

AD _____

Award Number: W81XWH-04-1-0477

TITLE: High Resolution Analysis of Copy Number Mutation
in Breast Cancer

PRINCIPAL INVESTIGATOR: Michael H. Wigler, Ph.D.

CONTRACTING ORGANIZATION: Cold Spring Harbor Laboratory
Cold Spring Harbor, NY 11724

REPORT DATE: May 2005

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

20051013 012

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 074-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503				
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE May 2005		3. REPORT TYPE AND DATES COVERED Annual (1 Apr 04 - 31 Mar 05)
4. TITLE AND SUBTITLE High Resolution Analysis of Copy Number Mutation in Breast Cancer			5. FUNDING NUMBERS W81XWH-04-1-0477	
6. AUTHOR(S) Michael H. Wigler, Ph.D				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Cold Spring Harbor Laboratory Cold Spring Harbor, NY 11724 E-Mail: wiggler@cshl.edu			8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012			10. SPONSORING / MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited				12b. DISTRIBUTION CODE
13. ABSTRACT (Maximum 200 Words) Cancer is a disease associated with both germline and somatic mutation, and undoubtedly evolves in the host through increasingly malignant states by changes in the genome. Many of these changes can be seen by making gene copy number measurements, wherein increased copies of genes are associated with oncogenes and decreased copy numbers are associated with tumor suppressor genes. It is our hypothesis that these copy number changes, if measured with sufficient accuracy and resolution, can be used for two important purposes: to define precisely the mutant genes that cause cancer, and to define molecular markers that correlate with malignant potential and response to therapy. The technique we have developed, ROMA (representational oligonucleotide microarray analysis), accomplished this. We have completed data acquisition of approximately 200 breast cancer biopsies and cell lines. We discern many loci that are commonly deleted or amplified in breast cancer but not in normal genomes. Many of these loci confirm previous knowledge, but many are as yet unexplained. Our studies further suggests that genome instability, the presence of amplifications and deletions, is a marker for poor survival, and we have developed mathematical measures of copy number profiles that accurately predict				
14. SUBJECT TERMS Genome Micro-Arrays, Representational Oligonucleotide Micro-Array Analysis (ROMA, Oncogenes Tumor Suppressor Genes, Comparative Hybridization, Copy Number Aberration, Early Detection.				15. NUMBER OF PAGES 75
				16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT Unlimited	

Table of Contents

Cover.....	1
SF 298.....	2
Table of Contents.....	3
Introduction.....	4
Body.....	4
Key Research Accomplishments.....	6
Reportable Outcomes.....	6
Conclusions.....	6
References.....	6
Appendices.....	7

1. Introduction

2. Body

Below is the original statement of work, which was submitted in October of 2003. We have made excellent progress on all phases of the work, but because of the costs to accomplish the work, some of the uncharted technical difficulties, new opportunities, and the limits of time, our work is incomplete.

- I.** **A.** We will catalog loci that are amplified or deleted in breast cancers using ROMA, and compare these to a database of normal copy number polymorphism in the human population. **B.** We will correlate the presence or absence of amplifications and deletions with the outcome and clinical staging of the disease, and with Herceptin responsiveness to develop ROMA-based disease markers. **C.** This information will also be used to generate FISH probes for cytological analysis of the disease at its earliest stage, when only a minimal number of cells are available for pathological analysis.
- II.** We will analyze frozen and formalin fixed tumor specimens as part of three studies: correlates of progression in ductal-carcinoma-in-situ; correlates of Herceptin responsiveness; and correlates of survival in advanced node negative cancer. We will analyze on the order of 300 specimens in each of two years, the duration of this study. The data will be publicly posted after the first year, and as it is collected thereafter, providing a database of recurring cancer abnormalities.
- III.** We will complete software development so that ROMA technology can be widely accessed by research and clinical oncologists, enabling, for example, the inclusion of ROMA data in clinical trials. Our work will include creation of relational databases, completion of segmentation algorithms, epicenter analysis algorithms, and gene-centric and event-centric querying systems. The genes in the epicenters of loci that are commonly amplified or deleted in breast cancers will be posted, so that they can serve as a starting point for the development of new cancer therapeutics.

I now describe the work accomplished and a projection of the road ahead.

- I.** **A.** We have completed data acquisition of approximately 200 breast cancer biopsies and cell lines, and approximately 300 normal genomes. We have developed a method to determine the loci that are commonly deleted and amplified in breast cancer but not in normal genomes, and are preparing a paper to publish much of our findings. Many of these loci confirm what is known, but many are new. At many loci, even very common ones, there are more than one candidate oncogene and tumor suppressor gene, and so the continued analysis of breast cancers will shed light on the exact genes driving cancer malignancy. Moreover, we have made the surprising discovery that there are regions in the breast cancer genome which are sometimes amplified but hardly ever deleted, even hemizygosly. These regions may contain the best targets for future chemotherapy. Analyses of further biopsies are required here as well to better delineate these regions.

The analysis of the normal genomes is an integral part of our campaign, because if we did not know which genome copy number changes were actually present in healthy genomes, we might mistake these when they are observed in cancers. The discovery of the normal copy number variation is found in Sebat et al., 2004.

B. We have examined the genome of diploid breast cancers, small at presentation, from node negative women. In general this group has a favorable prognosis, but some women nevertheless die. Our study suggests that genome instability, the presence of amplifications and deletions, is a marker for poor survival. Markers of poor prognosis would suggest more vigorous treatment for these women who are at higher risk, and so is an important clinical finding that will

impact survival. We wish to continue this study to more precisely quantify the risk, and develop a reduced set of markers that would form the basis of an affordable clinical trial.

Similar methodology can discriminate between diploid and aneuploid genomes, often useful markers for malignant potential, with aneuploid cancers being far more lethal.

C. Fluorescence in-situ hybridization is one of the most sensitive measures of copy number changes in single cells, and we have been working with one of the world's leaders in this technology, Anders Zetterberg, of the Karolinska Institute in Stockholm. He has confirmed many of our findings of copy number changes in breast cancer, and together we have established a new methodology for designing FISH probes. We use computational methods for repeat analysis to design PCR primers that generate relatively repeat free templates for hybridization. We are presently developing an even more powerful method to generate FISH probes by entirely synthetic chemical methods, using microarrays themselves to fabricate the primers for generating the probes. The result will yield more reproducible and interpretable results in a clinical laboratory setting.

- II. Analysis of formalin fixed material and ductal carcinoma in situ (DCIS) have presented some of our greatest technical challenges. Analysis of formalin fixed material is a critical step in making genome measurements commonplace in a clinical setting, because it is the favored way by which pathologist preserve specimens for histological inspection. Unfortunately, DNA is cross-linked in formalin fixed material and so is resistant to our standard protocol, which is based on amplification of fragments in the range of 200-1200 bp. However, we determined that fragments in the size range of 200-400bp can be amplified reliably from tissue fixed in formalin, and so have redesigned our protocol accordingly. To build a set of microarray probes that are able to detect these fragments is a long process, requiring the analysis of several million possible oligonucleotide candidates, and we are roughly half way through the stages of developing such microarrays. We will complete this process in the next grant period.

Analysis of the genome in DCIS is one of our most important goals, and one that is most aligned with the programmatic mission. If we can detect neoplasias early, we must be able to accurately assess their risk. This task is complicated because very few cells are available for analysis. We have therefore developed new protocols based on an initial amplification from DNA based on a highly processive DNA polymerase from bacteriophage phi 29. Early results indicate that we can obtain reasonably accurate analyses from as few as a hundred cells, which is more than enough to analyse DCIS. We propose to continue this protocol development and test it against DCIS in a collaboration with the pathologists at MSKCC in New York. One remarkable possibility that our new technology offers is the prospect of analysing the genome from the few epithelial cells circulating in blood or present in nipple lavage. If detected, abnormal genomes might serve as a most sensitive early indicator of the presence of cancers.

- III. We have completed many but not all of our original software development aims. In particular, we have completed segmentation algorithms that allow us to confidently detect segments undergoing copy changes in cancer and normal genomes (Sebat et al., 2004; Olshen et al., 2004; and Durawala et al., 2004), we have completed methods for interpreting the segments and performing epicenter analysis, that is the defining of the minimum recurrent loci which are probably driving the evolution of the cancer cell. We have completed the first drafts of another algorithm which we call "compression". Compression finds the "most informative" probes in our set, in terms of the changes observed in cancers. This algorithm accomplished two goals. First, it enables us to rank the loci in terms of a "significance" that may be equal or greater in value than "frequency". Second, it provides a guide in the development of arrays with a smaller number of probes, which are important when designing more affordable arrays for further data gathering and clinical trials.

In order to that our initial findings result in improvements in clinical practice, clinical trials and statistically verifiable results are needed. To perform such clinical trials, for example to determine markers that correlate with Herceptin response, we must make our technology affordable. This we are accomplishing by continued software and protocol development. In particular, we are perfecting the use of microarrays fabricated at higher probe densities, and assays performed in micro-chambers from one fabrication unit. This requires working out new algorithms for gridding images obtained from the highest resolution scanners that are commercially available (1 micron scanners) and from non-rectilinear probe grids. This translates into great cost reduction, because then up to twelve complete genome scans can be packed into each fabrication unit as twelve separate circular hybridization chambers. We expect to accomplish these goals within a year.

We have not completed our goal of creating the types of relational databases that enable an outsider to query all of our data in a web based format. This has proven to be almost as stubborn a problem as the English language itself. We are now in the process of hiring specialists in this type of software development now that we have a body of interpretable data that should be shared.

3. Key Research Accomplishments

We have completed copy number profiles on over 200 breast cancer genomes and over 300 genomes from apparently normal individuals at 35kb resolution (85K probes). We have validated our results by FISH (florescent-in-situ-hybridization). We have demonstrated that we can distinguish aneuploid and diploid tumor genomes based on copy number profile, and in diploid cancers, we can predict five year survival outcome with nearly 90% accuracy by copy number profiling alone. We have made strides with a variety of lab based protocols: performing ROMA on samples with as few as a hundred cells; developed a higher resolution microarray (400K probes rather than 85K) that should be suitable for more detailed genomic analysis and the analysis of formalin fixed specimens; we have discovered how to reuse our microarrays, thereby reducing costs by a factor of two or more; we have developed software and laboratory procedures for the design of inter-phase FISH primers. We have also made progress in developing database and data processing infrastructure and tools: design of sample and experimental relational databases to track and access data collection; segmentation protocols based on hidden markov models for defining the boundaries of segmental copy number changes; developed mathematical methods for epicenter mapping; and developed mathematical methods that allow classification of disease status based on genome profiles.

4. Reportable Outcomes

We published methods to design probes for gene copy number measurements (Healy et al., 2003), and demonstrated that ROMA can be used to measure copy number changes in breast cancers (Lucito et al., 2003). We have published methods for segmenting gene copy number measurements (Sebat et al., 2004; Olshen et al., 2004; and Daruwala et al., 2004). We have shown that ROMA can be used to detect segmental copy number polymorphisms in apparently healthy genomes (Sebat et al., 2004).

5. Conclusions

The major conclusion of our research so far is that the number, extent and position of the copy number abnormalities in breast cancer genomes can be correlated with disease status and outcome. As a practical matter, this means that in my opinion it is likely that in the near future genomic profiling will be a clinical diagnostic tool used to guide the physician in choosing the optimum treatment of the cancer patient, tailoring that treatment to the type of genomic abnormalities seen in her tumor. Moreover, we are discovering many new oncogenes and tumor suppressor loci that may reveal new drug targets for cancer therapies. The analysis of more cancer genomes by ROMA, and at higher resolution, will improve both our ability to make clinical correlations and define culprit gene mutations.

6. References

Healy, J., Thomas, E., Schwartz, J.T. and Wigler, M. (2003). Annotating large genomes with exact word matches. *Genome Research* 13: 2306-2315.

Lucito, R., Healy, J., Alexander, J., Reiner, A., Esposito, D., Chi, M., Rodgers, L., Brady, A., Sebat, J., Troge, J., West, J., Rostan, S., Nguyen, K.C.Q., Powers, S., Ye, K.Q., Olshen, A., Venkatraman, E., Norton, L. and Wigler, M. (2003). Microarray analysis of genome copy number variation. *Genome Research* 13: 2291-2305.

Olshen, A., Venkatraman, E.S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 4: 557-572.

Sebat, J., Muthuswamy, L., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T.C., Trask, B., Patterson, N., Zetterberg, A., Wigler, M. (2004) Large-scale copy number polymorphism in the human genome. *Science*, 305: 525-528.

Daruwala, R.-S, Rudra, A., Ostrer, H., Lucito, R., Wigler, M. and Mishra, B. (2004) A versatile statistical analysis algorithm to detect genome copy number variation. *Proc. Natl. Acad. Sci.*, 101: 16292-16297.

7. Appendices

Manuscripts included in appendix.

Healy et al 2003

Lucito et al 2003

Olshen 2004

Sebat et al 2004

Daruwala 2004

Methods

Annotating Large Genomes With Exact Word Matches

John Healy,^{1,3} Elizabeth E. Thomas,¹ Jacob T. Schwartz,² and Michael Wigler¹

¹Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA; ²Courant Institute of Mathematical Sciences, New York University, New York, New York 10003, USA

We have developed a tool for rapidly determining the number of exact matches of any word within large, internally repetitive genomes or sets of genomes. Thus we can readily annotate any sequence, including the entire human genome, with the counts of its constituent words. We create a Burrows-Wheeler transform of the genome, which together with auxiliary data structures facilitating counting, can reside in about one gigabyte of RAM. Our original interest was motivated by oligonucleotide probe design, and we describe a general protocol for defining unique hybridization probes. But our method also has applications for the analysis of genome structure and assembly. We demonstrate the identification of chromosome-specific repeats, and outline a general procedure for finding undiscovered repeats. We also illustrate the changing contents of the human genome assemblies by comparing the annotations built from different genome freezes.

Any genome can be conceptualized as a string of letters. Every word composed of those letters has a certain number of exact matches within the genome, its word count. Knowledge of word count is useful for probe design, discovery of repeat elements, genome annotation, and mathematical modeling of genome evolution. The tools available for sequence homology analysis, such as BLAST and FASTA (Pearson and Lipman 1988; Altschul et al. 1990) were not designed for this purpose, and are unnecessarily cumbersome. We therefore sought a new tool for finding the word counts for words of arbitrary length in any given genome.

Our interest in this problem has its origins in microarray hybridization analysis. We have developed methods using oligonucleotide probes for detecting gene copy number changes in mutant and normal genomes (Lucito et al. 2003). We require our probes to be highly unique in the genome. Our approach, like that others have taken, is to count the exact matches of probe substrings, or words, to the rest of the genome (Li and Stormo 2001). When such words have lengths below 16, this task can be accomplished using a simple tabulation of words and their counts. When the word length exceeds 15, such directly addressable structures become impractical. More robust data structures, such as the suffix array and suffix tree, could easily provide us with optimal or nearly optimal theoretical time bounds for word count determinations. However, in practice, these too proved to be impractical solutions for the case of the human genome for reasons that we will detail. We solved this problem by applying and building upon a Burrows-Wheeler transform of the entire human genome sequence.

The tool we created is capable of rapidly annotating any sequence, even the entire genome, with the counts of its constituent words. We quickly realized that this method has other applications beyond probe design. In this article we describe our algorithm, provide some implementation details, and then discuss the relationships between our implementation and pre-existing tools and data structures. Lastly, we illustrate some applications to the analysis of genome structure and assembly.

METHODS

Fundamentals

To determine word counts rapidly, we sought to minimize the number of computations per word and to eliminate time-consuming disk access operations. We achieve this by creating a data structure that we can efficiently query and that can also reside entirely in random access memory (RAM). Our solution depends upon the Burrows-Wheeler transform, a method used to create a reversible permutation of a string of text that tends to be more compressible than the original text (Burrows and Wheeler 1994). It is also strongly related to the suffix array data structure (Manber and Myers 1993) in ways that will be made apparent.

First, given a genome G of length K , we create a new string $G\$$ of length $K+1$ by appending a "\$" to the end of that genome. (We assume a single strand, reading left to right.) We then generate all $K+1$ "suffixes" of $G\$$, where the suffixes are the substrings that start at every position and proceed to the end. We next associate with each suffix the letter preceding it. In the case of the suffix that starts at the first position, we associate the new \$ character and assume that "\$" has the lowest lexicographical value in the genome alphabet. The string of antecedent letters, in the lexicographical order of their suffixes, is the Burrows-Wheeler transform of the $G\$$ string, which we refer to as the "B-W string" or the "BWT".

For example, if the genome were simply "CAT", our $G\$$ string would be "CAT\$". Then the suffixes of the genome in sorted order would be: "\$", "AT\$", "CAT\$", and "T\$". The Burrows-Wheeler transform of this particular $G\$$ would be the letters that "precede" each of those suffixes taken in the same order, specifically "TCSA". In practice, the sort operation is performed on the integer offsets, or pointers, into the original string based on the suffix that starts at that position. To continue the example, the list of pointers taken in the order of the sorted suffixes would be [3,1,0,2]. This list of pointers is in fact the suffix array for the string "CAT\$".

We could use the suffix array to compute word counts using a binary search (Manber and Myers 1993). However, the suffix array for the human genome, at approximately 12 gigabytes (3 billion, 4-byte integers), is too large to fit in RAM for any of our machines. We would also need access to the entire genome in order to perform such a binary search, adding another 3 gi-

³Corresponding author.

E-MAIL healy@cshl.edu; FAX (516) 367-8381.

Article and publication are at <http://www.genome.org/cgi/doi/10.1101/gr.1350803>. Article published online before print in September 2003.

gabytes uncompressed. On the other hand, the B-W string is alone sufficient to determine word counts. Recall that it is no larger than the original genome and, like any other string of characters, it can be compressed using any of a variety of text compression techniques. Furthermore, in our implementation, all but a negligibly small portion of the compressed form can remain so throughout execution. Together with auxiliary data structures that enable the B-W string to be rapidly queried, the entire structure for the human genome can be compressed into a little over 1 gigabyte of RAM.

The Basic Algorithm

Heuristically, the B-W string can be viewed as a navigational tool for a "virtual" Genome Dictionary, an alphabetical listing of all the suffixes of the human genome. Suppose we wish to know whether a substring occurs in the genome, and if so, in how many copies. Let us first consider the case where the substring is a single character, X . We can view all the occurrences of X in the Genome Dictionary as a block where F_X and L_X are the indices of its first and last occurrence, respectively. The size of this block, $k_X = L_X - F_X + 1$, is the number of occurrences of X , and is readily determined by counting the number of occurrences of X from the beginning to the end of the B-W string.

In order to consider the case for longer words, we first need to determine F_X , L_X , and k_X for each character X of the genome alphabet. The sizes of each block, the k_X 's, are easily determined by counting the instances of X in the B-W string. F_X is one plus the sum of the sizes of all antecedent blocks beginning with V , where V is any character occurring lexicographically before X . L_X is one less than the sum of k_X and F_X . We store the F_X and L_X for each letter X in an auxiliary data structure called "alphabounds".

We can now proceed inductively to find the count for a word Z . Suppose W is a suffix of Z , W exists in the genome, and we know the indices F_W and L_W of its block in the Genome Dictionary (Fig. 1A). To continue the induction we need to know whether XW exists as a substring, where X is the character preceding W in Z , and we need to know the indices of the XW block, F_{XW} and L_{XW} , in the Genome Dictionary.

