescence-labeled complementary DNA (cDNA) indicated that many of the known, related and predicted genes annotated across the chromosome are expressed. More importantly, this survey revealed twice as many transcribed bases as have been reported previously. Many of the newly discovered expressed fragments were verified by RNA blot analysis and a novel technique called differential hybridization mapping (DHM). Interestingly, a significant fraction of these novel fragments are expressed antisense to previously annotated introns. The coding potential of these novel expressed regions is supported by their sequence conservation in the mouse genome.

INTRODUCTION

As the sequencing phase of the human genome project nears completion, increasingly complete and accurate nucleotide-level data are becoming available (Lander *et al.* 2001, Venter *et al.* 2001). The next major challenge is to decipher the biological information encoded by the billions of ordered nucleotides. This goal requires identifying the various genes and proteins encoded in the DNA as well as how they function, how they are regulated, and how they work together to carry out complex biological processes. An essential step toward understanding the coding information of the human genome is to obtain a detailed knowledge of human transcriptional coding sequences on a genomic scale.

Current approaches for mapping mRNA-coding regions on a genomic scale have used a variety of techniques such as serial analysis of gene expression (SAGE), sequencing of expressed sequence tags (ESTs), STS mapping, radiation hybrid mapping, and full-length cDNA analysis (Saccone *et al.* 1996, Deloukas *et al.* 1998, Dunham *et al.* 1999, Caron *et al.* 2001). However, these techniques do not comprehensively interrogate all of the genomic coding information. Furthermore, these methods are not versatile for probing many tissue types and conditions, and consequently may fail to detect alternatively spliced messages or tissue-specific alterations in transcriptional activity.

Recently, new developments in microarray technology have made it possible for high-throughput mapping of the transcriptional activity of large segments of the genome (Shoemaker *et al*. 2001, Kapranov *et al*. 2002). Nucleic acids representing non-repetitive segments of a chromosome can be printed at high density and probed with labeled cDNAs prepared from various tissues (Hegde *et al*. 2000). In principle, this approach can be used to detect transcriptional activity of both protein-coding and non-protein-coding RNAs chromosome-wide.

This approach has been used recently in two complementary studies carried out by Shoemaker *et al*. (2001) and Kapranov *et al*. (2002). Shoemaker *et al*. (2001) developed oligonucleotide arrays to represent the known and predicted genes on human chromosome 22 and probed them with cDNA probes derived from RNA isolated from a number of tumor cell lines. Representative expression was observed for a majority of the known genes and a significant fraction of predicted genes, but the experiment did not comprehensively examine unannotated regions of the chromosome. Kapranov *et al*. (2002) developed a microarray containing 25-mer oligonucleotides for most of the non-repetitive DNA of human chromosome 22, and probed with double-stranded cDNA prepared from 11 different cell lines. The investigators observed RNA transcription in many unannotated regions. However, the expression of intron sequences (which comprise 36% of chromosome 22 DNA; Dunham *et al*. 1999) and the conservation of novel transcribed regions in other species were not reported.

This chapter describes the development of a microarray containing polymerase chain reaction (PCR) products encoding 17.4Mb of non-repetitive sequence on human chromosome 22, used to map transcribed regions from the entire chromosome. This array contains both coding and non-coding genomic DNA sequences. The non-repetitive regions of the chromosome were identified using the RepeatMasker program (Smit & Green, unpublished) and divided into 21,024 PCR fragments, ranging in size from 300bp to 1.4kb (mean size = 820bp). PCR primer sequences were designed and the fragments were amplified from HeLa genomic template DNA; 19,525 fragments, representing 93% of the targeted sequences, were successfully prepared. Fragments were printed in duplicate onto three glass slides using a contact microarrayer. A set of positive and negative control fragments was also included on each slide.

Following the construction phase, the microarrays were probed with cDNA reverse-transcribed from poly(A)$^+$ placental RNA in order to assess chromosome-wide transcriptional activity. Gene expression was detected from a significant fraction of the annotated regions of the chromosome using this approach. Moreover, a two-fold increase in transcription was observed over that expected by existing gene annotation. Interestingly, many of the novel transcribed sequences are located within introns of annotated genes. These findings suggest that a large fraction of the genome may be expressed as mRNA, and that there are likely to be many coding sequences that have not been annotated.

CONSTRUCTION OF A CHROMOSOME 22 MICROARRAY

Sequence analysis and primer selection. Chromosome 22q spans 34.5MB, of which 45% consists of repetitive elements (e.g., SINES, LINES, retroviral DNA, and low-complexity sequence) identified by the RepeatMasker program (see Appendix A for more information). The remaining sequence fragments of sufficient size to facilitate large-scale PCR (≥ 300bp) accounted for only 87% of the non-repetitive DNA; the sizes of many non-repetitive fragments fell below this threshold. To improve the sequence coverage, the dynamic programming algorithm described in Chapter 2 was applied to the chromosome sequence. The algorithm is designed to recover many of the smaller non-repetitive fragments by strategically incorporating short repetitive elements located between them, thereby joining the adjacent fragments into larger contiguous sequences amenable to PCR. This procedure generates an optimal tile path for the masked genomic sequence, simultaneously maximizing (1) the coverage of non-repetitive DNA from the target sequence and (2) the number of sequence fragments within a specified size range (in this case, 300bp–1.4kb), while minimizing the number of repetitive nucleotides included in the amplified sequences.

This procedure generated a set of target sequences amounting to 17.4MB, or 92% of the non-repetitive DNA of chromosome 22. PCR primer pairs were selected using the Primer3 software [written by S. Rozen and H.J. Skaletsky (1996); code available online at `http://www-genome.wi.mit.edu/genome_software/other/primer3.html`], and were designed to have similar melting temperatures in a 55–70°C range, low alignment scores, and preferably a 3′ C or G base for increased binding efficiency. Sequences exceeding 1.4kb were subdivided prior to the primer design stage, defining the upper bound of amplicon size. To ensure complete inter-fragment coverage between these adjacent sequences, the 5′ primer sequences for amplicons (2 . . *n*) from subdivided fragments were replaced with the reverse complement of the 3′ primer sequences from the amplicon directly preceding them. The modified primer pairs were examined for inter- and intra-oligo alignment, and the 3′ ends of problematic sequences were adjusted to reduce the potential for primer-dimer formation.

DNA preparation and array production. PCR reactions were performed using 2× QIAGEN MasterMix, 0.5μM of each primer, and 65ng HeLa genomic template DNA. Fragments were analyzed by agarose gel electrophoresis, and only those products that migrated as a single band of the predicted size were arrayed (Figure 3.1B). PCR products were precipitated with a 1:1 mixture of ethanol:isopropanol and dried and resuspended in 25μL H_2O. The fragments were mixed with an equal volume of DMSO for printing. Slides were printed in house with an SDDC-2 arrayer (ESI-Virtek) on Corning CMT GAPS slides. Arrays were cross-linked, and print quality was confirmed by staining for total DNA with POPO-3 (Figure 3.1A).

Several quality-control experiments were performed to assess the fidelity of the amplified sequence and the reproducibility of microarray hybridization results. First, 349 PCR fragments were sequenced with priority placed on those that hybridized to cDNA probes reverse-transcribed from placental poly(A)$^+$ RNA (see below). Sequences were compared with the entire human genome assembly using BLASTN (Altschul *et al*. 1997). Of the 349 fragments sequenced, 314 matched the expected chromosome 22 sequence. For the remaining 35 fragments, 15 matched a sequence very similar to that expected on chromosome 22 (mean = 95% sequence identity to that of chromosome 22), and 20 were from elsewhere in the genome.

The degree of non-specific hybridization to repetitive sequences was ascertained by probing the array with labeled Cot-1 DNA; approximately 6% of the array fragments were found to hybridize to the sample. This was reduced to 1%

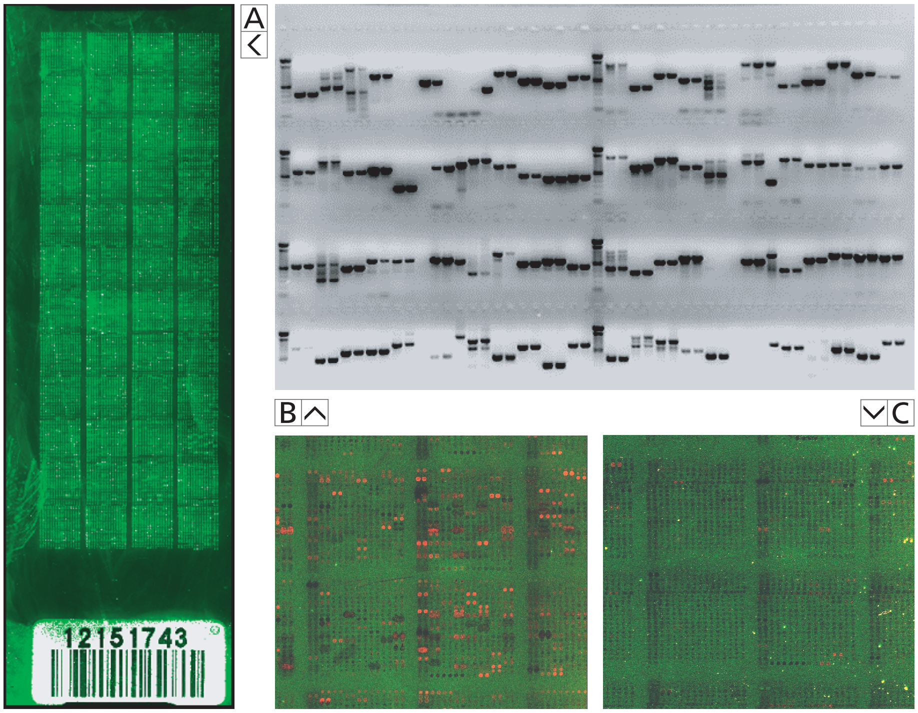


Figure 3.1. Assessment of chromosome 22 microarray sequences, printing consistency and repetitive DNA content. **A**) Microarray slide scanned for total DNA after staining with POPO-3 (Molecular Probes). **B**) Agarose gel electrophoresis of chromosome 22 PCR products. **C**) Hybridization of a subset of features to Cy5-labeled Cot-1 DNA (*left*); hybridization of control sample to the same region in the presence of unlabeled Cot-1 DNA (*right*).

when unlabeled Cot-1 DNA was added to the hybridizations (Figure 3.1C); unlabeled Cot-1 was therefore included in all subsequent experiments to reduce or eliminate non-specific hybridization signal.

CHROMOSOME 22 MICROARRAY DATABASE

Following the design of the chromosome 22 array and during the construction process, a web-accessible database was developed for chromosome-wide gene annotation and analysis of microarray data generated by the project. The system performs four central functions: (1) maintenance of array versions and printing layouts, (2) management of user spaces, (3) archival of microarray data, and (4) relating microarray features to genetic loci. The last point is of particular significance, as the chromosome 22 microarray design is not gene-oriented but instead based on a tile path over the entire non-repetitive sequence. It is therefore not obvious which microarray features are functionally significant from the experimental data alone. Instead, the locations of existing gene components must be considered in relation to each feature's location in genomic space. Furthermore, chromosomal sequence assemblies are periodically updated and gene annotation data is constantly in flux. A software system designed for tiling arrays must be able to consistently present information that accurately reflects the results of experiments performed against this changing background.

System architecture. To implement this functionality, the database architecture is designed around two basic systems (Figure 3.2A). First, a general-purpose microarray database serves as a data repository, slide design inventory and coordinate mapping system applicable to any microarray project. The format of the data is compliant with quantitated output from the GenePix densitometry software (Axon Systems). Among other functions, this component resolves the locations of individual features on the array according to a convolution map determined by the microspotting parameters used for a given slide design version. A second database contains the DNA sequence and gene annotation data specific to chromosome 22. These include known genes (Dunham *et al*. 1999, Hubbard *et al*. 2002), predicted exons (GenomeScan gene predictions contributed by Ru-Fang Yeh and Chris Burge, Massachusetts Institute of Technology), pseudogenes (Harrison *et al*. 2002), and snps (Balasubramanian *et al*. 2002).

The two systems function independently to manage user accounts and experimental data (microarray database) or keep the microarray sequences current with the latest sequence assemblies and gene information (annotation database). The databases interact via high-level software to align various genetic features to the positional coordinates of the microarray fragments in genomic space. The system is also linked to a separate platform for the numerical analysis of experimental data from any standard contact-printed array (Luscombe *et al*. 2003).

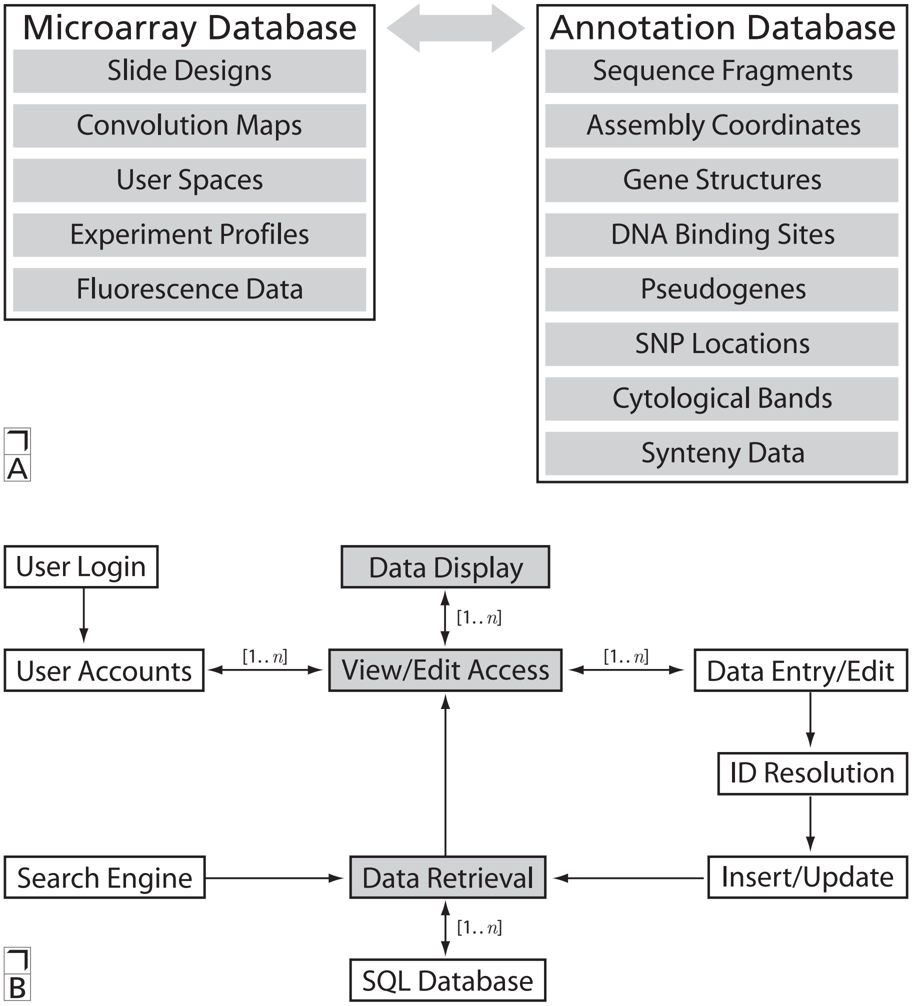
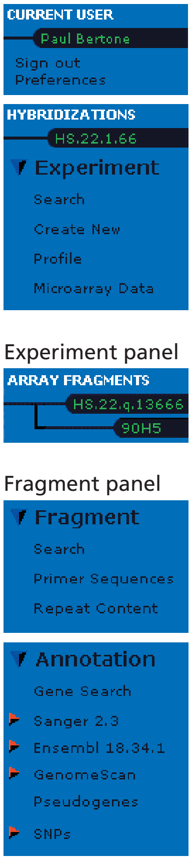


Figure 3.2. A) Architecture of the general-purpose microarray database (left) and project-specific chromosome 22 annotation database (right). The systems can function independently or interact to relate microarray data with annotated genetic elements. **B**) High-level dependency graph illustrating the relationships between various software components. Several core modules through which other processes typically operate are depicted in grey.

User interaction. Experimentalists interact with the system using a web interface to upload and process experimental data sets, where the microarray images have already been scanned and quantitated with the Axon GenePix system (Figure 3.3). The data is then aligned with the positional locations of genes,



User panel

This area is used to log registered users into and out of the system. A user account is not required to access the public areas of the system.

This section displays the current experiment. Once an experiment has been selected, detailed information about the experiment is presented, microarray fragments and associated gene annotation may be viewed, and the data becomes available for download.

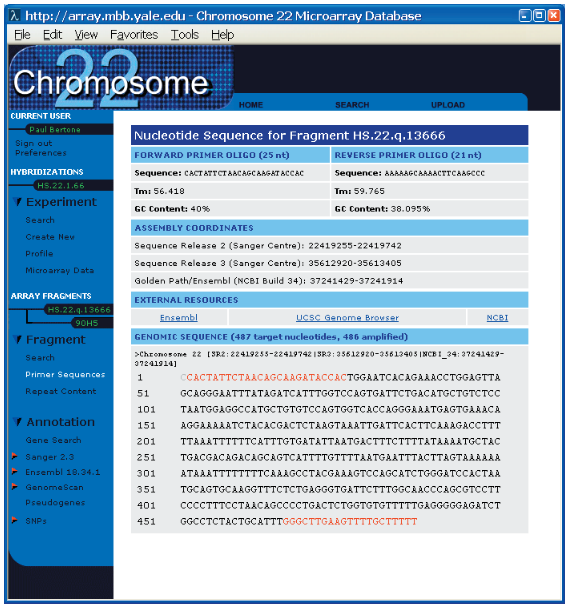
Experiment panel

Selected PCR fragments are displayed here, with the 96-well plate location appearing below the fragment identifier.

Fragment panel

When a microarray fragment has been selected, users can view its DNA sequence, PCR primers, genomic location in a number of sequence assemblies, and links to external web resources.

This area relates the genomic location of the fragment sequence to annotated genes, pseudogenes, or SNPs. When a fragment is selected and its genomic coordinates intersect these features, the corresponding data set will be highlighted with red arrows. Clicking these options in this menu will display the annotated features relative to the PCR fragment, along with additional annotation and links to external web resources, if available.



Nucleotide Sequence for Fragment HS.22.q.13666

FORWARD PRIMER OLIGO (25 nt)	REVERSE PRIMER OLIGO (21 nt)
Sequence: CACTATTCTAACAGCAAGATACCAC	**Sequence:** AAAAAGCAAAACTTCAAGCCC
Tm: 56.418	**Tm:** 59.765
GC Content: 40%	**GC Content:** 38.095%

ASSEMBLY COORDINATES

Sequence Release 2 (Sanger Centre): 22419255-22419742
Sequence Release 3 (Sanger Centre): 35612920-35613405
Golden Path/Ensembl (NCBI Build 34): 37241429-37241914

EXTERNAL RESOURCES

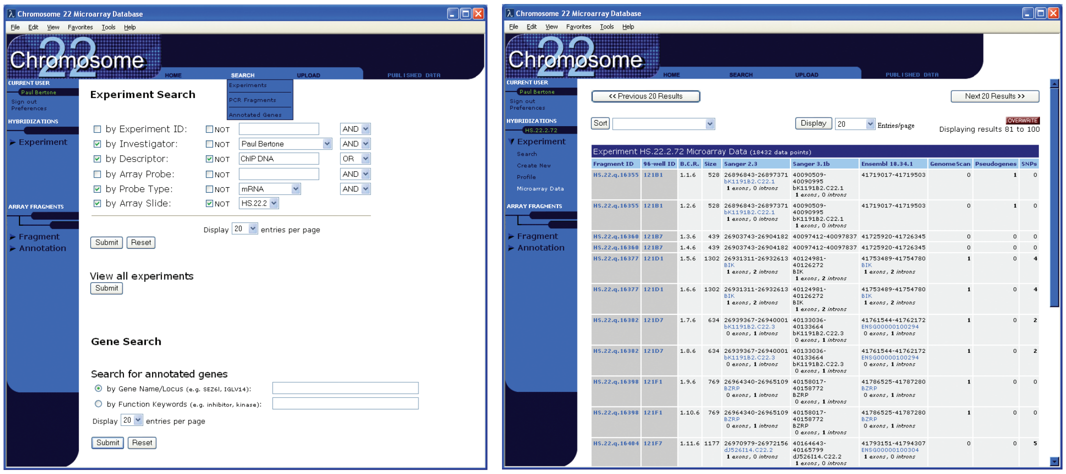
Ensembl | UCSC Genome Browser | NCBI

GENOMIC SEQUENCE (487 target nucleotides, 486 amplified)

```
>Chromosome 22 [SR2:22419255-22419742|SR3:35612920-35613405|NCBI_34:37241429-37241914]
1      CCACTATTCTAACAGCAAGATACCACTGGAATCACAGAAACCTGGAGTTA
51     GCAGGGAATTTATAGATCATTTGGTCCAGTGATTCTGACATGCTGTCTCC
101    TAATGGAGGCCATGCTGTGTCCAGTGGTCACCAGGGAAATGAGTGAAACA
151    AGGAAAAATCTACACGACTCTAAGTAAATTGATTCACTTCAAAGACCTTT
201    TTAAATTTTTTCATTTGTGATATTAATGACTTTCTTTTATAAAATGCTAC
251    TGACGACAGACAGCAGTCATTTTGTTTTAATGAATTTACTTAGTAAAAAA
301    ATAAATTTTTTTTCAAAGCCTACGAAAGTCCAGCATCTGGGATCCACTAA
351    TGCAGTGCAAGGTTTCTCTGAGGGTGATTCTTTGGCAACCCAGCGTCCTT
401    CCCCTTTCCTAACAGCCCCTGACTCTGGTGTGTTTTTGAGGGGGAGATCT
451    GGCCTCTACTGCATTTGGGCTTGAAGTTTTGCTTTTT
```

Figure 3.3. Breakdown of the main interface components of the microarray database software. The control panel is divided into three subsystem modules which function as independent units. Each module maintains its own internal state, allowing users to navigate through different areas of the site simultaneously.

pseudogenes and single-nucleotide polymorphisms (SNPs) in an automated fashion according to the desired sequence assembly (Figure 3.4A). PCR fragments or features of interest may then be explored in greater detail using a variety of graphical and text-based displays, with relevant links to external resources (Figure 3.4B). Specific genes or chromosomal regions may also be located on the array directly, using search functions that relate their nucleotide positions to the corresponding microarray fragments. Thus, researchers are able to correlate vast amounts of experimental data with existing knowledge in a rapid and intuitive way. At present the database contains nearly 1,000 experimental records comprising over 16 million individual data points.



Experiment Search

☐ by Experiment ID:	☐ NOT		AND
☑ by Investigator:	☐ NOT	Paul Bertone	AND
☑ by Descriptor:	☑ NOT	ChIP DNA	OR
☐ by Array Probe:	☐ NOT		AND
☑ by Probe Type:	☐ NOT	mRNA	AND
☑ by Array Slide:	☑ NOT	HS.22.2	

Display 20 entries per page

View all experiments

Gene Search

Search for annotated genes

● by Gene Name/Locus (e.g. SEZ6l, IGLV14):

○ by Function Keywords (e.g. inhibitor, kinase):

Display 20 entries per page

20 Entries/page

Displaying results 81 to 100

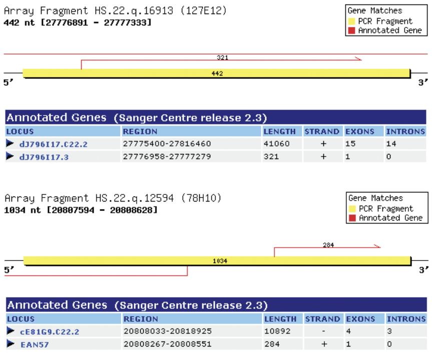
Experiment HS.22.2.72 Microarray Data (18432 data points)

Fragment ID	96-well ID	B.C.R.	Size	Sanger 2.3	Sanger 3.1b	Ensembl 18.34.1	GenomeScan	Pseudogenes	SNPs
HS.22.q.16355	121B1	1.1.6	528	26896843-26897371 bK1191B2.C22.1 1 exons, 0 introns	40090509-40090995 bK1191B2.C22.1 1 exons, 0 introns	41719017-41719503	0	1	0
HS.22.q.16355	121B1	1.2.6	528	26896843-26897371 bK1191B2.C22.1 1 exons, 0 introns	40090509-40090995 bK1191B2.C22.1 1 exons, 0 introns	41719017-41719503	0	1	0
HS.22.q.16360	121B7	1.3.6	439	26903743-26904182	40097412-40097837	41725920-41726345	0	0	0
HS.22.q.16360	121B7	1.4.6	439	26903743-26904182	40097412-40097837	41725920-41726345	0	0	0
HS.22.q.16377	121D1	1.5.6	1302	26931311-26932613 BIK 1 exons, 2 introns	40124981-40126272 BIK 1 exons, 2 introns	41753489-41754780 BIK 1 exons, 2 introns	1	0	4
HS.22.q.16377	121D1	1.6.6	1302	26931311-26932613 BIK 1 exons, 2 introns	40124981-40126272 BIK 1 exons, 2 introns	41753489-41754780 BIK 1 exons, 2 introns	1	0	4
HS.22.q.16382	121D7	1.7.6	634	26939367-26940001 bK1191B2.C22.3 0 exons, 1 introns	40133036-40133664 bK1191B2.C22.3 0 exons, 1 introns	41761544-41762172 ENSG00000100294 0 exons, 1 introns	1	0	2
HS.22.q.16382	121D7	1.8.6	634	26939367-26940001 bK1191B2.C22.3 0 exons, 1 introns	40133036-40133664 bK1191B2.C22.3 0 exons, 1 introns	41761544-41762172 ENSG00000100294 0 exons, 1 introns	1	0	2
HS.22.q.16398	121F1	1.9.6	769	26964340-26965109 BZRP 0 exons, 1 introns	40158017-40158772 BZRP 0 exons, 1 introns	41786525-41787280 BZRP 0 exons, 1 introns	1	0	0
HS.22.q.16398	121F1	1.10.6	769	26964340-26965109 BZRP 0 exons, 1 introns	40158017-40158772 BZRP 0 exons, 1 introns	41786525-41787280 BZRP 0 exons, 1 introns	1	0	0
HS.22.q.16404	121F7	1.11.6	1177	26970979-26972156 dJ526I14.C22.2 1 exons, 0 introns	40164643-40165799 dJ526I14.C22.2 1 exons, 0 introns	41793151-41794307 ENSG00000100304 1 exons, 0 introns	0	0	5

Figure 3.4. A) Search and display of microarray experiments and annotation data. SQL queries are constructed from web form input, returning tabular-formatted results. Various graphical displays (**B**) visualize gene structures, transcriptional start sites and single-nucleotide polymorphisms (SNPs).

A ┐

B ┘



Array Fragment HS.22.q.16913 (127E12)
442 nt [27776891 – 27777333]

Annotated Genes (Sanger Centre release 2.3)

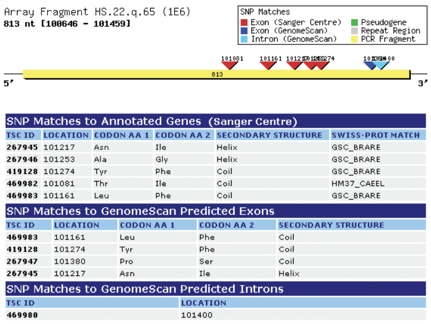
LOCUS	REGION	LENGTH	STRAND	EXONS	INTRONS
▶ dJ796I17.C22.2	27775400-27816460	41060	+	15	14
▶ dJ796I17.3	27776958-27777279	321	+	1	0

Array Fragment HS.22.q.12594 (78H10)
1034 nt [20807594 – 20808628]

Annotated Genes (Sanger Centre release 2.3)

LOCUS	REGION	LENGTH	STRAND	EXONS	INTRONS
▶ cE81G9.C22.2	20808033-20818925	10892	-	4	3
▶ EAN57	20808267-20808551	284	+	1	0

(**B**) Intersection of PCR fragments with various genetic features. Nested genes and pseudogenes (*top*), sequences likely to contain bi-directional promoters (▲ *above*), annotated gene structures (*right* ►), and SNPs (*below* ▼).