If and only if X occurs in the B-W string between F_W and L_W , then XW exists as a substring of the genome (Fig. 1A). Furthermore, the number of X 's in the "W block" of the B-W string, k_{XW} , is the copy number of the substring XW in the genome. Finally, the indices of XW are easily computed, namely:

1. $F_{XW} = F_X + b_{XW}$, and
2. $L_{XW} = F_{XW} + k_{XW} - 1$

where b_{XW} is the number of words beginning with X in the Genome Dictionary that occur before XW . Recall that F_X has been determined for each character X of the alphabet. b_{XW} can be determined by counting the number of X 's that occur before the W block of the B-W string (Fig. 1A).

We reiterate this procedure, lengthening the suffix one character at a time, stopping if the suffix does not exist in the Genome Dictionary. If the suffix W encompasses the entire word Z , k_W is the number of occurrences of Z in the genomic string. An outline of this procedure in pseudocode is displayed in Figure 1B.

The basic algorithm transforms a pattern matching problem into a counting problem. Counting thus becomes the rate-limiting step, and therefore we facilitate it in ways described below.

Preprocessing and Database Construction

To count exact matches of words our method requires only the B-W string, but to build the string we still need to create a suffix

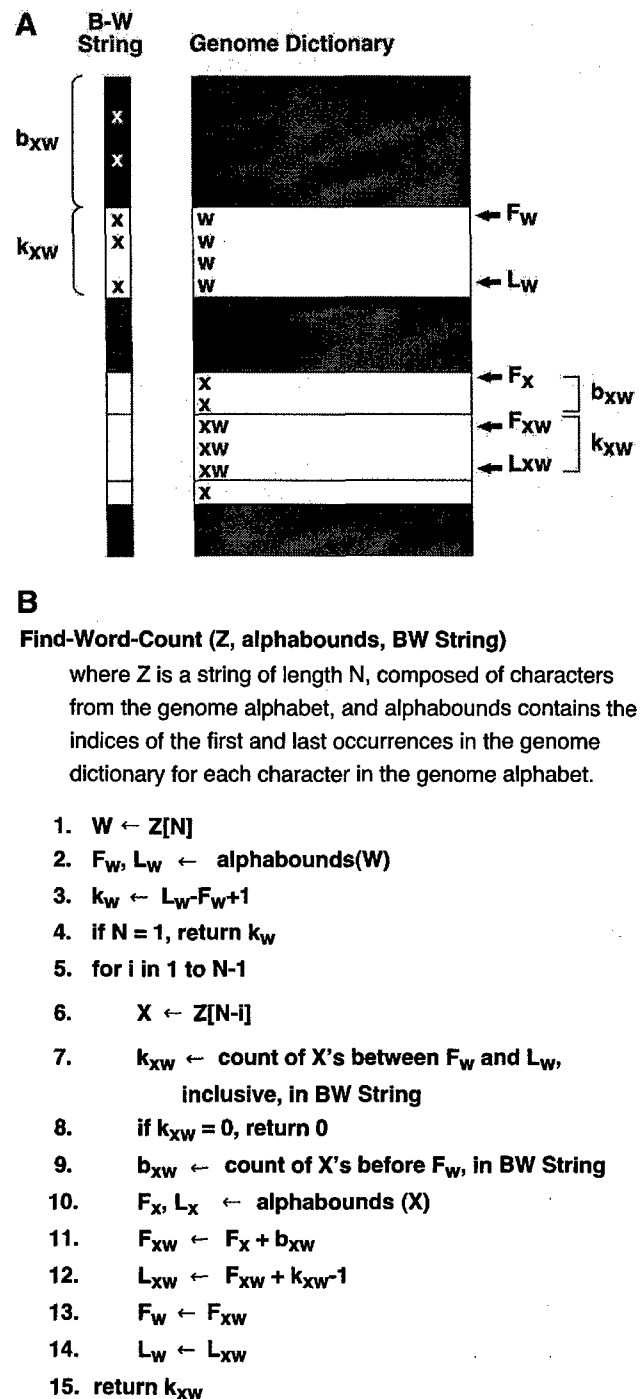


Figure 1 Our algorithm for rapidly determining the exact word counts in a large string for any length word. (A) Graphically defines the variables and data structures used in the algorithm. (B) A pseudocode representation of the algorithm itself.

array for the genome. Although the suffix array is not needed to determine word frequency, and is much too large to be held in RAM, it should be retained on disk, because it can also be used to find the coordinates of matches.

Building a suffix array can be reduced to a "sort in-place" operation. For a string of the size of the genome, we imple-

mented a parallel radix sort using a 16-node cluster. The genome was divided into 100 equal-size substrings, each overlapping by seven nucleotides. The offsets into the genome (i.e., the "genome" coordinate) within each substring were then assigned to one of 5^7 "prefix" bins according to the 7-mer at each offset. (The genome alphabet was A, C, G, T, and N.) The offsets within each prefix bin were then sorted based on the sequence following the 7-mer prefix, creating the suffix array.

For the human genome, we made a special case for N's. The human sequence contains about 6×10^8 N's, mainly in large blocks ranging from 200 kb to 30 Mb in length. The presence of these long blocks increased sort time by a factor of 10, so we decided not to sort coordinates with 7-mer prefixes containing N's. As long as the constituencies of blocks bounded by prefixes containing N's are correct, their internal order is irrelevant for determining counts of N-free words. Thus, all queries with sequences containing no N's are still valid. We refer to this variant as the "N-incomplete" Burrows-Wheeler transform.

The first character preceding each offset, taken in the order of the sorted offsets, constitutes the B-W string. The B-W string is still three gigabytes, too large for our workstations. To compress the string further, we used a simple dictionary-based compression scheme, where one of 125 distinct single byte codes represents one of each of the 5^3 possible three-letter substrings. We chose this compression scheme, even though greater compression can be achieved, because it has a constant compression ratio, 3:1, and allows us to count characters, for the most part, without decompressing.

In the pseudocode for our counting algorithm, all steps are rapid "look-ups" or simple computations except for steps 7 and 9 (Fig. 1B). These steps involve counting the B-W string over potentially large blocks of characters. In order to speed counting, we created an auxiliary data structure, the "K-interval counts", where K is an integer multiple of the compression ratio, by pre-counting on the B-W string. We determine the cumulative counts for each character and record them for every K^{th} position. To carry out counting steps, therefore, we need only count the particular character in the string from the relevant position to the nearest position that is a multiple of K. The number of characters that needs to be counted in any step is thus no more than $K/2$. In our implementation we set K equal to 300 characters, or, equivalently, 100 compressed bytes.

We have also experimented with the notion of subintervals of size $K^{\wedge}$ within each interval K. At every $K^{\wedge\text{th}}$ position within each K-interval, we record how many instances of each character we have seen since the beginning of the encompassing interval. If we limit the size of $K^{\wedge}$ to be $< 2^8$, for example, then the counts for each letter at every K-interval can be recorded using a single byte. This allows us to increase the "density" of the counting index by a factor of $K/K^{\wedge}$ while increasing the space requirements for the K-interval counts by a factor of only $[(K/K^{\wedge})/4]$. We have implemented a variant of our data structure that utilizes this hierarchical indexing scheme. Depending on the choices of K and $K^{\wedge}$, we have seen a three- to fivefold increase in query execution speed while maintaining a memory requirement of less than two gigabytes for the human genome. Full details of this variant are provided at our Web site (<http://mer-engine.cshl.edu>).

To further speed the counting process we introduce a final data structure, the "dictionary-counts". Recall our simple 3:1 compression scheme, where bytes 0 through 124 decompress to "AAA" through "TTT" respectively. The dictionary-counts structure is a small two-dimensional array that can be thought of as a matrix with 125 rows and five columns. Each row corresponds to one of the compression dictionary entries, and each column corresponds to each letter of the genome alphabet, A through T. Let us assume, for instance, that we are in the process of determining

the number of A's in step 9 of Figure 1B. Using the K-interval counts structure described above, we can "jump" to within at most 50 bytes of our current F_w in a single look-up. Let us also assume that this F_w is pointing to the third "T" in a compressed "ATT" which is in turn at the 49th byte of the interval. For each of the 48 preceding bytes, we simply use the byte as the row number in our dictionary-counts array and the letter of interest, "A", as our column "number". At those coordinates we will find the number of times that the letter "A" appears in that compressed byte. Therefore we must perform 48 look-ups in this small directly addressable table. Finally, we must decompress the 49th byte with another simple table look-up, and examine the first two letters "AT". The dictionary-counts may seem like a minor component. However, when it is combined with the K-interval counts structure, the act of counting any number of characters requires only $K/6 + 1$ table look-ups, plus two character comparisons in the worst case. In actuality this structure requires approximately 65 kilobytes of memory. It is also the data structure used for the majority of all computations in any single iteration of our algorithm.

We refer to the joint data structures and search protocols as the "mer search engine" or simply the "mer-engine".

Validation for the Human Genome

The most rigorous way to validate all the data structures and protocols we have just described is to perform a reverse transform. Starting with the position in the B-W string corresponding to the last character of a genome, and using the protocols and data structures just described, one should be able to reconstruct that genome sequence. However, the N-incomplete transform of the human genome is not a proper Burrows-Wheeler transform string, and hence the full genomic string cannot be reconstructed from it.

Therefore to validate our human mer search engine, we picked at random a million words of varying lengths, from three to 1000 characters. We determined the word count and coordinates of each by scanning the genome text. We compared the word counts with those returned by the "mer search engine", and in each instance there was complete concordance. We also referred to the suffix array to obtain coordinates, and they agreed perfectly as well.

Performance for Genomes

The time complexity of a query for a particular word is linearly proportionate to the length of the word and to the size of K for the K-interval counts. We have tested our implementation on a Dell PowerEdge 1650 with dual 1GHz Pentium III processors and 4GB of physical memory running Linux. Importing a human chromosome from disk, annotating with counts of all overlapping words of length 24 (for both sense and antisense), and writing the results to disk takes an average of 1 min per megabase. This hardware configuration is now over 2 yrs old, and we expect significantly faster execution times on machines purchased today. Furthermore, we expect that our "dictionary-counts" data structure, requiring a mere 65 kilobytes, will take advantage of the so-called Level 1 cache of present day CPU architectures. These statistics do not take into account the addition of the subintervals of size k . We have experienced reductions in time requirements of up to fivefold through this simple modification. The disadvantage of this variant is the additional space requirement of roughly 750 MB.

The time required for the preprocessing stage is dominated by the construction of the suffix array. This operation requires a sort of all of the cyclical permutations of the genome. Therefore

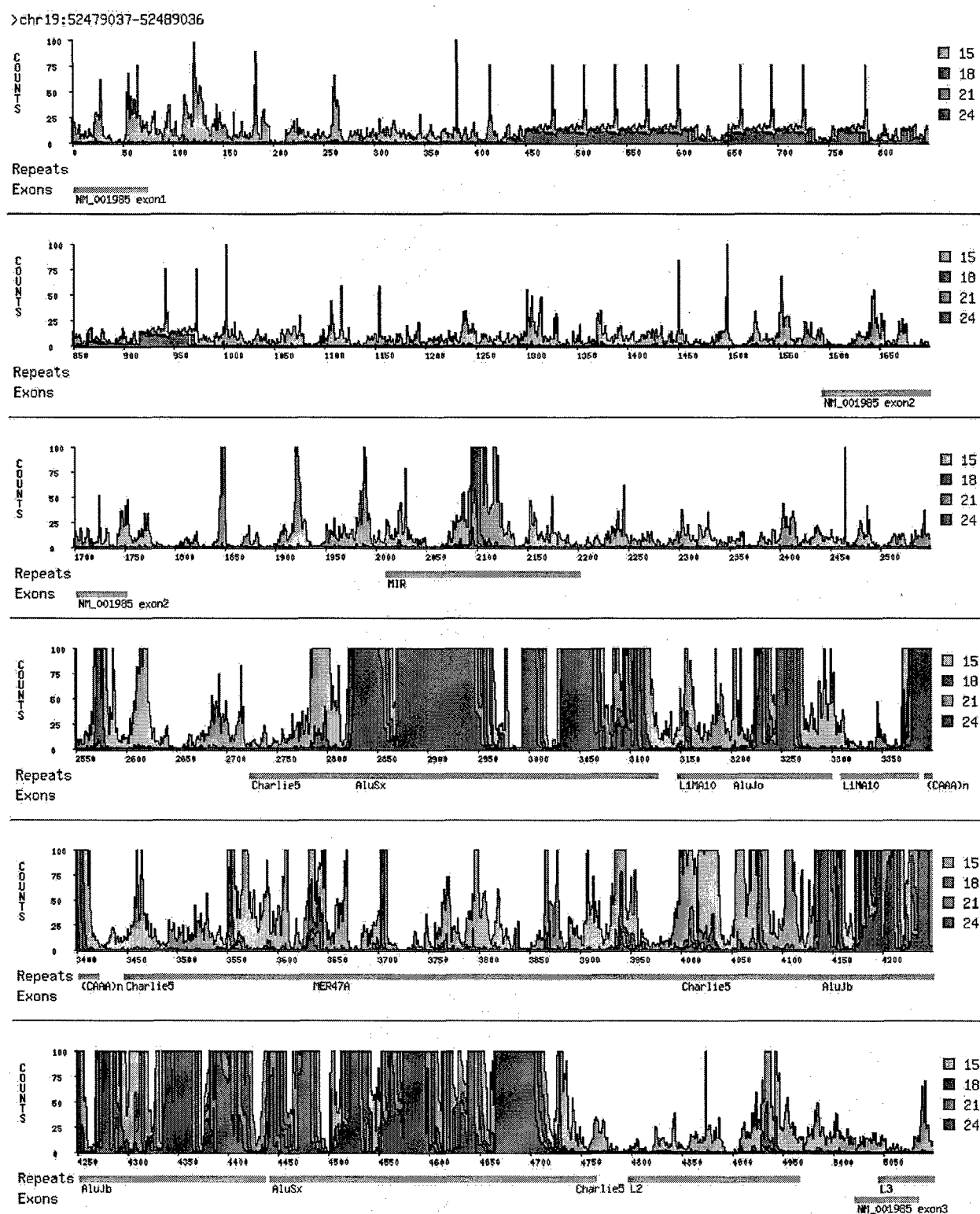


Figure 2 Word count terrain from a 5-kb region on chromosome 19. The coordinates of this region in the June 2002 assembly of the human genome are at the top. Along the x-axis is the relative position of a given word within the region; along the y-axis is the absolute word count, with counts for different lengths drawn in different colors, according to the legend. Word counts are capped at 100. Underneath the terrain, repeats detected by RepeatMasker are annotated in orange. Exons from the RefSeq data set are indicated in purple. In this case, the word counts are derived from the June 2002 assembly of the entire human genome.

the time complexity of the preprocessing stage in its entirety reduces to $O(n \lg n)$ where n is the size of the genome. For this process we make use of every node in our cluster, with a total execution time of approximately 6 h for a single assembly of the human genome.

Although there are a variety of ways for enhancing performance, the operating times both for preprocessing and annotation are reasonable with our current implementation.

Relation to Existing Tools

In the context of genome research, we are inclined to view our algorithm as a companion to methods or tools built around approximate homology searching such as BLAST and BLAT. In general we found that in pursuits such as repeat discovery and in particular probe design, our method reduces greater than 98% of the work to a simple and rapid "scan" operation; namely that of word-count annotation. The final analysis is then performed using a low-stringency approximate homology search on a vastly reduced set of "candidate" entities. In our probe design protocol described below, greater than 99% of our candidates already satisfied our full requirements prior to this final analysis. In this same sense, and rather appropriately, both of the approximate homology-based tools mentioned here use exact matches as their first-pass criteria before performing a more rigorous sequence alignment. It is feasible that our data structure could act as an alternative exact-matching "core" for variants of these tools.

We find that approximate homology tools alone are insufficient and impractical in such pursuits when performed at the whole-genome or multiple-genome scales. BLAST in particular, we find, tends to greatly multiply the amount of data that must be processed. For instance, when attempting to design unique probes within a large subset of the repeat-rich human genome, many of the candidate regions will have homology among themselves as well as within the entire genome. The resulting output contains the cross-products of these homologies. Furthermore, the best local alignments are reported, not all alignments. The output is therefore inadequate as an estimate of possible cross-hybridization in microarray experiments. BLAT, on the other hand, sacrifices completeness for speed; it cannot find matches for sequences that have a number of occurrences above a predetermined cutoff. Neither of these tools can readily yield a statistic that can be used as a measure of predicted cross-hybridization, such as an aggregate of counts for all constituent 15-mers.

REPuter (Kurtz and Schleiermacher 1999) is an existing program which can be used for repeat analysis and discovery as well as finding areas of uniqueness. It relies on exact pattern matching algorithms used for the traversal of its underlying data structure, which is a suffix tree. This program is a complete software solution for genome research in that it enables one to perform exhaustive repeat analysis, detection of unique substrings, and approximate alignments with statistics, all within a graphical user environment (Kurtz et al. 2001). Its usefulness in the context of the entire human genome and beyond, however, is limited due to tremendous memory requirements necessitated by the reliance on a suffix tree. We provide further detail of this issue in the following section.

Relation to Existing Data Structures

Our data structure could be described as a compressed index into a suffix array. The query process is essentially an attempt at a partial reverse Burrows-Wheeler transform of the query word within the context of the genome. A necessary component of this query process is a set of pointers into the suffix array, namely F_w

and L_w , which is carried through each iteration. In this way, the algorithm is an alternative to performing a binary search using the entire suffix array along with the entire genome. It is this freedom from the need to refer to coordinates of suffixes during search that allows us to achieve our tremendous space reduction. If there is further interest in retrieving the coordinates of every exact match, then the suffix array can be accessed as it normally would be; either from disk or active memory depending on available resources. It is worth noting, however, that there is a simple extension to our query algorithm that enables the retrieval of coordinates for all matches using only a small subset of the entire suffix array. Because our primary interest lies in the word count determinations alone, we refer the interested reader to our Web site for a full description of this extension (<http://mer-engine.cshl.edu>). In all comparisons made in this section, we assume this exclusive interest in word count queries within genomes.

A binary search through a suffix array can determine the count c of a word of length p within a genome of length n in $O(p \lg n)$ time while requiring $O(n \lg n) + O(n)$ bits of storage (Manber and Myers 1993). In practice, the suffix array for the human genome of length $n > 2^{31}$ requires a total of $5n$ bytes of storage; $4n$ bytes are required for the suffix array itself plus n bytes for the original string, all of which must be referenced throughout a search. If the hardware in use does not have sufficient RAM, then the search procedure is dominated by disk I/O operations. Disk retrievals are slower than active memory retrievals by many orders of magnitude. Our algorithm can perform a similar word count query in $O(pK)$ time requiring $O((n/K) \lg n) + O(n)$ bits of storage. In practice, our data structures for the human genome require $(n/3 + \{20 [n/(K/3)]\})$ bytes of storage where K is the size of the intervals in our K -interval counts structure. Herein lies the versatility of the mer-engine: K can be increased or decreased depending upon the requirements and available resources. If RAM is scarce then K can be increased by Q , resulting in a linear decrease in space requirements and similar increase in execution times, both proportional to Q .

Another data structure that is commonly used for exact pattern matching is the suffix tree. We refer the reader to Gusfield (1997) for a detailed description of suffix trees and the many possible variations on their construction and use in problems of exact and approximate pattern matching. A suffix tree requires $O(n \lg n) + O(n)$ bits of storage and $O(p)$ time to perform a word count for any word of length p which occurs c times within a genome of length n . Unfortunately these expressions, particularly the space requirement, do not translate directly to expected performance in modern computer architectures. Recall that the program REPuter uses a suffix tree as its underlying data structure (Kurtz and Schleiermacher 1999). The authors of that program present a method for reducing the space requirements of a suffix tree (Kurtz 1999), which is used by the REPuter program (Kurtz et al. 2001). However, REPuter is said to still require $12.5n$ bytes of storage for a suffix tree of a genome of length n (Kurtz and Schleiermacher 1999). This requirement is several times larger than the complete memory requirements of a suffix array. It is likely to be prohibitively large for all but the most expensive hardware platforms when applied to the entire human genome.

An "opportunistic data structure" based on the Burrows-Wheeler transform has been described (Ferragina and Manzini 2000) and is referred to as the "FM-Index". The core search algorithm for the FM-Index is nearly identical to the one described in our pseudocode and is used to perform word count queries. Through a very clever compression and indexing scheme, the FM-Index achieves space requirement bounds of $O((n / \lg n) \lg \lg n)$ bits of storage while being capable of performing word count

queries in $O(p)$ time for any word of length p within a genome of length n . This is true given the authors' assumption that their variable for the "bucket size" is assigned the value of $\lg n$. We'll refer to this parameter hereafter as b . This variable plays a role similar to that of our variable K in that it subdivides the transform string for better index performance. Note that for $K > (\lg^2 n / \lg \lg n)$ our implementation requires less space. If one increases the value for b beyond $\lg n$, particularly in the case where $n \geq 2^{32}$, they run the risk of dramatically increasing the space requirements for the FM-Index. More specifically, the structure referred to by the authors as " S " has space requirements bounded by the term $b2^{b^b}$, where b^b is the maximum size of any one of the (n/b) compressed buckets and has an upper bound of $c \lg n$ where $c < 1$. This means that the actual space requirements are dependent upon local properties of the transform string. Our space requirements are dependent only upon K and n (the alphabet size for genomes is negligibly small; however, it is a factor in practical space requirements for both the mer-engine and the FM-Index). If one decides to reduce b to avoid this risk, then our space requirement advantage increases.

The $O(p)$ time complexity for the FM-Index derives from the fact that within any iteration of the search procedure, where one iteration is performed for each of the p characters of the query word, counting is accomplished via look-ups within at least seven directly addressable data structures. Each of these look-ups requires constant theoretical time, so their combined time requirement reduces to $O(1)$. Recall the mer-engine variant in which subintervals of size K^b , where $K^b < K$ and $K < 2^8$ are introduced. Assume, for example, we choose values of $K = 240$ and $K^b = 15$. Then this mer-engine variant would require four table look-ups plus two character comparisons for each iteration of the search algorithm in the worst case. We believe this practical worst-case very nearly approximates a theoretical time complexity of $O(p)$ and has space requirements of roughly 60% of the original genome size n , including the compressed transform string, regardless of n . Furthermore, the last four steps of each iteration are isolated to accessing a structure that requires only 65 kilobytes of memory, again, regardless of n .

We could not locate any performance data for the implementation of the FM-Index referenced above. However, word count query performance for the *Escherichia coli* genome FM-Index has been analyzed for an implementation variant (Ferragina and Manzini 2001). The mer-engine for the human genome performs word count queries for words between eight and 15 nucleotides in length ~150 times faster than the *E. coli* implementation described therein. This does not take into account the speed-up that we observe with the introduction of subintervals to our K -interval counts structure. We believe this discrepancy may be accounted for by any combination of the following: difference in CPU clock speed, the fact that not all "buckets" remain in active memory for the duration of the test, and the requirement, in this particular variant, of complete decompression of buckets prior to the final counting stage. The authors Ferragina and Manzini (2001) do not mention any specific application of the FM-Index to genome research.

Alternative algorithms and data structures based on the Burrows-Wheeler transform have been defined (Miller 1996; Sadakane 1999). One algorithm relies heavily upon an additional "transformation matrix" which maps a character's position in the sorted list of all characters to its new position in the transform string (Miller 1996). The challenge with this strategy is finding a succinct way to store this transformation matrix, which starts out at exactly the same size as the suffix array for the same string. The other algorithm is simply a compressed form of a suffix array, which is entirely decompressed before performing a search (Sadakane 1999).

Availability

Our code for executing the Burrows-Wheeler transform is highly platform-dependent. That is to say, it was optimized for our particular cluster configuration and will likely require revision for general use. However, this code will be made available upon request, and all the information required for building the BWT is provided in the text above. To accommodate readers who wish to perform mer analyses without having to perform the Burrows-Wheeler transform, several preprocessed mer-engines are available. We have placed the BWT of the genomic strings for *S. pombe*, *C. elegans*, and *Fugu rubripes*, as well as the N-incomplete BWT of the June 2002 assembly of the genomic string for human chromosome 1 and the entire genome, and their auxiliary data structures, on our public Web site (<http://mer-engine.cshl.edu>) for downloading. Additionally, we have supplied C++ code that enables mer frequency queries from any of these strings residing either on disk or in RAM, and have provided the Perl code for visualizing the resulting C++ output (Fig. 2).

RESULTS AND DISCUSSION

Annotating Sequences With Word Counts

Using the above tools, any region of the genome can be readily annotated with its constituent mer frequencies. We have depicted annotations of a 5-kb region of chromosome 19 as a histogram in Figure 2, using four mer lengths, 15, 18, 21, and 24 bases. We call such annotations "terrains". For each coordinate and each word length, we determined the count of the succeeding word of the given length, in both the sense and antisense directions. We then plotted these counts on the y-axis, with each pixel on the x-axis corresponding to a coordinate. The heights of counts exceeding 100 are truncated, and each word length is color-coded (see Fig. 2 legend).

This region was picked somewhat arbitrarily, but it illustrates some major themes. We have taken repeat and exon annotations of this region from the human genome browser at UCSC (Karolchik et al. 2003) and aligned them to our terrain. There is significant discordance between annotated repeats and high terrain. We note that several regions annotated as repeats by the UCSC browser in fact have very low word counts, even with 15-mers. This is not unexpected, as our method is based on exact matches, and some repeats are very ancient and highly diverged. The relatively unique regions within repeats may nevertheless be useful for probe design, and the exact count method readily finds such regions.

To us, one of the most striking features of the terrain is the presence of narrow spikes in 15-mer counts. This is a virtually universal property of all regions of the human sequence we have examined, including coding exons. To develop a better understanding of this phenomenon, we decided to examine what the word count annotation of this region would look like if the genome were instead a randomly generated sequence, but with the same size and dinucleotide frequency distribution as the human genome. The terrain is still rough, but there are very few spikes. We hypothesize that these spikes result from the accidental coincidence of 15-mers in ordinary sequence with 15-mers present in high-copy-number repeats. Such high-copy-number sequences are not as frequently found in a random genome.

Computations on Subsets of the Genome

We also encounter regions of high terrain that are not annotated as repeats by RepeatMasker (<http://ftp.genome.washington.edu/RM/RepeatMasker.html>). RepBase (Jurka 2001), the database of repeats used by RepeatMasker, does not include region-specific or chromosome-specific repeats. With our method, such repeats are

Exons

NM_015383 18

Exons

NM_015383 17

GA-rich

Exons

NM_015383 16

Expos

NM_015383 15

GA-rich.

Exons

NM_015383 14

NM_015383 13

Expos

NM_015383.13

 $(GA)_n$

ET-rich
NM_015383.12

2312 **Genome Research**
www.genome.org

B

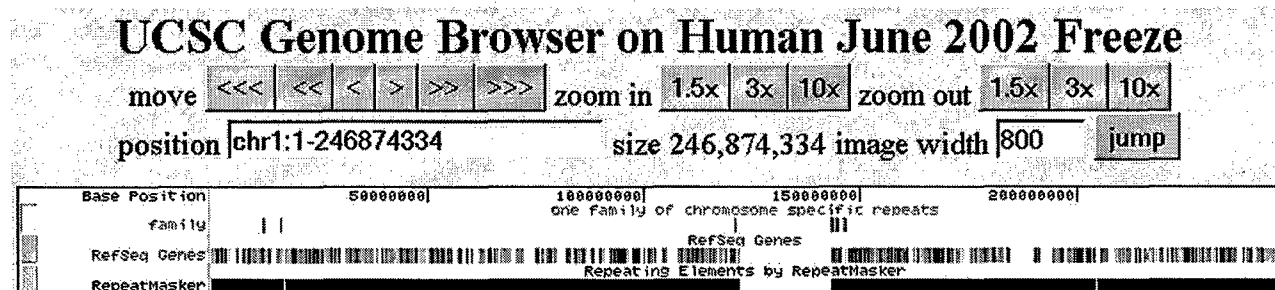


Figure 3 A chromosome 1-specific region. (A) This region was selected in the following way: 18-mers were identified whose chromosome 1 counts were $\geq 90\%$ of their whole genome counts. These 18-mers were strung together to create "chromosome-specific repeats" as long as the space between them was less than 100 bp. At the top of the figure are the coordinates of this region in the June 2002 assembly. Along the x-axis is the relative position of a given 18-mer within these coordinates. Along the y-axis is the absolute word count, with whole-genome counts drawn in gray and chromosome 1 counts in blue. Word counts are truncated at 100. Underneath the terrain, repeats detected by RepeatMasker are annotated in orange. Highlighted in purple are RefSeq exons that overlap this region from RefSeq gene NM_015383. (B) The chromosome-wide distribution of this family of chromosome-specific repeats, as viewed in the UCSC Genome Browser. The entire length of chromosome 1 is shown, with the purple "family" track indicating recurrences of this repeat. Below the family track are tracks that indicate both the positions of RefSeq genes and RepeatMasker annotations along chromosome 1.

easy to find because exact match counting can form the basis for a set algebra of the genome. In particular, we can make transform strings from subsets of the genome and examine the partition of words between these sets. Here we illustrate the use of this concept to find chromosome-specific repeats.

We made a transform string from chromosome 1 and annotated it with the word counts from itself and from the entire genome. We then looked for contiguous regions of chromosome 1, at least 100 bp in length, with high 18-mer counts in which the exact matches were found to derive mainly from chromosome 1. We readily found such regions, ranging in length from 100 bp to 35 kb. Focusing on one such region, we observed that its mer terrain was nearly a step function, composed of shorter sequences each with a signature modal frequency and length. We collected all of the chromosome-specific regions containing one of these signature regions and quickly identified a family of chromosome 1-specific sequences. Figure 3A illustrates the mer terrain for a portion of one of these family members; Figure 3B portrays the location of its recurrences on chromosome 1. At least one instance of this repeat has been annotated as overlapping a RefSeq gene (accession no. NM_015383), with many exons that together encode a large predicted protein sequence having low homology to myosin.

This is the first such repeat that we have investigated in any depth, and we expect to find other examples that merit attention. The same process by which we identify chromosome-specific repeats can be applied to finding repetitive DNA throughout the genome that is not recognized by RepeatMasker or other programs. One merely creates a mer-engine from the subsets of repeat sequences recognized by any pre-existing repeat analysis software of choice, and compares annotations from the whole genome mer-engine and the known repeat mer-engine to find unknown repeats.

Probe Design

Probes are generally useful for their ability to hybridize specifically to complementary DNA, and therefore one of the primary objectives in probe design is to minimize cross-hybridization. Some investigators have used repeat masking to exclude repeat regions from consideration for probe design. As we have described above, this is not a perfect solution, in that it does not protect the investigator from all regions that are repetitive, for

example, chromosome-specific repeats, and it excludes "repetitive" regions that are quite unique in actuality.

Although the rules for hybridization between imperfectly matched sequences are not well understood, it is clearly sensible to avoid probes that have exact "small" matches to multiple regions of the genome. Using a directly addressable data structure, such as a hash table, it would be a simple matter to store and retrieve counts for words as large as 14-mers. We could then attempt to minimize aggregate exact 14-mer match counts, but for genomic probes we think this method is inadequate. First, it is unclear that exact matches of 14-mers have any effect on hybridization under normally stringent annealing conditions. Nor do 14-mer counts predict homology, let alone uniqueness in the

Table 1. Size Distribution of Fragments Lost Between Assemblies of Chromosome 10

Fragment length interval (bp)	Percentage of total	Length of largest fragment in interval (bp)	Percentage of interval remapped
30–100	54	99	21
101–200	8	199	29
201–400	15.5	400	16
401–800	15.5	797	14
801–1600	5.3	1507	20
1601–3200	0.5	3008	100
3201–6400	0.6	5789	100
6401–12800	0.5	12293	100
12800+	0.1	21104	100

The fragments included in this distribution were chosen in the following way: the December 2001 assembly of Chromosome 10 was annotated with 18-mer counts within the entire December 2001 assembly as well as within the June 2002 assembly. We stored the coordinates of runs of at least 13 consecutive 18-mers whose counts transitioned from 1 to 0 between assemblies. These 18-mers were further clustered into "dropout fragments" as long as the gaps between them were not greater than 100 bp, and no more than 35% of the fragment length was composed of gaps. A homology search using BLAST was performed to compare the dropout fragments with the vector database; no homology to vector sequence was found. Approximately 800 dropout fragments were found, ranging from 30 bp to 21 kb in length with a combined length of approximately 300 kb.

genome. We have compared 16-mer counts to the geometric mean of counts from their constituent 14-mers, and we do not see a good correlation between the two for sequences that are essentially unique (data not shown).

We propose the following general protocol for probe design. First, choose the shortest length such that when the genome is annotated with mer-counts of exact matches of that length, sufficiently long stretches of uniqueness are found. Second, choose a shorter length such that exact matches of that length represent stable hybrids under the appropriate stringency conditions. Then, from the regions judged to be unique at the first length, choose probes that minimize the aggregate mer-counts of the second length. This protocol can be executed using the mer-engine tools we described in the previous section.

We followed this protocol in the accompanying article to select 70-mer probes from small BglII fragments (Lucito et al. 2003). We required uniqueness in the space of 21-mer counts, and then within these regions selected a 70-mer with the lowest sum of 15-mer counts, with a cut-off value of about 900. We added a few additional requirements, to eliminate runs of single nucleotides and severe base composition bias. Almost all probes picked by these protocols, and synthesized and printed on glass, in fact performed well under our microarray hybridization conditions.

We used BLAST to test whether probes picked by this protocol are indeed unique in the published genome sequence. We queried 30,000 such probes against the genome using the default parameters for MegaBLAST (filtration of simple sequence was turned off). More than 99% of our probes were unique over their entire length. However, for completeness, we suggest adding a final step to the probe design protocol, whereby all remaining candidates are subjected to a low-stringency approximate homology search against the genome in a best-first order.

Monitoring Genome Assemblage

As the human genome project progresses, new assemblies, based on freezes, are periodically released. We assume that each new assembly is an improvement upon the ones that came before. Because our probes derive from the December 2001 and April 2002 assemblies, downloaded from the UCSC Genome Browser, we have remapped the probe set to subsequent assemblies using BLAST and BLAT (Kent 2002). We also annotate them with the mer-engine built from each assembly because we have greater confidence in the hybridization ratios found for probes with stable copy number and map location.

An unexpected result of this process was that 1.2% of our probes vanished from the June 2002 assembly. That is to say that for 1.2% of our probes, all of their constituent 21-mers went from copy number one in their original assembly, to copy number zero in a subsequent assembly. Yet these probes behave as expected in our microarray experiments: They have good signal and hybridize to fragments with the restriction endonuclease profile predicted from their original assembly (Lucito et al. 2003).

Our surprise at this outcome prompted us to investigate the extent of this phenomenon on a larger scale. The mer-engine is the appropriate tool for this exploration. We decided to look for all unique sequences within a single chromosome within an assembly that were lost between that assembly and a subsequent one.

In particular, we annotated all of chromosome 10 from the December 2001 assembly with genomic 18-mer counts from both the original assembly and the June 2002 assembly. We observed a large number of *n*-to-*m* transitions in 18-mer counts, where "*n*-to-*m* transitions" refers to a mer that went from *n* copies in the original assembly to *m* copies in the subsequent assem-

bly. Although we describe the 1-to-0 transitions in this report, we note that they represent a small percentage of all transitions. We call 18-mers with 1-to-0 transitions "orphans". We stored the coordinates of runs of at least 13 consecutive orphans. We further clustered the orphans into "dropout fragments" as long as the gaps between them were not greater than 100 base pairs, and no more than 35% of the fragment length was composed of gaps.

We performed a homology search, using BLAST, to compare the dropout fragments with the vector database to eliminate any possible vector contaminants from our set. No homology to vector sequence was found. In total we found approximately 800 dropout fragments ranging from 30 bp to 21 kb in length, with a combined length of approximately 300 kb. Table 1 provides a list of the size distribution of the fragments.

At the time of this writing, we were able to perform a remapping of the fragments to the April 2003 assembly, and the percentage of fragments that returned to the assembly are provided. Although some returned, we found that new orphans were also created (data not shown). The coordinates of the dropout fragments in the original December 2001 assembly are available on our Web site.

We assume that many of the dropout fragments are indeed human sequence: They behave that way in our hybridization experiments; they have no homology to vector sequences; and some are conserved in mice. Although there may be technical reasons explaining the dropout of some of these fragments, such as difficulty in assembly or poor-quality sequence, it is also likely that, due to insertion/deletion and order-of-sequence polymorphisms in humans, no fixed linear rendition of the genome is feasible. It may initially strain credulity that a 21-kb region can be polymorphic, but such large-sized events have been documented (Robledo et al. 2002), and the data from our accompanying paper strongly suggest that much larger copy number polymorphisms are commonplace in the human gene pool (Lucito et al. 2003).

ACKNOWLEDGMENTS

This work was supported by grants and awards to M.W. from the NIH and NCI (5R01-CA78544; 1R21-CA81674; 5R33-CA81674-04), Tularik Inc., 1 in 9: The Long Island Breast Cancer Action Coalition, Lillian Goldman and the Breast Cancer Research Foundation, The Miracle Foundation, The Marks Family Foundation, Babylon Breast Cancer Coalition, Elizabeth McFarland Group, and Long Islanders Against Breast Cancer. M.W. is an American Cancer Society Research Professor. E.T. is a Farish-Gerry Fellow of the Watson School of Biological Sciences and a predoctoral fellow of the Howard Hughes Medical Institute.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**: 403-410.
- Barrows, M. and Wheeler, D.J. 1994. A block sorting lossless data compression algorithm. Technical Report 124, Digital Equipment Corporation, Palo Alto, CA.
- Ferragina P. and Manzini G. 2000. Opportunistic data structures with applications. In *41st IEEE Symposium on Foundations of Computer Science*, pp. 390-398.
- . 2001. Experimental study of an opportunistic index. In *Proceedings 12th ACM-SIAM Symposium on Discrete Algorithms*, pp. 269-278.
- Gusfield, D. 1997. *Algorithms on strings, trees, and sequences*. Cambridge University Press, NY.
- Jurka, J. 2001. Repbase update: A database and an electronic journal of repetitive elements. *Trends Genet.* **16**: 418-420.
- Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al.