Array Fragment HS.22.q.65 (1E6)
813 nt [100646 – 101459]

SNP Matches to Annotated Genes (Sanger Centre)

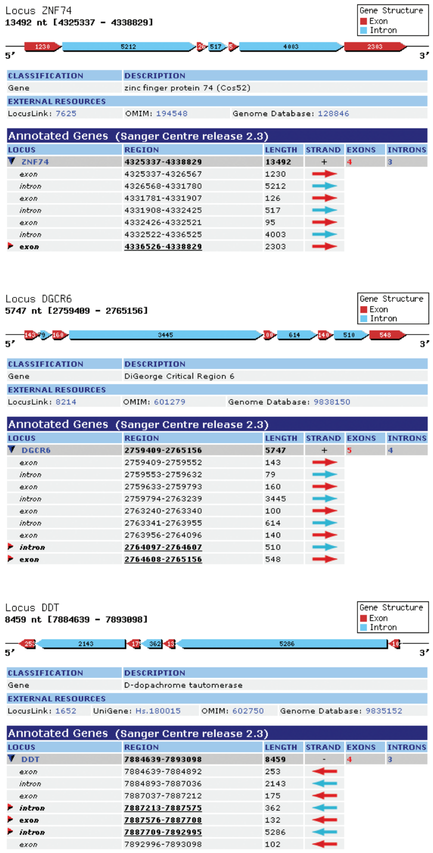
TSC ID	LOCATION	CODON AA 1	CODON AA 2	SECONDARY STRUCTURE	SWISS-PROT MATCH
267945	101217	Asn	Ile	Helix	GSC_BRARE
267946	101253	Ala	Gly	Helix	GSC_BRARE
419128	101274	Tyr	Phe	Coil	GSC_BRARE
469982	101081	Thr	Ile	Coil	HM37_CAEEL
469983	101161	Leu	Phe	Coil	GSC_BRARE

SNP Matches to GenomeScan Predicted Exons

TSC ID	LOCATION	CODON AA 1	CODON AA 2	SECONDARY STRUCTURE
469983	101161	Leu	Phe	Coil
419128	101274	Tyr	Phe	Coil
267947	101380	Pro	Ser	Coil
267945	101217	Asn	Ile	Helix

SNP Matches to GenomeScan Predicted Introns

TSC ID	LOCATION
469980	101400



Locus ZNF74
13492 nt [4325337 – 4338829]

CLASSIFICATION	DESCRIPTION	
Gene	zinc finger protein 74 (Cos52)	
EXTERNAL RESOURCES		
LocusLink: 7625	OMIM: 194548	Genome Database: 128846

Annotated Genes (Sanger Centre release 2.3)

LOCUS	REGION	LENGTH	STRAND	EXONS	INTRONS
▼ ZNF74	4325337-4338829	13492	+	4	3
exon	4325337-4326567	1230	→		
intron	4326568-4331780	5212	→		
exon	4331781-4331907	126	→		
intron	4331908-4332425	517	→		
exon	4332426-4332521	95	→		
intron	4332522-4336525	4003	→		
▶ **exon**	**4336526-4338829**	2303	→		

Locus DGCR6
5747 nt [2759409 – 2765156]

CLASSIFICATION	DESCRIPTION	
Gene	DiGeorge Critical Region 6	
EXTERNAL RESOURCES		
LocusLink: 8214	OMIM: 601279	Genome Database: 9838150

Annotated Genes (Sanger Centre release 2.3)

LOCUS	REGION	LENGTH	STRAND	EXONS	INTRONS
▼ DGCR6	2759409-2765156	5747	+	5	4
exon	2759409-2759552	143	→		
intron	2759553-2759632	79	→		
exon	2759633-2759793	160	→		
intron	2759794-2763239	3445	→		
exon	2763240-2763340	100	→		
intron	2763341-2763955	614	→		
exon	2763956-2764096	140	→		
▶ **intron**	**2764097-2764607**	510	→		
▶ **exon**	**2764608-2765156**	548	→		

Locus DDT
8459 nt [7884639 – 7893098]

CLASSIFICATION	DESCRIPTION		
Gene	D-dopachrome tautomerase		
EXTERNAL RESOURCES			
LocusLink: 1652	UniGene: Hs.180015	OMIM: 602750	Genome Database: 9835152

Annotated Genes (Sanger Centre release 2.3)

LOCUS	REGION	LENGTH	STRAND	EXONS	INTRONS
▼ DDT	7884639-7893098	8459	-	4	3
exon	7884639-7884892	253	←		
intron	7884893-7887036	2143	←		
exon	7887037-7887212	175	←		
▶ **intron**	**7887213-7887575**	362	←		
▶ **exon**	**7887576-7887708**	132	←		
▶ **intron**	**7887709-7892995**	5286	←		
exon	7892996-7893098	102	←		

RNA TRANSCRIPT MAPPING

Microarray hybridization. To experimentally map the transcriptionally-active regions of chromosome 22, placental poly(A)$^+$ RNA was hybridized to the array (Figure 3.5A). RNA from placenta was chosen because it is (1) a normal tissue (i.e., not cancerous or from cell lines), (2) a complex tissue composed of many cell types, and (3) easily obtained in large quantities from a single source. Each chromosome fragment was probed in six independent experiments with cDNA derived from triple-selected poly(A)$^+$ placental RNA.

Using amino-allyl cDNA labeling reagents (Ambion), 1.5µg of placental poly(A)$^+$ RNA that had been purified three times through an oligo(dT) matrix was reverse-transcribed via the M-MLV RNA-dependent DNA polymerase. The reactions were primed with both oligo(dT) and random decamers in an equimolar mix in the presence of an amino-allyl-modified cytosine. After reverse transcription, the template RNA was degraded in the presence of NaOH at 70°C. The cDNAs were ethanol-precipitated and resuspended in 0.1 M $NaHCO_3$ to facilitate coupling of the Cy5 mono-amine dye (Amersham) to the amino-allyl functional group. After the coupling reaction, fluorescence-labeled cDNAs were separated from unincorporated dye with Sephadex gel filtration columns. The cDNA samples were then ethanol-precipitated and resuspended in 5× SSC, 25% formamide, and 15µg of Cot-1 DNA (Invitrogen) to block non-specific hybridization. Microarrays were hybridized at 42°C as described (Hegde *et al*. 2000).

DATA ANALYSIS

Determination of hybridization positives. Microarrays were scanned with an Axon 4000A scanner, and images were analyzed with GenePix Pro 3.0 software. To identify fragments exhibiting significant hybridization signal, a statistical data analysis scheme was devised specifically for microarrays probed with a single color fluor. To identify hybridized fragments, the raw GenePix output was processed as follows: (1) Spots with aberrant morphology, or those with intensities below the threshold of detection were discarded. (2) Within individual experiments, spot pairs (fragments printed in duplicate side by side) were excluded from further analysis if the variation (= $I_1 - I_2/I_1 + I_2$) between them was greater than three standard deviations of the error distribution of the data points. (3) The six replicate experiments were normalized with one another to scale the Cy5 intensity spreads to a common range. A resampled variance was calculated for each experiment and the distributions were scaled to have equal variances. Different scale factors were calculated for each block of spots on the

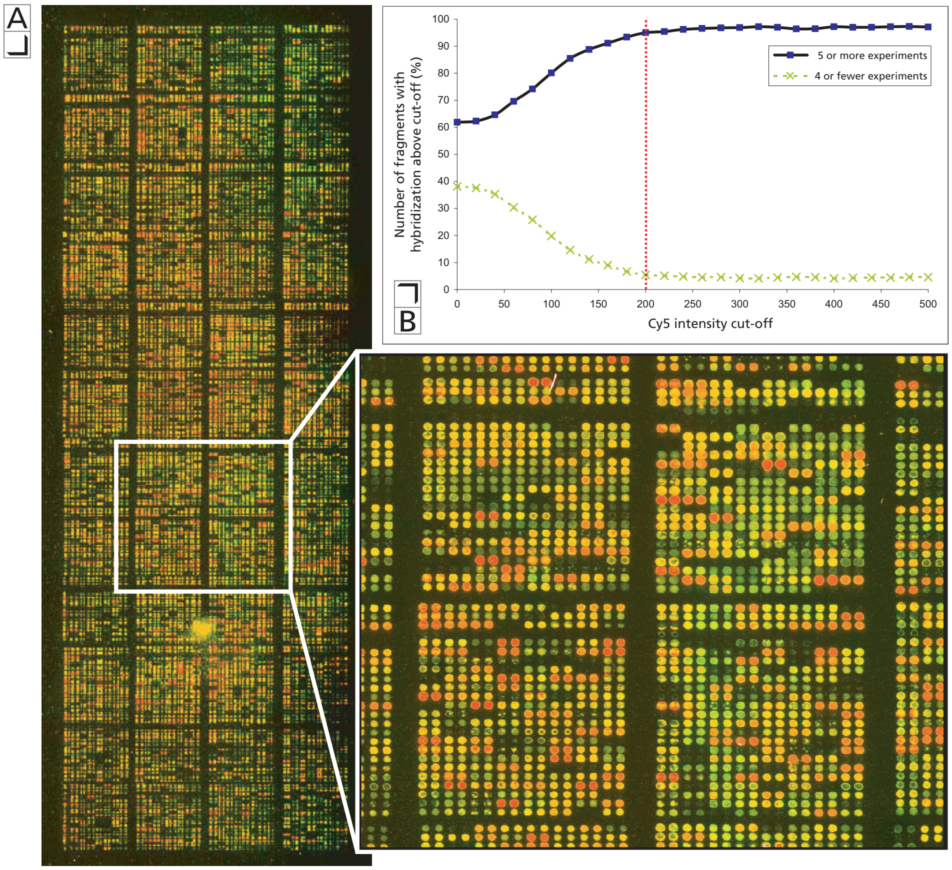


Figure 3.5. A) Fluorescence micrograph of hybridized chromosome 22 array. **B**) Determination of intensity cutoff in determining positive hybridizing fragments. There is a clear leveling of consistency after 200 intensity units. The plot demonstrates that fragments with an intensity >200 were present with that intensity or higher in 5 out 6 replicate experiments. Fragments that hybridized in 4 or less of the 6 replicate experiments and with an intensity ≥ 200 were summed to give a false-positive rate of 5%.

Annotated feature	Total	Exon-containing	Intron-containing
Gene	946 (11.9%)	428 (15.8%)	518 (9.8%)
Related gene	135 (11.4%)	66 (13.6%)	69 (9.9%)
Predicted gene	87 (9.9%)	50 (15.2%)	37 (6.8%)
Unannotated	1,302 (12.2%)		

Table 3.1. Distribution of positive hybridizing fragments and their respective gene annotations from the Sanger 2.3 data release. Parentheses indicate the percentage of total microarray fragments in the annotation category that showed positive hybridization. An equal magnitude of transcription was observed in previously unannotated regions.

slide to correct for intensity variations dependent on slide location (Goryachev *et al.* 2001, Yang *et al.* 2002). (4) The final Cy5 intensity for each chromosome 22 fragment was obtained as the mean for duplicate spots within an experiment and the median value across replicate experiments. The number of experiments in which each fragment exhibited statistically significant hybridization was also recorded. (5) The number of fragments that hybridized in 1..*n* replicate experiments was calculated, and only those fragments that hybridized in five or more replicates were retained for further analysis. Here, fragments that hybridized in fewer than five experiments were considered to be false positives.

Figure 3.5B plots the percentage of fragments that hybridized in five or more experiments against different Cy5 intensities. The Cy5 intensity cutoff of 200 fluorescence units for positive hybridized fragments was determined empirically from the plot, on which a sharp rise in the proportion of fragments in five or more experiments can be observed; at this intensity a total of 2,504 positive hybridized fragments with a false-positive rate of 5% were identified. Figure 3.6 depicts the transcriptional activity and density of hybridizing chromosome 22 sequence fragments in relation to Sanger Centre annotated genes, and Table 3.1 summarizes the annotation distribution of these fragments.

Detection of annotated genes. To compare these results with known features of chromosome 22, annotated genes corresponding to the version 2.3 data release from the Sanger Centre were aligned to the sequence coordinates of the 21,024 microarray fragments. The genes in the Sanger Centre collection fall into three categories: (1) known genes, which are well-characterized genes with a

Annotated feature	Identified	Total	Identified/Total (%)
Gene	206	339	60.8
Related gene	45	112	40.2
Predicted gene	35	109	35.8

Table 3.2. Genes in the three Sanger 2.3 annotation categories that were represented by at least one hybridizing exon. A total of 60.8% of known genes were detected using only one tissue type, as well as detecting expression from a large fraction, 40.2% and 35.8%, of the related and predicted genes, respectively. This success rate is similar not only to other studies using microarrays to annotate human chromosome 22, but also to studies using ESTs.

sequenced full-length cDNA; (2) related genes, which are homologous to other known genes; and (3) predicted genes, which are posited by homology to EST clusters. For the 339 known genes in the Sanger annotation data, at least one exon was found to hybridize in 206 (60.8%) cases (Table 3.2). In addition to detecting expression of the known genes, 40.2% and 35.8% of the related and predicted genes were found to be expressed, respectively (Table 3.2). This result demonstrates that a majority of the chromosome 22 genes can be detected using a single tissue type, and that this approach can globally detect known, related, and predicted genes simultaneously.

Detection of novel transcribed sequences in unannotated regions of chromosome 22. Hybridization of cDNA probes to known and predicted exons was accompanied by an equal amount of hybridization to previously unannotated sequences. A total of 1,302 (12.2%) of 10,693 fragments lacking prior annotation were observed to be expressed in placental tissue (Table 3.1). This amount is similar to the 946 (11.8%) of the 7,967 microarray fragments intersecting known genes. Figure 3.6, box A, illustrates a large amount of transcriptional activity in a region of chromosome 22 that was previously unannotated. Positions with spikes in the density plot and low frequency of red bars indicate regions of novel transcriptional activity. Box B shows a peak in transcriptional activity corresponding to known gene annotations. Viewed together, these results indicate that there are as many transcribed sequences in unannotated regions as in annotated regions.

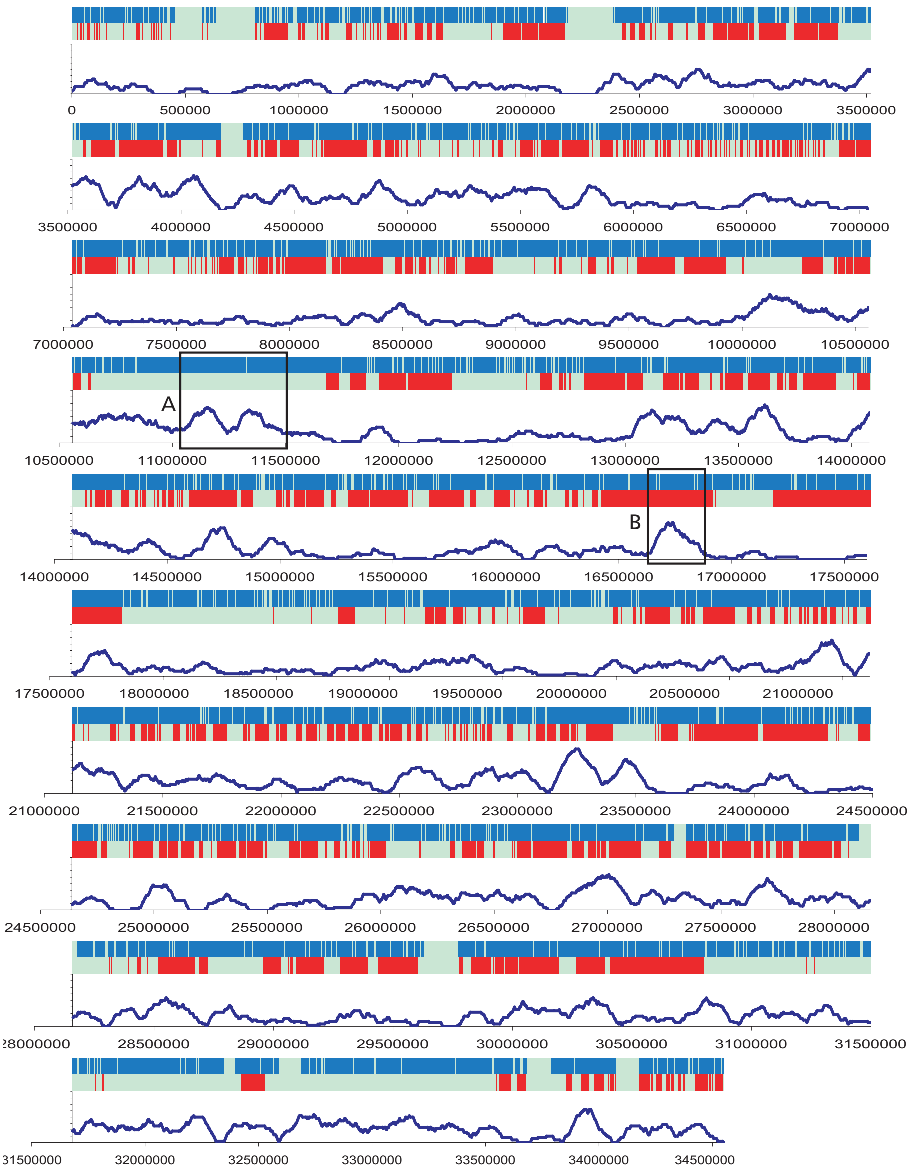


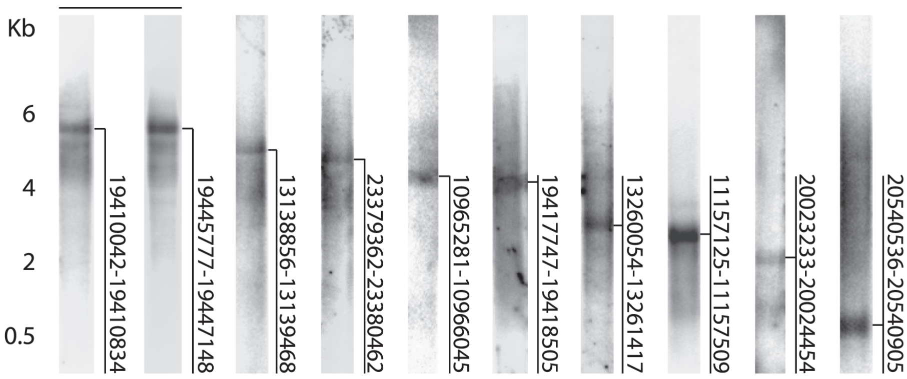
Figure 3.6. Transcriptional activity over chromosome 22 as measured in placental RNA. Dark blue bars indicate regions that are represented on the DNA microarray. Red bars indicate the positions of Sanger Centre release 2.3 annotated genes. The magnitude of the density plot represents the number of positive hybridizing fragments divided by the total number of fragments in a 100-kb window. (A) A high level of transcriptional activity in a previously unannotated region of chromosome 22. (B) A peak in transcriptional activity corresponding to known gene annotations. Coordinates are given from centromere to telomere (starting at band 22q11.1).

RNA BLOT VERIFICATION OF NOVEL TRANSCRIBED SEQUENCES

To confirm that the unannotated transcribed sequences (defined here as transcriptionally-active regions, or TARS) are expressed as mRNA, 118 RNA blots of placental poly(A)$^+$ RNA were probed with randomly selected TAR sequences. Northern blots of triple-selected poly(A)$^+$ placental RNA were obtained from Ambion. Five blots were cut into a total of 50 single-lane strips. Each strip was pre-hybridized in ULTRAhyb buffer (Ambion) for 2 hours and then hybridized to radiolabeled probes prepared from novel TAR sequences using Strip-EZ labeling reagents (Ambion). Hybridizations were carried out for 16 hours at 42°C. Strips were washed twice in Northern Max (Ambion) high-stringency buffer followed by three washes in Northern Max low-stringency buffer. Single-lane filters were stripped according to the Strip-EZ protocol.

Three fragments containing exons of known genes were also used to probe the RNA blots as a control; all three identified transcripts of the appropriate size. Thirty (25%) unannotated fragments hybridized to mRNA transcripts ranging in size from 600 nt to >10kb. Several probes hybridized with multiple isoforms, perhaps indicating the presence of alternate splice products. Interestingly, two

Figure 3.7. Northern blot analysis of 118 fragments that were expressed in previously unannotated regions of chromosome 22. Thirty (25.4%) showed discrete bands. Ten are shown above and labeled with the corresponding chromosomal location of the probe used in the Northern hybridization. Bar indicates two probes separated by ~30kb in genomic space that hybridize to the same 6-kb transcript.



probes separated by 30kb in genomic space hybridized to the same 6-kb transcript, further indicating that this region encodes a novel gene (Figure 3.7).

To ensure that transcripts were not homologous to coding sequence elsewhere in the genome, all probes producing transcripts were searched using blastn (Altschul *et al.* 1990). The results indicated that 26/30 matched only the chromosome 22 genomic sequences and 4 probes have potential homology ($E <$ $1e^{-5}$) to other genomic coding sequences. Thus, most of the transcribed sequences identified by the RNA blot analysis are derived solely from chromosome 22. The lower than expected success rate of the RNA blot analysis (30/118) was also noted in a similar study (Kapranov *et al.* 2002); it is speculated that many novel tars are expressed in low copy number, explaining why most have eluded prior detection using less sensitive methods.

differential hybridization mapping

To precisely map the expressed regions as well as determine the DNA strand of the hybridizing sequence, a novel strategy termed differential hybridization mapping (dhm; Kumar *et al.* 2002) was used. Briefly, a 60-nt oligomer and its complement were selected from regions within the hybridizing pcr fragments, spotted on the array, and probed with labeled poly(A)$^+$ placental cDNAs. The cDNA will hybridize to the oligonucleotide that the message derived from and not to its complement. Thus, differential hybridization of the two oligonucleotides maps the expression to one strand.

Prediction of potential exons. To find potential exons in the 1,302 unannotated tars, their sequences were analyzed using four commonly used gene prediction methods: Genscan (Burge & Karlin 1997), Grailexp (Guigó *et al.* 1992), Geneid (Xu & Uberbacher 1997), and by homology to known genes. Candidate sequences were searched against the non-redundant sequence database (nrdb) and Ensembl protein sequence databases using tblastx with six-frame translation (Altschul *et al.* 1997).

The matches then were filtered for repetitive sequences with the RepeatMasker program. To eliminate overlapping results, homology matches were filtered such that lower-scoring matches that overlapped with a higher-scoring match by more than 40nt were discarded. The three gene prediction programs Genscan, GrailEXP and GeneID were also applied to each amplicon sequence.

For each resulting set of exon predictions, a non-redundant list was made such that better-scoring predictions were chosen in preference to those with lower scores. GrailEXP makes predictions using a large database of ESTs, cDNAs, and mRNAs; these predictions are chosen in preference to any other prediction. The remaining exon predictions were chosen in the following order of preference: (1) Genscan with exon probability ≥ 0.1, (2) GrailEXP, (3) GeneID, (4) Genscan with exon probability < 0.1. Any additional potential exons produced from the homology searches detailed above were also included. The final non-redundant collection of exon predictions was then used to derive 60-nt probes by selecting unique internal sequences from each predicted exon region using the Primer3 software.

For the top 381 exon predictions, a 60-mer oligonucleotide was selected to represent a unique sequence from each predicted exon and its complement. In this way, oligonucleotide selection is expected to be biased toward potential coding sequences. Oligonucleotides were synthesized by Illumina, resuspended in 50% DMSO at 50µM and printed and hybridized as described above. The oligonucleotide pairs were spotted on a separate area of the chromosome 22 array and probed with labeled poly(A)$^+$ placental cDNAs.

Differential hybridization analysis. DHM oligonucleotide arrays were scanned and processed using the same method as for the chromosome 22 arrays. When one of the oligonucleotides in the pair hybridized and the complement did not, they were considered to hybridize differentially to one strand. Those pairs exhibiting differential expression on the same strand in three of four replicate experiments were scored as positive expressed sequences. To identify positive hybridized oligonucleotides, Cy5 fluorescence signals for oligonucleotide pairs (strand and antistrand) were compared to each other, providing a measure of pairwise differences in hybridization ($= I_{\text{strand}} - I_{\text{antistrand}}/I_{\text{strand}} + I_{\text{antistrand}}$). Oligonucleotides that had no detectable signal or that were filtered from the data set were assigned an intensity value of zero. The distribution of the pairwise differences approximated a normal distribution, and a set of 119 outlier pairs was selected as being differentially hybridized ($P < 0.001$). For each pair, the oligonucleotide with the higher Cy5 signal was identified as exhibiting positive hybridization.

Significant differential hybridization was observed in 53 of the 381 pairs, indicating that the hybridizing region and strand could be identified in many cases. As a control, multiple oligonucleotides were included to map a region that contains an exon sequence on one strand representing a gene known to be expressed in placental tissue, and an intron sequence on the opposite strand. As expected, only the exon strand hybridized to the poly(A)$^+$ RNA (Figure 3.8C). Presumably, in the cases that did not exhibit differential hybridization the expressed region was not represented by the 60-mer oligonucleotides or both strands were transcribed. In summary, the RNA blot analysis and oligonucleotide DHM data independently verified that a significant number of the unannotated hybridizing regions are expressed as mature mRNA transcripts.

NOVEL TRANSCRIPTION WITHIN ANNOTATED INTRONS

Analysis of the hybridizing fragments that intersect annotated introns revealed the unexpected finding that many introns contain transcribed sequences. In fact, of the 5,264 fragments located entirely within annotated introns, 518 (9.8%) were found to be expressed in five of six experiments. There are three possible explanations for this observation: (1) a novel transcribed sequence is encoded on the strand opposite the intron; (2) there is an unannotated exon located within the intron that had not been discovered previously; or (3) expressed intron sequences were detected.

To distinguish among these possibilities, the DHM technique was used as described above. For this, 119 60-mer oligonucleotides representing various intron regions and their complementary sequences were spotted onto a microarray and probed with labeled poly(A)$^+$ placental cDNAs. Of the 119 oligonucleotide pairs, 23 (19.3%) showed significant differential hybridization. Expression from the same strand as the intron was detected in 13 cases, indicating that sequences from within the introns are transcribed. In five of these cases, an exon was pre-

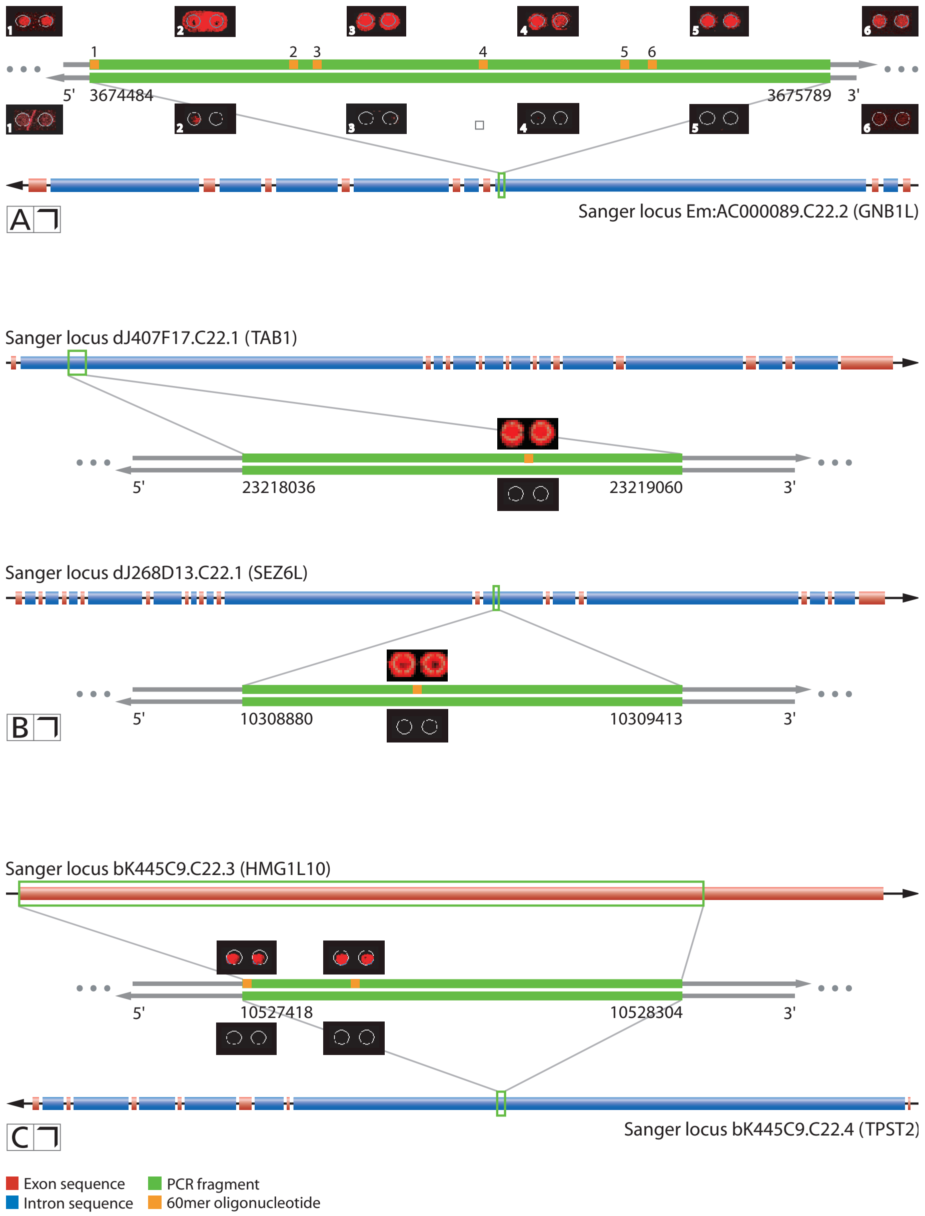

1
2
3
4
5
6
1 2 3 4 5 6
5'
3674484
3675789
3'
A
Sanger locus Em:AC000089.C22.2 (GNB1L)
Sanger locus dJ407F17.C22.1 (TAB1)
5'
23218036
23219060
3'
Sanger locus dJ268D13.C22.1 (SEZ6L)
5'
10308880
10309413
3'
B
Sanger locus bK445C9.C22.3 (HMG1L10)
5'
10527418
10528304
3'
C
Sanger locus bK445C9.C22.4 (TPST2)
🟥 Exon sequence
🟦 Intron sequence
🟩 PCR fragment
🟧 60mer oligonucleotide

Figure 3.8. Differential hybridization mapping within positive PCR fragment sequences. (A) Hybridization to multiple 60-nt oligonucleotides positioned opposite an intron sequence annotated on the antisense strand. (B) Hybridization to oligonucleotides representing a predicted exon within an annotated intron on the sense strand. (C) Control spots showing differential hybridization to a known exon (1) located on the strand opposite an annotated intron and (2) whose expression was previously verified.

dicted to lie within the intron; one example is presented in Figure 3.8B. In ten cases, transcription originated from the opposite strand of the intron, suggesting that a novel transcribed sequence overlaps with the intron. In total, nearly half of the hybridizing fragments that intersect intron regions were shown to contain expressed sequences antisense to their respective introns.