2003. The UCSC Genome Browser Database. *Nucl. Acids Res.* **31**: 51–54.
- Kent, W.J. 2002. BLAT—The BLAST-like alignment tool. *Genome Res.* **12**: 656–664.
- Kurtz, S. 1999. Reducing the space requirement of suffix trees. *Software—Practice and Experience* **29**: 1149–1171.
- Kurtz, S. and Schleiermacher, C. 1999. REPuter: Fast computation of maximal repeats in complete genomes. *Bioinformatics* **15**: 426–427.
- Kurtz, S., Choudhuri, J.V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. 2001. REPuter: The manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* **29**: 4633–4642.
- Li, F. and Stormo, G.D. 2001. Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics* **17**: 1067–1076.
- Lucito, R., Healy, J., Alexander, J., Reiner, A., Esposito, D., Chi, M., Rodgers, L., Brady, A., Sebat, J., Troge, J., et al. 2003. Microarray analysis of genome copy number variation. *Genome Res.* (this issue).
- Manber, U. and Myers, E.W. 1990. Suffix arrays: A new method for on-line string searches. *Proc. 1st ACM-SIAM SODA*, 319–327.
- Miller, J.W. 1996. Computer implemented methods for constructing a compressed data structure from a data string and for using the data structure to find data patterns in the data string. United States Patent 6,119,120, Microsoft Corporation.
- Pearson, W.R. and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **85**: 2444–2448.
- Robledo, R., Orru, S., Sidoti, A., Muresu, R., Esposito, D., Grimaldi, M.C., Carcassi, C., Rinaldi, A., Bernini, L., Contu, L., et al. 2002. A 9.1-kb gap in the genome reference map is shown to be a stable deletion/insertion polymorphism of ancestral origin. *Genomics*. **80**: 585–592.
- Sadakane, K. 1999. A modified Burrows-Wheeler transformation for case-insensitive search with application to suffix array compression. In *DCC: Data Compression Conference*, IEEE Computer Society TCC, Snowbird, UT.

WEB SITE REFERENCES

- <http://ftp.genome.washington.edu/RM/RepeatMasker.html>; Smit, A.F.A. and Green, P., RepeatMasker documentation.

Received March 19, 2003; accepted in revised form August 1, 2003.

Representational Oligonucleotide Microarray Analysis: A High-Resolution Method to Detect Genome Copy Number Variation

Robert Lucito,^{1,5} John Healy,¹ Joan Alexander,¹ Andrew Reiner,¹ Diane Esposito,¹ Maoyen Chi,¹ Linda Rodgers,¹ Amy Brady,¹ Jonathan Sebat,¹ Jennifer Troge,¹ Joseph A. West,¹ Seth Rostan,¹ Ken C.Q. Nguyen,² Scott Powers,^{1,2} Kenneth Q. Ye,³ Adam Olshen,⁴ Ennapadam Venkatraman,⁴ Larry Norton,⁴ and Michael Wigler¹

¹Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA; ²Tularik Inc., Genomics Division, Greenlawn, New York 11740, USA; ³Department of Applied Math and Statistics, SUNY at Stony Brook, Stony Brook, New York 11794, USA;

⁴Memorial Sloan-Kettering Cancer Center, New York, New York 10021, USA

We have developed a methodology we call ROMA (representational oligonucleotide microarray analysis), for the detection of the genomic aberrations in cancer and normal humans. By arraying oligonucleotide probes designed from the human genome sequence, and hybridizing with "representations" from cancer and normal cells, we detect regions of the genome with altered "copy number." We achieve an average resolution of 30 kb throughout the genome, and resolutions as high as a probe every 15 kb are practical. We illustrate the characteristics of probes on the array and accuracy of measurements obtained using ROMA. Using this methodology, we identify variation between cancer and normal genomes, as well as between normal human genomes. In cancer genomes, we readily detect amplifications and large and small homozygous and hemizygous deletions. Between normal human genomes, we frequently detect large (100 kb to 1 Mb) deletions or duplications. Many of these changes encompass known genes. ROMA will assist in the discovery of genes and markers important in cancer, and the discovery of loci that may be important in inherited predispositions to disease.

[The photoprint arrays were a kind gift of NimbleGen Systems Inc. and were fabricated to our design.]

Cancer is a disease caused, at least in part, by somatic and inherited mutations in genes called oncogenes and tumor suppressor genes. It is likely that we know only a minority of the critical genes that are commonly mutated in the major cancer types. The identification of these genes can lead to rational targets for chemotherapy. Moreover, in many cases, the knowledge of which genes have been mutated can predict the course of neoplasias, including their therapeutic vulnerabilities, if any. This knowledge is likely to become increasingly important as cancers, or suspected cancers, are detected at earlier and earlier stages.

Methods for finding cancer genes date back to the early 1980s, but general methods have only recently been developed. This problem is being addressed by a variety of evolving techniques, some capable of detecting the genetic losses and amplifications that often accompany the mutation of tumor suppressor genes or oncogenes, respectively. We describe here our success with ROMA (representational oligonucleotide microarray analysis), a technique that evolved from an earlier method, RDA (representational difference analysis; Lisitsyn et al. 1993). Like RDA, ROMA detects differences present in cancer genomes. ROMA also has applications to the identification of genetic variation in individuals caused by gene deletions or duplications, some of which may be related to inherited disease.

We developed RDA as one general approach to the cancer problem. RDA compares two genomes by subtractive hybridiza-

tion. To apply RDA, the complexity of the two genomes must first be reduced so that hybridization can go nearly to completion. To achieve this, we use low-complexity representations, a PCR-based method (Lisitsyn et al. 1993; Lucito et al. 1998). To compare genomes, they are cleaved in parallel with a restriction endonuclease, ligated to oligonucleotide adapters, and amplified by PCR. The shorter restriction endonuclease fragments are preferentially selected after many cycles of PCR, resulting in the reduced nucleotide complexity that is the essential characteristic of representations.

RDA has been successfully used to detect deletions and amplifications in tumors, and its use has led to the discovery of several candidate tumor suppressor genes and oncogenes (Li et al. 1997; Hamaguchi et al. 2002; Mu et al. 2003). However, RDA does not lend itself to the high-throughput genomic profiling of hundreds to thousands of cancer samples that can then be analyzed in parallel. Such vast parallel analysis is likely to be needed if the majority of complex genetic causes of cancer are to be identified.

Microarray analysis is a high-throughput method that has been widely used to profile gene expression in cancers (DeRisi et al. 1996; Golub et al. 1999; Van't Veer et al. 2002), and three groups, including ours, have adapted microarrays to detect genomic deletions and amplifications in tumors. Pinkel et al. (1998) have used arrays of BAC DNAs as hybridization probes; Pollack et al. (1999) have used cDNA fragments as probes; and in our first implementation, we used microarrays of fragments from representations as probes to analyze genomic representations (Lucito et al. 2000). All three methods use the comparative "two-color" scheme, in which simultaneous array hybridization de-

⁵Corresponding author.

E-MAIL rlucito@cshl.org; FAX (516) 367-8381.

Article and publication are at <http://www.genome.org/cgi/doi/10.1101/gr.1349003>. Article published online before print in September 2003.

tests a "normal" genome at one fluorescent wavelength and a pathological genome at another.

We previously demonstrated that complexity reduction of samples by representation improves signal-to-noise performance, and diminishes the amount of sample required for analysis, relative to other microarray hybridization methods (Lucito et al. 2000). However, useful interpretation of genomic array hybridization data requires that the arrayed probes be mapped, and this was a daunting task when we used fragments as probes. Moreover, in our previous implementation we used random fragment libraries, and we therefore could not create arrays focused in certain regions of the genome at will.

Adopting microarrays of oligonucleotide probes solves these problems. Representations are based on amplification of short restriction endonuclease fragments, and hence are predictable from the nucleotide sequence of the genome. Therefore, with the publication of the rough draft of the human genome (Lander et al. 2001), we can now design oligonucleotide probes that will hybridize to representations, and map them computationally. We developed algorithms for choosing from each predicted short fragment a 70-mer ("long") oligonucleotide probe with a minimal degree of sequence overlap to the rest of the genome. Through computation on the published human sequence, we can design almost any distribution of probes within the genome.

There are many other advantages to oligonucleotide-based microarrays. Based on our experience with the earlier implementation of this method using fragment arrays, the quality and reproducibility of printed oligonucleotide arrays ("print format") are superior. Although there is a large initial capital outlay to purchase large sets of oligonucleotides, the printed arrays are very inexpensive per unit when costs are amortized, and laborious and expensive replication of an underlying collection is not required. Furthermore, "long" oligonucleotide probes can be synthesized directly on an array surface (photoprinted arrays), and we demonstrate herein the equivalence of the two formats. In the photoprint format, there is no underlying physical collection at all (Singh-Gasson et al. 1999). In either case, whether printed or photoprinted, the composition of the array can be absolutely specified and hence is completely reproducible by others.

We show results from two array formats. The printed arrays are a format that is readily achievable. The regions that are represented on the array can be changed to suit the user. A whole-genome array can be printed with the desired resolution. Smaller ROMA arrays can be designed and printed to focus on specific regions of the genome if wished, the advantage being that less capital outlay would be required for a smaller set of oligonucleotides. Results from the second format used, photoprint array, were presented to demonstrate the power of high-resolution copy number analysis.

In this paper, we demonstrate our system, illustrating results and analytical techniques, present high-resolution analysis of cancer genomes, and provide initial evidence for widespread copy number polymorphism in humans. We discuss applications of our method, compare our method to other methods for global genomic analysis, and outline likely future developments.

RESULTS

Overview

This paper describes a complex procedure, observations, and methods of analysis that are highly interactive. We therefore give here an overview of our results to guide the reader through a sensible reading of this portion of the manuscript. The first section reviews the technique of representations, and in particular

"depleted" representations. Next we describe the design and selection of probes selected to hybridize well to representations. We introduce the two array formats that we use. The third section illustrates how to use hybridization to depleted representations to validate the composition of an array design, and the fourth section illustrates the use of such hybridization data to characterize probes and model overall array performance. Next we view essentially raw data of tumor and normal genomes, using two very different array formats, and show that the data from both formats are highly comparable. In the next section, we demonstrate a new statistical approach to gene copy number analysis based on segmentation analysis, and apply the method to two cancer genomes. The clonal nature of the cancers appears evident, as does the highly turbulent nature of their genomic rearrangements. The concordance between copy number analysis and our mathematical model is re-examined. Then we look more closely at several genetic lesions detected by our arrays following our statistical processing. Several distinct types of lesions are illustrated, including large regions of amplification and very narrow regions of homozygous and hemizygous deletion. Different types of inferences that can be made by the method are demonstrated. In the final section, we find a surprising abundance of "normal" variation in copy number between two individuals, and illustrate the need to coordinate data about such variation with interpretation of cancer data.

Representations

Representations reduce the complexity of samples in a reproducible way, thereby increasing signal to noise during hybridization to arrayed probes. Representations also provide a means to amplify the quantity of sample, and allow a very convenient way to validate and simulate array performance.

In our present studies, we have limited ourselves to the use of representations made with *Bgl*III, an enzyme with a typical 6-bp recognition site. *Bgl*III is one of many restriction enzymes that satisfy these useful criteria: It is a robust enzyme; its cleavage site is not affected by CpG methylation; it leaves a four-base overhang; and its cleavage sites have a reasonably uniform distribution in the human genome. After cleavage with *Bgl*III, we ligate adaptors, and use the resulting product as a template for a PCR reaction. Because PCR selects small fragments, *Bgl*III representations are made up of the short *Bgl*III fragments, generally smaller than 1.2 kb, and we estimate that there are ~200,000 of them, comprising ~2.5% of the human genome, with an average spacing of 17 kb.

For array characterization, we use "depleted" *Bgl*III representations. These are representations made according to the usual protocol, but prior to PCR (to selectively amplify small *Bgl*III fragments), the adaptor-ligated *Bgl*III fragments are cleaved with a second restriction endonuclease. Cleavage destroys the capacity of some fragments to be exponentially amplified. For example, a *Bgl*III representation-depleted by *Eco*RI would consist of all small *Bgl*III fragments of the genome that do not contain within them *Eco*RI sites. Depleted representations are used for probe validation and modeling performance because we can remove a known subset of fragments from the representation, and observe the consequence upon hybridization to those probes complementary to the depleted fragments.

In all of the experiments described herein, we have used comparative hybridization of representations prepared in parallel. Our approach works best if the DNA from two samples being compared is prepared at the same time, from the same concentration of template, using the same protocols, reagents and thermal-cycler. This diminishes the "noise" created by variable yield upon PCR amplification.

Design and Selection of Probes, and Composition of Probes for Microarrays Formats

We describe the design (length and composition) and selection of probes using two very distinct formats for the synthesis of arrayed probes.

Our probes are derived from the short *Bgl*III restriction endonuclease fragments that we predict to exist from analysis of the human genome sequence. We initially evaluated probes of length 30 through 70, using methods described in the next section. The signal-to-noise ratio was maximal for probes of 70 nt in length, and we chose that length as our standard.

We selected our probes to be as unique as possible within the human genome, and tried to minimize short homologies to all unrelated sequences. We devised algorithms by which we could annotate any sequence of the genome with its frequency of exact matches in the genome (Healy et al. 2003). These algorithms were used to choose regions within the predicted *Bgl*III fragments that are unique for their constituent 18-mers or 21-mers, and then within these regions, choose 70-mers with the minimal arithmetic mean of their constituent 15-mer exact matches. Subsets of the 70-mers were then tested for uniqueness in the human genome by a low homology search using BLAST.

We used two formats for constructing microarrays. In the first of these, the "print" format, we purchased nearly 10,000 oligonucleotides made with solid-phase chemistry, and printed them with quills on a glass surface. In the second format, "photoprint arrays," oligonucleotides were synthesized directly on a silica surface using laser-directed photochemistry by NimbleGen Systems Inc. The photoprint arrays were a gift of NimbleGen Systems Inc., and fabricated to our design. Many more probes can be synthesized per array with laser-directed photochemistry, and in these experiments our arrays contained 85,000 oligonucleotide probes.

The probe composition for the 85K set was determined by a combination of design and selection, as described below. Unlike oligonucleotide probes synthesized by standard phosphoramidite solid-phase chemistry, certain oligonucleotides synthesized by laser-directed photochemistry are made in poor yield. However, unlike probes synthesized by the solid-phase chemistry and then printed, the cost of testing a set of probes synthesized directly on a chip is no more than the cost of the chip itself. Therefore, we tested ~700,000 unique 70-nt probes (see Methods) predicted to be complementary to small *Bgl*III fragments, arrayed on eight chips. These were hybridized with standard *Bgl*III and *Eco*RI-depleted *Bgl*III representations, and we picked the 85,000 with the most intense signal when hybridized to a single normal human DNA, "J. Doe." These 85,000 were then arrayed on a single chip.

In both our 10K and 85K formats, probes are arrayed in a random order, to minimize the possibility that a geometric artifact during array hybridization will be incorrectly interpreted as a genomic lesion.

Validation of Printed Arrays With Depleted Representations

We should be able to observe a very clear and predictable pattern to arrays hybridized with depleted representations if and only if these conditions are met: The available human genome sequence assembly is accurate; our method of probe design and selection is valid; our hybridization conditions are sufficiently robust to give a good signal-to-noise ratio for our probe population; and we have correctly deconvoluted the probe addresses on our arrays during data processing. We put all our array designs through such tests. Moreover, the data we collect can further be used for probe calibration and to create simulations that predict the

power of the array hybridization to detect various genomic lesions, as will be described in a following section.

To illustrate this process with a 10K array, we show in Figure 1 results obtained with *Bgl*III representations depleted by *Hind*III. In Figure 1A, we graph the ratios of hybridization intensity of each probe along the Y-axis. (See Methods for a description of how we process raw scanned data. We perform no background subtraction, as that only increases noise.) Each experiment is performed in color reversal, and the geometric mean of ratios from the separate experiments is plotted. Probes that we predict to detect fragments in both the full and depleted representations, based on the published human sequence, are grouped on the left. There are ~8000 probes that are predicted to be present in both depleted and nondepleted representations. Probes that we predict will not detect fragments in the depleted representation are grouped on the right. There are ~1800 probes predicted as being depleted.

From the experiment shown in Figure 1A, we can infer that the promise of the method is largely fulfilled: The restriction profile of representational fragments is correctly predicted, the probes are correctly arrayed, and the probes detect the predicted fragment with acceptable signal intensity.

To calculate the data shown in Figure 1A, each hybridization was performed in color reversal, and the geometric mean of ratios from the separate experiments was plotted. In Figure 1B, the agreement between the ratios of the color reversal experiments is graphed, as a log-log scatter plot, showing excellent correlation of the data regardless of the labeling choice.

Modeling Array Hybridization

Variation in the ratio of intensities is evident from Figure 1A. Some probes fail to exhibit the predicted elevated ratios. There are several possible explanations for this. For example, the oligonucleotide probe may not have been correctly or completely synthesized, or the respective *Bgl*III fragment may not be present in the representation as predicted. The latter can happen, for example, if the public genome sequence is in error, or if there is a polymorphism at one of the *Bgl*III sites in the sample genome resulting in a longer *Bgl*III fragment than expected.

When, as here, there is significant variation in measurements, statistical methods need to be used for the most accurate interpretation of data. It is also often useful to construct a mathematical model that can simulate measurement. Moreover, a good model can help predict the limits of detection, and be of assistance in the design of experiments. In this section, we describe a mathematical model that fits the data, and in a later section we describe statistical methods for data analysis. The mathematical model is useful for individual probe characterization, a clearer interpretation of the data, and the sharpening of statistical tools.

There is always more than one way to model data, and various enhancements can be added, but for our arrays we have found that a simple equation and sampling technique creates a model with great predictive power. This model will be described in detail in a subsequent manuscript, but it is based on an equation for the intensity of the *i*-th probe in a given channel, $I[i]$:

$$I[i] = \alpha * (\gamma * A[i] * c[i] + \beta).$$

In this equation, $c[i]$ is the concentration of *Bgl*III fragment complementary to the *i*-th probe prior to representation; and $A[i]$ is the combined "performance character" of the probe and its complementary *Bgl*III fragment. The parameters of the equation are elements of distributions. α is a multiplicative system noise; β is an additive system noise that encompasses background hybridization; and γ is the multiplicative noise created during parallel

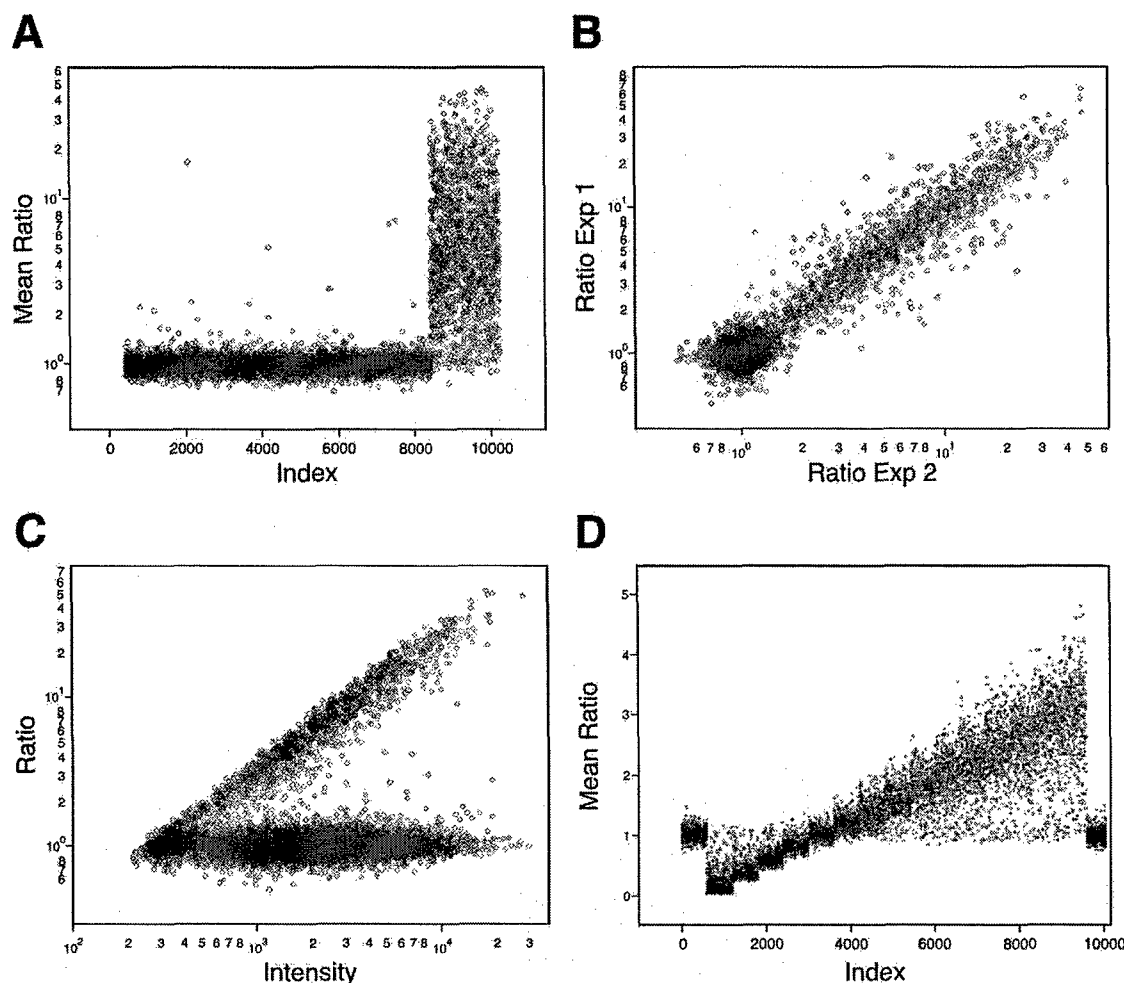


Figure 1 The predictability of informatics and accuracy of the array measurements using 10K microarrays. (A) The results, where the samples hybridized are *Bgl*II representation and *Bgl*II representation depleted of fragments with a *Hind*III cleavage site. The Y-axis (Mean Ratio) is the mean measured ratio from two hybridizations of depleted representation to normal representation plotted in log scale. The X-axis (Index) is a sorted index, such that those probes that derive from fragments that do not have an internal *Hind*III restriction cleavage site sort first and those with an internal *Hind*III site sort last. This allows the separation of these two subsets for visualization of the cleavage results. (B) The reproducibility of the duplicate experiments used to generate the average ratio in A. The Y-axis (Ratio Exp1) is the measured ratio from experiment 1, and the X-axis (Ratio Exp2) is the measured ratio of experiment 2. Both axes are plotted in log scale. (C) Graph of the normalized ratio on the Y-axis as a function of intensity of the sample that was not depleted on the X-axis. Both the ratio and intensity were plotted in log scale. (D) Data generated by simulation. The X-axis (Index) is a false index. Probes, in groups of 600, detect increasing copy number, from left to right; 600 flanking probes detect normal copy number. The Y-axis (Mean Ratio) is the mean ratio calculated from two hybridizations.

representation and labeling. By definition, both α and γ have a mean of 1, and for a diploid genome, $c[i] = 1$.

$A[i]$ can be viewed as the "brightness" of the i -th probe, and is a major determinant of the signal-to-noise ratio. In principle, $A[i]$ should depend on at least two factors: the proportionate amplification of the fragment complementary to the probe during representation; and the purity of the probe. For example, a probe that is complementary to a poorly amplified fragment will have a low A value. Conversely, a probe complementary to a well-amplified fragment should be "bright" and have a high signal-to-noise ratio. Similarly, a probe that is synthesized with poor yield will have a low intensity and a poor signal-to-noise ratio. Other factors may influence A , such as the secondary structure of the probe and its base composition.

In the actual data, the highest ratios are observed from the most intense probes (see Fig. 1C). According to the model, this is explained by a fairly constant nonspecific signal for most probes.

That is, β is independent of the probe. Thus, the "brightest" probes also have the highest specific to nonspecific signal. This observation was the basis for our selection of the probes of the 85K set in the photoprint format (see above).

The model makes additional predictions: First, actual ratios are linearly related to measured ratios, and second, the standard deviation of probe measurement is a strong function of ratio, being a minimum for ratios of unity. Using parameters derived from the experiments displayed in Figure 1, we illustrate these relationships in Figure 1D. We assume 15 sets of 600 probes with various copy numbers $n/4$, with $n = 0-14$, bracketed by 600 probes of diploid copy number ($4/4$) on either end, measured against a diploid genome ($c[i] = 1$), and measured in duplicate. Note that the mean measured ratio of a set of probes is a linear function of the "true" copy number, the number of gene copies per cell, and the mean measured ratio, R_M , of a subset of probes reflects their true ratio, R_T , by the following equation:

$$R_M = (R_T * S_N + 1) / (S_N + 1).$$

This is one general form of a linear equation in which $R_M = 1$ when $R_T = 1$. S_N is an experimental character, which we think of as "specific to nonspecific" noise. We can solve for S_N from any pair of nonunitary R_M and R_T values. We use this tool below to analyze two cancer genomes, below.

Views of Tumor Genomes at 10K and 85K Resolution

Array hybridization data can be readily viewed, after deconvolution of probes into genomic order, without any model. In particular, genomic lesions, whether deletions or amplifications, are visually obvious. We show in the matrix of panels of Figure 2 the array hybridization data for three genomic comparisons. Figure 2, A1–A3, shows breast cancer (aneuploid) versus "normal" (diploid) data from the same biopsy of a patient (CHTN159). Figure 2, B1–B3, shows a breast cancer cell line (SK-BR-3) derived from a patient of unknown ethnicity versus an unrelated normal male ("J. Doe") of mixed European and African parentage. Figure 2,

C1–C3, shows a normal male (African pygmy) versus the same J. Doe. In each case, the samples were hybridized twice, with color reversal, and the geometric mean ratio (on a log scale) is plotted versus the genome order of the probes.

The samples from Figure 2A were derived by flow sorting the nuclei of a surgical biopsy into aneuploid and diploid fractions, and making representations from as few as 15,000 nuclei (~100 ng of DNA). We estimate that the aneuploid fraction has perhaps 10% contamination from diploid nuclei, whereas the diploid fraction is not expected to be completely normal. Nevertheless, highly interpretable data result.

These data are in two formats: the 10K print format (Fig. 2A1,B1,C1) and an 85K photoprint format (Fig. 2A2,B2,C2). Unlike the 10K format, probes of the 85K format were also selected for performance, as described and justified in earlier sections. This selection procedure produces a slight bias, in that no probe from the 85K set will detect a small *Bgl*II fragment that is homozygously missing in J. Doe. The consequences of this bias can be seen in comparisons of the 10K print format with the 85K pho-

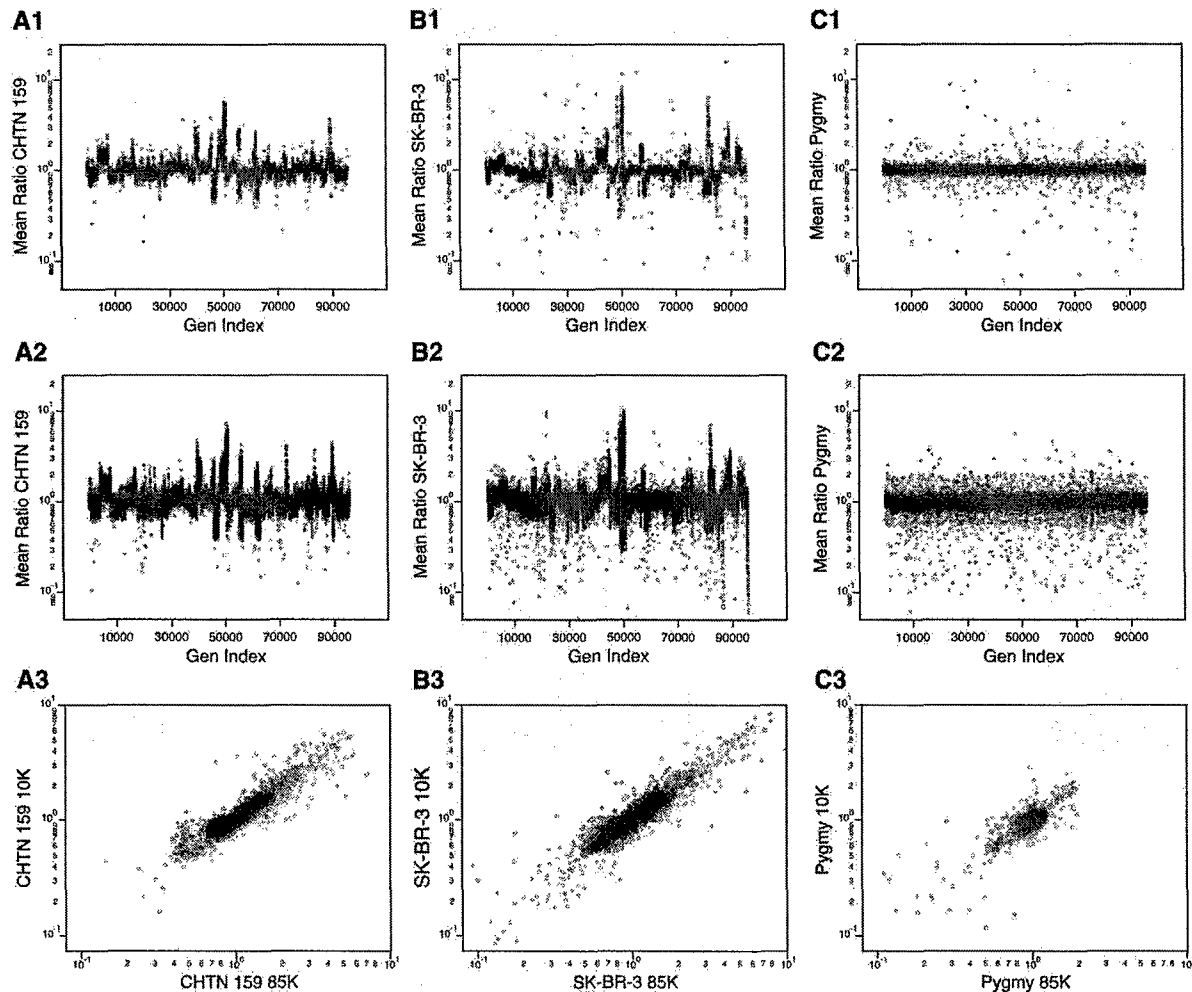


Figure 2 The genomic profiles for (A) a primary breast cancer sample (CHTN159), with aneuploid nuclei compared with diploid nuclei from the same patient; (B) a breast cancer cell line compared with a normal male reference; and (C) a normal male compared with a normal male reference, using the 10K printed array (A1,B1,C1) and the 85K photoprint array (A2,B2,C2). In each case (rows 1 and 2), the Y-axis is the mean ratio, and the X-axis (Gen Index) is an index of the probes' genomic order based on the June 2002 assembly, that is, NCBI Build 30. The probes were put into genomic order concatenating Chromosomes 1 through Y. (A3,B3,C3) The correspondence of the ratios measured from "brother" probes (see text for details) present in the 10K and the 85K microarrays. The Y-axis is the measured ratio from the 10K microarray, and the X-axis is the measured ratio from the 85K microarray.

toprint format. In results from the 10K print format, there are roughly equal numbers of extreme "singlets" above and below a copy number of 1 (most apparent in Fig. 2C1). In contrast to this, using the 85K format, more extreme singlets are below rather than above a copy number of 1 (Fig. 2C2).

In Figure 2, A1, A2, B1, B2, C1, C2, increased copy number is indicated by a ratio above 1, and decreased copy number by a ratio below 1. Even at this global view, with all probes displayed, several interesting observations can be made. There are clearly profiles to the cancer genomes, large regions of amplification, some quite high, and large regions of deletion (Fig. 2A,B). The profiles of the cancer genomes are varied. In contrast, the profile of the normal-normal appears to be flat, although some features can be seen. These will be examined more closely below.

There are, in all three genomes, many stand-alone probes detecting minor losses and gains, which we attribute to heterozygous *Bgl*II polymorphism. These are manifest in the normal-normal comparison (Fig. 2C2) as a "shell" of probes that approach ratios of 0.5 and 2.0 throughout the genome.

In contrast, in the tumor-normal comparison, wherein the normal is matched, there is only one stand-alone probe detecting major gains, and the stand-alone probes detecting major losses are more or less confined to extensive regions showing minor loss. This pattern is consistent with a hypothesis of allelic polymorphism and loss of heterozygosity (LOH). For a patient with heterozygosity at a *Bgl*II fragment, with a large and a small fragment, loss of the small allele will result in the virtual loss of specific signal because the large allele will not be abundant in the representation. This will present as an apparent major loss. On the other hand, a loss of the large allele, for example, by gene conversion, would at most result in a twofold increase in ratio, appearing as a minor gain.

It is evident, looking at the results of the 10K print and the 85K photoprint formats in Figure 2, A1, A2, B1, B2, C1, C2, that the two systems capture a similar view of the larger genomic features. A correspondence between the two formats can be seen quantitatively. We call probes "brothers" if they share complementarity to the same *Bgl*II fragment. Brothers do not necessarily have overlapping sequence, or may be complementary across their entire length. In Figure 2, A3, B3, C3, we plot the ratios of brothers from one format to ratios of their brothers from the other format. There are in excess of 7000 brother probes. For all three experiments, in spite of the fact that the probe sequences differ between formats, the order of arraying is different, the hybridization conditions differ, and the surfaces of the array are different, there is remarkable concordance between the ratios of brother probes regardless of format.

Automated Segmentation and Whole-Genome Analysis

Because of the extent of the data, and its statistical nature, automated tools for feature recognition that are statistically based are extremely useful. One part of our group has developed a statistical segmentation algorithm termed circular binary segmentation (CBS) that parses the probe ratio data into segments of similar mean after taking variance into account (Olshen et al. 2002). The algorithm works by analyzing one chromosome at a time and, within that chromosome, recursively identifying the best possible segmentation. Each proposed split is accepted or rejected based on the probability that the difference in mean could have arisen by chance. This probability is determined using a randomization method. The algorithm is a novel modification of binary segmentation (Sen and Srivastava 1975). Because of its nonparametric nature, the algorithm cannot identify aberrations with fewer than three probes. We discuss detecting smaller lesions below.

Figure 3 illustrates some of the output for the analysis of the cancer cell line SK-BR-3 at 85K resolution. We show four chromosomes, the highly turbulent Chromosome 8, a somewhat less active Chromosome 17, Chromosome 5, and the X-chromosome. The segmentation profiles and segment means for the 10K and 85K sets are very similar (data not shown), but clearly are not identical. More features are seen with the 85K set. In the next section, we inspect some of the data more closely. The full data, and that for the other two genomes, can be viewed at our Web site (<http://roma.cshl.org/>).

Once segmented, we can assign to every probe the mean ratio of the segment to which it belongs, and then view the assigned mean ratios in sorted order. We do this for the two cancer genomes in Figure 4, A (CHTN159) and C (SK-BR-3). It is evident from the figure the segment mean ratios within each genome are quantized, with major and minor plateaus of similar value. In fact, it is likely that we can deduce the copy number by counting. As determined by flow analysis, the tumor is subtriploid, and the cell line is tetraploid. Assuming each sample is roughly monoclonal, then the two major plateaus in the tumor would be two and three copies per cell, and the major plateaus in the cell line are likely to be three and four copies per cell.

We can then use the copy number assumptions of the major plateaus to solve the ploidy and S_N for each experiment. Our method is to use a version of equation 2 for each plateau. We select R_M , the mean measured ratio, as the average of the probes of the segments in the plateau. We first set R_T to C_N/P , where C_N is the "true" copy number. C_N is the number of gene copies per cell, assumed to be known and equal for the plateau. P is the ploidy of the tumor genome. The result is two equations and two unknowns, with the unknowns being P and S_N . For the tumor biopsy experiment (Fig. 4A), we calculate the ploidy P to be 2.60, and S_N to be 1.13. For the cell line experiment (Fig. 4C), we calculate that P is 3.93, and S_N is 1.21. We can then use equation 2 again to calculate what mean ratios would be expected for higher and lower copy numbers. These expectations are marked on the respective graphs, from zero to a copy number of 12, with horizontal lines forming a "copy number lattice." The assigned mean-segment values for probes are displayed in genome order, embedded with the expected copy number lattice (Fig. 4B,D).

The copy number lattice fits remarkably well the minor plateaus of the data, especially for the higher copy numbers. However, there appears to be error in the expected ratios for probes detecting loss. The assigned mean-segment ratios of probes detecting loss cluster around values somewhat below the predicted values. In other words, the array appears to perform better for deletions than predicted based on the major plateaus and our present model. This deviation might be explained if we reexamine our assumption of clonality, and will be investigated further.

Specific Illustrative Examples

There is clearly too much data to be described in a printed paper, and the reader is invited to visit our Web page (<http://roma.cshl.org/>). In this section, we discuss a few examples taken from the array data of SK-BR-3 that illustrate several aspects of our system.