To further investigate this observation, DHM was used with multiple oligonucleotide probe pairs to completely cover a subset of the hybridizing fragments located within annotated introns. In one case, six oligonucleotide pairs from a 1.3-kb region showed differential hybridization to the strand antisense to an annotated intron (Figure 3.8A). In another example, two positive 60-nt probes hybridized within a 400-bp region opposite a known intron. In these cases, the regions that are transcribed on the opposite strand of introns are not short in length because multiple probes detect expression throughout the segment. In summary, novel transcribed sequences were identified hybridizing to regions both internal to annotated introns and to the strand opposite introns.

ASSESSMENT OF EVOLUTIONARY CONSERVATION

Many of the positive hybridizing fragments whose sequences lie outside those of known genes are likely to correspond to novel exons. It follows that a percentage of these are likely to be homologous to other mammalian genes, providing supporting evidence of putative coding regions. To assess the degree of sequence conservation with other mammalian sequences, a homology comparison of unannotated TARs with the mouse genome was performed using BLASTN and BLASTP using published criteria.

Positive fragments intersecting genes known to be mouse orthologs were identified as follows. A comprehensive set of annotated human genes on chromosome 22 with established homology to mouse genes was compiled using five data sets obtained from the NCBI [National Center for Biotechnology Information, Human/Mouse Homology Maps (May 2002); `http://www.ncbi.nlm.nih.gov/Homology`]. These consist of human-mouse orthologs identified by homology between the genetic map represented in the Mouse Genome Database (MGD; Blake *et al.* 2002) and the Whitehead/MRC radiation hybrid map (Hudson

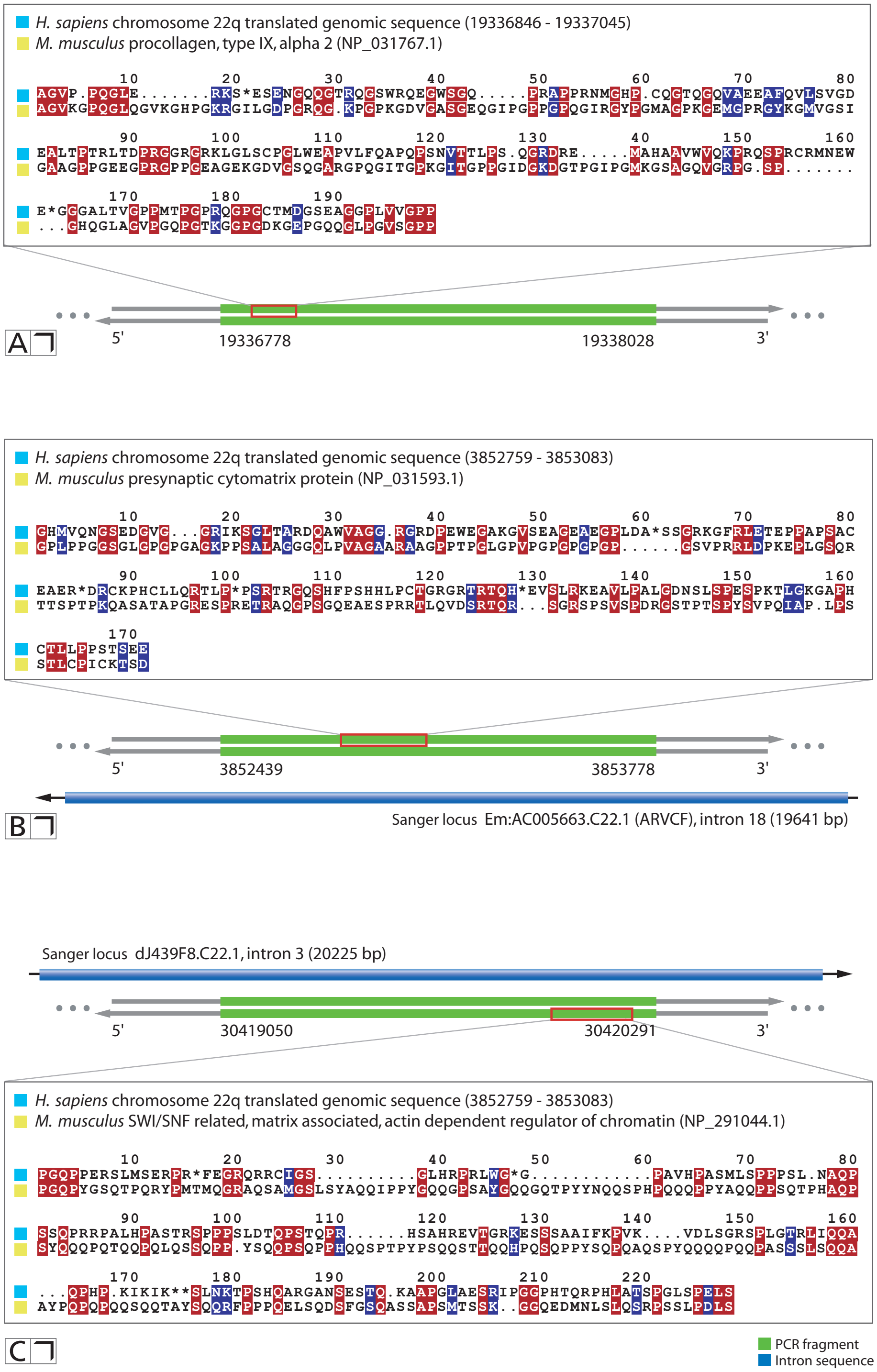

🟦 H. sapiens chromosome 22q translated genomic sequence (19336846 - 19337045)
🟨 M. musculus procollagen, type IX, alpha 2 (NP_031767.1)
10 20 30 40 50 60 70 80
AGVP.PQGLE.......RKS*ESENGQQGTRQGSWRQEGWSGQ.....PRAPPRNMGHP.CQGTQGQVAEEAFQVLSVGD
AGVKGPQGLQGVKGHPGKRGILGDPGRQG.KPGPKGDVGASGEQGIPGPPGPQGIRGYPGMAGPKGEMGPRGYKGMVGSI
90 100 110 120 130 40 150 160
EALTPTRLTDPRGGRGRKLGLSCPGLWEAPVLFQAPQPSNVTTLPS.QGRDRE.....MAHAAVWVQKPRQSPRCRMNEW
GAAGPPGEEGPRGPPGEAGEKGDVGSQGARGPQGITGPKGITGPPGIDGKDGTPGIPGMKGSAGQVGRPG.SP.......
170 180 190
E*GGGALTVGPPMTPGPRQGPGCTMDGSEAGGPLVVGPP
...GHQGLAGVPGQPGTKGGPGDKGEPGQQGLPGVSGPP
5'
19336778
19338028
3'
A
🟦 H. sapiens chromosome 22q translated genomic sequence (3852759 - 3853083)
🟨 M. musculus presynaptic cytomatrix protein (NP_031593.1)
10 20 30 40 50 60 70 80
GHMVQNGSEDGVG...GRIKSGLTARDQAWVAGG.RGRDPEWEGAKGVSEAGEAEGPLDA*SSGRKGFRLETEPPAPSAC
GPLPPGGSGLGPGPGAGKPPSALAGGGQLPVAGAARAAGPPTPGLGPVPGPGPGP......GSVPRRLDPKEPLGSQR
90 100 110 120 130 140 150 160
EAER*DRCKPHCLLQRTLP*PSRTRGQSHFPSHHLPCTGRGRTRTQH*EVSLRKEAVLPALGDNSLSPESPKTLGKGAPH
TTSPTPKQASATAPGRESPRETRAQGPSGQEAESPRRTLQVDSRTQR...SGRSPSVSPDRGSTPTSPYSVPQIAP.LPS
170
CTLLPPSTSEE
STLCPICKTSD
5'
3852439
3853778
3'
Sanger locus Em:AC005663.C22.1 (ARVCF), intron 18 (19641 bp)
B
Sanger locus dJ439F8.C22.1, intron 3 (20225 bp)
5'
30419050
30420291
3'
🟦 H. sapiens chromosome 22q translated genomic sequence (3852759 - 3853083)
🟨 M. musculus SWI/SNF related, matrix associated, actin dependent regulator of chromatin (NP_291044.1)
10 20 30 40 50 60 70 80
PGQPPERSLMSERPR*FEGRQRRCIGS..........GLHRPRLWG*G............PAVHPASMLSPPPSL.NAQP
PGQPYGSQTPQRYPMTMQGRAQSAMGSLSYAQQIPPYGQQGPSAYGQQGQTPYYNQQSPHPQQQPPYAQQPPSQTPHAQP
90 100 110 120 130 140 150 160
SSQPRRPALHPASTRSPPPSLDTQPSTQPR......HSAHREVTGRKESSSAAIFKPVK....VDLSGRSPLGTRLIQQA
SYQQQPQTQQPQLQSSQPP.YSQQPSQPPHQQSPTPYPSQQSTTQQHPQSQPPYSQPQAQSPYQQQQPQQPASSSLSQQA
170 180 190 200 210 220
...QPHP.KIKIK**SLNKTPSHQARGANSESTQ.KAAPGLAESRIPGGPHTQRPHLATSPGLSPELS
AYPQPQPQQSQQTAYSQQRFPPPQELSQDSFGSQASSAPSMTSSK..GGQEDMNLSLQSRPSSLPDLS
C
🟩 PCR fragment
🟦 Intron sequence

Figure 3.9 (◂ *previous*)**.** Mouse protein homology within translated PCR fragment sequences. (**A**) Homology match between mouse sequence and a positive microarray fragment whose sequence coordinates lie outside annotated genes. (**B**,**C**) Examples of mouse protein matches to human genomic sequences that are opposite annotated introns. In both cases the homology match is antisense to the intron.

et al. 2001) with the NCBI Build 28/UCSC HG10 human genome assembly (UCSC Human Genome Project Working Draft, December 2001 assembly; `http://genome.cse.ucsc.edu`).

Each homologous gene found on chromosome 22 was cross-referenced with Sanger-annotated genes, and the positive fragments that intersect them were identified. To assess the degree of sequence similarity between the remaining positive microarray fragments and mouse sequences, the fragment sequences were queried against the draft mouse genome (NCBI Mouse Genome Release 27) using BLASTN for nucleotide-nucleotide comparisons (Altschul *et al.* 1990), and to the RefSeq repository of mouse protein sequences (Pruitt & Maglott 2001) using BLASTX for six-frame translational nucleotide-protein comparisons. In each case a threshold *e*-value of 0.0001 was used to select significant matches, with the additional restriction that only matches exceeding 200nt were considered significant for the mouse genomic DNA comparison.

Of the 1,231 positive microarray fragments intersecting Sanger-annotated genes, 541 (~44%) intersect an ortholog in the mouse genome. Interestingly, 90 (7%) positive fragments that do not intersect with annotated genes potentially encode proteins that are homologous to mouse proteins (82) or genomic sequence (8), although the occurrence of stop codons in some reading frames may contribute to nonsense-mediated decay if so translated. For instance, an unannotated fragment was found to share a significant degree of sequence similarity with a mouse procollagen protein, as depicted in Figure 3.9A. Of the 90 DNA fragments that contain internal sequences similar to mouse proteins, 25 are located in introns, and many are on the antisense strand of the annotated introns (Figures 3.9B and C). Thus, it appears that a large fraction of the novel TARS are evolutionarily conserved and may encode functional proteins or noncoding RNAs.

COMPARISON TO OTHER METHODS

Detection of annotated and predicted genes. Transcription was observed in a single tissue type from many known and predicted genes that have been identified previously. This includes the majority (60.8%) of all the known genes on the chromosome. In addition, 40.2% and 35.8% of the related and predicted genes were detected, respectively. This success rate is similar to that of other studies (de Souza *et al.* 2000, Shoemaker *et al.* 2001). However, those studies used the earlier annotation data accompanying the original chromosome 22 sequence, which contained many more related and predicted genes. In contrast, the study described here used the latest Sanger 2.3 annotation, in which many related and predicted genes have now been classified as known (98 and 50, respectively, relative to the initial Sanger Centre data release for chromosome 22). Nonetheless, microarray analysis was able to verify the few remaining predicted and related genes, demonstrating the sensitivity of this approach.

Transcription in unannotated regions. In addition to the annotated regions, transcription was detected in many chromosome 22 regions that have not been detected previously. There are probably two reasons for this. First, the unannotated TARs may be expressed at low abundance. Only 25% of the 118 hybridizing fragments from unannotated regions detected discrete transcripts using RNA blot analysis. Those fragments that did not detect mRNAs using RNA blot analysis may therefore encode low-abundance transcripts. The second reason novel TARs may have been detected is that the use of a chromosome-wide tiling array interrogates most of the unique sequences of the chromosome, and is thus more comprehensive than most other methods.

Several hypotheses may explain the biological functions of the novel transcribed regions. It is likely that in many cases these encode low-abundance proteins of new genes. It is also possible that the transcribed regions correspond to previously missed exons of known genes. A third possibility is that they may function as non-coding RNAs (i.e., siRNAs, snoRNAs, hnRNAs, or other small RNAs); in this capacity they might serve in a structural, catalytic or regulatory capacity. For instance, if the novel coding segments produce antisense transcripts, they might control the levels, export, or translation of genes encoded on the opposite strand. Regardless of their functions, these newly discovered expression regions are clearly an important source of new biological information, as many of them are conserved among mammals.

The microarray approach is comprehensive. A variety of other studies have been used to annotate chromosome 22. SAGE, ESTS, and ORESTES have identified a number of coding segments on the chromosome (Saccone *et al*. 1996, Deloukas *et al*. 1998, de Souza *et al*. 2000, Liang *et al*. 2000, Caron *et al*. 2001). However, these studies are biased toward detecting the most abundant transcripts, and they are often limited by the short stretches of DNA that are sequenced. The microarray approach is more suitable for expression profiling because several different tissue types can be analyzed in parallel to determine tissue-specific abundance. Also, this approach can be used to elucidate other annotation features, whereas the previously mentioned techniques cannot; for example, identification of transcription factor binding sites via hybridization of chromatin immunoprecipitated DNA.

Two independent microarray studies have also investigated the transcriptional activity of the chromosome. Shoemaker *et al*. (2001) synthesized oligonucleotide probes to represent many predicted exons from Genscan. Although the method was able to detect transcripts for 185 (57%) of the 325 Genscan-predicted genes, their study did not examine the majority of non-repetitive sequence on chromosome 22. Although a limited tiling survey examined a 113-kb region to investigate exon-intron boundaries, the microarray was designed primarily to detect annotated and predicted genes. The approach described here is more comprehensive and universally applicable to a wide range of experiments.

Mapping of Affymetrix probes. An independent study by Kapranov *et al*. (2002) interrogated transcriptional activity using high-density oligonucleotide arrays containing 25-nt oligonucleotide probes spaced, on average, 10nt apart to cover most of the non-repetitive DNA of chromosomes 21 and 22. These 25-nt probes were developed using the original chromosome 22 contig sequences, corresponding to the initial Sanger Centre data release (Dunham *et al*. 1999).

To relate the present transcription data to the results of this study, a procedure was developed to map positive oligonucleotide sequences to the current assembly of chromosome 22 on which the microarray described here was constructed. The original contig sequences were obtained, and each was subdivided into 500-bp fragments. These subsequences were aligned with the current assembly of chromosome 22q with BLASTN (Altschul *et al*. 1990), using a long word length of 400bp to obtain a single optimal match for each fragment. The center positions of the positive oligonucleotides were known relative to the original contig sequences; an offset could therefore be computed for each 25-mer oligonucleotide with the offset shifting its coordinates according to the chromosomal location of the contig fragment on which the oligonucleotide was originally placed. Using this method, short oligonucleotide sequences could be accurately located on the updated chromosome assembly, while avoiding the many spurious homology matches that would result from comparing each 25-nt sequence with the entire chromosome directly.

The experiment by Kapranov *et al*. prepared cRNA probes from 11 cell lines that were hybridized to the oligonucleotide arrays. This study also found that many unannotated regions of the chromosome are expressed as polyadenylated RNA; however, there are a number of differences between the two studies. First, they did not report that expression is observed from within annotated introns, nor did they assess the degree of homology between expressed sequences and those in other genomes to establish evidence for conserved regions. Second, cDNA probes from different sources were used. Kapranov *et al*. used probes from RNA isolated from 11 cancer cell lines, whereas the present study used normal placental poly(A)$^+$ RNA. Third, double-stranded cDNA probes were generated from the RNA; thus, they could not determine which strand is transcribed in the oligonucleotide hybridizations.

Although no microarray is entirely comprehensive, the PCR-based approach has several advantages. First, it contains large regions of contiguous sequence information, ensuring that no information is omitted. However, PCR-based arrays are of lower resolution, and the exact hybridizing region must be determined by other methods such as DHM. Despite the inherent differences in the two approaches, a thorough comparison of transcribed sequences reveals extensive overlap between the two studies. Of the 2,504 hybridizing fragments identified in the present study, 10% (250) were not detected in the Affymetrix investigation, indicating that the two methodologies are complementary.

Another advantage of PCR-based microarrays is that they can be produced in an academic lab and at high throughput. Thus, the approach is easily amenable to serially hybridizing many tissue types to determine tissue-specific transcripts. This array is also a versatile tool for many other purposes such as identifying transcription-factor-binding sites in conjunction with chromatin immunoprecipitation methods (discussed in Chapter 5). Eventually, transcription maps derived from tiling array experiments may also serve a comparative evolutionary function. Typically, whole genome sequences are compared to find similarities that have been preserved through evolution. Although this is a valid and useful approach, experimentally-derived transcription data may also be used to discover individual conserved transcripts between related species. The latter may be a useful way to determine functional differences across genomes as well as fine-resolution evolutionary changes in chromosomes.

SUMMARY

This study embodied a multifaceted approach to provide a detailed transcriptional map of human chromosome 22. A microarray containing most of the unique sequence was developed and subsequently hybridized to probes prepared from human placental poly(A)$^+$ RNA to identify transcriptionally active regions across the chromosome. In addition to detecting known and predicted coding regions, an equal level of transcription was measured in previously unannotated regions. Many of these novel coding segments were verified to produce bona fide messages via RNA blot analysis. A comparison of novel transcribed regions with mouse sequences revealed that many are evolutionarily conserved in mammals. These regions were defined at a higher resolution using differential hybridization mapping (DHM) with oligonucleotide arrays. DHM analysis verified that a significant fraction of expression originates within introns and antisense to introns. These studies indicate that a considerable portion of the human genome may be transcribed, often in unexpected places. ⊞

Global Identification of Transcribed Sequences in the Human Genome

4

Elucidating the transcribed regions of the genome constitutes a fundamental aspect of human biology, yet this remains an outstanding problem. To comprehensively identify these sequences, a series of high-density oligonucleotide tiling arrays were constructed representing sense and antisense strands of the entire non-repetitive sequence of the human genome. Transcribed sequences were located across the genome via hybridization to complementary DNA samples, reverse-transcribed from polyadenylated RNA obtained from human liver tissue. In addition to identifying many known and predicted genes, this experiment found 10,595 transcribed sequences not detected by other methods. A large fraction of these are located in intergenic regions distal from previously annotated genes and exhibit significant homology to other mammalian proteins.

introduction

The prevailing gene structures in many organisms consist primarily of coding sequences with few and short intervening regions, and thus their characterization is largely straightforward. In contrast, mammalian genes often contain many short exons interspersed with very large introns, making the identification of coding sequences difficult; a comprehensive and accurate map of human coding sequences therefore does not exist. Functional assays are expected to be essential for the identification of coding segments and verification of predicted genes.

In principle, tiling microarrays offer the ability to comprehensively investigate the transcribed regions of any species using an unbiased approach. Recently, various microarray technologies have been applied to assess genome-wide transcription in bacterial and plant genomes (Selinger *et al.* 2000, Tjaden *et al.* 2002, Yamada *et al.* 2003), as well as transcription over human chromosomes 21 and 22 (Kapranov *et al.* 2002, Rinn *et al.* 2003). Each of these methods identified many previously unannotated features, noting a high degree of novel transcription beyond that expected by existing gene annotation data. These studies clearly demonstrated the merit of the microarray approach to the problem of large-scale transcript mapping; however, until now the large size of mammalian genomes has precluded the construction of genome-wide high-resolution tiling arrays.

A HUMAN GENOME TILING ARRAY

Using maskless photolithographic DNA synthesis technology, 134 high-density oligonucleotide microarrays were constructed to represent 1.5Gb of non-repetitive genomic DNA from each strand of the human genome. A total of 51,874,388 36-nucleotide (nt) probes, positioned every 46nt on average, were selected to interrogate sense and antisense strands of the genome and synthesized at a feature density of 390,000 probes per array (Figure 4.1). To measure global transcriptional activity, the arrays were hybridized to fluorescence-labeled cDNA reverse-transcribed from triple-selected polyadenylated [poly(A)$^+$] liver tissue RNA pooled from several individuals.

A pilot study was first performed to test the reproducibility of the platform. Multiple arrays were probed with cDNA samples derived from identical and independent labeling reactions, producing technical replicates having r^2 correlations between 0.90 and 0.95, indicating that the experiments are highly reproducible. To further reduce the effect of potential variation across individual cDNA samples, pooled reverse transcription products of 20 separate labeling reactions were used to probe the genome.

Sequence processing. The template sequences used to design the microarrays were derived from NCBI Build 31/UCSC HG13 (UCSC Human Genome Project Working Draft, November 2002 assembly) of the human genome sequence assembly (Lander *et al.* 2001, Venter *et al.* 2001). Each chromosome sequence was screened for repetitive elements and low-complexity DNA with RepeatMasker (Smit & Green, unpublished) in sensitive mode, in conjunction with the RepBase collection of repetitive sequence elements (Jurka 2000). Additional low-complexity sequence filtering was performed with the NSEG program (Wooton & Federhen 1996) using a minimum segment length of 21 nucleotides, trigger complexity of 1.4, and extension complexity of 1.6.

Oligonucleotide probe selection. Following sequence processing, the remaining 1.5Gb of non-repetitive DNA was analyzed to identify optimal hybridization probes according to several criteria: 1) nucleotide frequency informa-

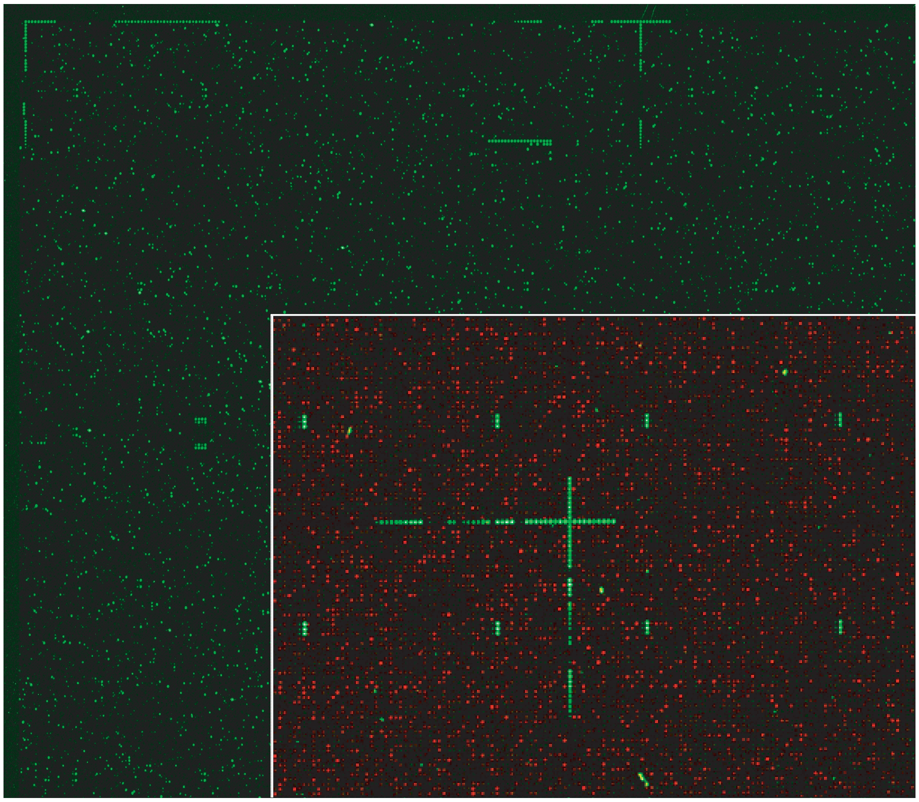


Figure 4.1. Fluorescence micrograph of high-density maskless photolithographic tiling arrays The arrays are designed to contain 396,000 36mer features in a 14 × 17.4 mm^2 area, where each feature occupies approximately 15μm. A parallel sample comprising end-labeled 24mer control oligos, shown in contrasting color, are hybridized to an embedded set of features (inset). These are used to align the array image in software, facilitating the automated extraction of experimental fluorescence data.

tion, calculated to determine the uniqueness of every 36mer in the genome; this measure was used to select probes that occur less than five times on average to reduce the potential for cross-hybridization; 2) intra-oligo alignment scores, used to exclude sequences that could form a loop with a stem greater than seven bases; 3) other sequence-dependent factors such as length, extent of complementarity and overall base composition. A total of 51,874,388 36mer oligonucleotide probes were selected to represent both sense and antisense strands of the non-repetitive DNA at an average resolution of 46nt (probes are spaced every 10nt on average). For DNA synthesis purposes, each selected probe was designed not to use more than 75% of the maximum number of cycles required to synthesize an oligomer of that length. For 36mer oligonucleotides this cutoff was 108 cycles. Virtual masks describing the layout of each microarray design were developed with the ArrayScribe program (NimbleGen Systems).

Microarray construction. Glass slides were incubated in 10% sodium hydroxide and silanized in a 2% bis(2 hydroxyethyl)-aminopropyltriethoxysilane solution (v/v; United Chemical Technologies). High-density oligonucleotide arrays were fabricated as described (Sing-Gasson *et al*. 1999, Nuwaysir *et al*. 2002, Albert *et al*. 2003). Briefly, 2-nitrophenyl propoxycarbonyl (NPPOC) photolabile DNA synthesis reagents (Hasan *et al*. 1997) were used in conjunction with Expedite DNA synthesizers (Applied Biosystems). Maskless Array Synthesis (MAS) units (NimbleGen Systems) were connected to the DNA synthesizers to manufacture custom arrays using photolabile phosophoramidites (NPPOC-D-adenosine (N6-tac) β-cyanoethylphosphoramidite, NPPOC-D-cytidine (N4-Isobutyryl) β-cyanoethylphosphoramidite, NPPOC-D-guanosine (N2-ipac) β-cyanoethylphosphoramidite, NPPOC-D-thymidine-β-cyanoethylphosphoramidite) obtained from Proligo. After synthesis on the MAS was completed, the base-protecting groups were removed in a solution of ethylenediamine:ethanol (1:1 v/v) (Aldrich) for two hours. The arrays were rinsed with water, dried and stored desiccated until use.