The first example is a closer inspection of a region of a break in the X-chromosome, seen in Figure 3D. SK-BR-3, which derives from a female, has been compared to an unrelated male. The expectation is that probes in the X-chromosome will have elevated ratios. This is the case through much of the long arm of Chromosome X. In the midst of Xq13.3, over a region spanning 27 kb, there is a sharp break in copy number, and for the remainder of the chromosome, ratios near 1 are observed (Fig. 5A). This example demonstrates the boundaries that can be drawn

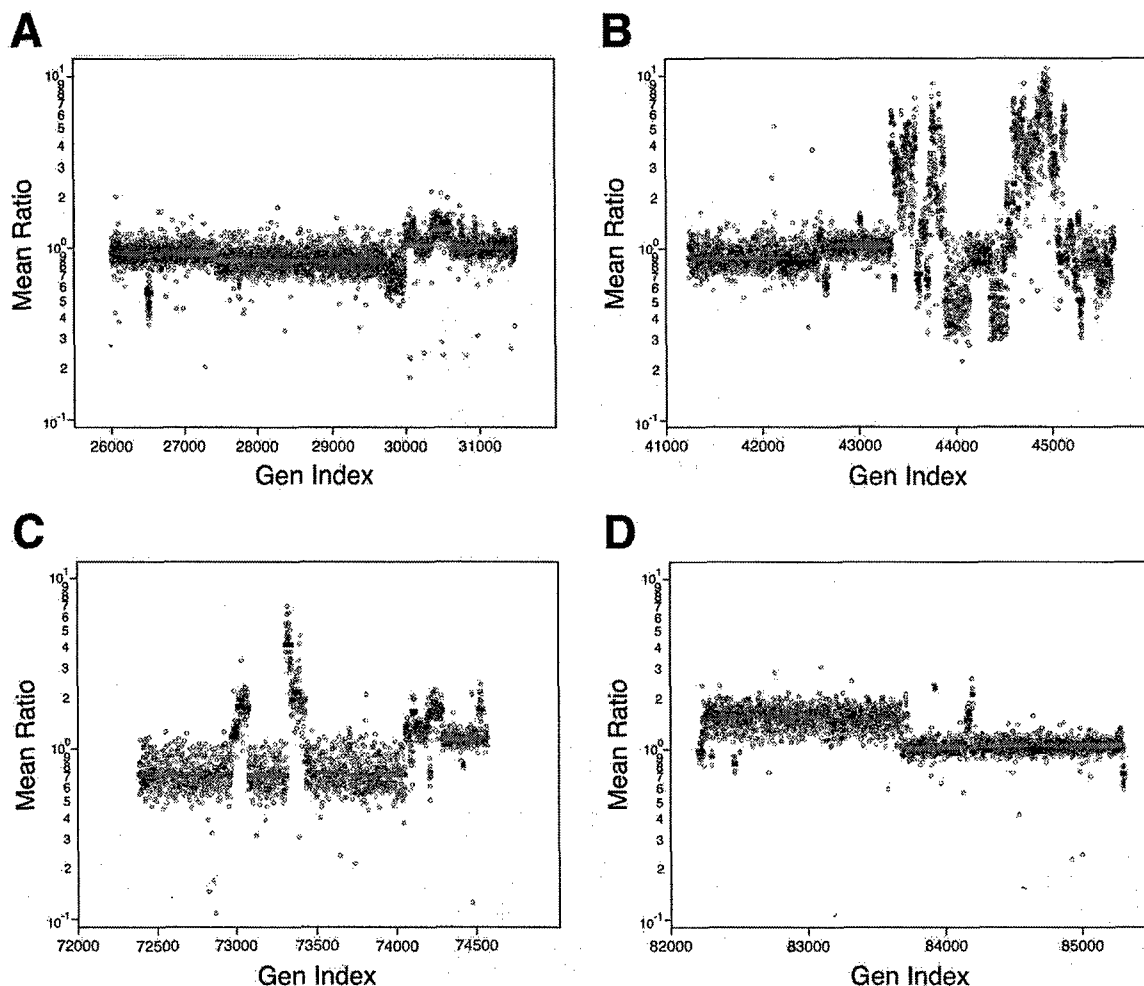


Figure 3 Several chromosomes with varying copy number fluctuations from analysis of the tumor cell line SK-BR-3 as compared with the normal reference. The Y-axis (Mean Ratio) represents the mean ratio of two hybridizations in log scale. The X-axis (Gen Index) is an index of the genomic coordinates, as described above. (A) Copy number fluctuations identified for Chromosome 5, (B) for Chromosome 8, (C) for Chromosome 17, and (D) for the X-chromosome.

from the array data by segmentation. In our data there are other examples of sharp copy number transitions that must break genes.

There are three to four narrow amplifications in SK-BR-3, each containing two or fewer genes, among which are transmembrane receptors. But broad amplifications can also be informative. The second example comes from the highly turbulent Chromosome 8 (see Fig. 3B). Despite the abundance of aberrations, we can clearly discern distinct regions of amplification. One such region is shown in Figure 5B. The rightmost peak is approximately a 1-Mb stretch, comprised of 37 probes (probe coordinates 45099–45138, June 2002 assembly, or NCBI build 30 genome coordinates 126815070–128207342). Yet it contains a single RefSeq gene, *c-myc*.

There is a second very broad peak in SK-BR-3, ascending to the left of the *c-myc* peak, and off the graph. This broad peak has a broad shoulder on its right (probe coordinates 44994–45051, June 2002 assembly, or NCBI build 30 genome coordinates 123976563–125564705), with a very narrow peak in its midst. We can overlay on this the segmentation data from the tumor genome, CHTN159, which has an even broader peak encompassing *c-myc* (probe coordinates 44996–45131, June 2002 assembly,

or NCBI build 30 genomic coordinates 124073565–127828283). The peak in CHTN159 also encompasses the shoulder of the second SK-BR-3 peak (Fig. 5B). Thus, the shoulder may contain candidate oncogenes that merit attention. Within that region, at the narrow peak, we find *TRC8*, the target of a translocation implicated in hereditary renal carcinoma (Gemmill et al. 1998). This example illustrates the value of coordinating data from multiple genomes, and the need for automated methods for analyzing multiple data sets.

We next show an example of a narrow deletion that highlights the need for high-resolution arrays, and also raises additional questions. The lesion occurs on Chromosome 5. In Figure 5C, we show a combined 10K (red) and 85K (blue) view. We do not show segmentation, but show the copy number lattice. A deletion is evident at both 10K and 85K resolutions (probe coordinates 26496–26540, June 2002 assembly, or NCBI build 30 genome coordinates 14231414–15591226), one we judge to be hemizygous loss, but which may represent the presence of one copy in a tetraploid genome. The boundaries are much more clearly resolved at 85K. This region contains *TRIO*, a protein having a GEF, SH3, and serine threonine kinase domain (Lin et al. 2000); *ANKH*, a transmembrane protein (Nurnberg et al. 2001);

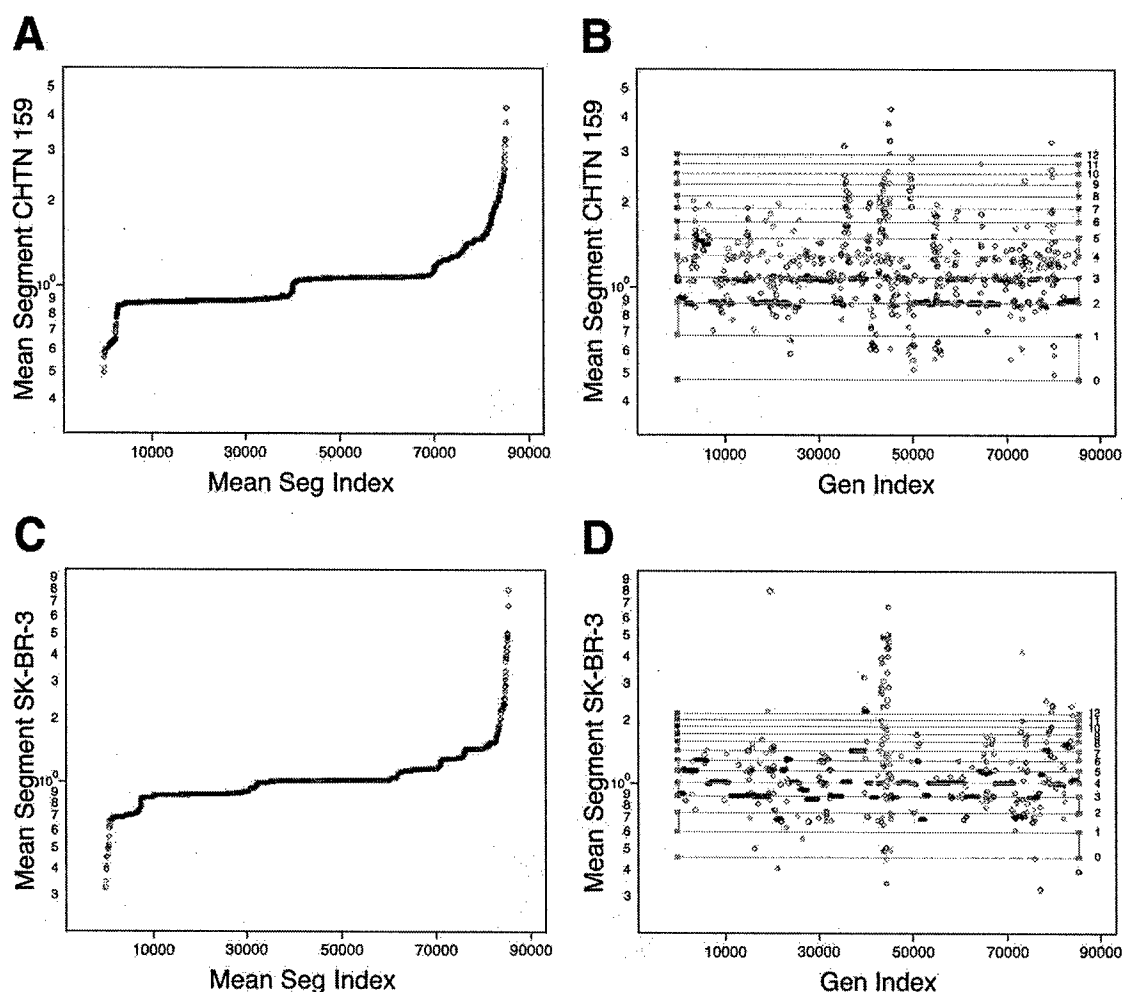


Figure 4 The mean segmentation calculated from the analysis of SK-BR-3 compared with (A,B) the normal reference and (C,D) CHTN159. In all panels, the Y-axis is the value of the mean segment for each probe in log scale. In A and C, the X-axis (Mean Segment Index) is in ascending value of the assigned mean segment. In B and D, the X-axis (Gen Index) is the genomic index, as described above. Plotted on top of the mean segment data is a copy number lattice extrapolated from the array data using formulas within the text (horizontal lines). The calculated copy number for each horizontal line is to the right of the lattice.

and *FBXL*, a component of the ubiquitin ligase mediated protein degradation pathway (Ilyin et al. 2000).

It is also clear from the data that the lesion does not appear "neat." In the middle of the deletion are four or five probes that report ratios near 1. We can consider several explanations for this result. First, the hybridization to those probes may have failed for a variety of reasons. For example, the probes might not have been completely synthesized, or their complementary *Bgl*III fragments might not have amplified well. However, the intensities of these probes are in the middle range for all probe intensities, which diminishes the likelihood of this hypothesis. Second, the human assembly may be in error, and the outlier probes have been incorrectly posted at this location. Third, the deletion event may indeed be complex, the result of a localized genomic instability.

Our last example is a region of homozygous loss (Fig. 5D). In this example, a cluster of zinc-finger proteins on Chromosome 19 is affected (probe coordinates 77142–77198, June 2002 assembly, or NCBI build 30 genomic coordinates 21893948–24955961). These genes, having zinc-finger domains, may encode transcription factors, whose deletion may have a role in tumorigenesis.

There are an abundance of narrow hemizygous and homozygous lesions. These are seen both in the analysis of the cancer cell line and the cancer biopsy. However, as described below, we must take caution in their interpretation. Our next examples will all be in the context of normal-normal variation.

Examining Normal Genomic Variation

In this section, we demonstrate the need to coordinate cancer genome analysis with a knowledge base of normal genomic variation.

When the tumor DNA cannot be matched against normal DNA, and an unrelated normal DNA is used as a reference, the differences observed may be the result of polymorphic variation. This variation can be of two sorts, the run-of-the-mill point sequence variation, of the sort that creates or destroys a *Bgl*III fragment, SNPs for example, or actual copy number fluctuation present in the human gene pool. The former is relatively harmless, as it will produce scattered noise that can largely be filtered by statistical means.

We illustrate the application of a very mild filtration algorithm: If a ratio is the most deviant of the surrounding four, we

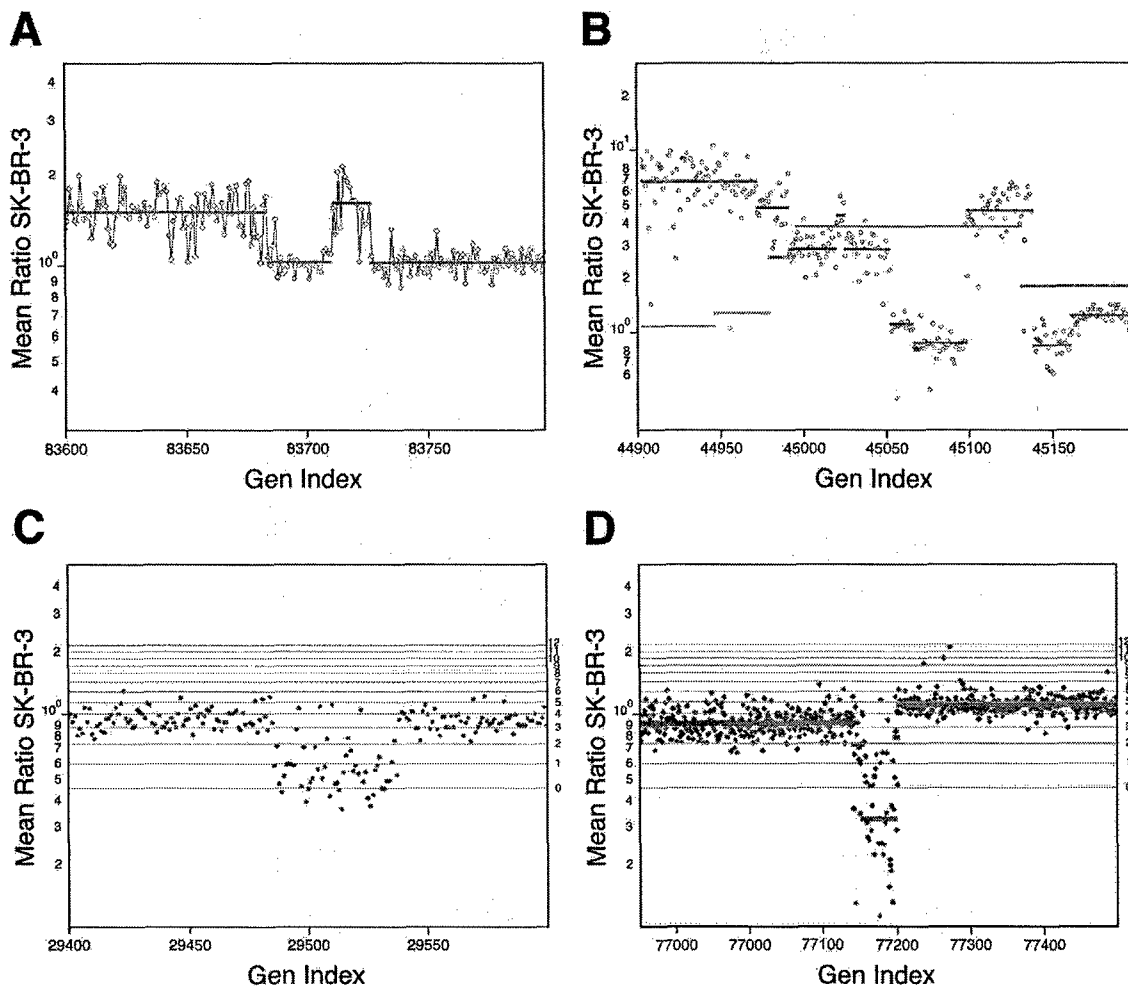


Figure 5 In all panels, the Y-axis (Mean Ratio SK-BR-3) is the mean ratio of two hybridizations of SK-BR-3 compared with a normal reference in log scale. The X-axis (Gen Index) is the genomic index, as described. (A) A region from the X-chromosome with a region of loss. Plotted over the measured array ratio is the calculated segmentation value. (B) A region of Chromosome 8 (*c-myc* located to the right of the center of the graph) from results of SK-BR-3 in comparison to normal reference. Plotted on top of the data are the segmentation values for SK-BR-3 in comparison to the normal reference in red and the segmentation values for the primary tumor CHTN159 in green. (C) A lesion on Chromosome 5 demonstrating the resolving power of the 85K as compared with the 10K array. Results are from SK-BR-3 compared with a normal reference. Spots in red are from the 10K printed microarray, and spots in blue are from the 85K photoprint array. Horizontal lines are copy number estimates, based on modeling from mean-segment values. (D) Comparison of SK-BR-3 to normal reference, displaying a region of homozygous deletion on Chromosome 19. The mean-segment value is plotted as a red line, and horizontal lines are copy number estimates as described.

replace it with the closer ratio of its two neighbors. In Figure 2C2, we showed a normal-normal comparison. The data look flat, with a cloud of scattered polymorphisms. In Figure 6A (combined 10K and 85K sets), we have applied filtration. The data no longer look so flat, and the cloud of scattered polymorphism is lifted, revealing nonrandom clusters of deviant probe ratios. These clusters reflect large-scale genomic differences between normal individuals, and we will say more of this presently.

Polymorphic variation of the scattered variety can also be filtered by serial comparison of experiments. We illustrate such a process in Figure 6B. In this figure, we display data from SK-BR-3 compared with normal donor J. Doe, the 85K ratios displayed in blue circles, and the 10K in red. On the same graph we display the ratios of J. Doe compared with another normal, DNA from an African pygmy, in green triangles. This is a fairly typical field of view. We see three probes of extreme ratio in the SK-BR-3-normal hybridization that can be identified as polymorphisms by comparison to hybridization between the two normal individuals.

The simplest interpretation is that J. Doe is $+/-$, pygmy $+/-$, and SK-BR-3 $-/-$, where + designates the presence of a small *Bgl*III fragment and designates the absence of a fragment (most likely a SNP at a *Bgl*III site). In general, pairwise comparisons of three genomes allow interpretable calls of allele status. Hence, we suggest that when a malignant genome cannot be paired to a matched normal, or perhaps even when it can, such genomes should be compared with a single reference normal donor, whose allele status can be firmly established by extensive comparisons against other normals.

Polymorphism in copy number, however, presents a different sort of problem. In this case, many probes within a region will show a deviation from a ratio of unity, and the pattern will appear coherent, not scattered. Statistical means will not suppress this signal. But do such variations commonly exist, and are they likely to be a source of misinterpretation if ignored? The perhaps surprising answer is emphatically, yes.

Figure 6A indicates that there are gross regional differences

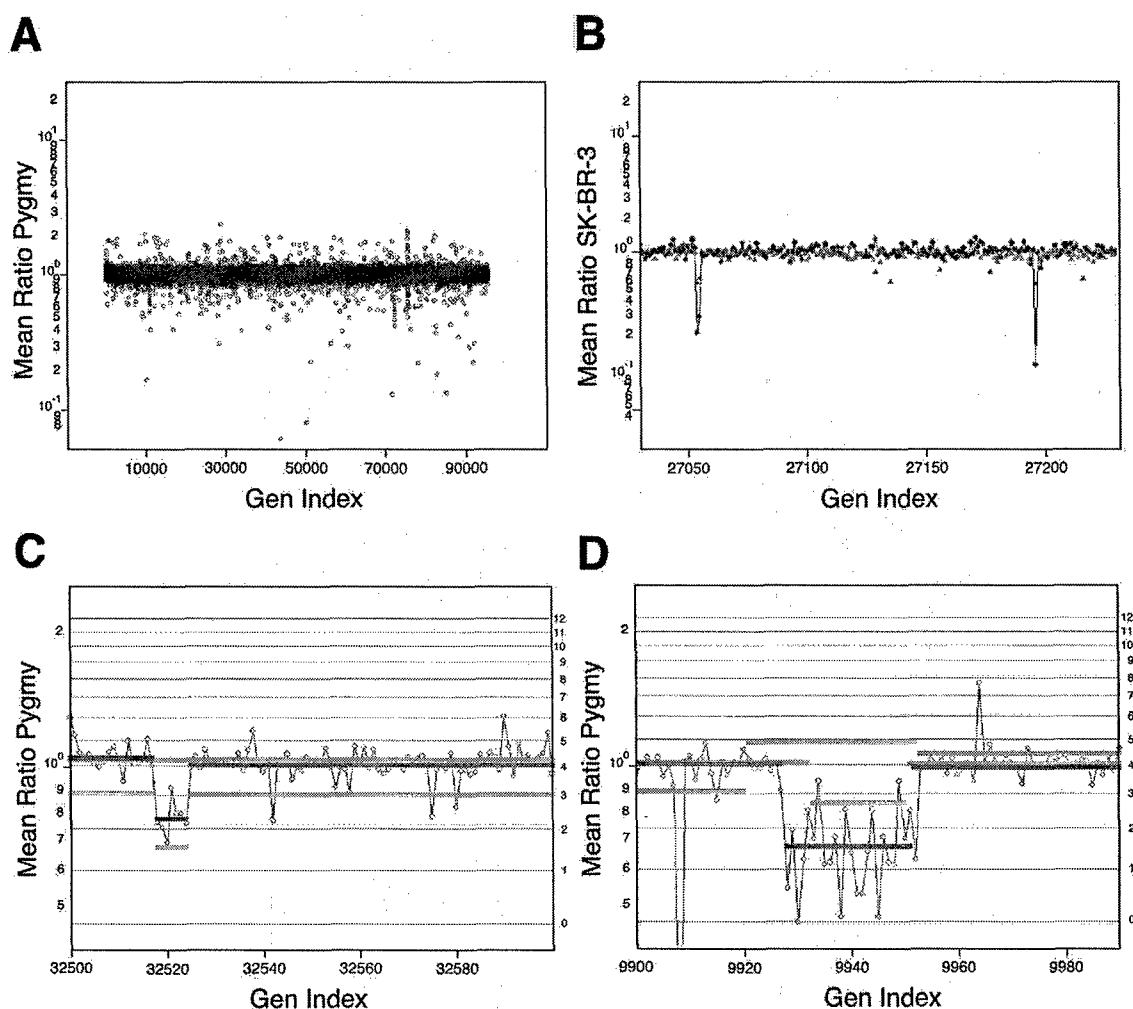


Figure 6 (A) The results of a normal genomic profile compared with a normal, identical to that displayed in Figure 2C2 with the exception that singlet probes have been filtered as described in the text. (B) The serial comparison of experiments for a small region from Chromosome 4. The Y-axis is the mean ratio in log scale. The X-axis is the genomic index, as described. The blue (85K) and red (10K) spots are from the comparison of SK-BR-3 to normal. The green is a comparison of a pygmy to the normal reference. (C) A lesion found in the normal population on Chromosome 6. The blue spots are plotted by mean ratio for analysis of the pygmy to the normal reference. The red line is the mean-segment value for the pygmy-to-normal reference comparison. The green line is the mean-segment value for the SK-BR-3-to-normal reference comparison. The blue line is the segment value from the primary tumor (CHTN159 aneuploid to diploid) comparison. (D) A region of Chromosome 2. The data shown in blue circles are from the comparison of SK-BR-3 to the normal reference. The mean-segment line for this comparison is shown in green. The mean-segment line for the comparison of a pygmy to the normal reference is shown in red and for the primary tumor CHTN159 in blue. For C and D, the calculated copy number for the horizontal lines is found to the right of the panel.

in the normal-normal comparison. Indeed, many regions that display altered copy number between the two normal individuals are revealed upon segmentation analysis. Close inspections of two such regions are displayed in Figure 6, C and D, with ratios as connected blue dots and copy number lattice values in orange. In Figure 6C, the abnormal region is 135 kb on Chromosome 6p21 (probe coordinates 32518–32524, June 2002 assembly, or NCBI build 30 genomic coordinates 35669083–35804705), and encompasses three known genes. In Figure 6D, the region is a 620-kb region from Chromosome 2p11 (probe coordinates 9927–9952, June 2002 assembly, or NCBI build 30 genomic coordinates 88787694–89385815) that contains a number of heavy chain variable regions.

We observe on the order of a dozen such regions in any normal-normal comparison. They range from 100 kb to >1 Mb in length and are more frequently observed near telomeres and cen-

tromeres, but can apparently occur anywhere. They often encompass known genes. We are presently investigating this phenomenon more fully, and will report on them subsequently. For now, we show how they impact the interpretation of cancer-normal data.

In Figure 6, C and D, we have overlain the segmentation values from the analysis of SK-BR-3 in green. The copy number lattice for SK-BR-3 is plotted as orange lines. Figure 6C illustrates a region in SK-BR-3 that would be called a deletion in comparison to the normal. In SK-BR-3 compared to normal, the flanking region occurs at a copy number that we judge to be two copies per cell, and within that region, copy number becomes reduced to one. But the same region appears in the comparison of pygmy DNA to the same normal. In Figure 6D, we observe an analogous condition on Chromosome 2p11. In this panel, we have also plotted segmentation data from the tumor. This region is evidently abnormal there as well.

Hence, we are inclined to view this "lesion" as pre-existing in the normal cells of the patient.

DISCUSSION

Comparison of Methodologies for Global Genomic Analysis

We have described a method, representational oligonucleotide microarray analysis, or ROMA, that is useful for detecting amplifications and deletions and sites of breakage in cancer and normal genomes. Detection of these events can in principle be used to discover genes involved in cancer and other diseases of genetic origin, and serve as markers or guides for the diagnosis and treatment of such diseases. Because our method is sensitive to even single nucleotide polymorphisms at restriction endonuclease sites, it could in principle also be used as a high-density array for detecting SNPs.

There are other methods for global analysis of cancers. Most well known is the gene expression microarray (Chee et al. 1996; DeRisi et al. 1996). This method does not find the primary lesions in a cancer, but rather the sequels of mutation. Gene expression microarrays are based on RNA extracted from tumors, and RNA is a very unstable molecule, difficult to extract in a reliable manner. Moreover, the outcome of expression array analysis will be extremely dependent on difficult-to-control factors such as sample handling, and other complicating physiological variables such as tumor infiltration by normal stroma and inflammatory cells. Our method is based on DNA, a very stable molecule, easily extracted even from tissue that has been mishandled. The DNA is the repository of the causative molecular events, and the presence of normal infiltrating stroma and inflammatory cells dilutes the signal but does not change it. We do not intend for our method to exclude RNA analysis, and in fact the two together would be more valuable than either alone.

There are other DNA-based methods for measuring changes in copy number in cancers. The oldest of these is fluorescent in situ hybridization (FISH), which is used clinically to evaluate amplification at the *ErbB-2* locus in breast cancer, for example (Tkachuk et al. 1990; Bartlett and Mallon 2003). In work in progress, we have shown that our method is essentially equivalent for evaluating amplification at *ErbB-2*, but, of course, our method evaluates the entire genome, not just a single locus that may be important in selecting cancer therapy. The major advantage of FISH is that it is essentially a single-cell assay that can thus be performed on very few cells, such as might be available upon needle biopsy. Our method requires perhaps ~2000 cells, and is a mass measurement, not a single-cell assay. However, our method points to loci that may be converted into FISH-based assays, and that is a major strength.

Another DNA-based method is the BAC array, which is a method that is more commonly known, and more widely practiced, than our method (Pinkel et al. 1998; Snijders et al. 2003). Present BAC arrays suffer from much lower resolution, on the order of 3000 probes. At their maximum, 30,000 member arrays, there are still fewer probes into the genome, and the size of the BAC, 150 to 200 kb, ultimately obscures high resolution. For example, we can observe very small deletions and amplifications that would be entirely missed with even high-density BAC arrays. Additionally, because our method is based on representations, our sample size can be smaller than is needed for the standard BAC array protocol. (However, users of BAC arrays may use our representational approaches to diminish their need for large sample sizes.) Furthermore, BAC arrays cannot be fabricated to industrial standards, as can our arrays. The composition of our arrays is precisely specified, nucleotide for nucleotide, and a

highly reproducible standard product can be made available for wide usage. Again, each and every one of our probes can be readily calibrated for performance, a property that cannot be readily done with BAC probes. Finally, our arrays are based on oligonucleotides derived from the human sequence assembly, the lingua franca of human genetics, and can therefore be precisely and automatically mapped into all the databases of all mapped genes and genetic disorders. This cannot be done with BACs, which can be unstable under propagation and can be chimeric. The one advantage of BAC arrays is that they are presently cheaper, but that is likely to be a short-lived advantage.

cDNA arrays have also been used for measuring copy number mutations (Pollack et al. 1999; Hyman et al. 2002), whereby whole genomic DNA is hybridized to a cDNA expression array. These are presently insensitive. Moving averaging of the measured probe ratios is used to decrease system noise, and this results in a decrease in resolution. Therefore this methodology is useful for the detection of larger amplifications and deletions. However, detecting deletions is problematic because of overall signal-to-noise issues of single fragment or oligonucleotide probes. ROMA has overcome this problem by decreasing the complexity of the genome, thereby increasing the signal-to-noise ratio for each probe.

Is Our Knowledge of Cancer Complete?

Science has identified many of the commonly mutated genes in cancer, and we know many of the cellular pathways on which they act. Some think a basic theory of cancer is comprised of only a few basic principles, sufficient to explain the nature of the disease. However, it is a poor and unnecessary gamble to act as though our theory is correct, or that our knowledge of specific facts is nearly complete. Future progress in detection, prognosis, and treatment of cancer will depend on the accuracy and completeness of our understanding of its specific molecular causes.

There are simple tests for the completeness of our understanding and knowledge of how cancers survive in and kill their hosts. If our knowledge of the genes were complete, we would see a plateau in the number of common mutant genes found in all cancers. If our understanding of the principles were complete, even advanced cancers with a large number of accumulated genetic lesions would show only a small number of commonly affected pathways. It follows from this, that if mutation in a single gene were sufficient to affect a given pathway, even advanced cancers would show only a small number of commonly affected genes, the remainder of lesions being highly sporadic.

The microarray-based method we have just described can partially address these issues. We can readily identify loci in the genome that undergo amplification, deletion, and imbalanced breaks. Although there are many other possible mechanisms that alter critical genes, such as point mutations, balanced translocations, and possibly stable epigenetic changes, many if not most oncogenes and tumor suppressor genes will eventually be found in the types of lesions that we can readily detect. Moreover, if a region is commonly found altered in cancers, that region harbors a good candidate cancer gene. Therefore, the application of our method to a large series of cancers, and the comprehensive comparative analysis of such data, should reveal the existence and number of candidate cancer genes in cancers.

Sources of Cancer Genomes

We have demonstrated the application of our method to two types of sample: a tumor and a cancer cell line. There are advantages and problems associated with each type. Cancer cell lines are "universal" reagents. They are self-replenishing, and can be passed between investigators. There is always ample material for

analysis, and they tend to be monoclonal. They are suitable for further functional analysis, whether by gene expression profiling, genetic manipulation to restore or block a suspected tumor suppressor gene or oncogene, or by tumorigenicity studies. Almost always there is no matched normal to control for scattered polymorphic variation, but as we have described above, this is not a serious limitation, as long as the unmatched normal can be characterized. The significant disadvantages of cell lines are that they can drift genetically, and they have undergone selection by virtue of their survival in tissue culture. There is a limited repertoire of such cell lines, and no correlations between clinical presentation and copy number can be made.

The direct analysis of tumor material offers many opportunities. There is a virtually unlimited source of different samples, and they can often be matched to the same normal, easing somewhat the analytical burden of interpretation. It is in principle possible to determine whether there are clinical parameters, such as survival and drug responsiveness, that correlate with specific gene amplification, deletions, and breakage, or overall patterns of genomic instability. These correlations may find utility in the treatment of patients. The disadvantages of tumor material are also clear. Tumors are always contaminated with stroma, can be oligoclonal, poorly preserved, and available in limiting amounts. Fortunately, our method seems to be highly sensitive, and does not require vast amounts of starting material. We routinely start from 50 ng of sample, which corresponds to ~10,000 nuclei, and the method can be practiced with as few as 2000 nuclei or less. Either flow sorting or microdissection can enrich tumor purity, but amplifications and many deletions will be observable even with material that is only 50% tumor (reconstruction experiments; data not shown).

Technological Critique

Our method rests on three pillars: complexity reduction by representations, the human genome assembly, and oligonucleotide microarrays.

Because of the success of the human genome sequencing project, and the reproducibility of representations, we are able to design oligonucleotide probes that are complementary to a given representation, such as the *Bgl*III representations that we have used here. Because the human genome sequence is very reliable, at least locally, we are able to experimentally validate our computationally derived designs by exploiting the known restriction endonuclease sites in our fragments (see Fig. 1). In principle, we can thus calibrate every probe's performance. The detection of these ~1800 predicted probes validates the ability of this method to detect and identify copy number fluctuations. There are ~10% of the probes that are poor performers in the pin printed format. By calibrating the probes, performance can be accounted for during further analysis. Performance improves with the photo-print format because of the empirical selection of the oligonucleotides.

Of the 8000 probes predicted to hybridize to fragments not cleaved by *Hind*III (see Fig. 1), ~16 appear to hybridize to *Bgl*III fragments that are in fact cleaved. We estimate that these 16 detect homozygous and heterozygous *Hind*III sites, in equal proportion. We attribute this to a divergence of about one nucleotide in 300 between our sample and the published human sequence, which could result from either polymorphism or sequencing errors. If this number were mainly caused by polymorphism, then roughly one in 30 *Bgl*III fragments would also be polymorphic. From other experiments, we estimate that the rate of *Bgl*III polymorphism between unrelated individuals is more on the order of one in 60, corresponding to a divergence from the published human sequence of 1 in 600. Because the

public human sequence is reasonably well assembled, we automatically have associated map positions for every probe that are as accurate as the genome assembly. The algorithms we use for designing these probes are in part described here, and in part in Healy et al. (2003). Our approach allows us to design probes that have minimal cross-reaction to the remainder of the genome. Microarrays for any species, for example, mouse, can be built in short order once a reliably complete and assembled genome sequence is publicly available.

There are many advantages to an oligonucleotide microarray format. The composition of the microarray is precisely formulated, and hence entirely reproducible by others. The work presented here demonstrates the equivalence of measurements achieved by the printed and light directed microarray formats. Using printed arrays we can achieve densities of 30,000 probes per slide, and using in situ light-directed synthesis, we have achieved densities of 190,000, although only 85K data are illustrated here. The latter technique has many advantages over the printed array. Besides achieving higher density, the layout of probes and the choice of probes are flexible. Although the unit cost of printed arrays is presently below the costs of light-directed microarrays, with the latter there is no need for a large initial capital expenditure for the purchase of oligonucleotides.

Our method is dependent on representations. Without complexity reduction, which increases the concentration of DNA complementary to the probes, signal intensity from specific hybridization is too weak to measure above background. Dependence on representations is a mixed blessing. Representations use PCR both for the amplification of sample and complexity reduction. As a consequence, very little sample is required. However, PCR does introduce noise, and this requires that the test sample be compared with a control sample that is prepared exactly in parallel. We find that if the starting DNAs of test and control are of comparable quantity and quality, then subsequent parallel sample preparation, from PCR to labeling, is usually sufficient to give data of the type that is illustrated in this report.

There are a finite number of repeat-free 70-mer-long oligonucleotide probes in the genome that are useful for measuring *Bgl*III representations. We estimate that there are on the order of 120,000 of these scattered about the genome in a Poisson-like distribution, and the distribution of probes does not reflect the distribution of genes. At present we only array ~85,000 probes. Although the average distance between these 85,000 probes is ~30 kb, there are regions of the genome that are very poorly represented. We are therefore designing other types of representations, and other formats of probes, that will give us even higher coverage of the genome. In principle, any desired density of coverage is possible.

Data Interpretation

All array-based data require interpretation using statistical tools of varying sophistication. Ours is no exception, but our system is relatively unique. First, unlike cDNA expression profiling, there are clear theoretical expectations of copy number measurements. When comparing a test sample to a normal genome, there is a clear expectation of how normals, except for polymorphisms, will behave. Moreover, if the test sample is clonal, we expect probe ratios to be clustered, reflecting discrete integral copy numbers per cell. Second, because the restriction endonuclease profile of fragments is known, virtually all probes can be calibrated, and array performance can be very accurately modeled. Third, because the probes are ordered in the genome, and lesions are expected to be regional, with defined starts and stops, the expectation is that consecutive probe ratios within these regions will share a distribution. Thus, we have developed "segmenta-

tion" algorithms that are designed to parse the data into regions with similar distributions.

Our present segmentation algorithm requires a minimum of three probes to define a lesion, but clearly this is conservative. For example, when our tumor sample is compared with a matched normal, polymorphisms are controlled, and even a single probe with an elevated copy number in the tumor is likely to be meaningful. Other approaches to data analysis should be pursued, and we are attempting to integrate polymorphism data, probe calibration data, and probe intensity data into a more comprehensive model. Our present methods are not finished, but they are clearly already useful. We expect that the borders of regions can be drawn very sharply, most often to within a single probe, and this is confirmed in modeling experiments (data not shown).

We will report on our progress in statistical methods in subsequent publications. In the end, however, no statistical interpretation of a single experiment is certain, and only the accumulation of larger data sets and molecular confirmation can increase confidence in a conclusion.

Normal Polymorphic Variation

Scattered polymorphism is evident in comparison of normal individuals, and even in the comparison of a single individual in a "depletion" experiment (see Fig. 1). Most of these likely arise from single nucleotide polymorphisms in the human population. For example, loss of a *Bgl*III site may cause a fragment to be absent in a *Bgl*III representation. Such events can interfere in data interpretation in several ways. Except for the case of increased copy number in a matched tumor-normal, the ratio from a single probe outlier cannot be considered a somatic lesion, as it may represent a genetic polymorphism, with or without loss of heterozygosity. Similarly, the boundaries of a segment may not be accurately called if the bounding probe is complementary to a polymorphic fragment. Lastly, a string of probes that by chance are all complementary to polymorphic fragments may give rise to the appearance of a consistent lesion. Fortunately, the frequency of these polymorphisms is low, less than one fragment in 30, so most boundaries are not obscured, and runs of polymorphisms with the appearance of a lesion will occur rarely. Much of the informatic "damage" caused by polymorphisms can be contained, either by filtering out scattered outliers, or by accumulating data on normal genomes used for comparisons.

There is another type of "polymorphism" that we see, which for now we call "copy number" polymorphism. This type is much more interesting, and more pernicious, than scattered polymorphism, and it is documented in Figure 6. A series of regionally clustered probes may display a consistently altered ratio in the comparison of one normal sample against another. We see these regions in every normal-normal comparison that we have made, and many of these lesions appear in cancer-normal comparisons. In fact, some of these regions may be prone to genomic instability (see Fig. 6D). They vary in size from <100 kb to in excess of 1 Mb, and in most cases encompass genes. Creating a large database of normal-normal comparisons may mitigate the misinterpretation of these lesions as somatic events occurring in cancer, and this is something we intend to do.

Our present hypothesis is that these normal-normal variations are in fact copy number polymorphisms, genetic in origin, but this is by no means proven here, nor is it the only plausible hypothesis. For example, these variant regions might be caused by locally high sequence divergence, or the consequence of highly altered chromatin structure, affecting the yield of DNA during purification from nuclei. Additional experimentation is needed to resolve these questions, and work in progress strongly

indicates that the majority of these normal variations are, indeed, alterations in the gene pool. If there is, in fact, widespread copy number variation in humans, such variation might well contribute to human traits, including disease susceptibility and resistance.