Sample preparation and labeling. Triple-selected human liver tissue poly (A)$^+$ RNA pooled from several individuals was obtained from Ambion. First-strand cDNA was generated using M-MLV reverse transcriptase (RNase H$^-$) with equimolar concentrations of oligo(dT) primers and random decamers. The reactions were carried out at 42°C for 2 hours in the presence of amino allyl-dUTP to facilitate the secondary labeling of an amine-reactive fluorescent conjugate. Following reverse transcription the products were heated at 95°C for 5 minutes to denature the RNA:DNA hybrids and heat-inactivate the reverse transcriptase, after which the RNA template was hydrolyzed via incubation with NaOH at 65°C for 15 minutes. Reverse transcription products from 20 separate reactions were produced in this manner and pooled to reduce technical variability between samples. The cDNAs were precipitated in ethanol:isopropanol (1:1 v/v) and resuspended in 0.1M $NaHCO_3$ to facilitate coupling of Alexa Fluor 555 NHS esters (Molecular Probes) to the reactive groups of the amino allyl-dUTPs. Following incubation at room temperature for 1 hour, the labeled cDNAs were purified with CyScribe GFX glass fiber spin columns (Amersham Bioscience) and isopropanol-precipitated.

Microarray hybridization. Microarrays were probed with 2µg–3µg labeled cDNA in 320µL hybridization buffer (50mM MES, 0.5M NaCl, 10mM EDTA, and 0.005% (v/v) Tween-20) for 20 hours at 50°C. Hybridizations were performed in disposable adhesive chambers (Grace BioLabs) in a hybridization oven with constant agitation. After hybridization, the arrays were washed on an orbital platform in non-stringent buffer (6× SSPE, 0.01% [v/v] Tween-20) for 10 minutes at room temperature, then in stringent buffer (100mM MES, 0.1M NaCl, 0.01% Tween-20) for 30 minutes at 45°C. This was followed by a 5-minute wash in non-stringent buffer and a 4-minute wash in 0.2× SSC. Microarrays were dried with compressed nitrogen gas. Images were acquired with an Axon 4000B laser scanner at 5µm resolution and intensity data were extracted with NimbleScan software (NimbleGen Systems). A more detailed hybridization protocol is included as Appendix E.

DATA ANALYSIS

Detecting annotated genes. To correlate fluorescence intensity values with meaningful chromosomal features, the oligonucleotide probe coordinates were aligned with current gene annotation data, using the RefSeq (Pruitt *et al.* 2000) and Ensembl (Hubbard *et al.* 2002, Birney *et al.* 2004) databases. Alignment of the fluorescence intensities to the chromosomal coordinates of many known genes shows strong agreement between hybridization signals and annotated exons (Figure 4.2).

To systematically determine the number of annotated genes detected with this approach, a simple statistical method was devised for scoring the observed transcriptional activity of annotated genes. Each probe is assigned a value of one

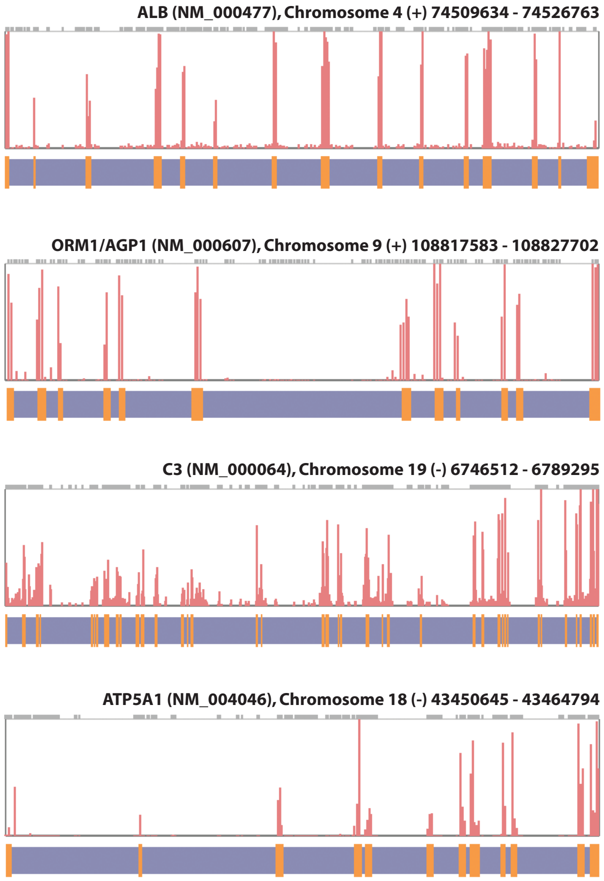


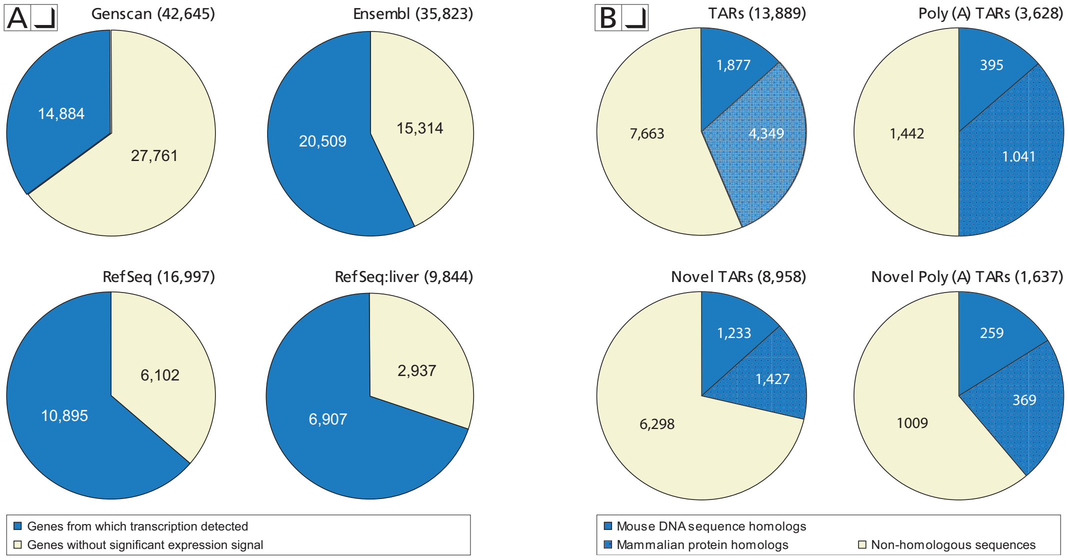
Figure 4.2. Annotated genes aligned with microarray fluorescence intensities. Comparison of the gene structures with intensity data shows strong agreement with expected exon-intron boundaries. The upper two examples illustrate uniform representation across the entire gene, whereas the lower two examples show a slight 3′ bias inherent in reverse-transcription labeling of RNA. Grey segments at the top of each graph indicate the coverage of oligonucleotide probes tiled across non-repetitive regions of each respective chromosome.

if its fluorescence intensity is greater than the median intensity of all probes on the array, and zero otherwise. For a given gene, the expected count of one's within annotated exons follows a binomial distribution; an unusually high count of one's therefore yields low *P*-values (sign test). Genes having *P*-values < 0.05 were regarded as demonstrating positive hybridization. This measurement essentially compares the fluorescence intensity of each probe within a gene against the median probe intensity across the entire microarray to determine whether they are significantly different.

A total of 16,997 annotated genes from RefSeq, 35,823 genes from Ensembl, and 42,645 genes predicted by Genscan (Burge & Karlin 1997) were scored in this manner. Based on these criteria, transcription was detected from 64%

(10,895), 57% (20,509), and 35% (14,884) of genes in each data set, respectively (Figure 4.3A). These results agree with the expectation that fewer genes should be experimentally detected from annotation data sets that include putative genes predicted by homology or ab initio methods, as opposed to a curated collection of characterized genes. Nonetheless, these results provide the first genome-wide experimental confirmation that many of the predicted genes are transcribed, suggesting that they are functional. A subset of 9,844 RefSeq genes with corresponding UniGene (Wheeler *et al.* 2003) annotations that indicate transcription in liver tissue was also examined; 70% (6,907) of these were detected by tiling array analysis.

Figure 4.3. A) Proportion of genes detected from each of four annotation sources. The percentages of genes detected from each data set increase as the annotation shifts from solely ab initio predictions (Genscan) to fully characterized genes (RefSeq). **B**) Conservation between transcriptionally-active regions (TARS) and other mammalian sequences. Forty-one percent of TARS and 50% of poly(A)-associated TARS were found to be homologous, as were 29% and 39% of novel TARS from each category. A large number of TARS show significant similarity to known proteins (BLAST *e*-values 10^{-5}), suggesting that many of these may be functional elements. A subset of these exhibited sequence similarity to regions of the mouse genome when restricted to similar e-values (solid blue sections).



Mapping unannotated transcripts. In addition to detecting known and predicted genes, a primary goal of this study was to identify regions of the genome exhibiting novel transcription. Transcribed sequences outside of previously annotated exons are expected to correspond primarily to (1) unannotated exons from alternatively spliced messages, (2) under-represented 3′- and 5′-untranslated regions, (3) non–protein-coding RNA transcripts, and (4) novel transcripts coding for functional proteins.

To scan for these sequences, aggregate transcription units were considered consisting of at least five consecutive probes exhibiting fluorescence intensities in the top 90th intensity percentile, and the genomic coordinates of which lay within a 250-nt window. These were compiled from throughout the genome and their locations compared relative to those of annotated gene components (Figure 4.4).

A total of 13,889 transcription units, ranging in size from 209nt to 3,438nt, were identified in the genome by these criteria; 400 are expected under the null hypothesis of zero transcription. One-third (4,931) correspond to previously annotated exons; the remaining 8,958 are new transcribed sequences that are referred to as transcriptionally active regions, or TARs (see Chapter 3). A total of 1,566 TARs are located within previously annotated introns on the same strand, raising the possibility that they correspond to previously unknown exons. How-

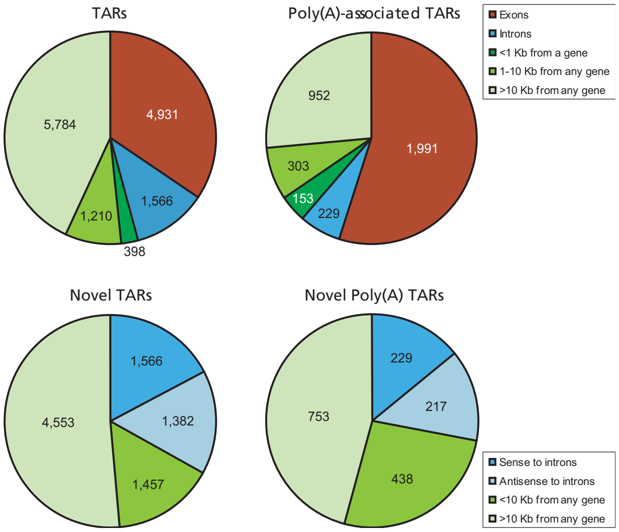


Figure 4.4. Distribution of TARs relative to annotated genes. Occupancy within gene components and proximity to known genes are depicted for all TARs (upper charts) and for novel TARs that lie outside annotated exons (lower charts). Most of the novel TARs are located more than 10kb from any previously annotated gene, suggesting that these correspond to distinct transcribed sequences.

ever, an equal number of TARs (1,529) lie on the antisense strand of introns, indicating that many of the intronic TARs likely represent novel transcription units. Over half of all TARs were found to be distal to annotated genes (defined here as greater than 10kb from any gene), indicating the presence of an additional 5,784 transcribed elements that are apparently unrelated to known genes.

An independent set of criteria was used to identify TARs in which probe hybridization intensities were correlated with the presence of a polyadenylation signal 3′ of the active region. Transcription units of (exactly) five consecutive probes were considered with fluorescence intensities in the top 80th intensity percentile, appearing in windows of 250nt where the 3′ region contains or lies near a polyadenylation signal. These are required to appear downstream of the 15th nucleotide of the 3′ oligonucleotide in the transcribed region; an additional 51 (46 + 5) downstream nucleotides are included in the calculation to ensure full coverage of the sequence. Instances of "AATAAA" sequences are designated type I, and "ATTAAA" type II.

An additional 3,628 TARs were identified using this method; 100 such instances are expected to occur at random in the genome. Most (1,991) lie within annotated exons, whereas 952 are located more than 10kb from any annotated gene. Of the 1,371 type I and 674 type II poly(A) sequences identified within exons of known genes, 94% (1,289) of type I and 90% (607) of type II instances occur in the 3′ exon of the gene in question, a strong indication of the effectiveness of this approach. The fraction of poly(A) TARs distinct from annotated exons (1,637), combined with the 8,948 novel TARs identified above, yields a total of 10,585 new transcribed sequences throughout the genome.

confirmation of novel transcripts

To validate the transcription of identified TARs with an independent method, reverse transcription polymerase chain reaction (RT-PCR) assays were performed using human liver poly(A)$^+$ RNA, targeting 48 poly(A)-associated and 48 non–poly(A)-associated TARs. PCR primer pairs were designed for each region and the reactions were carried out with the RETROscript system (Ambion); 0.25μg poly(A)$^+$ RNA was reverse transcribed for each sample. An identical aliquot of

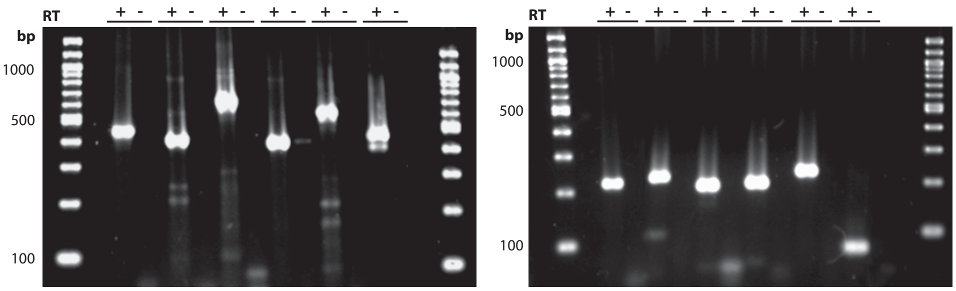


Figure 4.5. RT-PCR validation of TAR sequences. A group of variable-length TARS (*left*) between 400bp and 650bp is shown opposite a group of approximately equal-length poly(A)-associated TARS (*right*). PCR products are loaded adjacent to their corresponding negative control samples.

each reaction mixture was used as a minus-reverse transcriptase control. PCR products were electrophoresed on 2.5% fine-resolution agarose gels; 90 of the 96 (90%) displayed bands of the appropriate size in a single-pass assay with no detectable band in the negative control, and 10 of the positive amplicons produced multiple bands in addition to the targeted product (Figure 4.5).

As a further validation, the novel TARS were compared against data derived from the second phase of the Kapranov *et al*. (2002) transcript mapping experiment on chromosomes 21 and 22. Over 41% of TARS match the transcribed fragments, or "transfrags," identified in their study. Because of the highly stringent selection of TARS described here, many low-abundance transcripts are not identified by these criteria and an appreciable false-negative rate is expected.

EVIDENCE OF CODING POTENTIAL BY EVOLUTIONARY CONSERVATION

Novel TARS were then compared with other mammalian DNA sequences to assess their potential for coding functional elements. BLAST (Altshul *et al*. 1990) searches revealed that many TARS are homologous to sequences in the mouse genome. Of the 8,958 novel TARS, 24% (2,185) produced BLAST alignments with *e*-values less than 10^{-5}, with most of these (1,486) having *e*-values less than 10^{-20}. This compares with 39% (5,419) of the initial set of 13,889 TARS (i.e., novel TARS and those corresponding to exons of known genes) that produced BLAST

	Total	Exons	Introns	< 1 kb	1-10 kb	> 10 kb
TARS	13,889	4,931	1,566	398	1,210	5,784
Poly(A)-associated TARS	3,628	1,991	229	153	303	952
Type I (AATAAA)	2,393	1,371	137	105	187	593
Type II (ATTAAA)	1,325	674	101	51	123	376
	BLAST: mouse genome			BLAST: mammalian proteins		
	$1e^{-5}$	$1e^{-10}$	$1e^{-20}$	$1e^{-5}$	$1e^{-10}$	$1e^{-20}$
TARS	5,419	4,747	3,761	4,349	4,008	3,311
Poly(A)-associated TARS	1,515	1,247	936	1,307	1,198	995
Type I (AATAAA)	1,044	862	637	905	830	685
Type II (ATTAAA)	517	423	328	436	401	340

Table 4.1. Distribution of TARS relative to annotated genes and conserved sequences. Many TARS (40%) correspond to known exons; however, a significant fraction (38%) are located more than 10kb from any previously annotated gene. BLAST results compare TARS to mammalian protein sequences and to the mouse genome. A total of 6,934 (40%) of all TARS are homologs to the mouse genome (*e*-value $\leq 10^{-5}$), with 5,656 (32%) homologous to protein sequences (25 to 30% of TARS belong to both categories), providing evidence for possible functional roles in humans.

scores with *e*-values less than 10^{-5}; of these, total of 3,761 had *e*-values less than 10^{-20}. Similarly, 32% (532) of the 1,637 novel poly(A)-associated TARS yielded BLAST alignments with *e*-values less than 10^{-5}, with 342 less than 10^{-20} (Figure 4.3B). Of the initial set of 5,419 TARS and 1,515 poly(A)-associated TARS found to be homologous to sequences in the mouse genome, a respective 27% (1,488) and 21% (321) from each category are located more than 10kb from any previously annotated gene.

In addition to assessing the degree of genome conservation, mouse proteins were aligned with TAR sequences that were translated in all possible reading frames (Table 4.1). A total of 16% (1,427) and 12% (1,091) of novel TARS produced BLAST matches less than 10^{-5} and 10^{-20}, respectively, compared with 31% (4,329) and 24% (3,311) of the total number of TARS with matches below these *e*-values. Higher percentages of poly(A)-associated TARS were found to be homologous to mouse proteins: 23% (369) of the novel subset and 36% (1,307) of the total set of poly(A) TARS matched protein sequences with *e*-values less than

10^{-5}, with 19% (305) and 27% (995) in each category having *e*-values less than 10^{-20}. Thus, although many TARs are expected to encode proteins, novel TARs generally exhibit a lesser degree of sequence conservation than those intersecting known genes. This is particularly true for poly(A)-associated TARs owing to the higher degree of conservation of protein-coding sequences relative to 3′-untranslated regions.

To estimate the number of TARs potentially arising from the cross-hybridization of mRNA transcripts to sequences elsewhere in the genome, 9,408 novel TARs that additionally do not lie antisense to annotated exons were compared to the library of human cDNA sequences in the Ensembl database. Only 11%

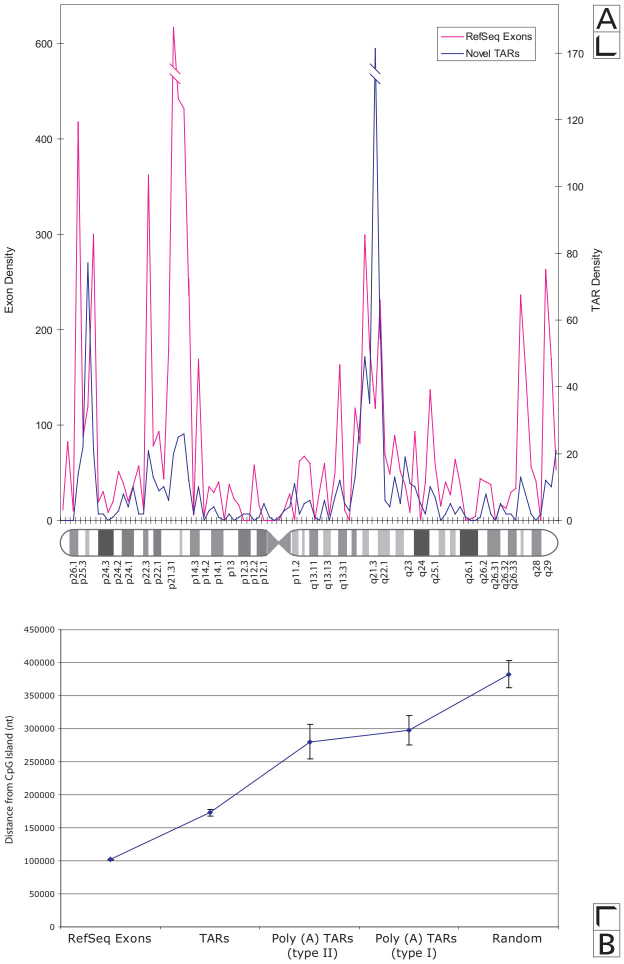


Figure 4.6. (**A**) Density plot of RefSeq-annotated exons across human chromosome 3 compared with the density of novel transcriptionally active regions (TARs). The distribution of novel TARs is similar to that of annotated exons, indicating that they are colocated with genes on a global scale. Units on the abscissa are given in 1-Mb intervals.

(**B**) Average distances to the nearest upstream CpG island for all RefSeq exons, novel TARs, novel poly(A)-associated TARs, and 1000 randomly selected locations in the genome. The distribution of novel TARs is similar to that of RefSeq exons, whereas the random locations are the most distant from CpG islands. As expected, poly(A)-associated TARs are located at intermediate distances because they correspond primarily to 3′ exons.

(1,034) were found to share at least 95% identity over a stretch of 150nt. Of the remaining 8,374 nonhomologous novel TARS, 347 intersect the genomic coordinates of processed pseudogenes (Harrison *et al*. 2002, Zhang *et al*. 2003), providing evidence for possible pseudogenic transcription.

GLOBAL COMPARISON TO CHROMOSOMAL FEATURES

Finally, the distribution of TARS was compared to the locations of known genes and CpG islands. A density plot relating TARS and RefSeq-annotated exons along chromosome 3 (Figure 4.6A) revealed that TARS are located in the same regions as known genes. The density of TARS is correlated with the distribution of RefSeq-annotated genes along each chromosome (Pearson correlation coefficient $r^2 = 0.35$, $P < 0.002$). Comparison of distances to the nearest upstream CpG island indicates that the relative locations of novel TARS distal to annotated genes are similar to those of RefSeq exons, whereas the distal poly(A)-associated TARS are located farther away, which is expected because most of these should correspond to the 3′ ends of genes (Figure 4.6B). The distances of all distal TARS to CpG islands were found to be significantly less than those of randomly selected locations ($P < 0.0001$).

SUMMARY

These findings demonstrate that it is possible to use high-resolution oligonucleotide microarrays for the comprehensive analysis of the human genome. Because many transcribed sequences are located in distinct intergenic regions distant from known genes, their precise mapping can only be accomplished using genomic tiling arrays in which nearly all of the non-repetitive DNA is available for hybridization to RNA transcripts. Several bacterial artificial chromosome (BAC) clone-based genomic tiling arrays have been developed for comparative genomic hybridization (CGH) studies in humans (Buckley *et al*. 2002, Iskanian *et al*. 2004); however, the identification of short transcription units requires interrogating the genome sequence at a resolution of tens of base pairs, a measurement that is not possible to obtain with BAC technology.

This analysis identified thousands of new transcribed regions throughout the human genome and confirmed the transcription of predicted genes on a global scale. These results provide a draft expression map for the entire genome, revealing a much more extensive and diverse set of expressed sequences than was previously annotated. Conservation between many of the novel transcribed sequences and well-characterized mouse proteins provides strong evidence that a large number of them are likely to encode functional transcripts. Many conserved transcribed sequences are located in regions distal to known genes, and a notable fraction of these are of sufficient length to encode proteins of 300 or more amino acids. The remainder may encode small proteins, untranslated exons, or RNAs whose functions have yet to be elucidated (Mattick 2003, Kampa *et al*. 2004). These latter RNAs may serve alternate regulatory or structural roles and await detailed characterization. ⊞

Conclusion and Future Directions: Regulatory Pathway Discovery 5

The studies described thus far exploit the parallel nature of microarray experiments to address the problem of human genome annotation. Through the development of genomic DNA tiling arrays, large-scale transcript mapping experiments can refine existing gene models and reveal the presence of thousands of novel transcribed sequences. A second emerging application of microarrays involves the identification of regulatory elements in genomic loci. This approach attempts to map the DNA sequences recognized by transcriptional regulators via hybridization of chromatin bound by the cognate proteins in vivo. Through the identification of transcription factor binding sites, tiling arrays can be applied to the study of gene regulation on a global scale. This study concludes by reviewing current results in this area, then demonstrating how transcription factor binding analysis can be linked with gene expression data to synthesize large-scale portraits of genetic function. In examining the complex integration of various genomic data, the focus will shift from human genetics to the yeast model *Saccharomyces cerevisiae*; here, the ability to perform such analyses is greatly enhanced by the existence of many experimental data sets. Ultimately, the approaches described at the end of the chapter will be amenable to the human system as more comprehensive data becomes available.

global identification of regulatory elements

Transcription factors are regulatory proteins that bind DNA to modify chromatin or recruit components of the transcriptional apparatus, ultimately manifesting or repressing the expression of their target genes. Identifying the genes regulated by an organism's complement of transcription factor proteins is central to our understanding of diverse cellular processes. It is therefore highly desirable to attain a comprehensive inventory of the *cis*-regulatory sequences that constitute the promoter elements to which a given transcription factor binds. Although the in vitro DNA-binding sequences of many factors have been established to varying degrees of accuracy, in vivo binding can be affected by a multitude of complex determinants. These include variations in local chromatin structure and accessibility, interaction of transcriptional activators with remote enhancer elements, and involvement of ancillary proteins. Thus, a given factor can bind to different locations in vivo to coordinate the transcription of different sets of genes, depending on the cellular conditions in which it is expressed.

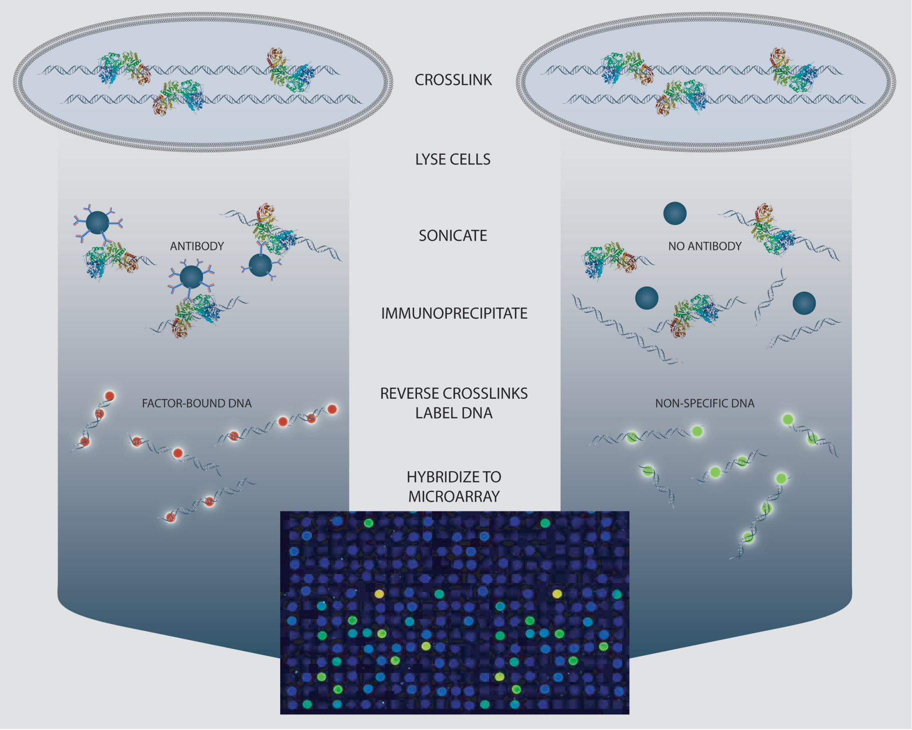
Until recently, characterizing the interactions of DNA-binding proteins with the genome was possible only on a single-gene basis, primarily through in vivo footprinting studies. In concert with other experimental protocols, microarrays can now be used to study the behavior of transcriptional activators in a manner analogous to gene expression analysis. Since functional binding sites of transcription factors are expected to occur primarily within intergenic regions, microarray-based analyses of gene regulation have emerged with the tandem development of genomic tiling arrays. In shifting the selection of DNA sequences away from the exclusive representation of genes, tiling arrays facilitate the unbiased mapping of transcription factor binding sites on an unprecedented scale.

The most widely adopted procedure involves the hybridization of chromatin immunoprecipitated (ChIP) DNA to a genomic DNA tiling array, commonly referred to as ChIP-chip (Horak & Snyder 2002a, Lieb 2003). In this approach, protein-DNA interactions in cells expressing the factor of interest are fixed in situ with a crosslinking agent, typically formaldehyde (Solomon & Varshavsky 1985). Nuclear extracts are isolated and the transcription factor-bound DNA is sonicated to reduce the size of the fragments to be analyzed. The factor is then immunoprecipitated, either with antibodies against the native protein or via an epitope tag fused to the transcription factor gene. The crosslinks are reversed with heat treatment and fluorescence-labeled samples are prepared following the purification of the immunoselected chromatin fragments (Figure 5.1).

The labeled DNA is then hybridized to a microarray in parallel with a negative control sample. This can be derived from genomic DNA, or consist of an identical sample precipitated either in the absence of antisera or with pre-immune sera. The resulting data can therefore be treated like those generated by a two-channel differential gene expression experiment, where fluorescence intensity ratios are computed after normalizing the signals from the two channels (Figure 5.2A). The main analytical difference between a differential expression experiment and a ChIP-chip experiment is that in the latter case, statistical outliers are expected to occur only in the fluorescence channel corresponding to the immunoprecipitated sample (Figure 5.2B). A significant increase in fluorescence intensity therefore corresponds to the enrichment of a specific population of DNA fragments in excess of those represented in the control sample, and are assumed to have hybridized to chromatin fragments containing transcription factor-bound sequences.

Once identified, transcription factor-bound sequence fragments can be mapped to their genomic loci and their positions compared with existing gene

Figure 5.1. ChIP-chip protocol for microarray-based chromatin profiling. Protein-DNA interactions within cells expressing a transcription factor of interest are treated with formaldehyde to promote in vivo crosslinking. This is followed by lysis, shearing of the genomic DNA, and immunoselection of protein-DNA complexes from nuclear extracts using antibodies against the transcription factor. The immunoprecipitated DNA is purified, fluorescence-labeled, and hybridized to a tiling or intergenic microarray in parallel with a negative control sample. The control may be derived either from immunoprecipitations performed in the absence of antibodies or with control antibodies, from a deletion strain or cell line, or from genomic DNA.



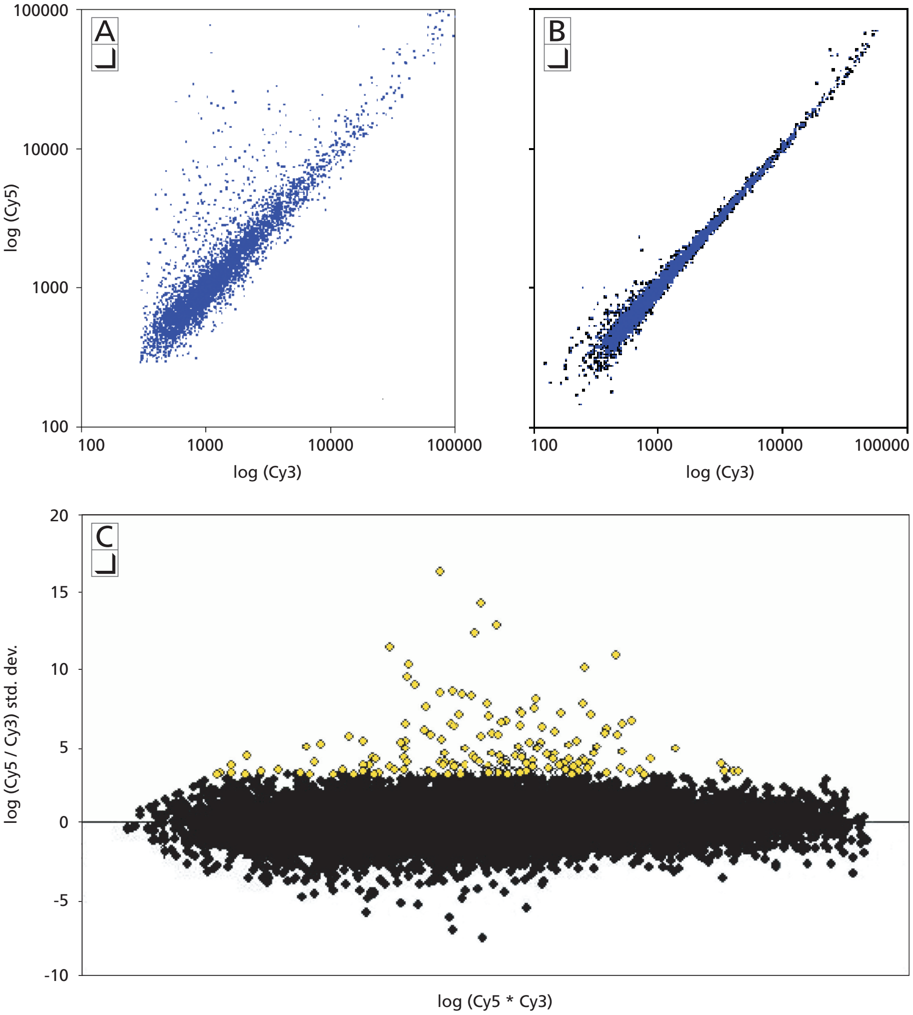


Figure 5.2. Scatter plots of chromatin-immunoprecipitated DNA versus a negative control sample (**A**), and a negative-versus-negative control experiment (**B**). As seen in the first example, the enrichment of transcription factor-bound DNA produces an increase in fluorescence intensity at hybridizing microarray features. **C**) Statistical outliers are typically identified as features whose $\log_2$ intensity ratios exhibit fold change increases above a given threshold, or exceed several standard deviations from the normalized intensity distribution (Quackenbush 2002, Luscombe *et al*. 2003).

annotation (Figure 5.3A). The total number of DNA fragments enriched via immunoprecipitation is usually a superset of those involved in gene regulation. Some factors recognize highly specific promoter sequences and associate with chromatin infrequently, whereas others may bind constitutively to many sites throughout the genome. A number of transcription factors have been observed to bind to promoter regions in clusters, such that several binding events constitute a fewer number of regulatory loci. Additionally, since the immunoprecipitated chromatin fragments are double-stranded, either strand of the denatured sample becomes available to anneal with complementary array sequences. It is therefore impossible to distinguish on which strand the factor's promoter sequence lies from this experiment alone. Instead, one must consider both strand orientations equally when analyzing the data, observing the proximity of binding sites to annotated genes to determine which are likely to be involved in regulatory function.

Because transcription factor binding alone does not necessarily indicate the locations of functional promoters, evidence to support regulatory function must be accumulated by integrating other experimental data. Differential gene expression, easily observed through microarray analysis, can reveal which genes are affected in response to the stimuli under which a transcription factor is induced. This information is superimposed with binding site data to reveal where DNA binding occurs on the chromosome relative to the locations of differentially-expressed genes (Figure 5.3B). Ultimately, careful consideration is required to interpret the results of these experiments in a biologically meaningful way.

The ChIP-chip approach was first explored in the yeast model. Ren *et al*. (2000) used a microarray of PCR products representing 6,361 yeast intergenic regions to map the genome-wide binding locations of Gal4 and Ste12. Their analysis revealed 3 novel gene targets in addition to those previously known to be regulated by Gal4, and 29 genes specifically regulated by Ste12. Shortly thereafter, Iyer *et al*. (2001) developed a similar approach, constructing a PCR-product array of approximately 6,700 intergenic and promoter regions to map the genome-wide binding locations of the transcription factors SBF and MBF during the G1/S transition of the mitotic cell cycle. They identified over 200 genes regulated by the factors, finding SBF and MBF to be implicated in cell wall biogenesis and DNA replication, respectively. Lieb *et al*. (2001) then used the ChIP-chip method to map the binding sites of Rap1, previously associated with telomere modification and mating-type transcriptional repression. As an essential gene,

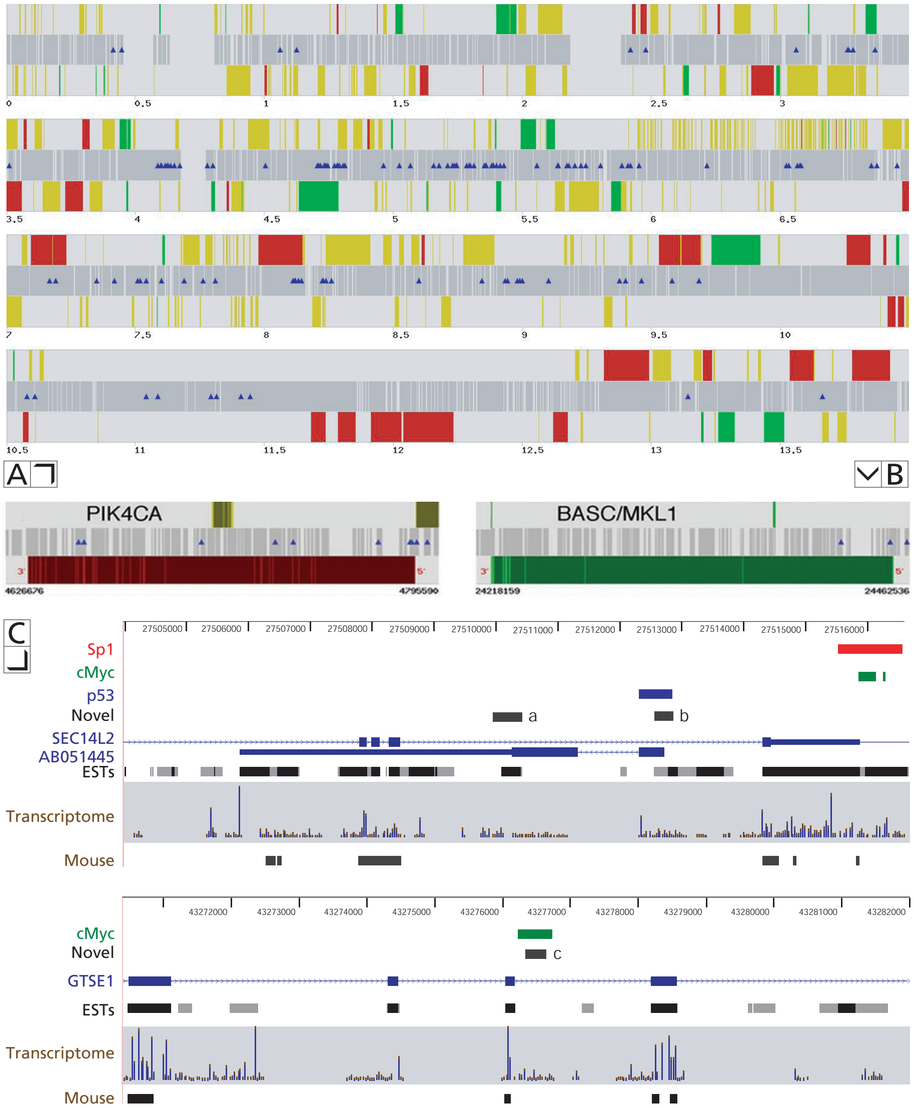


Figure 5.3. A) Binding distribution of CREB over a segment of human chromosome 22 illustrating transcription factor binding within coding and intergenic regions as well as clusters of binding sites upstream of annotated genes (Euskirchen *et al*. 2004). Binding sites are marked as blue triangles across the chromosome; up-regulated, down-regulated, and non-differentially-expressed genes appear in red, green and yellow, respectively. **B**,**C**) Examples of NF-κB (**B**), Sp1, c-Myc and p53 (**C**) binding adjacent to differentially expressed genes on chromosome 22 (Martone *et al*. 2003, Cawley *et al*. 2004). Although some DNA binding sites are located in or near canonical promoter regions 5′ of annotated genes, others lie in gene-dense regions where a single regulatory element may control the expression of multiple targets (**B**) as well as novel transcribed sequences (**C**; unannotated transcription units are labeled a–c).

mutations to Rap1 that affect DNA binding are lethal and thus the regulatory characterization of the factor is recalcitrant to conventional genetic analyses. Binding site location analysis identified approximately 5% of yeast genes regulated by Rap1, implicating the factor in key cellular processes such as protein biosynthesis and energy metabolism.

An advantage of these experiments is that since protein-DNA interactions are fixed in vivo, the experiment can be performed under varying cellular conditions to assess regulatory activity in different environmental contexts. For example, the Ren *et al*. study measured enrichment in transcription factor binding site occupancy in response to changes in carbon source and mating pheromone, relating the proximity of these sites to genes whose expression levels changed under similar conditions (i.e., genes whose promoters were bound by Gal4 and induced in galactose, and those bound by Ste12 and induced by α factor).

ChIP-chip analysis was first extended to a mammalian system by Horak *et al*. (2002b). Using a PCR product microarray representing the 75-kb human β-globin locus, the binding distribution of the hematopoietic lineage-specific transcription factor GATA-1 was measured in erythroleukemic K562 cells. Only a single region within the β-globin locus had been previously known to contain GATA-1 binding sites; however, the factor was observed to bind a region upstream of the γG gene in addition to confirming the results of previous observations. In a study focused on 5′ regulatory sequences, Ren *et al*. (2002) developed a promoter-proximal microarray containing PCR-amplified genomic loci directly upstream of 1,444 human genes. The array was used to identify ChIP-enriched sequences bound by the transcriptional activator E2F1 during the G1/S phase transition of the cell cycle, and the repressor E2F4 during quiescence.

Subsequent to these analyses, ChIP-chip has been used to survey transcription factor binding over entire human chromosomes. Martone *et al*. (2003) mapped the binding distribution of NF-κB (p65) across chromosome 22 in HeLa cells induced in the presence of tumor necrosis factor (TNF-α). Using the same microarray platform, Euskirchen *et al*. (2004) investigated CREB binding in the cAMP-inducible JEG-3 choriocarcinoma cell line. Both studies revealed a wide distribution of binding sites across the chromosome relative to annotated genes (Figure 5.3A). Particularly interesting was the finding that many binding sites are located proximal to 3′ ends of genes and within annotated introns, challenging the traditional view that transcription factors act exclusively in promoter regions directly upstream of transcriptional start sites. Using oligonucleotide arrays, Cawley *et al*. (2003) surveyed the binding of c-Myc, Sp1 and p53 in Jurkat and HCT-1116 cells over chromosomes 21 and 22. As was reported in the NF-κB and CREB studies, transcription factor binding was observed at many locations upstream of 5′ ends, proximal to 3′ ends, and internal to annotated genes (Figure 5.3C). Coincident binding of Myc and Sp1 was also found to occur at numerous locations, suggesting the possibility that some of their target genes are co-regulated by the two factors.

Unlike the yeast experiments which employed an intergenic array to assess transcription factor binding, the human chromosome studies surveyed binding over all of the non-repetitive DNA in an unbiased fashion. The arrays were designed to represent both coding and intergenic regions irrespective of existing gene annotation, as was the case for the previous chromosome-wide surveys of RNA transcription. In comparing the locations of enriched ChIP fragments to annotated genes as illustrated in Figure 5.3A, it becomes clear that a complete representation of the genome sequence is required to fully characterize the binding distribution of a given transcription factor. Although ChIP-chip experiments performed with arrays that represent promoter-proximal regions (Ren *et al*. 2002, Li *et al*. 2003, Gao *et al*. 2004, Odom *al*l. 2004) or CpG islands (Weinmann *et al*. 2002, Mao *et al*. 2003, Wells *et al*. 2003) can provide a wealth of valuable information about transcription factor association with canonical regulatory loci, the resulting data is likely to be incomplete. A given factor may bind alternate promoters, remote enhancers or other locations that are quite distant from transcriptional start sites. This is a particularly significant issue when such experiments are applied to mammalian genomes, which exhibit an unusually small percentage of annotated coding sequence relative to the amount of intergenic DNA.

An alternative technique developed to analyze DNA binding in *Drosophila* is known as DNA adenine methyltransferase identification, or DamID (van Steensel & Henikoff 2000, van Steensel *et al*. 2001). In this approach, a transcription factor gene is fused to *E. coli* DNA adenine methyltransferase (Dam), which methylates the N^6 position of the adenine nucleotide in the sequence GATC. Methylation will occur at or around these sites in vivo, marking the locations of transcription factor binding. The genomic DNA is then subjected to DpnI endonuclease digestion and unmethylated chromatin fragments are removed by incubating with DpnII. The remaining DNA is amplified, labeled and hybridized to a genomic DNA tiling array. Sun *et al*. (2003) used this technique to map the DNA binding locations of gaf and the heterochromatin protein hp1, using a pcr-product tiling array representing approximately 3mb of chromosome 2 containing the *Adh–cactus* region as well as the 85-kb *82F* locus on chromosome 3.

synthesis of transcriptional regulatory networks

Naturally, some targets of transcription factors are themselves genes that encode regulatory proteins. If the target genes of each successive transcription factor in a regulatory cascade are determined, these relationships can be linked to form a circuit whose topology describes their combined activity. Recently, graph theoretic methods have been applied to associate transcription factors with their target genes in complex regulatory networks. In this model a directed graph is produced having a scale-free topology, where transcription factors tend to localize in hubs of regulatory control (Shaw 2003). Some transcriptional regulators have been shown to modulate the expression of a disproportionately large number of genes, following power-law behavior with respect to the number of outgoing connections originating from a given factor (Babu *et al*. 2004). Conversely, the number of genes regulated by multiple factors has been shown to decrease exponentially relative to the number of transcriptional regulators involved (Guelzim *et al*. 2002). Key transcription factors are therefore likely to be essential genes whose deletion would produce a lethal phenotype (Yu *et al*. 2004) and may constitute points of vulnerability in complex regulatory systems.

Lee *et al*. (2002) explored the construction of gene regulatory networks after performing Chip-chip analysis on 106 yeast transcription factors to determine their genome-wide binding sites using an intergenic array. By observing common patterns in the data they were able to identify several basic regulatory motifs that describe transcription factor-target relationships (Figure 5.4A). These include single-input, multi-input and autoregulatory motifs, feedforward loops, multicomponent loops, and regulatory chains. The binding site data produced by these experiments was later integrated with gene expression data by Bar-Joseph *et al.* (2003) to identify 106 distinct regulatory modules, based on the classification of 68 transcription factors and 655 genes. In an extension of the study by Iyer *et al.*, Horak *et al.* (2002c) investigated the gene targets of nine transcription factors regulated by SBF during the G1/S cell cycle transition using the ChIP-chip approach, using the data to build a transcription factor network (Figure 5.4B). Functional annotation linked to the transcription factor-target relationships revealed a complex regulatory cascade governing cell growth and differentiation.

Once derived from experimental data, transcription factor-target relationships can be combined with gene expression profiles to analyze complex functional pathways. Where transcription factor-target relationships are available, known associations can be incorporated from public databases such as TRANSFAC (Matys *et al.* 2003); others can be derived from experimental data or predicted by comparing gene expression profiles between transcription factors and putative target genes. For example, Qian *et al.* (2003) was able to use support vector machines (SVMs) to predict the regulatory targets of 36 yeast transcription factors based on gene expression data. A total of 3,419 regulated genes were predicted through observation of both co-expressed and time-shifted expression profiles. Yu *et al.* (2003) integrated yeast gene expression profiles with an extensive transcriptional regulatory network constructed from ChIP-chip and other experimentally-derived transcription factor binding data. They used the network to identify global expression patterns in the relationships between transcription factors and the genes they regulate, accounting for inverted and time-shifted behavior. Genes belonging to the same regulatory motif were often found to be co-expressed, exhibiting higher expression levels when multiple transcription factors were involved.

Gene regulatory networks are not static entities but dynamic structures that are expected to undergo significant topological changes in response to variations in cellular physiology. Luscombe *et al.* (2004) integrated gene expression and transcription factor binding data from a variety of sources to construct an elabo-

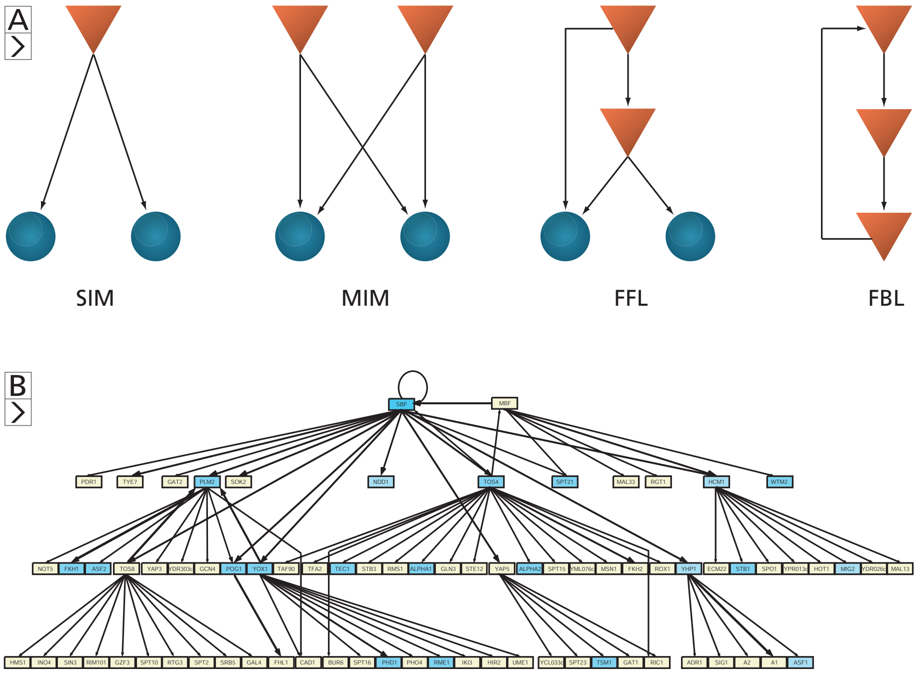
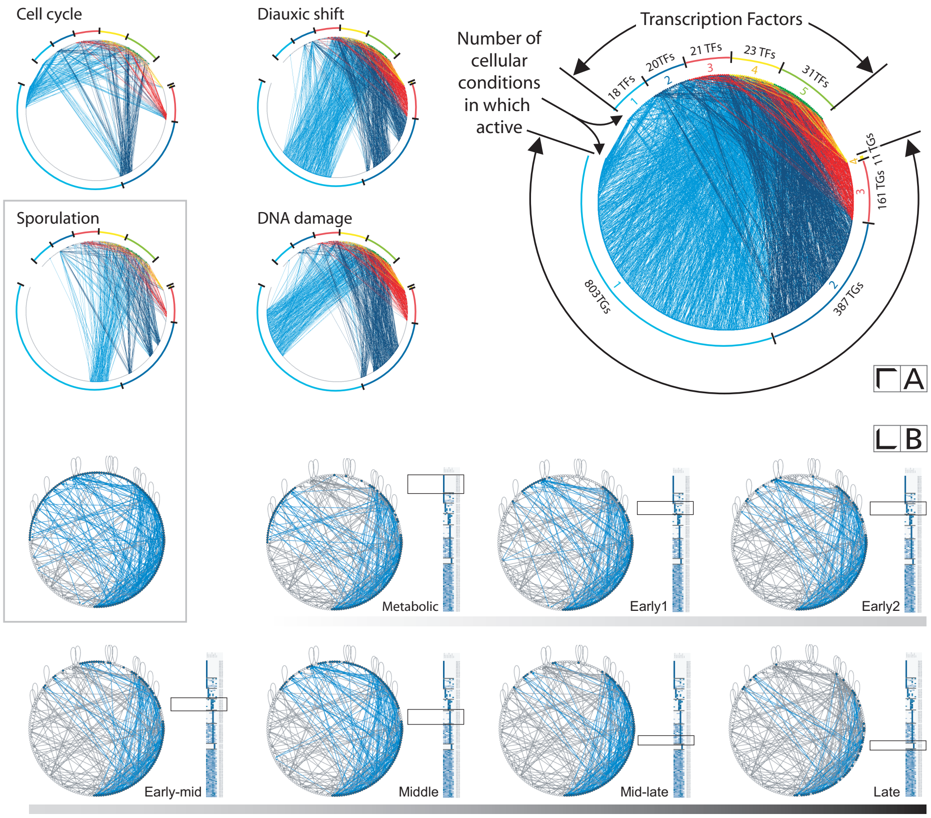


Figure 5.4. A) Several common gene regulatory network motifs, identified through genome-wide investigation of transcription factor binding (after Lee *et al*. 2001). Transcriptional regulators are represented as triangles and target genes as spheres. Depicted from left to right are the single-input motif (SIM), multiple-input motif (MIM), feed-forward loop (FFL), and feedback loop (FBL). Not pictured are the autoregulatory motif and regulatory chain, which are derivative of the single-input motif. **B**) Transcription factor network describing a cascade of regulatory control downstream of the cell cycle regulators SBF and MBF during the G1/S transition (Horak *et al*. 2002c).

rate network comprising 7,074 regulatory interactions in yeast. Examining the occurence of the motifs defined by Lee *et al*., they further analyzed expression profiles to determine which regulatory subnetworks are active under different environmental conditions such as the cell cycle, diauxic shift, sporulation, DNA damage and stress response (Figure 5.5A). During response to external stimuli, regulatory cascades were shown to be fairly simple and involve few feedback interactions. More complex circuitry was observed during the cell cycle and sporulation, which appear to require multiple regulatory stages involving highly interconnected transcription factor relationships (Figure 5.5B). The study also characterized the influence of regulatory hubs in the system, finding that many hubs involve multi-functional transcription factors that regulate essential cellu-

Figure 5.5. A) Complex transcriptional regulatory networks derived from yeast ChIP-chip and gene expression data, illustrating the global static network as well as condition-specific sub-networks. Transcription factors and target genes appear as nodes on the outer perimeter of each graph, respectively. Edges denote regulatory interactions and are colored according to the number of cellular conditions in which they have been identified (Adapted from Luscombe *et al*. 2004). **B**) Detailed analysis of dynamic gene regulation during the multi-stage transcriptional program of sporulation (N. Luscombe, personal communication). The complete set of sporulation-associated interactions is represented in the upper leftmost graph (boxed), followed by a series of graphs highlighting the specific regulatory sub-networks activated in successive stages of the pathway. Combinatorial transcription factor usage occurs in distinct subsections of the network, as evidenced by differential gene expression patterns observed at each stage.



lar processes. Despite their biochemical significance, the majority of regulatory hubs were observed to be transient in nature, influencing widespread transcriptional activity in some conditions but not others. Through the examination of local and global regulatory pathways at several levels of complexity, this work presents a seminal perspective of the large-scale temporal dynamics of genetic control.

summary

The limited feature density of early microarray platforms led to an initial focus on gene-based sequence representation and, consequently, on comparative gene expression profiling. As array fabrication technology continues to improve, the commensurate increase in feature density has enabled the construction of microarrays able to cover large regions of eukaryotic chromosomes, spanning intergenic as well as coding sequences. Novel applications of tiling arrays are constantly emerging for the large-scale characterization of chromosome dynamics. White *et al*. (2004) recently used tiling arrays to measure DNA replication timing across human chromosome 22 during S phase of the cell cycle, through differential hybridization of early- and late-replicating chromatin from lymphoblast and fibroblast cells. The study identified 24–26 regions of early and late DNA replication, ranging in size from 100kb to 2Mb and generally associated with defined cytological bands. A total of nine chromosomal regions exhibited differential replication timing between the two cell types. Additionally, a strong correlation was observed between early replication and the expression of novel transcribed regions having low coding potential.

Unlike gene-directed approaches, tiling array experiments enable the discovery of novel genetic elements. In particular, they are becoming increasingly important for the identification of previously unannotated transcribed sequences and the large-scale analysis of gene regulation via the unbiased interrogation of the genome. Various tiling array platforms have recently been adopted as primary discovery tools by the ENCODE Project Consortium, in an effort to provide an in-depth transcriptional and regulatory characterization of 44 select regions of the human genome (Feingold *et al*. 2004).

Computational methods to relate gene expression with transcription factor binding can produce complex networks from which higher-order regulatory mechanisms may be derived. It is expected that our ability to elucidate functional relationships from high-throughput genomic data will be enhanced through the combination of these experimental and computational techniques. This integrated approach represents a powerful analysis methodology, able to generate an unprecedented view of the transcriptional regulatory program of the cell. ⊞

Repetitive Elements in Eukaryotic Genomes

A

Repeat sequences in eukaryotes are roughly classified as moderately and highly repetitive. Moderately repetitive elements generally belong to one of three broad categories: retroelements, transposable elements, and functional genes, while highly repetitive elements comprise telomeres, satellites, minisatellites and microsatellites.

Retroelements, or retroposons, comprise endogenous retroviral and nonretroviral sequences that have been integrated into the genome and mobilize via an RNA intermediate. Those belonging to the viral family include Ty elements in *Saccharomyces cerevisiae*, the copia elements in *Drosophila melanogaster*, and long interspersed nucleotide elements (LINES) in mammalian genomes. Most LINES belong to the L1 family, and encode both an endonuclease and a reverse transcriptase to mediate the copying and insertion of new L1 elements following transcription by RNA polymerase II. Interestingly, 89% of functional human genes contain at least one LINE element.

Occasionally, endogenous mRNA transcripts may be reverse-transcribed and inserted elsewhere in the genome by the L1 proteins. These sequences usually exhibit the features of mature mRNA, including spliced intron-less sequences and poly(A) tails; for this reason they are referred to as processed pseudogenes.. Because they are integrated outside the genomic context of their original promoters, pseudogenes are usually inactive.

The nonviral family is characterized by noncoding sequences that apparently originated as reverse-transcribed RNA, similar to the events that give rise to processed pseudogenes. One prominent category consists of short interspersed nucleotide elements (SINES), which differ from LINES in that they originate from the products of RNA polymerase III. A significant fraction of the moderately repetitive DNA in the human genome – accounting for approximately 21% of the total genomic DNA – consists of nonviral sequences belonging to the Alu family. Alu repeats consist of approximately 300bp sequences interspersed with nonrepetitive DNA, named after the restriction enzyme that cleaves the DNA at a single site 170bp into the sequence.

Transposable elements, or DNA transposons, are mobile sequences able to propagate themselves from one site in the genome to another. Transposons code for proteins that are able to mediate the copying of the sequence directly, without the use of an RNA intermediate. Functional gene families may also occur in high copy number, organized as tandem arrays or dispersed throughout the genome. Typical examples of tandem arrays of functional genes are histone, rRNA and tRNA genes; dispersed gene families include the globins and immunoglobins.

Telomeric repeats and satellite sequences constitute highly repetitive DNA. Telomeres comprise 250–1,000 copies of short repeats at the ends of chromosomes appended enzymatically by telomerase, protecting the ends of chromosomes from degradation during replication. Satellite DNA is a generic description for tandem arrays of very highly repetitive sequences spanning up to several kilobases in length. They are typically located in heterochromatic regions and centromeres, appearing in clusters up to 1mb in size. Minisatellites are tandemly repeated arrays of moderately-sized (9bp to 100bp, but usually about 15bp) sequences, and generally have mean array lengths of 500bp to 30kb. They are found in euchromatic regions of the genomes of vertebrates, fungi and plants and are highly variable in array size. Microsatellites are moderately repetitive sequences, composed of arrays of short (2bp–6bp) repeats found in vertebrate, insect and plant genomes. Human chromosomes collectively contain at least 30,000 microsatellite loci found in euchromatic regions of the genome.

B Proof of Optimality for the Linear-time, Constant-space Algorithm

To see why this algorithm produces an optimal partitioning, we proceed by induction on the length of the inspected sequence and assume that the algorithm has been correct prior to the ith element (i.e., the partitioning up to k is optimal, and no decision can be made so far on the window between k+1 and i). We will show only one case of the proof; the rest is very similar.

Without loss of generality assume that the last known region R, currently extending up to k, is an included region. Consider the case when $D < -C$, in which the algorithm will terminate R at i and start an excluded region at i+1. Suppose however that there is an optimal partitioning P with score s_P which extends R at least up to position i, contrary to what the algorithm yields. Define a new partitioning N, identical to P except for the window between k+1 and i, which in N is part of an excluded region, and let us compute its score s_N. There are two possibilities: if in P the included region ends at i and an excluded region starts at i+1, then N has the same number of partitions as P, but one region boundary has been shifted from i in P to k in N. Hence s_N is equal to s_P plus the difference in the scores on the window between k+1 and i; these scores are exactly $V[I_{(k+1)..i}]$ under the partitioning P and $V[X_{(k+1)..i}]$ under N, therefore:

$$\mathrm{s_N} = \mathrm{s_P} - V[I_{(k+1)..i}] + \mathrm{V}[X_{(k+1)..i}] = s_P - D > s_P \ ,$$

since the difference D is negative by our assumption. The second possibility is that the included region starting in P extends after i; this means that in N this region is subdivided into two regions by the excluded region from k+1 to i, so N contains one more included region than P. Hence

$$s_N = s_P - V[I_{(k+1)..i}] + V[X_{(k+1)..i}] - 2C \ > s_P$$

again by the assumption that $D < -C$. Thus, in both cases we have $s_N > s_P$, which contradicts the assumption of optimality of the partitioning P.

Other partitionings which terminate the included region earlier than i can be shown similarly suboptimal by the following observation. Since by assumption the algorithm postponed the decision until i, the difference D must be between $-C$ and 0 at all intermediate points. For the case when the algorithm postpones the partitioning decision, the proof of correctness is to construct two sequences sharing the same prefix up to i but requiring different optimal partitionings of the window from k+1 to i, which shows that indeed no decision guaranteeing optimality can be made at i. In other words, a partitioning solution not satisfying the tests in the algorithm cannot be optimal. ■

Optimal Sequence Tiling

C

This part considers several variations of the following basic tiling problem: given a sequence of real numbers with two size bound parameters, we want to find a set of tiles of maximum total weight such that each tiles satisfies the size bounds. A solution to this problem is important to a number of computational biology applications such as selecting genomic DNA fragments for PCR-based amplicon microarrays (Chapter 2) and performing homology searches with long sequence queries. The goal is to design efficient algorithms with linear or near-linear time and space in the normal range of parameter values for these problems. For this purpose, the solution to a basic online interval maximum problem is introduced via a sliding window approach; it is then shown how this solution can be used in a non-trivial manner for many of the tiling problems introduced. NP-hardness results are also discussed.

PROBLEM STATEMENTS AND NEW ALGORITHMS

This section defines several tiling problems and describes algorithms for their solution. Typical parameter values are outlined for DNA microarray design and homology search applications.

Problem statements. Based on the applications discussed previously, several related tiling problems are introduced and formalized. The following notations are used:

- $[i, j)$ denotes the set of integers $\{i, i+1, \ldots, j-1\}$;
- $[i, j] = [i, j+1)$;
- $f[i, j)$ and $f[i, j]$ denote the elements of an array f with indices in $[i, j)$ and $[i, j]$, respectively.

The tiling problems presented here build upon the basic genome tiling algorithm developed in Chapter 2, referred to as the genomic sequence tiling (*GTile*) *problem* and described as follows. The input consists of an array $c[0, n)$ of real numbers and two integer size parameters l and u. A subarray $B = c[i, j)$ is called a *block* of *length* $j - i$ and *weight* $w(B) = \Sigma_{k=i}^{j-1} c_k$, the weight of a set of blocks is the sum of their weights and a block is called a *tile* if its length belongs to $[l,$

$u]$. The goal is to find a set of pairwise disjoint tiles with the maximum possible weight. The tiling problems of interest in this paper are variations, restrictions and generalizations of the GTile problem specified by a certain combinations of the following items:

Compressed versus uncompressed input data:

This is motivated by a simple binary classification of the non-repetitive regions of the genome sequence from their repetitive or low-complexity counterparts. Now all entries of $c[0,\ n)$ is either x or $-x$ for some *fixed* $x > 0$. Hence, the input sequence can be more efficiently represented by simply specifying beginnings and endings of *blocks of identical values*[8]. In other words, we can compress the input sequence $c[0,n)$ to a sequence of integers (indices) $S[0,\ m + 1)$ such that

- $S_0 = 0,\ S_{\mathrm{m}} = n + 1,\ S_1 \geq S_0$ and $S_{\mathrm{i}} > S_{i-1}$ for all $i \in [1,\ m]$;
- each element of $c[S_{2j},\ S_{2j+1})$ is x for all j;
- each element of $c[S_{2j-1},\ S_{2j})$ is $-x$ for all j.

The input size $m + 1$ of such a compressed input data is typically *significantly smaller* than n. As a result, we can get significantly faster algorithms if we can design an algorithm for compressed inputs with a running time nearly linear in m. Furthermore, this also allows one to develop efficient hybrid approach to solving the tiling problems: first use a crude binary classification of the regions to quickly obtain an initial set of tiles and then refine the tiles taking into consideration the relative importances of the high-complexity elements.

Unbounded versus bounded number of tiles:

Another important item of interest is when the number of tiles that may be used is at most a given value t, which could be considerably smaller than the number of tiles used by a tiling with no restrictions on the number of tiles. This is motivated by the practical consideration that the capacity of a microarray as obtainable by current technology is bounded.

Overlapping versus non-overlapping tiles:

To enhance searching sequence databases for sequence homology searches to allow for the case when potential matches can be found at tile boundaries, it may be useful to relax the condition of disjointness of tiles by allowing two tiles to share at most p elements for some given (usually small) $p > 0$. However, to ensure that we do not have too many overlaps, we need to *penalize* them by subtracting the weight of each overlapped region from the sum of weights of all tiles, where the *weight* of each overlapped region is the sum of the elements in it. In other words, if $\mathcal{T}$is the set of tiles and $\Re$ is the set of elements of **C** that belong to more than one tile in $\mathcal{T}$, then the weight is $\Sigma_{T \in \mathcal{T}} w(T) - \Sigma_{c_i \in \Re} c_i$.

One dimensional versus *d*-dimensional:

Generalization of the GTile problem in d dimensions has applications in database designs and related problems [,,,][2]. In this case, we are given a d-dimensional array **C** of size $n_1 \times n_2 \times \cdot \times n_d$ with $2d$ size parameters l_1, $l_2, \ldots, l_d$, u_1, $u_2, \ldots, u_d$, a tile is a rectangular subarray of **C** of size $p_1 \times p_2 \times \ldots \times p_d$ satisfying $l_i \leq p_i \leq \mathrm{u}_i$ for all i, the weight of a tile is the sum of all the elements in the tile and the goal is again to find a set of tiles such that the sum of weights of the tiles is maximized.

Only those combinations of the above four items which are of importance in tiling applications are examined. To simplify exposition, unless otherwise stated explicitly, the GTile problem we consider is *1-dimensional* with *uncompressed* inputs, *unbounded* number of tiles and *no overlaps*. In addition to the previously defined notations, unless otherwise stated, the following notations and variables are used with their designated meanings throughout the rest of the Appendix: $n + 1$ is the number of elements of the (uncompressed) 1-dimensional input array $c[i,j]$, $n_1 \leq n_2 \leq \ldots \leq n_d$ are the sizes of the dimensions for the d-dimensional input array, $w(\mathcal{T})$ is the weight for a *set* of tiles $\mathcal{T}$, t is the given number of tiles when the number of tiles is bounded and p is the maximum overlap between two tiles in 1-dimension. Finally, all logarithms are in base 2 unless stated otherwise explicitly.

Typical parameter values for microarray design and homology search.

$n+1$ (the DNA sequence length):

Although the sizes of sequenced eukaryotic genomes range from 12 Mb (for the budding yeast *Saccharomyces cerevisiae*) to 3.4 Gb (*H. sapiens*), these exist as separate chromosomes that are treated as individual sequence databases by the tiling algorithms. Eukaryotic chromosomes range in size from approximately 230 kb (*S. cerevisiae* chromosome I) to 256 Mb (human chromosome 1), with the average human chromosome being 150 Mb in size.

l and u (lower and upper bounds for tile sizes):

In computing an optimal set of tiles for microarray design, tile sizes can range from 200 bp to 1.5 kb. Sequence fragments below 200 bp become difficult to recover when amplified in a high-throughput setting. An upper bound of 1.5 kb balances two factors: (1) obtaining maximal sequence coverage with a limited number of tiles, and (2) producing a set of tiles which are small enough to achieve sufficient array resolution. In practice the average tile size is 800 when l and u are set to 300 and 1500, respectively. For some instances of the homology search problem it may be desirable to extend the upper bound from 1.5 kb to 2 kb, representing the typical size of processed eukaryotic messenger RNA transcripts.

p (maximum overlap between two tiles):

For microarray applications, tiles are disjoint; that is, the overlap parameter p is 0. However, searching sequence databases for homology matches can be enhanced by introducing a maximum overlap of $p \leq 100$ nucleotides for the case when potential matches can be made at tile boundaries.

t (maximum number of tiles, when the number of tiles is bounded):

In selecting tiles for microarray applications, t can be specified to limit the number of sequence fragments considered for PCR amplification.

For mammalian DNA where repeat content (and subsequent sequence fragmentation) is high, we can expect the non-repetitive sequence nucleotides to cover $n/2$ sequence elements; the desired number of tiles to be computed will thus be $[n/2]$ divided by $[(u+l)/2]$ (the average of u and l). For homology search problems t is unbounded.

m (size of compressed input):

It is difficult to give an accurate estimate of the number of non-repetitive sequence fragments in the target sequence following repeat screening since it varies greatly with the organism. Typically, human chromosomes end up having between 2 to 3 times as many non-repetitive sequence fragments (before processing) as there are final tiles (after processing), that is, m is roughly between $2t$ and $3t$. In other words, in practice m may be smaller than n by a factor of at least 600 or more.

Algorithm synopsis. Table C.1 summarizes the main algorithms; all of these are either new or direct improvements of any previously known algorithms. All of the methods use simple data structures such as a double-ended queues and are therefore easy to implement. The techniques used for many of these tiling problems in one dimension use the solution of an online interval maximum (OLIM) problem. Section 3 discusses the OLIM problem together with an efficient solution using a windowing scheme reminiscent of that in Datar *et al*. (2002). However, the main consideration in that study was the reduction of space because of the online nature of their problems; here the primary concern is time complexity since sequence processing is off-line in nature (and hence space for storing the entire input is always used). Moreover, the windowing scheme used here is somewhat different from that in Datar et al. since we need to maintain multiple windows of different sizes and data may not arrive at evenly spaced time intervals.

Algorithm version	Time $O()$	Space $O()$	Approximation ratio	Theorem
Basic	n	n	Exact	2
Overlap is from a s-subset of $[0, \delta]$, $\delta < \frac{\ell}{2}$	sn	n	Exact	4
Compressed input	$m\frac{\ell}{u-\ell}$	$m\frac{\ell}{u-\ell}$	Exact	5
Number of tiles given	$\min\{n \log \frac{n}{\ell}, nt\}$	n	Exact	7
d-dimensional	$\left(\left(\frac{u}{\ell}\right) \varepsilon\right)^{4\left(\frac{u}{\ell}\right)^2 \varepsilon^2} M\varepsilon^2$	M	$\left(1 - \frac{1}{\varepsilon}\right)^d$	16
d-dimensional, number of tiles given	$tM + dM \log^\varepsilon M$ $+dN\frac{\log N}{log \log N}$	M	$\left(\Pi_{i=1}^{d-1} \left(\lfloor 1 + \log n_i \rfloor\right)\right)^{-1}$	16
	$M^{(2^\varepsilon-1)^{d-1}+1}\ dt$	$M^{(2^\varepsilon-1)^{d-1}+1}\ dt$	$\left(\Pi_{i=1}^{d-1} \left(\lfloor 1 + \frac{\log n_i}{\varepsilon} \rfloor\right)\right)^{-1}$	16

Table C.1. Tiling algorithm summary. The parameter $\varepsilon > 1$ is any arbitrary constant. An s-subset is a subset of s elements. For the d-dimensional case, $M=\Pi_{i=1}^{d} n_i(u_i-l_i+1)$, $N=\max_{1 \leq i \leq d} n_i$ and $[u/l]=\max_i[u_i/l_i]$. For biological applications $p \leq 100 < [l/2] \ll n$, $t \cong [n/(u+l)]$, $m \ll n$ and $[l/(u-l)] < 6$. The column labeled "Approximation Ratio" indicates whether the algorithm computes the optimal solution exactly or, for an approximation algorithm, the ratio of the total weight of our tiling to that of the optimum.

COMPUTING THE ONLINE INTERVAL MAXIMUM VIA SLIDING WINDOW

This section describes an online interval maximum (OLIM) problem which is used to design many of the remaining algorithms. Datar et al. (2002) considered a restricted version of the OLIM problem in the context of maintaining stream statistics in the sliding window model and briefly mention a solution for this problem. The problem in its most general form can be stated as follows.

Input: (1) a sequence $a[0, n)$ of real values in increasing order where each value a_i is an *argument* or a *test* (possibly both), (2) 2α real numbers $l_1, u_1, l_2, u_2, \ldots, l_\alpha, u_\alpha$ with $0 < l_1 < u_1 < l_2 < u_2 < \ldots < l_\alpha < u_\alpha$ and (3) a real value function g defined on the *arguments*.

Output: for every test number a_k compute the maximum b_k of the α quantities $b_{k,1}, b_{k,2}, \ldots, b_{k,\alpha}$, where $b_{k,i}$ is given by $b_{k,i} = \max g(a_j) : a_k - u_i \leq a_j < a_k - i$ and a_j is an argument.

Online limitations: read the elements of the sequence $a[0, n)$ one at a time from left to right and compute b_k (if a_k is a test) *before* computing $g(a_k)$.

Theorem 1 *The OLIM problem can be solved in* $O(n_1\beta + n\alpha)$ *time using* $O(n_1 + \alpha)$ *space, where* n_1 *and* n_2 *are respectively the numbers of arguments and tests in the input and* β *is the maximum time to compute* $g(x)$ *for any* x.

Proof. A queue Q_i is maintained for each $i \in [1,\alpha]$. When the pair $(a_k,g(a_k))$ is computed for each argument a_k, it is stored in the abovementioned queues such that the following invariant is satisfied for each Q_i: Q_i stores a *minimal* set of argument-value pairs such that for some future test a_m that has not been read yet, it is possible to have $b_{m,i}=g(x)$ for some $(x, g(x))$ in Q_i. After reading each a_k, Q_i can be maintained using the following two rules:

Rule 1: Remove from Q_i every $(x, g(x))$ such that $x < a_k - u_i$. The validity for this rule is obvious from the definition of $b_{m,i}$ and the fact that the sequence $a[0, n)$ is in increasing order.

Rule 2: Let p be the *smallest index* of an argument such that $a_k - u_i \le a_p < a_k - l_i$ and $(a_p, g(a_p)) \notin Q_i$. Remove from Q_i every $(x,g(x))$ such that $g(x) \le g(a_p)$ and then insert $(a_p, g(a_p))$ in Q_i. Rule 2 is valid because for $m \ge k$ if $b_{m,i}$ is computed as the maximum value of a set that contains a removed $(x, g(x)$, then this set must also contain $(a_p, g(a_p))$. This is true because $x < a_p$ and therefore rule 1 would remove $(x, g(x))$ earlier.

If all the needed insertions are performed to Q_i using Rule 2, then the following holds: if $j < m$ and $(a_j, g(a_j))$ and $(a_m, g(a_m))$ are simultaneously present in Q_i, then $g(a_j) > g(a_m)$. Consequently, the maximum of the g-values in Q_i is contained in the *oldest* pair in Q_i. These observations allow each Q_i to be maintained as a double-ended queue where $\texttt{front}(Q_i)$ stores the maximum of all the g-values of the elements in Q_i (needed to compute $b_{m,i}$, and to perform Rule 1), while $\texttt{tail}(Q_i)$ has the minimum g-value (needed to perform Rule 2). The following is a high-level pseudocode of the main parts of the proposed algorithm for Q_i when reading a_k assuming all the parameters have been appropriately intialized:

```
 (* windowing scheme for Qi *)
 (* notations *)
 Let q be the least index of an argument of a that has
been read
  but has not considered for insertion to $Q_i$ yet;
 for each argument $a_i$, $g(a_i)$ is calculated once and stored;

 (* algorithm for $Q_i$ *)
 (* currently read number is $a_k$ *)

 (* Execute Rule 1 *)
 $(x,g(x))$ ← front($Q_i$);
 while $((Q_i \neq \varnothing)$ and $(x < a_k - u_i)$
  remove front($Q_i$)
 endwhile;

 (* Execute Rule 2 *)
 while $((a_p < a_k - l_i))$
  $(y,g(y))$← tail($Q_i$)
  while $((Q_i \neq \varnothing)$ and $(g(y) \leq g(a_p))$
   remove tail($Q_i$)
  endwhile;
  add $(a_p,g(a_p))$ to $Q_i$ at its tail;
  $p$ ← $p+1$
 endwhile;

 (* calculate $b_{k,i}$ if necessary *)
 if ($a_k$ is a test) then $(z,g(z))$ ← front($Q_i$); $b_{k,i}=g(z)$
```

For each queue Q_i and each a_k it is necessary to check if Q_i must be updated using either of the two rules. This takes $O(n\alpha)$ time. Because each argument is inserted to and deleted from a queue exactly once, these updates takes $O(n_1\alpha)$ total time. For every test the maximum of the maxima of the queues is computed; this takes $O(n_2\alpha)$ time. ■

BASIC GENOMIC SEQUENCE TILING (GTILE)

Theorem 2 *The genomic sequence tiling (GTile) problem can be solved in $O(n)$ time using $O(n)$ space.*

Proof. Dynamic programming is used to reduce the GTile problem to OLIM. Subproblem k has $c[0,\ k)$ as input. Let m_k be the sum of weights of the tiles and $c[d_k,\ e_k)$ be the last tile in an optimum solution of subproblem k. If $m_k = m_{k-1}$, then subproblem k has the same solution as subproblem $k - 1$, otherwise this solution consists of tile $c[d_k,\ k)$ and the tiles in the solution of subproblem d_k. Let $s_k = w(c[0,\ k))$, hence $w(c[i,\ j)) = s_j - s_i$. It is trivial to compute s_k for all $k \in [0,\ n]$ in $O(n)$ time and space. Obviously, $m_k = 0$ for $0 \le k \le l$. For $k > l$, m_k and d_k can be computed recursively as follows:

```
let $i \in [k-u,k-l]\cap[0,\infty)$ be an index that maximizes $v_i=m_i+s_k-s_i$;
if $v_i > m_{k-1}$, then $m_k=v_i$, $d_k=i$ and $e_k=k$
 else $m_k=m_{k-1}$, $d_k=d_{k-1}$ and $e_k=e_{k-1}$
```

To complete the proposition on time and space complexity, it suffices to show how to compute m_k for every $k > l$ in a total of $O(n)$ time and space. For each k we can first search for $i \in [k-u,k-l]\cap[0,\infty)$ that maximizes $y_i=m_i-s_i$; then we know $v_i=y_i+s_k$. This is the OLIM problem with input array $a[0,n)$, $a_i=i$, each a_i is both an argument and a test, $\alpha = 1$, $l_1=l+1$, $u_1=u$ and $g(a_i)=m_i-s_i$. It is easy to recover an optimal tiling via the d_k and e_k values. ■

GTILE WITH OVERLAPS

This section consider the GTile problem when the overlap p between two tiles is an element of some s-subset A of $[0,\ \delta]$ with $\delta < [l/2]$. The constraint $\delta < [(l)/2]$ holds for biological applications since typically $p \le 100$ and $l \cong 300$. An important consequence of this constraint is the following observation:

Observation 3 *No c_i can belong to more than two tiles.*

Proof. Suppose that some c_i belongs to three tiles $c[b_1,\ e_1)$, $c[b_2,\ e_2)$ and $c[b_3,\ e_3)$ with $e_1 \le e_2 \le e_3$ and $b_3 < e_1$. Since each tile is of length at least l and the length of overlap between any two tiles is less than $[l/2]$, $b_1 \le b_2 \le b_3$. Now, $e_2-b_2=(e_2-e_1)+(\mathrm{e}_1-b_2) < (e_2-b_3)+(e_1-b_2) < [(l)/2]+[(l)/2]=l$, hence the tile $c[b_2,e_2)$ does not satisfy the size bounds. ■

Using the above observation, we can prove the following result.

Theorem 4 *The GTile problem with overlaps as described above can be solved in $O(sn)$ time using $O(n)$ space.*

Proof. Let $A \subseteq [0,p]$ be the s-subset for this problem. Some of the notations in the proofs of Theorems 3 and 4 are reused here. Let m_k' be the sum of weights of the tiles and $c[d_k', k)$ was the last tile in an optimum solution of subproblem k in which c_k was the ending of the last tile. For $k \leq l$, $m_k' = 0$. For $k > l$, m_k' and d_k' can be computed in the following manner:

```
let $i \in [k-u, k-l] \cap [0,\infty)$ be an index that
 maximizes $v_i = m_i + s_k - s_i$;
$m_k' = v_i$, $d_k' = i$
```

This can again be solved in $O(n)$ time and space via the OLIM problem in the same manner as in Theorem 4. Let $[h_k, f_k)$ be tile previous to the last tile $[d_k, k)$ in an optimum solution of subproblem k in which c_k was the ending of the last tile. If the last tile was overlapped by the previous to last tile by $a \in A$ elements in an optimum solution of subproblem k, then the total weight of the solution is $(s_k - s_{dk}) + m_{dk+a}' - (s_{dk+a} - s_{dk}) = s_k + m_{dk+a}' - s_{dk+a}$. Now, for each $k > l$, we can compute m_k, d_k, e_k and f_k in the following manner:

```
($v_i$ is the solution when the last tile $[i, k)$ was not overlapped)
 let $i \in [k - u, k - l] \cap [0, \infty)$ be an index that maximizes $v_i = m_i + s_k - s_i$

 ($\mu_{j,a}$ is the solution when the last tile $[j,k)$ was overlapped by $a$ elements)
 (computation of $\mu_{j,a}$ takes $O(s)$ time)
 let $j \in [k - u, k - l] \cap [0, \infty)$ and $a \in A$ be the indices that maximize $\mu_{j,a} =$
$s_k + m_{j+a}' - s_{j+a}$;

 if $m_{k-1} \geq \max\{v_i, \mu_{j,a}\}$ then $m_k = m_{k-1}$, $d_k = d_{k-1}$, $e_k = e_{k-1}$, $f_k = f_{k-1}$
 else
  if $v_i > \mu_{j,a}$ then $m_k = v_i$, $d_k = i$, $e_k = k$, $f_k = e_i$
  else $m_k = \mu_{j,a}$, $d_k = j$, $e_k = k$, $f_k = j + a$
```

By Observation 5, the last three tiles in the solution of subproblem k cannot have a common element and hence the solution is correct. Obviously, it suffices to find that $i \in [k{-}u,\ k{-}l] \cap [0, \infty)$ that maximizes $m_i - s_i$ and that $j \in [k - u + a,\ k - l + a] \cap [0, \infty)$ that maximizes $m_j' - s_j$. Each of them is can again be solved via OLIM problem. The tiles in an optimal solution can be recovered recursively via the d_i, e_i and f_i values. ■

GTILE WITH COMPRESSED INPUT

Theorem 5 *The GTile problem with compressed input data can be solved in* $O(\alpha m)$ *time using* $O(\alpha m)$ *space where* $\alpha = \lceil l/(u - l) \rceil$.

The remainder of this section outlines a proof the above theorem, reducing this problem to OLIM. A key idea in this reduction is to extend the definition of a tile to include unions of adjacent tiles. Then the problem can be redefined by requesting that tiles in an optimal solution are separated by entries that do not belong to any tile. Notice that the set of lengths that an extended tile may have is $\cup_{i=1}^{\infty}[il,\ iu] = (\cup_{i=1}^{\alpha-1}[il, iu]) \cup [\alpha l, \infty)$, where α is the smallest positive integer such that $(\alpha + 1)l \leq \alpha u + 1$. The legal length of any tile in any solution can be of α different kinds: the i^{th} kind (for $1 \leq i < \alpha$) is $[il,\ iu]$ and the α^{th} kind is $[\alpha l, \infty)$. The following lemma shows how to further restrict the beginnings and endings of extended tiles.

Lemma 6 *There is an optimum solution for the GTile problem with compressed input in which the (extended) tiles are only of the following kinds:*

$c[S_{2i},\ S_{2j+1})$, i.e., starting and ending with a full block of x's;

$c[S_{2i+1} - kl,\ S_{2i+1})$ or $c[S_{2i+1} - ku,\ S_{2i+1})$ for some positive integer $k \in [1, \alpha]$, i.e., ending with a full block of x's and the length of the tile is either a multiple of l or a multiple u;

$c[0,\ kl)$ or $c[0,\ ku)$ for some positive integer $k \in [1, \alpha]$.

Proof. If the length of a selected tile is neither a multiple of l nor a multiple of u, then it must be of type (i) since otherwise a tile of greater weight could be obtained by moving one of the ends. Now suppose that the length of a selected tile is a multiple of l or u and it does not end at some S_{2i+1}. If it ends with an x, the tile to the right can be shifted until it ends with the last x of this block of x's, or it coalesces with another tile in the solution and then the shifting process is again applied to this coalesced tile. If it ends with a $-x$, it can be shifted to the right with similar results. This type of shifting produces tiles of type (ii) unless after shifting the tile to the ledt it ends at c_0, producing tiles of type (iii). If the length of a tile of type (ii) or (iii) produced by the above shifting is greater than αu, then there are two possibilities: if the tile starts in a block of x's then we extend the tile until the begining of this block producing a tile of type (i); however if the tile ends in a block of $-x$'s the tile can be reduced in size by moving its begining to the right until either the length is αu or the tile ends with a full block of x's, whichever occurs earlier. ■

In this algorithm, we first need to generate all possible beginnings and endings for the extended tiles. This can be done in a straightforward manner:

- Generate the monotonically increasing sequence of indices of *all possible beginnings* of tiles: S_{2i} such that $2i \leq m$ [type (i) and iii)] and $S_{2i+1} - kL$ such that $0 < k \leq \alpha$ and $L \in \{l, u\}$ [type (ii)]. The total number of such beginnings is at most $\lceil [m/2] \rceil + \alpha m$. Duplicates are removed from the list.

- Generate the monotonically increasing sequence of indices of *all possible endings* of tiles: S_{2j+1} such that $2j + 1 \leq m$ [type (i) and ii)] and kL such that $0 < k \leq \alpha$ and $L \in \{l, u\}$. The total number of such endings is at most $\lceil [m/2] \rceil + 2\alpha$. Duplicates are removed from the list.

- These monotonically increasing sequences of beginnings and endings are then merged into one increasing sequence X of beginnings and endings.

The above steps can be carried out in $O(m\alpha)$ time and space. As in Theorem 4, define s_i to be $w(c[0, i))$ for $i \in \{S_0, S_1, \ldots, S_m\}$. It is easy to calculate all such s_i's in $O(m)$ time and space. The list X is then processed from left to right; the following is maintained when processing each element x of the list X:

(a)

If x is a beginning, then we maintain m_x, $v_x = m_x - s_x$ and *previous*$(x) = c[d_x, e_x)$, where m_x the maximum possible sum of entries of set of disjoint tiles of legal length that end *before* x and *previous*(x) is the last tile from such a solution. If z is the entry in X before x, then we need to set $m_x = m_z$, $v_x = m_z - s_x$ and *previous*(x)=*previous*(z); if x is the first element of X then we set $m_x = 0$ and $v_x = s_x$.

(b)

If x is an ending, then $v_z = y \in (\cup_{i=1}^{\alpha-1}[x-il, x-iu]) \cup [x-\alpha l, 0]\{v_y\}$. We set $m_x = (s_x - s_z) + m_z = s_x + v_z$, $v_x = m_x - s_x$ and *previous*$(x) = c[c_x, d_x) = c[z, x)$. Let x' be the entry in X previous to x (if such an entry does not exist, then simply set $m_{x'} = -\infty$ in the following formula). Then, if $m_x < m_{x'}$ then we set $m_x = m_{x'}$.

It is clear that the only nontrivial step is the computation of m_z in **(b)** above. This is however again the OLIM problem: the sequence $a[0, n)$ is the sorted list of beginnings and endings X $= (x_1, x_2, \ldots)$, each beginning is an argument, an ending is a test, $g(x) = v_x$ for argument x, $[l_i, u_i] = [il+1, iu]$ for $1 \leq i < \alpha$ and $[l_\alpha, u_\alpha] = [\alpha l + 1, n]$. The final solution has a total weight of m_n; to recover the tiling we start with *previous*(m_n) in the collection and recursively look at the previous values for the part of the optimal solution that ends before the beginning of *previous*(m_n).

Because we have $O(m\alpha)$ arguments, $O(m+\alpha)$ tests and α queues, a straightforward application of Theorem 3 shows that we use $O(m\alpha^2)$ time and $O(m\alpha)$ space. However, the algorithm can be designed with $O(m\alpha)$ running time as explained below.

For ease of counting in the analysis, assume that duplicates are not removed from the list of begining; this can only increase the running time. For analysis, the beginings will be partitioned into two sets: Γ_1 contains the $O(m)$ beginings of type S_{2i+1} and Γ_2 contains the remaining $m\alpha$ beginings. The endings are also partitioned into two sets: Δ_1 contains the 2α endings of the form kL for $0 < k \leq \alpha$ and $L \in \{l, u\}$ and Δ_2 contains the remaining $O(m)$ endings.

First, let us calculate the total time that we will take to check if one of the α queues in the OLIM problem need to be updated.

(a)

The $m\alpha$ beginings in Γ_2 can be partitioned into groups of 2α beginings, where each such group of 2α beginings are of the form $S_{2i+1}-kL$ for $0 < k \leq \alpha$, $L \in \{l, u\}$ and a distinct S_{2i+1}. However, notice that the two endings $S_{2i+1} - kL$ for a particular k need to be checked for insertion in only the queue Q_k in the algorithm for the OLIM problem. With this minor modification in the code of the algorithm for the OLIM problem, the total time to check if all appropriate queues need to be updated for the beginings in Γ_2 is $O(m\alpha)$.

(b)

For each begining in Γ_1 and each ending in Δ_2, we may need to check each of the α queues for update, hence the total time taken for these endings in $O(m\alpha)$.

(c)

For a pair of endings $kL \in \Delta_1$, we need to check only one possible entry v_0 for insertion into its k^{th} queue Q_k; the remaining queues need not be checked for update for these pair of endings. Hence, the total time taken for the endings in Δ_1 is $O(\alpha)$.

The total time taken for insertions and deletions of the beginings in the queues can now be calculated. Since each of the $m\alpha$ beginings in Γ_2 can be inserted and/or deleted from at most one queue, the total time for all these insertions and deletions is $O(m\alpha)$. Each of the remaining $O(m)$ beginings can be inserted or deleted from each of the α queues at most once, hence the total time taken for this part is also $O(m\alpha)$.

GTILE WITH BOUNDED NUMBER OF TILES

This section considers the case when the maximum number of tiles t is given.

Theorem 7 *The GTile problem with bounded number of tiles can be solved in* $O(\min\{n\log n, nt\})$ *time using* $O(n)$ *space.*

An algorithm with $O(nt)$ time and space can be developed using the approach ofTheorem 4, maintaining separate queues for each possible value of number of tiles. In the rest of this section, we will use a different approach to reduce the space to $O(n)$ which is significant since t could be large. This section also describes another algorithm that runs in $O(n \log n)$ time using $O(n)$ space, which is significant since typically for these applications $\log n \ll t$.

Sets and sequences of block ends. Recall that a *block* is contiguous subsequence $c[p,q)$ of the given input sequence, a block of length at least l and at most u is a tile and the solution consists of a set of disjoint tiles. A set of blocks $\boldsymbol{S}$ can be uniquely characterized by the set of endpoints of its blocks by using the following two quantities (where the first component of an ordered pair is λ or ρ depending on whether the endpoint is the left or the right endpoint of the block, respectively):

$$ends(c[a,\ b)) = \{(\lambda,\ a),\ (\rho,\ b)\};\ \ ends(\boldsymbol{S}) = \bigcup_{T\in\boldsymbol{S}} ends(T).$$

A *block end* $e = (\rho,\ m)$ has *side* $side(e){=}\rho$ and *position* $pos(e) = m$. A set of ends E is *consistent* if $E = ends(\boldsymbol{S})$ for some set of non-empty blocks $\boldsymbol{S}$. A partial order $\prec$ among the block ends is introduced as follows: $e \prec f$ if $pos(e) < pos(f)$ or if $pos(e) = pos(f)$, $side(e) = \rho$ and $side(f) = \lambda$. A set E of $m + 1$ ends ordered according to $\prec$ is the sequence $E^{\rightarrow} = (e_0,\ e_1, \ldots,\ e_m)$.

The test for consistency of E is obvious: the number of endpoints $m + 1$ in E has to be even and the sequence $side(e_0)$, $side(e_1)$, $\ldots$, $side(e_m)$ has to be $(\lambda, \rho, \ldots, \lambda, \rho)$. The requirement that $(\rho,\ k) \prec (\lambda,\ k)$ reflects the fact that empty blocks (i.e., blocks of the form $c[k,\ k)$) are not allowed.

In this subsection we will assume that $\boldsymbol{S}$ and $\boldsymbol{T}$ are sets of blocks with $A = ends(\boldsymbol{S})$ and $A' = \text{ends}(\boldsymbol{T})$; hence both A and A' are consistent. We also assume that $B = A \oplus A' = (A - A') \cup (A' - A)$, $C = A \cup A'$ and $B^{\rightarrow} = (b_0, \ldots, b_{2k-1})$. If $A \oplus D$ is consistent, D is referred to as an *alteration* of $\boldsymbol{S}$, and $S \oplus D$ is the set of blocks $\boldsymbol{U}$ such that $ends(\boldsymbol{U}) = A \oplus D$. Obviously, B is an alteration of $\boldsymbol{S}$. We want to characterize the subsets of B that are alterations as well. For every $i \in [0, k)$ we say that b_{2i} and b_{2i+1} are *dyads* in B.

Lemma 8 *Dyads in B are adjacent in $C^{\rightarrow}$.*

Proof. For the sake of contradiction, suppose that in $C^{\rightarrow}$ entries b_{2i} and b_{2i+1} are separated by another entry, say a, not in B. Then a is preceded by an odd number of elements of B. Consequently, if a is preceded by an odd (respectively, even) number of elements of A, then it is preceded by an even (respectively, odd) number of elements of $A \oplus B$. Thus if the consistency of A dictates that $side(a) = \lambda$ (respectively, $side(a) = \rho$) then the consistency of $A \oplus B$ dictates that $side(a) = \rho$ (respectively, $side(a) = \lambda$), a contradiction. ■

Lemma 9 *Assume that $D \subset B$ does not separate any pair of dyads of B (*i.e.*, for each pair of dyads in B, D either has either both or none of them). Then $A \oplus D$ is consistent.*

Proof. Each $a \in A - D$ is preceded by an even number of elements of D, because if it is preceded by some $b \in D$ then, by Lemma 7.1 and the fact that D does not separate any pair of dyads in B, it is also preceded by the dyad of b which is also in D. Thus the parities of the positions of a in A and $A \oplus D$ are the same. Because $B - D$ also does not separate any pairs of dyads in B and $A \oplus D = A' \oplus (B - D)$, the same reasoning shows that for any $d \in D - A \subset A' - (B - D)$ the parities of the positions of d in A' and $A \oplus D$ are the same. As a result, in the ordering of $A \oplus D$ every λ is on an even position and every ρ is on an odd position. ■

Modifying a Set of Tiles. The assumptions of the previous subsection are now revised to assume that $\boldsymbol{S}$ and $\boldsymbol{T}$ are two sets of tiles (i.e., they satisfy the size bounds), and the notion of alteration is redefined as follows: D is an *altera-*

tion of $\boldsymbol{S}$ if $\boldsymbol{S} \oplus D$ is a set of tiles. Again, we want to characterize the alterations of $\boldsymbol{S}$ that are subsets of B.

If $g < h < i < j$ and $c[g, i]$, $c[h, j) \in \boldsymbol{S} \cup \boldsymbol{T}$ we say that (λ, h) and (ρ, i) are adjuncts. The following lemma is required.

Lemma 10 *Two adjuncts must be adjacent in $C^{\leftrightarrow}$, they must both belong to B and they are not dyads.*

Proof. Without loss of generality assume that $c[g, i) \in \boldsymbol{S}$ and $c[h, j) \in \boldsymbol{T}$. Clearly, $(\lambda, g) \prec (\lambda, h) \prec (\rho, i) \prec (\rho, j)$. No block end in A is between (λ, g) and (ρ, i) and no block end in A' is between (λ, h) and (ρ, j). This shows that (λ, h) and (ρ, i) are adjacent in $C^{\leftrightarrow}$ and that they are both in B. To see that they are not dyads, note that $A \oplus \{(\lambda, h), (\rho, i)\}$ cannot be consistent and use Lemma 7.1. ■

Note that a pair of adjuncts is easy to recognize: it must be a pair of the form $\{b_{2i-1}, b_{2i}\} = \{(\lambda, g), (\rho, h)\}$ where either $b_{2i-1} \in A - A'$ and $e_{2i} \in A' - A$, or $b_{2i-1} \in A' - A$ and $b_{2i} \in A - A'$. Let G_B be the graph with the vertex set as the set of block ends B and with two kinds of edges: between pairs of dyads and between pairs of adjuncts. By Lemmas 7.1 and 7.2 these sets of edges form two *disjoint* matchings of G_B. The following lemma is crucial to this approach.

Lemma 11 *If $D \subseteq B$ is the set of vertices in a connected component of G_B, then D is an alteration of $\boldsymbol{S}$.*

Proof. Because D does not separate any pair of dyads, by Lemma 7.1 $\boldsymbol{U} = \boldsymbol{S} \oplus D$ is a set of disjoint blocks. Suppose that $c[g, h) \in \boldsymbol{U}$ is not a tile (i.e., $h - g \notin [l, u]$) we will then obtain a contradiction. Obviously $c[g, h) \notin \boldsymbol{S}$ and $c[g, h) \notin \boldsymbol{T}$. Hence exactly one of (λ, g) and (ρ, h) is in A and the other one is in A'. Without loss of generality assume that $c[g, i) \in \boldsymbol{S}$ and $c[j, h) \in \boldsymbol{T}$. We can exclude the cases when $j \leq g < i \leq h$ or $g \leq j < h \leq i$ because then $h - g \in [l, u]$. The case when $g < j < i < h$ can also be excluded, because then the adjuncts (λ, j) and (ρ, i) are separated by D (because D contained (ρ, i) but did not contain (λ, j)) and hence D did not contain all the vertices in a connected component of G_B. Similarly, if $j < g < h < i$ then the adjuncts (λ, g) and (ρ, h) are separated by D. Thus it remains to consider the case when $g < i \leq j < h$. But then $(\rho, i) \in D$, $(\lambda, j) \notin D$, $(\rho, h) \in D$. This contradicts the assumption that D is a connected component of G_B. ■

Alterations that are vertices in a connected component of G_B will be called *atomic*. Obviously any alteration can be expressed as a union of one or more disjoint atomic alterations and two disjoint atomic alterations can be applied in any order on a given set of tiles to obtain the same set of tiles. We will say that an atomic alteration is *increasing, neutral* or *decreasing* if $|\boldsymbol{S} \oplus D| - |\boldsymbol{S}|$ equals 1, 0 or −1, respectively.

Lemma 12 *If D is an atomic alteration of $\boldsymbol{S}$, then* $-1 \leq |\boldsymbol{S} \oplus D| - |\boldsymbol{S}| \leq 1$.

Proof. Except for its first and last elements, D is covered by disjoint pairs of adjuncts. Applying a pair of adjuncts removes one element of $ends(\boldsymbol{S})$ and inserts another, thus leaving the total number of block ends unchanged. Hence the net change in the number of block ends can come only from the first and the last elements of D, and because the number of block ends changes by at most 2, the number of blocks changes by at most 1. ■

Computing $\boldsymbol{S}_t$ in $O(nt)$ time using $O(n)$ space. Let $\boldsymbol{S}_0 = \varnothing$ and $\boldsymbol{S}_{t+1}$ be a set of $t + 1$ tiles of *maximum* weight that can be obtained by applying a *minimal* alteration (i.e., an alteration that is not properly contained in another alteration) to $\boldsymbol{S}_t$.

Lemma 13 *If $\boldsymbol{S}_{t+1} = \boldsymbol{S} \oplus D$ then D is an atomic alteration.*

Proof. Suppose that D is *not* an atomic alteration and thus it is a union of more than one *disjoint* atomic alterations. Disjoint atomic alternations can be applied in any order. If one of them, say D_0, is neutral, then one of the following two cases occur:

- D_0 changes the weight of the set of tiles. Then $w(\boldsymbol{S}_t \oplus D_0) > w(\boldsymbol{S}_t)$ or $w(\boldsymbol{S}_t + 1 \oplus D_0) > w(\boldsymbol{S}_{t+1})$, contradicting the definition of $\boldsymbol{S}_t$ or $\boldsymbol{S}_{t+1}$.

- D_0 does not change the weight of the set of tiles. Then we could apply $D - D_0$ rather than D to get $\boldsymbol{S}_{t+1}$ (i.e., D is not minimal).

If D contains a decreasing atomic alteration D_0, then it must also contains an increasing alteration, say D_1, and we can use the neutral alteration $D_0 \cup D_1$ to obtain a contradiction similar to above. Hence every atomic alterations contained in D is increasing and thus D can contain only one such alteration. ■

Based on the above results, a simple version of the algorithm is as follows.

```
$S_0 = \varnothing,\ w(S_0) = 0$
for $p = 1$ to $t$ do
    compute $S_p = S_p - 1 \oplus D$ by finding the increasing atomic
    alteration $D$ that produces maximum gain in total weight
    if $w(S_p) \leq w(S_p - 1)$ then
    output $S_p - 1$, exit ($S_p$ is the best solution)
output $St$
```

The proposition on the time and space complexity of the above algorithm follows from the following lemma.

Lemma 14 *Given S_{p-1}, an atomic alteration D such that $\boldsymbol{S}_p = \boldsymbol{S}_{p-1} \oplus D$ can be found $O(n)$ time and space.*

Proof. For simplicity, $\boldsymbol{S}$ will denote $\boldsymbol{S}_{p-1}$. As in Theorem 4, $s_k = w(c[0,\ k))$ is computed for all $k \in [0,\ n]$ in $O(n)$ time and space. For a possible tile end e let D_e={$f \in D$: $f \prec e$}. Note that D_e does not include e itself. This algorithm scans all possible tile ends in $\prec$ order from left to right. In other words, the elements of $c[0,\ n)$ are examined in left-to-right order, and each element will be first considered as a possible tile ending and then as a possible tile begining. When the tile end e is considered, we need to know what D_e could be. The following case analysis must be performed. The default value of the quantity $A(i)$ defined below is $-\infty$ for all $i \in [0,\ n)$.

Case 1:

e=$(\lambda,\ i)$.

Case 1.1:

for some $c[g,\ h) \in \boldsymbol{S}$ we have $g < i < h$.

It follows that $D_e \neq \varnothing$ since otherwise D would not be an increasing atomic alteration and that $E = D_e \cup \{(\rho,\ h)\}$ is a neutral alteration. Let $\mathbf{A}(i)$ be the set of neutral atomic alterations that contain $(\rho,\ h)$ as the only element that does not precede e and let

$A(i) = \max_{E \in \mathbf{A}(i)} w(S \oplus E) - w(S)$.

Because $w(\boldsymbol{S})$ cannot be increased by a neutral atomic alteration, $A(i) \leq 0$. Later when we need to consider e as a possible left end of a new tile, we will consider it with a priority of $v_i = A(i) - s_i$.

Case **2.1** shows how to compute the value of $A(i)$ when the last new tile ends at i using the OLIM problem. Note that since any element is first considered as a possible ending and then a possible ending, Case **2** will occur before Case **1** for that element. Hence, we simply set $A(i)$ to be the maximum of $A(i-1)$ and current $A(i)$, and update v_i accordingly if necessary.

Case 1.2:

Case **1.1** does not hold. Then e would have to be the first tile end in D and thus the left end of the leftmost new tile. We should now consider e as a possible left end with priority $v_i = -s_i$. We should allow for e being the left end of an old tile (i.e., $i = g$ for some $c[g,\ h) \in \boldsymbol{S}$) which indicates an alteration where the leftmost new tile shares its left end with an old tile.

Case 2:

$e = (\rho,\ i)$.

Case 2.1:

for some $c[g,\ h) \in \boldsymbol{S}$ we have $g < i < h$.

We consider e to be the right end of a new tile. The matching left end, $(\lambda,\ j)$, must satisfy $j \leq g$ (i.e., it must be to the left of $c[g,\ h)$). We choose j as the index of the matching left end that satisfies $v_j = \max_{i-u\ \leq\ z\ \leq}$

$_{\min\{i-l,g\}}\{v_z\}$, and this allows the consideration of a new alteration to the set $A(i)$; the total gain of this alteration is $v_j - w(c[g, h)) + s_i$. Then, if this gain is greater than the current value of $A(i)$, we replace $A(i)$ with this new gain and update v_i accordingly. Notice that computation of the v_j's can be done via the OLIM problem in the same manner as in the proof of Theorem 4.

Case 2.2:

Case **2.1** does not hold and $c[g, h)$ is the last tile of $\boldsymbol{S}$ to the left of i. Then e would be the rightmost tile end of an increasing atomic alteration (or, if $i = h$, the right end of a tile whose left end is the rightmost element of the alteration). The matching left end (λ, q) is selected such that $v_q = \max_{i-u \le z \le i-1} v_z$ and the gain of this alteration is $s_i + v_q$. Then, if this gain is greater than the current value of $A(i)$, we replace $A(i)$ with this new gain and update v_i accordingly. Notice that the computation of the v_q's can again be done via the OLIM problem. ■

Remark 1 *The claim of Lemma 7.3 holds even if D is a neutral or decreasing alteration (which produces a minimum total weight decrease) by using a very similar algorithm. Moreover we actually also compute, for each possible tile end e, the optimum alteration of the prescribed type in which all elements precede e.*

Computing S_t in $O(n \log[n/l])$ time using $O(n)$ space. The algorithm proceeds in phases. Before a phase, we computed a set of t disjoint tiles, say $\boldsymbol{S}$, that has the largest weight under the constraint that each tile is contained in one of the blocks $c[a_0, a_1), c[a_1, a_2), \ldots, c[a_{k-1}, a_k)$. For this phase, some a_i is selected such that, after the phase, $\boldsymbol{S}$ is replaced with some $\boldsymbol{S} \oplus B$ that maximizes the sum of weights under the constraint that each tile in $\boldsymbol{S} \oplus B$ is contained in $c[a_0, a_1), c[a_1, a_2), c[a_{i-2}, a_{i-1}), c[a_{i-1}, a_{i+1}), c[a_{i+1}, a_{i+2}), \ldots, c[a_{k-1}, a_k)$ (i.e., the new set of blocks is obtained from the old set of blocks by coalescing the two blocks $c[a_{i-1}, a_i)$ and $c[a_i, a_{i+1})$ into one block $c[a_{i-1}, a_{i+1})$).

We can start with $a_i = il$; each block is a tile of minimum length, thus the weight of each tile can be computed to select the tiles with t largest weights. Using any linear time algorithm for order statistics, the first phase can be completed in $O(n)$ time and space. In Lemma 15 below we show that a single phase can be performed in $O(M + \log[n/l])$ time and $O(n)$ space, where $M = \max_{i=1}^{k}\{a_i - a_{i-1}\}$. This will be sufficient to satisfy the proposition on the time and space complexity of the complete algorithm by the following analysis. We first coalesce adjacent pairs of blocks of length l into blocks of length $2l$ (unless there is only one block of length l remaining). This requires $O(n/l)$ phases and each of them takes $O(l)$ time, because during these phases the longest block has length $2l$. Hence the total time and space complexity for these n/l phases is $O(n)$. Repeating the same procedure for blocks of length $2l$, $4l$,..., it follows that that if all blocks but one have the maximum length then we can half the number of blocks in $O(n)$ time and space, and again, all blocks but one will have the maximum length. Obviously, the current phase is complete when one block remains. Since each phase can be carried out independently of any other phase, the space complexity is $O(n)$, and the total time complexity is $O(\Sigma_{i=1}^{\log(n/l)} ((2^{i+1}l + \log[n/l]) \ \ [n/(2^i l)]))$ $= O(n \log[n/l])$. Hence it suffices to prove the following lemma.

Lemma 15 *A single phase can be performed in* $O(M + \log n)$ *time and* $O(n)$ *space.*

Proof. For an increasing (respectively, decreasing, neutral) atomic alteration B, let B be called the *best* increasing (respectively, decreasing, neutral) atomic alteration if $w(\boldsymbol{S} \oplus B)$ is maximized. The following data structures are maintained throughout all phases. For each block, we store that part of the current solution that is contained in that block, the best increasing atomic alteration of that part and the best decreasing one. Moreover, we will have two priority queues Q_1 and Q_2 of blocks, in which the *priority* of a block is the *gain* in total weight of its best increasing and decreasing atomic alterations, respectively. These data structures can be initialized in $O(n)$ time and space using Lemma 7.3 and Remark 7.3. Note that both Q_1 and Q_2 contain at most $O(n/l)$ entries.

We need to show how to find the desired alteration B which consists of an union of one or more atomic alterations. All possible structures of B are considered. If the new solution contains no tile $c[g, h)$ such that $g < a_i < h$, then $B = \varnothing$. Otherwise, assume that such a tile exists. Under this assumption, the number of end points of tiles contained in each of $c[a_{i-1}, a_i)$ and c[a_i, $\mathrm{a}_{\mathrm{i}+1}$) changes from even to odd, which shows that B contains a pair of dyads b, b' such that $b \prec (\rho, a_i)$ and $(\lambda, a_i) \prec b'$. Let D be the atomic alteration such that $\{b, b'\} \subseteq D \subseteq B$. We have the following 3 cases.

Case 1:

$|\boldsymbol{S} \oplus D| - |\boldsymbol{S}|$

D is a neutral atomic alteration. Because $|\boldsymbol{S} \oplus (B - D)| = |\boldsymbol{S}|$, we could alter $\boldsymbol{S}$ with $B - D$ before the current phase, hence this alteration cannot increase the weight, and therefore it is not needed. Thus in this case $B = D$. Hence, for this case, we need to find a best neutral atomic alteration in $c[a_{i-1}, a_{i+1})$ (i.e., one that yields the maximum increase of weight). Using Lemma 7.3 and Remark 7.3 such an alteration can be found in $O(a_{i+1} - a_{i-1})$ time and space. Subsequently, in similar time and space we can find in $c[a_{i-1}, a_{i+1})$ a best increasing and decreasing atomic alterations and update Q_1 and Q_2 in $O(\log[n/l])$ time.

Case 2:

D is an increasing atomic alteration. $C = B - D$ does not contain a nonempty atomic alteration F that does not change the size of $\boldsymbol{S}$; otherwise since we could alter $\boldsymbol{S}$ with F before the current phase, F cannot increase the weight of $\boldsymbol{S}$ and thus we do not need it. Hence C is a decreasing atomic alteration.

To find B, we consider two subcases.

Case 2.1:

C is contained in a block different than c[a_{i-1}, a_{i+1}). We can find a best D using the algorithm from Lemma 7.3, in $O(M)$ time and a best C from the priority queue Q_2 in $O(n/l)$ time, both using $O(n)$ space. Subsequently, we need to compute a best increasing and a best decreasing atomic alterations for both $c[a_{i-1}, a_{i+1})$ and the block that contains C (using Lemma 7.3 and Remark 7.3) and update Q_1 and Q_2 in a total of $O(M + \log[n/l])$ time and $O(n)$ space.

Case 2.2:

C is contained in $c[a_{i-1},\ a_i)$. We first compute for each $j \in [a_{i-1},\ i)$ the best decreasing atomic alteration contained in $c[a_{i-1},\ j)$ and set $X(j)$ to be its *gain* (i.e.,the increase in total weight produced by it (perhaps negative)). As noted in Remark 7.3, this can be done in $O(M)$ time and $O(n)$ space. Now, to find the best increasing atomic alteration D, we proceed as in Lemma 7.3 except that when we consider some $(\lambda,\ k)$ to be the leftmost left end of a tile introduced by D we set its priority to be $X(j) - s_k$. Hence such a D can again be found in $O(M)$ time and $O(n)$ space. At the end, we must update Q_1 and Q_2 as in Case **1** in $O(\log[n/l])$ time.

Case 2.3:

C is contained in $c[a_i,\ a_{i+1})$. This case is mirror-symmetric of Case **2.2**, so we can apply the same methods, except that we will be scanning the merged block from right to left.

Case 3:

D is a decreasing atomic alteration. Then $C = B - D$ is an increasing atomic alteration by an argument very similar to that in Case **2**. Case **3** is then symmetric of Case **2** if we replace increasing (respectively, decreasing) by decreasing (respectively, increasing) in the remainder of Case **2**. ■

Tiling Sequences with Binary Morphological Operations

D

A number of established methods can be applied to repeat-masked DNA sequences to approximate an optimal tiling solution. By treating a genome sequence as a vector of nucleotide "pixels", we can use image segmentation techniques such as region growing and other relaxation processes to close small repetitive elements in the genomic sequence, thereby merging the adjacent high-complexity sequences into contiguous tiles. This approach can be expressed using standard binary morphological algebra (Serra 1980). We first assign all nrDNA and rpDNA elements from a target genome sequence S to sets A and B, respectively:

$$A = \bigcup_{a \in S} a \in \text{nrDNA} \ , \ B = \bigcup_{b \in S} b \in \text{rpDNA}$$

Given this conceptual distinction, we can operate on the sequence using binary morphological operations. In order to apply binary operations to repeat-masked genomic DNA, it is necessary to first reduce the mixed nucleotide sequence to a bilevel representation. Thresholding assigns a new binary value $b(x)$ to each nucleotide in the original "greylevel" sequence image, thereby generating a mask of the original sequence. Repeat and high-complexity nucleotides are assigned binary values by the following thresholding operation:

$$b(n) = \begin{cases} G_{nr}, & g(n) > Z \\ G_{rp}, & g(n) \leq Z \end{cases}$$

where $b(x)$ is equal to one of two possible assignments $[G_{nr}, G_{rp}]$. This segmentation method discretizes the initial greylevel value g of any nucleotide n according to a predetermined threshold Z, thereby converting nrDNA and rpDNA sequence positions to the binary values prescribed by the parameters G_{nr} and G_{rp}, respectively.

The converted bilevel image can now be processed in several ways to yield an expansion of the nrDNA regions into the rpDNA regions. The nrDNA elements in set A can be transformed depending on how they relate to the "background" component of the sequence, comprising the rpDNA elements in B and referred to as the structuring element. The dilation of an input image A by a structuring element B is then described by:

$$A \oplus B = \bigcup_{b \in B} (A + b)$$

where $(A{+}b)$ indicates the translation of A by b. Essentially, this implies that in order to dilate set A by the structuring element B, B is first translated by all elements in A. The dual operation to dilation is erosion:

$$A \oplus B = \bigcup_{a \in A} (B + a) \text{ , or } \; A \ominus B = (A^C \oplus B^C)$$

where A^c and B^c indicate the complements of A and B, respectively, and $(B{+}a)$ represents the translation of B by a. The union of these translations constitutes $A{\ominus}B$. In cases where a the lengths of repetitive elements exceed the degree of dilation, an equal number of erosion operations will restore the repeat and the adjacent nrDNA regions will remain separate. However, if two dilated nrDNA regions meet, the repeat region will be closed and erosion will have no effect in that local area. This can be accomplished with a closing operation where we first dilate the nrDNA structuring element, then erode by the same amount. Closing comprises a dilation operation followed by erosion:

$$\text{close}(A, B) = (A \oplus B) \ominus B$$

In this manner, rpDNA regions whose lengths are less than the number of dilation cycles are closed, and the adjacent nrDNA fragments are effectively merged into larger tiles. Although this approach describes a simple approximation to the tiling problem, it is dependent on the use of a threshold constant for dilation-erosion cycles which corresponds to a fixed maximum number of nucleotides that each repetitive element can span.

Microarray Hybridization Protocols E

The solutions and protocols below correspond to the procedure used to probe the human genome arrays described in Chapter 4. Each microarray is designed to include a number of control features complementary to a common DNA sequence. When the arrays are hybridized, an end-labeled 24mer oligo is used to illuminate these features, forming a grid to which the scanned image will be aligned in software. An appropriate concentration of this control oligo must be spiked into the hybridization solution and should fluoresce in the same emission spectrum as the labeled DNA probe. Thus, a two-channel experiment must include two control oligos, each fluorescence-labeled to coincide with the excitation and emission spectra used for the two experimental samples.

Fluor-5′ TTCCTCTCGCTGTAATGACCTCTA 3′

HYBRIDIZATION SOLUTIONS

Volumes appropriate for hybridizations in Secure-Seal adhesive modules
(Grace BioLabs SA-200, adhesive seal tabs ST-200)

Pre-hybridization solution	400µl
2X MES hybridization buffer	200µl
Nuclease-free water	192µl
Herring or salmon sperm DNA, 10mg/ml	4µl
BSA, 50mg/ml	4µl

Hybridization solution	320µl
2X MES hybridization buffer	160µl
Cot-1 DNA, 1mg/ml	30µl
BSA, 50mg/ml	3µl
Labeled control oligo, 100nM	3µl
(*multiple channel experiments require additional control oligos*)	
Labeled cDNA probe	≥ 2µg/sample
Water	to 320µl

Volumes appropriate for hybridizations using standard coverslips and HybChambers (Gene Machines HYB-03 or HYB-04)

Pre-hybridization solution	300µl
2X MES hybridization buffer	100µl
Nuclease-free water	198µl
Herring or salmon sperm DNA, 10mg/ml	1µl
BSA, 50mg/ml	1µl

Hybridization solution	40µl
2X MES hybridization buffer	20µl
Cot-1 DNA, 1mg/ml	4µl
BSA, 50 mg/ml	1µl
Labeled control oligo, 100nM	1µl
(*multiple channel experiments require additional control oligos*)	
Labeled cDNA probe	≥ 2µg/sample
Nuclease-free water	to 40µl

References to water = Millipore or nuclease-free water. Do not use DEPC-treated water.

BUFFER COMPONENTS

12X MES stock buffer	500ml
(1.22M MES, 0.89M $[Na^+]$)	
MES, free acid monohydrate	35.2g
MES, sodium salt	96.65g
Water	to 500ml

Final pH should be 6.5 - 6.7

Store protected from light at 4°C

2X MES hybridization buffer	250ml
(100mM MES, 1M [Na$^+$], 20mM EDTA, 0.01% Tween-20)	
12X MES stock buffer	41.5ml
5M NaCl	88.5ml
0.5M EDTA	20.0ml
10% Tween-20	0.5ml
Water	99.5ml

Store protected from light at 4°C

Non-stringent wash buffer (NSWB)	1L
(6X SSPE, 1mM DTT, 0.01% Tween-20)	
20X SSPE	300ml
10% Tween-20	1.0ml
1M DTT	1.0ml
Water	to 1L

Store at room temperature

Stringent wash buffer (SWB)	1L
(100mM MES salt and free acid solution, 0.1M [Na$^+$], 1mM DTT, 0.01% Tween-20)	
12X MES stock buffer	83.3ml
5M NaCl	5.2ml
10% Tween-20	1.0ml
1M DTT	1.0ml
Water	to 1L

Store protected from light at 4°C.

20X SSPE	1L
Combine:	
NaCl	175.3g
NaH_2PO_4-H_2O	27.6g
EDTA	7.4g

in 800 ml distilled water.

Boil until salt is fully dissolved.

Adjust the pH to 7.4 with 10M NaOH.

Adjust the volume to 1L with distilled water.

20X SSC	1L
Combine:	
NaCl	175.3g
Sodium citrate	88.2g

in 800 ml distilled water.

Boil until salt is fully dissolved.

Adjust the pH to 7.0 with 1M HCl.

Adjust the volume to 1L with distilled water.

Filter all solutions with 0.2 µm vacuum filter units.

hybridization protocol

(for Secure-Seal gasket hybridization modules)

1. Pipette 400μl of pre-hyb solution into hybridization module and incubate at 45°C for 20 minutes.
2. Remove pre-hyb solution and replace with 320μl hybridization solution. A small air bubble will remain in the module and will agitate the solution during hybridization.
3. Carefully dry both input ports on hybridization module with Kimwipes. This will guarantee complete adhesion of the sealing tabs to the clean surface.
4. Using forceps, apply adhesive tabs to both input ports. The adhesive tabs should be laid completely flat so that a water-tight seal is formed.
5. Insert each array into a 50ml tube covered with aluminum foil. Place tubes into hybridization oven and enable wheel rotation.
6. Incubate for 16-20 hours at 50°C.

Following hybridization:

1. Heat swb to 45°C.
2. Briefly disable the rotating wheel in the hybridization oven, remove one Falcon tube and enable rotation again. It is important to keep the arrays completely hydrated to prevent probe from drying on any part of the printed area. A small volume of liquid will have evaporated during the hybridization cycle, so the arrays should be constantly agitated until the hybridization solution can be removed.
3. Using forceps, carefully remove both adhesive tabs. Minimize the exposure of arrays to ambient light.
4. Remove hybridization solution with 1000μl pipette and immediately replace with 400μl nswb. Gently pipette up and down 1-2 times and discard nswb, replacing with a fresh aliquot.
5. Return the array to its original foil-wrapped 50ml tube, oriented horizontally in a tube rack to prevent buffer from leaking out of the module.

6. When all arrays have been processed, remove the adhesive hybridization modules from the slides and place each array into light-tight microscope slide box containing NSWB.
7. Wash in NSWB by agitating on rotating platform for 10 minutes, changing the buffer once.
8. Replace buffer with 45°C SWB. Wash in SWB for 30 minutes, changing the buffer every 5 minutes.
9. Replace buffer with NSWB. Wash in NSWB for 5 minutes, changing the buffer twice.
10. Wash in 0.2X SSC for one minute to remove residual salt from NSWB. Repeat this step for a total of 4 washes.

Dry arrays with compressed nitrogen gas. Scan as soon as possible at ≤ 5 µm resolution.

HYBRIDIZATION PROTOCOL
(*for HybChambers/slide coverslips*)

1. Apply pre-hyb solution to arrays and incubate at 45°C for 15 minutes.
2. Remove pre-hyb solution, rinse with distilled water.
3. Apply hybridization solution, place coverslip onto array and seal in HybChamber.
4. Incubate in water bath for 16-20 hours at 50°C.

Following hybridization:

1. Heat SWB to 45°C.
2. Remove coverslips and immediately place arrays into light-tight microscope slide box containing NSWB.
3. Wash in NSWB by agitating on rotating platform for 10 minutes, changing the buffer once.
4. Replace buffer with 45°C SWB. Wash in SWB for 30 minutes, changing the buffer every 5 minutes.

5. Replace buffer with nswb. Wash in nswb for 5 minutes, changing the buffer twice.
6. Wash in 0.2X SSC for one minute to remove residual salt from nswb. Repeat this step for a total of 4 washes.

Dry arrays with compressed nitrogen gas. Scan as soon as possible at ≤ 5 µm resolution.

equipment and reagents

Gene Machines (www.genemachines.com)

hyb-04	HybChamber Mica tinted dual microarray hybridization chambers

Grace BioLabs (www.gracebio.com)

sa-200	Secure-Seal gasketed hybridization chambers
st-200	Adhesive seal tabs

Invitrogen (www.invitrogen.com)

15279011	Human Cot-1 DNA

Sigma-Aldrich (www.sigmaaldrich.com)

69896	2-Morpholinoethanesulfonic acid sodium salt
69892	4-Morpholineethanesulfonic acid monohydrate

Bibliography

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF *et al*. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**: 1651–1656.

Albert TJ, Norton J, Ott M, Richmond T, Nuwaysir K, Nuwaysir EF, Stengele KP, Green RD (2003) Light-directed 5'→3' synthesis of complex oligonucleotide microarrays. *Nucleic Acids Res*. **31**: e35.

Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA. (2004) Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol*. **14**: 283–291.

Bao Z, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res*. **12**: 1269–1276.

Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK (2003) Computational discovery of gene modules and regulatory networks. *Nat Biotechnol*. **21**: 1337–1342.

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. **27**: 573–580.

Berman P, DasGupta B, Muthukrishnan S (2002) On the exact size of the binary space partitioning of sets of isothetic rectangles with applications. *SIAM J Dis Math*. **15**: 252–267.

Berman P, DasGupta B, Muthukrishnan S, Ramaswami S (2001) Improved approximation algorithms for tiling and packing with rectangles. *Proc 12th Annual ACM-SIAM Symposium on Discrete Algorithms*, 427–436.

Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**: 2242–2246.

Blake JA, Richardson JE, Bult CJ, Kadin JA, Eppig JT (2002) The Mouse Genome Database (MGD): The model organism database for the laboratory mouse. *Nucleic Acids Res*. **30**: 113–115.

Buckley PG, Mantripragada KK, Benetkiewicz M, Tapia-Paez I, Diaz De Stahl T *et al*. (2002) A full-coverage, high-resolution human chromosome 22 genomic microarray for clinical and research applications. *Hum Mol Genet*. **11**: 3221–3229.

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol*. **268**: 78–94.

Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA *et al*. (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291**: 1289–1292.

Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, Wheeler R, Wong B, Drenkow J, Yamanaka M, Patel S, Brubaker S, Tammana H, Helt G, Struhl K, Gingeras TR (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**: 499–509.

Cormen TH, Leiserson CL, Rivest RL (1990) *Introduction to Algorithms*. Cambridge, MA: MIT Press.

Datar M, Gionis A, Indyk P, Motwani R (2002) Maintaining stream statistics over sliding windows. *Proc 13th Annual ACM-SIAM Symposium on Discrete Algorithms*, 635–644.

Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, Rodriguez-Tome P, Hui L, Matise TC, McKusick KB, Beckmann JS *et al*. (1998) A physical map of 30,000 human genes. *Science* **282**: 744–746.

de Souza SJ, Camargo AA, Briones MR, Costa FF, Nagai MA, Verjovski-Almeida S, Zago MA, Andrade LE, Carrer H, El-Dorry HF *et al*. (2000) Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. *Proc Natl Acad Sci USA*. **97**: 12690–12693.

DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**: 680–686.

Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, Bruskiewich R, Beare DM, Clamp M, Smink LJ *et al*. (1999) The DNA sequence of human chromosome 22. *Nature* **402**: 489–495.

Euskirchen G, Royce TE, Bertone P, Martone R, Rinn JL, Nelson FK, Sayward F, Luscombe NM, Miller P, Gerstein M, Weissman S, Snyder M (2004) CREB binds to multiple loci on human chromosome 22. *Mol Cell Biol*. **24**: 3804–3814.

Ewing B, Green P (2000) Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet*. **25**: 232–234.

Feingold EA, Good PJ, Guyer MS, *et al*. (2004) The ENCODE (ENCyclopedia of DNA Elements) project. *Science* **306**: 636–640.

Fodor SP, Rava RP, Huang XC, Pease AC, Holmes CP, Adams CL (1993) Multiplexed biochemical assays with biological chips. *Nature* **364**: 555–556.

Gao Y, Li J, Strickland E, Hua S, Zhao H, Chen Z, Qu L, Deng XW (2004) An *Arabidopsis* promoter microarray and its initial usage in the identification of HY5 binding targets in vitro. *Plant Mol Biol*. **54**: 683–699.

Gelfand MS, Roytberg MA (1993) A dynamic programming approach for predicting the exon-intron structure. *Biosystems* **30**: 173–182.

Giegerich R (2000) A systematic approach to dynamic programming in bioinformatics. *Bioinformatics* **16**: 665–667.

Goryachev AB, Macgregor PF, Edwards AM (2001) Unfolding of microarray data. *J Comput Biol*. **8**: 443-461.

Gotoh O (1982) An improved algorithm for matching biological sequences. *J Mol Biol*. **162**: 705–708.

Guelzim N, Bottani S, Bourgine P, Kepes F (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet*. **31**: 60–63.

Guigo R, Knudsen S, Drake N, Smith T (1992) Prediction of gene structure. *J Mol Biol*. **226**: 141–157.

Harrison PM, Hegyi H, Balasubramanian S, Luscombe NM, Bertone P, Echols N, Johnson T, Gerstein M (2002) Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res*. **12**: 272–280.

Hasan A, Stengele KP, Giegrich H, Cornwell P, Isham KR, Sachleben RA, Pfleiderer W, Foote RS (1997) Photolabile protecting groups for nucleosides: synthesis and photodeprotection rates. *Tetrahedron* **53**: 4247–4264.

Hedrick SM, Cohen DI, Nielsen EA, Davis MM (1984) Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. *Nature* **308**: 149–153.

Hegde P, Qi R, Abernathy K, Gay C, Dharap S, Gaspard R, Hughes JE, Snesrud E, Lee N, Quackenbush J (2000) A concise guide to cDNA microarray analysis. *Biotechniques* **29**: 548–556.

Hochbaum DS (1997) *Approximation Algorithms for NP-Hard Problems*. Boston, MA: PWS Publishing.

Horak CE, Snyder M (2002a) ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Methods Enzymol*. **350**: 469–483.

Horak CE, Mahajan MC, Luscombe NM, Gerstein M, Weissman SM, Snyder M (2002b) GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIp-chip analysis. *Proc Natl Acad Sci USA*. **99**: 2924–2929.

Horak CE, Luscombe NM, Qian J, Bertone P, Piccirrillo S, Gerstein M, Snyder M (2002c) Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. *Genes Dev*. **16**: 3017–3033.

Hubank M, Schatz DG (1994) Identifying differences in mRNA expression by representational difference analysis of cDNA. *Nucleic Acids Res*. **22**: 5640–5648.

Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T *et al*. (2002) The Ensembl genome database project. *Nucleic Acids Res*. **30**: 38–41.

Hudson TJ, Church DM, Greenaway S, Nguyen H, Cook A, Steen RG, Van Etten WJ, Castle AB, Strivens MA, Trickett P *et al*. (2001) A radiation hybrid map of mouse genes. *Nat Genet*. **29**: 201–205.

Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, Kobayashi S, Davis C, Dai H, He YD, Stephaniants SB, Cavet G, Walker WL, West A, Coffey E, Shoemaker DD, Stoughton R, Blanchard AP, Friend SH, Linsley PS (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol*. **19**: 342–347.

Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, Coe BP, Snijders A, Albertson DG, Pinkel D, Marra MA, Ling V, MacAulay C, Lam W (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nat Genet*. **36**: 299–303.

Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**: 533–538.

Johnson JM, Edwards S, Shoemaker D, Schadt EE (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet*. **21**: 93–102.

Jurka J (2000) Repbase Update: a database and an electronic journal of repetitive elements. *Trends Genet*. **9**: 418–420.

Jurka J, Klonowski P, Dagman V, Pelton P (1996) CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. *Comput Chem*. **20**: 119–121.

Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammana H, Gingeras TR (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res*. **14**: 331–342.

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**: 916–919.

Kawai J, Shinagawa A, Shibata K, et al. (2001) Functional annotation of a full-length mouse cDNAs collection. *Nature* **409**: 685–690.

Khanna S, Muthukrishnan S, Paterson M (1998) On approximating rectangle tiling and packing. *Proc 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, 384–393.

Khanna S, Muthukrishnan S, Skiena S (1997) Efficient array partitioning. *Proc 24th International Colloquium on Automata, Languages, and Programming*, 616–626.

Kumar A, Harrison PM, Cheung KH, Lan N, Echols N, Bertone P, Miller P, Gerstein MB, Snyder M (2002) An integrated approach for finding overlooked genes in yeast. *Nat Biotech*. **20**: 58–63.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**: 799–804.

Li Z, Van Calcar S, Qu C, Cavenee WK, Zhang MQ, Ren B (2003) A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc Natl Acad Sci USA* **100**: 8164–8169.

Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J (2000) Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat Genet*. **25**: 239–240.

Liang P, Pardee AB (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **257**: 967–971.

Lieb JD, Liu X, Botstein D, Brown PO (2001) Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat Genet*. **28**: 327–334.

Lieb JD (2003) Genome-wide mapping of protein-DNA interactions by chromatin immunoprecipitation and DNA microarray hybridization. *Methods Mol Biol*. **224**: 99–109.

Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ (1999) High density synthetic oligonucleotide arrays. *Nat Genet*. **21**: 20–24.

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*. **14**: 1675–1680.

Luscombe NM, Royce TE, Bertone P, Echols N, Horak CE, Chang JT, Snyder M, Gerstein M (2003) ExpressYourself: A modular platform for processing and visualizing microarray data. *Nucleic Acids Res*. **31**: 3477–3482.

Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431**: 308–312.

Mao DY, Watson JD, Yan PS, Barsyte-Lovejoy D, Khosravi F, Wong WW, Farnham PJ, Huang TH, Penn LZ (2003) Analysis of Myc bound loci identified by CpG island arrays shows that Max is essential for Myc-dependent repression. *Curr Biol*. **13**: 882–886.

Maruyama K, Sugano S (1994) Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene* **138**: 171–174.

Martone R, Euskirchen G, Bertone P, Hartman S, Royce TE, Luscombe NM, Rinn JL, Nelson FK, Miller P, Gerstein M, Weissman S, Snyder M (2003) Distribution of NF-kappaB binding sites across human chromosome 22. *Proc Natl Acad Sci USA* **100**: 12247–12252.

Mattick JS (2003) Challenging the dogma: the hidden layer of nonprotein-coding RNAs in complex organisms. *Bioessays* **25**: 930–939.

Mattick JS (2004) RNA regulation: a new genetics? *Nat Rev Genet*. **5**: 316–323.

Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*. **31**: 374–378.

Muthukrishnan S, Poosala V, Suel T (1999) On rectangular partitions in two dimensions: algorithms, complexity and applications. *Proc 7th International Conference on Database Theory*, 236–256.

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*. **48**: 443–453.

Nuwaysir EF, Huang W, Albert TJ, Singh J, Nuwaysir K, Pitas A, Richmond T, Gorski T, Berg JP, Ballin J, McCormick M, Norton J, Pollock T, Sumwalt T, Butcher L, Porter D, Molla M, Hall C, Blattner F, Sussman MR, Wallace RL, Cerrina F, Green RD (2002) Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Res*. **12**: 1749–1755.

Odom DT, Zizlsperger N, Gordon DB, Bell GW, Rinaldi NJ, Murray HL, Volkert TL, Schreiber J, Rolfe PA, Gifford DK, Fraenkel E, Bell GI, Young RA (2004) Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303**: 1378–1381.

Ota T, Suzuki Y, Nishikawa T, et al. (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat Genet*. **36**: 40–45.

Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP (1994) Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci USA*. **91**: 5022–5026.

Pruitt KD, Tatusova T, Maglott DR (2003) NCBI Reference Sequence project: update and current status. *Nucleic Acids Res*. **31**: 34–37.

Qian J, Lin J, Luscombe NM, Yu H, Gerstein M (2003) Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics* **19**: 1917–1926.

Quackenbush J (2002) Microarray data normalization and transformation. *Nat Genet*. **32** (Suppl): 496–501.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA (2000) Genome-wide location and function of DNA binding proteins. *Science* **290**: 2306–2309.

Ren B, Cam H, Takahashi Y, Volkert T, Terragni J, Young RA, Dynlacht BD (2002) E2F integrates cell cycle progression with DNA repair, replication, and G2/M checkpoints. *Genes Dev*. **16**: 245–256.

Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM, Hartman S, Harrison PM, Nelson FK, Miller P, Gerstein M, Weissman S, Snyder M (2003) The transcriptional activity of human chromosome 22. *Genes Dev*. **17**: 529–540.

Rychlik W, Rhoads RE (1989) A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA. *Nucleic Acids Res*. **17**: 8543–8551.

Saccone S, Caccio S, Kusuda J, Andreozzi L, Bernardi G (1996) Identification of the gene-richest bands in human chromosomes. *Gene* **174**: 85–94.

SantaLucia J (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci USA*. **95**: 1460–1465.

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**: 467–470.

Selinger DW, Cheung KJ, Mei R, Johansson EM, Richmond CS, Blattner FR, Lockhart DJ, Church GM (2000) RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. *Nat Biotechnol*. **18**: 1262–1268.

Serra J (1980) *Image Analysis and Mathematical Morphology*. New York: Academic Press.

Shaw S (2003) Evidence of scale-free topology and dynamics in gene regulatory networks. *Proc. ISCA 12th Int. Conf. on Intelligent and Adaptive Systems and Software Engineering* 37–40.

Shoemaker DD, Schadt EE, Armour CD, He YD, Garrett-Engele P, McDonagh PD, Loerch PM, Leonardson A, Lum PY, Cavet G, Wu LF, Altschuler SJ, Edwards S, King J, Tsang JS, Schimmack G, Schelter JM, Koch J, Ziman M, Marton MJ, Li B, Cundiff P, Ward T, Castle J, Krolewski M, Meyer MR, Mao M, Burchard J, Kidd MJ, Dai H, Phillips JW, Linsley PS, Stoughton R, Scherer S, Boguski MS (2001) Experimental annotation of the human genome using microarray technology. *Nature* **409**: 922–927.

Singh-Gasson S, Green RD, Yue YJ, Nelson C, Blattner F, Sussman MR, Cerrina F (1999) Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat Biotechnol*. **17**: 974–978.

Smith A, Suri S (1999) Rectangular tiling in multi-dimensional arrays. *Proc 10th Annual ACM-SIAM Symposium on Discrete Algorithms*, 786–794.

Smit AFA, Green P. RepeatMasker; `http://ftp.genome.washington.edu/RM/RepeatMasker.html`.

Smith TF, Waterman MS (1981) Comparison of biosequences. *Adv Appl Math*. **2**: 482–489.

Snyder EE, Stormo GD (1993) Identification of coding regions in genomic DNA: an application of dynamic programming and neural networks. *Nucleic Acids Res*. **21**: 607–613.

Solomon MJ, Varshavsky A (1985) Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. *Proc Natl Acad Sci USA* **82**: 6470–6474.

Strausberg RL, Feingold EA, Klausner RD, Collins FS (1999) The mammalian gene collection. *Science* **286**: 455–457.

Sun LV, Chen L, Greil F, Negre N, Li TR, Cavalli G, Zhao H, Van Steensel B, White KP (2003) Protein-DNA interaction mapping using genomic tiling path microarrays in *Drosophila*. *Proc Natl Acad Sci USA*. **100**: 9428–9433.

Suzuki Y, Yoshitomo-Nakagawa K, Maruyama K, Suyama A, Sugano S (1997) Construction and characterization of a full length-enriched and a 5′-end-enriched cDNA library. *Gene* **200**: 149–156.

Stolc V, Gauhar Z, Mason C, Halasz G, van Batenburg MF, Rifkin SA, Hua S, Herreman T, Tongprasit W, Barbano PE, Bussemaker HJ, White KP (2004) A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* **306**: 655–660.

Tjaden B, Saxena RM, Stolyar S, Haynor DR, Kolker E, Rosenow C (2002) Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. *Nucleic Acids Res*. **30**: 3732–3738.

van Steensel B, Henikoff S (2000) Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat Biotechnol*. **18**: 424–428.

van Steensel B, Delrow J, Henikoff S (2001) Chromatin profiling using targeted DNA adenine methyltransferase. *Nat Genet*. **27**: 304–308.

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. *Science* **270**: 484–487.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*. (2001) The sequence of the human genome. *Science* **291**: 1304–1351.

Vernon D (1991) *Machine Vision: Automated Visual Inspection and Robot Vision*. New York – London: Prentice Hall.

Wang X, Seed B (2003) Selection of oligonucleotide probes for protein coding sequences. *Bioinformatics* **19**: 796–802.

Weinmann AS, Yan PS, Oberley MJ, Huang TH, Farnham PJ (2002) Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarry analysis. *Genes Dev*. **16**: 235–244.

Wells J, Yan PS, Cechvala M, Huang T, Farnham PJ (2003) Identification of novel pRb binding sites using CpG microarrays suggests that E2F recruits pRb to specific genomic sties during S phase. *Oncogene* **22**: 1445–1460.

White EJ, Emanuelsson O, Scalzo D, Royce T, Kosak S, Oakeley EJ, Weissman S, Gerstein M, Groudine M, Snyder M, Schübeler D (2004) DNA replication-timing analysis of human chromosome 22 at high resolution and different developmental states. *Proc Natl Acad Sci USA*. **101**: 17771–17776.

Wootton SC, Federhen S (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comp Chem*. **17**: 149–163.

Xu Y, Uberbacher EC (1997) Automated gene identification in large-scale genomic sequences. *J Comput Biol*. **4**: 325–338.

Yamada K, Lim J, Dale JM et al. (2003) Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* **302**: 842–846.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*. **30**: e15.

Yelin R, Dahary D, Sorek R, Levanon EY, Goldstein O, Shoshan A, Diber A, Biton S, Tamir Y, Khosravi R, Nemzer S, Pinner E, Walach S, Bernstein J, Savitsky K, Rotman G (2003) Widespread occurrence of antisense transcription in the human genome. *Nat Biotechnol*. **21**: 379–386.

Yu H, Luscombe NM, Qian J, Gerstein M (2003) Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet*. **19**: 422–427.

Yu H, Greenbaum D, Xin Lu H, Zhu X, Gerstein M (2004) Genomic analysis of essentiality within protein networks. *Trends Genet*. **20**: 227–231.

Zhang Z, Berman P, Miller W (1998) Alignments without low-scoring regions. *J Comp Biol*. **5**: 197–210.

Zhang Z, Harrison PM, Liu Y, Gerstein M (2003) Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res*. **13**: 2541–2558.

Index

14
BAUHAUSBÜCHER

L. MOHOLY-NAGY