METHODS

Reagents

Oligonucleotides were synthesized by Illumina Inc. Human Cot-1 DNA (15279-011) and yeast tRNA (15401-029) were supplied by Invitrogen Inc. Restriction enzymes, ligase, and Klenow fragments (M0212M) were supplied by New England Biolabs. The Megaprime labeling kit, Cy3-conjugated dCTP, and Cy5-conjugated dCTP were supplied by Amersham-Pharmacia. Taq polymerase was supplied by Eppendorf. Centricon YM-30 filters were supplied by Amicon (42410), and formamide was supplied by Amresco (0606-500). Phenol:chloroform was supplied by Sigma (P2069). NimbleGen photoprint arrays were a gift from NimbleGen Systems Inc.

Representation

*Bgl*III representations, in general, were prepared as previously described (Lucito et al. 2003b). A major change is that amplification was carried out in an MJ Research Tetrad. Sixteen 250- μ L tubes were used for amplification of the representation. The cycle conditions were 95°C for 1 min, 72°C for 3 min, for 25 cycles, followed by a 10-min extension at 72°C. The contents of the tubes were pooled when completed. Representations were cleaned by phenol:chloroform extraction, precipitated, resuspended, and the concentration determined. Representations depleted of specific fragments by restriction enzyme were prepared in the same manner with the following modification. After ligation of adaptor, the mixture was cleaned by phenol:chloroform extraction, precipitated, and resuspended. The ligated fragments were then digested with the second chosen enzyme. In the text, *Hind*III was used. This material was then used as template in the PCR reaction.

Probe Selection

We performed an in silico *Bgl*III digestion of the human genome by locating all *Bgl*III restriction sites within the present draft assembly and storing all sequences of *Bgl*III fragments that are between 200 and 1200 bp in length. Fragments were annotated with the counts of their substituent, overlapping 15-mers and 21-mers using the "mer-engine" constructed from the same draft assembly (see accompanying manuscript by Healy et al. 2003). For each fragment, the following attributes were determined for every substituent, overlapping 70-mer: maximum 21-mer count, arithmetic mean of 15-mer counts, percent GC content, the quantity of each base, and the longest run of any single base. All 70-mer probes that possess any of the following characteristics were eliminated: maximum 21-mer count >1, GC content <30% or >70%, a run of A/Ts >6 bases, a run of G/Cs >4 bases. From the remaining set of 70-mers, the one (or more) that has a GC/AT proportionality closest to that of the genome as a whole as well as a minimal mean 15-mer count were selected. As a final check for overall uniqueness, the optimal probes for each fragment were compared with the entire genome using BLAST (default parameters were used with the exception of filtration of low complexity sequence, which was not performed). Any probe found to have any degree of homology along 50% or more of its length to any sequence other than itself was eliminated.

Printed Arrays

We used the Cartesian PixSys 5500 (Genetic Microsystems) to array our probe collection onto slides. We are presently using a 4 $\times$ 4 pin configuration. The dimension of each printed array was roughly 2 cm². Our arrays were printed on commercially prepared silanated slides (Corning ultraGAPS #40015). Pins used for the arrayer are from Majer Precision.

Labeling

DNA was labeled as described (Lucito et al. 2003a). Briefly, place DNA template (dissolved in TE at pH 8) in a 0.2-mL PCR tube. Add 10 μ L of Primers from the Amersham-Pharmacia Megaprime labeling Kit and pipette up and down several times. Bring volume up to 100 μ L with dH₂O, and mix. Place tubes in Tetrad at 100°C for 5 min, then place on ice for 5 min and add 20 μ L of labeling buffer from the Amersham-Pharmacia Megaprime labeling Kit, 10 μ L of label (Cy3-dCTP or Cy5-dCTP), and 1 μ L of NEB Klenow fragment. Place the tubes in a Tetrad and incubate at 37°C for 2 h. Combine the labeled samples (Cy3 and Cy5) into one Eppendorf tube and add 50 μ L of 1 μ g/ μ L human Cot 1 DNA, 10 μ L of 10 mg/mL stock yeast tRNA, and 80 μ L of Low TE (3 mM Tris at pH 7.4, 0.2 mM EDTA). Load all into a Centricon Filter and centrifuge for 10 min at 12,600 rcf. Discard flowthrough and wash with 450 μ L of Low TE. Centrifuge at 12,600 rcf and repeat twice. Collect the labeled sample by inverting the centricon column into a new tube and centrifuging for 2 units at 12,600 rcf. Transfer labeled sample to a 200- μ L PCR tube and adjust volume to 10 μ L of Low TE.

Slide Preparation

Slides were prepared as in Lucito et al. (2003a) with the following changes. Prehybridization buffer for printed microarrays consisted of the following, 25% deionized formamide, 5 $\times$ SSC, and 0.1% SDS. Pour into a coplin jar or other slide processing chamber and preheat to 61°C. UV cross-link DNA to slide (using a Strategene Statalinker, set Energy to 300 mJ, rotate slide 180°, keeping the slide in the same spot in the cross-linker, and repeat). NimbleGen photoprinted arrays do not require UV cross-linking. Wash slides in the following solutions: 2 min in 0.1% SDS, 2 min in milliQ H₂O, 5 min in milliQ H₂O that has boiled, and finally in ice cold 95% benzene-free EtOH. Dry slides by placing in a metal rack and spin at 75 rcf for 5 min. Printed microarray slides were incubated in the 61°C prehyb solution. After 2 h, wash slides in milliQ H₂O for 10 sec. Dry slides by placing in a metal slide rack and spin for 5 min at 75 rcf. NimbleGen photoprinted arrays do not require prehybridization.

Hybridization

The hybridization solution for printed slides consisted of 25% formamide, 5 $\times$ SSC, and 0.1% SDS. The hybridization solution for NimbleGen photoprinted arrays consisted of 50% formamide, 5 $\times$ SSC, and 0.1% SDS. For each, 25 μ L of hybridization solution was added to the 10 μ L of labeled sample and mixed. Samples were denatured in an MJ Research Tetrad at 95°C for 5 min, and then incubated at 37°C for 30 min. Samples were spun down and pipetted onto a slide prepared with lifter slip and incubated in a hybridization oven such as the Boekel InSlide Out oven set at 58°C for printed arrays or 42°C for NimbleGen photoprinted arrays for 14 to 16 h. After hybridization, slides were washed as follows: brief wash in 0.2% SDS/0.2 $\times$ SSC to remove the coverslip, 1 min in 0.2% SDS/0.2 $\times$ SSC, 30 sec in 0.2 $\times$ SSC, and 30 sec in 0.05 $\times$ SSC. Slides were dried as before by placing in a rack and spinning at 75 rcf for 5 min, and then scanned immediately. An Axon GenePix 4000B scanner was used setting the pixel size to 10 μ m for printed arrays and 5 μ m for NimbleGen photoprinted arrays. GenePix Pro 4.0 software was used for quantitation of intensity for the arrays. Array data were imported into S-PLUS for further analysis. Measured intensities without background subtraction were used to calculate ratios. Data were normalized using an intensity-based lowest curve fitting algorithm similar to that described in Yang et al. (2002). Data obtained from color reversal experiments were averaged and displayed as presented in the figures.

ACKNOWLEDGMENTS

We thank Emile Nuwaysir and Todd Richmond of NimbleGen Systems Inc. for providing slides and support, and Masaaki Hamaguchi for critical comments on the manuscript. We also thank Joe Derisi and Michael Eisen for technical comments on

the printing of oligonucleotides. Tumor samples were supplied by the Cooperative Human Tissue Network, which is funded by the National Cancer Institute. Other investigators may have received samples from these same tissues. This work was supported by grants awards to M.W. from the National Institutes of Health and NCI (5R01-CA78544; 1R21-CA81674; 5R33-CA81674-04); Tularik Inc.; 1 in 9: The Long Island Breast Cancer Action Coalition; Lillian Goldman and the Breast Cancer Research Foundation; The Miracle Foundation; The Marks Family Foundation; Babylon Breast Cancer Coalition; Elizabeth McFarland Group; and Long Islanders Against Breast Cancer. Support was granted to R.L. from the National Institutes of Health and NCI (K01 CA93634-01). M.W. is an American Cancer Society Research Professor.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

- Bartlett, J. and Mallon, E.C.T. 2003. The clinical evaluation of HER-2 status: Which test to use? *J. Pathology* **199**: 418–423.
- Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhard, D.J., Morris, M.S., and Fodor, S.P. 1996. Accessing genetic information with high-density DNA arrays. *Science* **274**: 610–614.
- DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su, Y.A., and Trent, J.M. 1996. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* **14**: 457–460.
- Gemmell, R.M., West, J.D., Boldog, F., Tanaka, N., Robinson, L.J., Smith, D.I., Li, F., and Drabkin, H.A. 1998. The hereditary renal cell carcinoma 3;8 translocation fuses FHIT to a patched-related gene, TRC8. *Proc. Natl. Acad. Sci.* **95**: 9572–9577.
- Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., et al. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**: 531–537.
- Hamaguchi, M., Meth, J.L., von Klitzing, C., Wei, W., Esposito, D., Rodgers, L., Walsh, T., Welch, P., King, M.-C., and Wigler, M. H. 2002. *DBC2*, a candidate for a tumor suppressor gene involved in breast cancer. *Proc. Natl. Acad. Sci.* **99**: 13647–13652.
- Healy, J., Thomas, E.E., Schwartz, J.T., and Wigler, M.H. 2003. Annotating large genomes with exact word matches. *Genome Res.* (this issue).
- Hyman, E., Kauraniemi, P., Hautaniemi, S., Wolf, M., Mousses, S., Rozenblum, E., Ringner, M., Sauter, G., Monni, O., Elkhouloun, A., et al. 2002. Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Res.* **62**: 6240–6245.
- Ilyin, G.P., Rialland, M., Pigeon, C., and Guguén-Guillouzo, C. 2000. cDNA cloning and expression analysis of new members of the mammalian F-box protein family. *Genomics* **67**: 40–47.
- Lander, E.S., Linton, L.M., Birren, B., Nussbaum, C., Zody, M.C., 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.
- Li, J., Yen, C., Liaw, D., Podsypanina, K., Bose, S., Wang, S.I., Puc, J., Millaresis, C., Rodgers, L., McCombie, R., et al. 1997. PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science* **275**: 1943–1947.
- Lin, M.Z. and Greenberg, M.E. 2000. Orchestral maneuvers in the axon: Trio and the control of axon guidance. *Cell* **101**: 230–242.
- Lisitsyn, N., Lisitsyn, N., and Wigler, M. 1993. Cloning the differences between two complex genomes. *Science* **258**: 946–951.
- Lucito, R., Nakimura, M., West, J.A., Han, Y., Chin, K., Jensen, K., McCombie, R., Gray, J.W., and Wigler, M. 1998. Genetic analysis using genomic representations. *Proc. Natl. Acad. Sci.* **95**: 4487–4492.
- Lucito, R., West, J., Reiner, A., Alexander, J., Esposito, D., Mishra, B., Powers, S., Norton, L., and Wigler, M. 2000. Genetic alterations in cancer detected by hybridization to micro-arrays of genomic representations. *Genome Res.* **10**: 1726–1736.
- Lucito, R. and Wigler, M. 2003a. Preparation of Slides and Hybridization. In *Microarray-based representational analysis of DNA copy number* (eds. D. Bowtell and J. Sambrook), pp. 394–399. Cold Spring Harbor Press, Cold Spring Harbor, NY.
- Lucito, R. and Wigler, M. 2003b. Preparation of Target DNA. In *Microarray-based representational analysis of DNA copy number* (eds. D. Bowtell and J. Sambrook), pp. 386–393. Cold Spring Harbor Press, Cold Spring Harbor, NY.

- Mu, D., Chen, L., Zhang, X., See, L.-H., Koch, C.M., Yen, C., Tong, J.J., Spiegel, L., Nguyen, K.C.Q., Servoss, A., et al. 2003. Genomic amplification and oncogenic properties of the *KCNK9* potassium channel gene. *Cancer Cell* **3**: 297–302.
- Nurnberg, P., Thiele, H., Chandler, D., Hohne, W., Cunningham, M.L., Ritter, H., Leschik, G., Uhlmann, K., Mischung, C., Harroop, K., et al. 2001. Heterozygous mutations in ANKH, the human ortholog of the mouse progressive ankylosis gene, result in craniometaphyseal dysplasia. *Nat. Genet.* **28**: 37–41.
- Olshen, A.B. and Venkatraman, E.S. 2002. *Change-point analysis of array-based comparative genomic hybridization data*. American Statistical Association, Alexandria, VA.
- Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W. L., Chen, C., Zhai, Y., et al. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* **20**: 207–211.
- Pollack, J.R., Perou, C.M., Alizadeh, A.A., Eisen, M.B., Pergamenschikov, A., Williams, C.F., Jeffrey, S.S., Botstein, D., and Brown, P.O. 1999. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* **23**: 41–46.
- Sen, A. and Srivastava, M.S. 1975. On tests for detecting change in mean. *Ann. Stat.* **3**: 98–108.
- Singh-Gasson, S., Green, R.D., Yue, Y., Nelson, C., Blattner, F., Sussman, M.R., and Cerrina, F. 1999. Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat. Biotech.* **17**: 974–978.
- Snijders, A.M., Nowee, M.E., Fridlyand, J., Piek, J.M., Dorsman, J.C., Jain, A.N., Pinkel, D., van Diest, P.J., Verheijen, R.H., and Albertson, D.G. 2003. Genome-wide-array-based comparative genomic hybridization reveals genetic homogeneity and frequent copy number increases encompassing CCNE1 in Fallopian tube carcinoma. *Oncogene* **22**: 4281–4286.
- Tkachuk, D.C., Westbrook, C.A., Andreeff, M., Donlon, T.A., Cleary, M.L., Suryanarayan, K., Homge, M., Redner, A., Gray, J., and Pinkel, D. 1990. Detection of bcr-abl fusion in chronic myelogenous leukemia by in situ hybridization. *Science* **250**: 559–562.
- Van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**: 530–536.
- Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J., and Speed, T.P. 2002. Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**: e15–15.

WEB SITE REFERENCES

<http://roma.cshl.org/>; ROMA.

Received March 20, 2003; accepted in revised form August 1, 2003.

March 2004

Circular Binary Segmentation for the Analysis of Array-based DNA Copy Number Data

Adam B. Olshen, E. S. Venkatraman

Department of Epidemiology and Biostatistics

Memorial Sloan-Kettering Cancer Center

1275 York Avenue, New York, NY 10021

e-mail: olshena@mskcc.org

Robert Lucito, Michael Wigler

Cold Spring Harbor Laboratory

Cold Spring Harbor, NY 11724

Abstract

DNA sequence copy number is the number of copies of DNA at a region of a genome. Cancer progression often involves alterations in DNA copy number. Newly developed microarray technologies enable simultaneous measurement of copy number at thousands of sites in a genome. We have developed a modification of binary segmentation, which we call *circular binary segmentation*, to translate noisy intensity measurements into regions of equal copy number. The method is evaluated by simulation and is demonstrated on cell line data with known copy number alterations and on a breast cancer cell line data set.

Key Words: change-point, binary segmentation, array CGH, ROMA

1 Introduction

The DNA copy number of a region of a genome is the number of copies of genomic DNA. In humans the normal copy number is two for all the autosomes. Variations in copy number are common in cancer and other diseases. These variations are a result of genomic events causing discrete gains and losses in contiguous segments of the genome. For this reason, efforts have been made over the last ten years to make whole genome copy number maps from a single study. Technologies to accomplish this have included comparative genomic hybridization (CGH) (Kallioniemi et al., 1992) and representational difference analysis (RDA) (Lisitsyn et al., 1993). In order to increase the resolution of the resulting maps, both techniques have been modified for use with microarrays, the laboratory techniques of which are similar to cDNA gene expression experiments. Each microarray consists of thousands of genomic targets or probes, which we will sometimes refer to as *markers*, that are spotted or printed on a glass surface. In a copy number experiment a DNA sample of interest, called the *test sample*, and a diploid *reference* sample are differentially labelled with dyes, typically Cy3 and Cy5, and mixed. This combined sample is then hybridized to the microarray and imaged which results in test and reference intensities for all the markers.

The modification of conventional CGH to obtain high resolution data is called array CGH (aCGH) (Pinkel et al. 1998; Snijders et al. 2001). Here the genomic targets are bacterial artificial chromosomes (BACs), which are large segments of DNA, typically 100 – 200 kilobases. Representational Oligonucleotide Microarray Analysis (ROMA) (Lucito et al., 2000; Lucito et al., 2003) is the high resolu-

tion version of RDA. In ROMA, the test and reference samples are based on “representations,” (Lisitsyn et al., 1993), which are subsets of a genome. To create representations, genomic DNA is first shattered using an enzyme. The DNA pieces of proper size, less than 1.2 kilobases, are then selectively amplified by PCR. Importantly, the positions of shearing and pieces that amplify are the same every time. Typically, a representation contains less than 5% of the original sample. The reduction in complexity of a representation compared to the original sample leads to a reduction in hybridization to the wrong marker, which is termed *cross-hybridization*. Thus the DNA segments on a ROMA array can be much smaller than for other types of copy number arrays. A third technique to estimate copy number is to simply employ the same cDNA arrays used for gene expression studies (Pollack et al., 1999; Pollack et al., 2002).

The data from array based copy number experiments are the test and reference sample intensities for each marker. Since we assume that the reference sample does not have any copy number aberrations, markers with normalized test intensities significantly greater than the reference intensities are indicative of copy number gains in the test sample at those positions. Similarly, significantly lower intensities in the test sample are signs of copy number losses. The statistical methods for analyzing copy number data are thus aimed at identifying locations of gains or losses of copy numbers.

The most common method of analysis for these data is to identify gains and losses using thresholds, such as in Weiss et al. (2003). These thresholds are often based on the variability of data from experiments where the test sample and

the reference sample are the same normal tissue. Sometimes, the data are first smoothed via local averaging. A variant of the typical analysis can be seen in Pollack et al. (2002). Here, the data were smoothed and a statistic was calculated for each marker in normal-normal experiments by averaging over an optimally determined window size. Then, a threshold was determined for gains or losses based on the false discovery rate (Benjamini and Hochberg 1995). A model-based approach for array copy number data is due to Hodgson et al. (2001). They fit a three-component normal mixture model to mouse islet tumor data. In this model, there is one component for “decreased” copy number, one for “normal” copy number, and one for “increased” copy number. Autio et al. (2003) developed CGH-Plotter which combines filtering, 3-means clustering and dynamic programming to split CGH data into three groups as above. A promising approach is due to Snijders et al. (2003). They developed heuristic methods for fitting a Gaussian hidden Markov model to array copy number data. Linn et al. (2003) used a change-point model on RNA expression data to obtain the maximum likelihood estimate of the location of a copy number change, which they then compared to estimates from array DNA copy number data.

The model underlying our work is that gains or losses of copy number are discrete. These aberrations occur in contiguous regions of the chromosome that often cover multiple markers up to whole chromosome arms or chromosomes. In addition, the array copy number data can be noisy, so that some markers will not reflect the true copy number in the test sample. Therefore, we seek a method to split the chromosomes into regions of equal copy number that accounts for the

noise in the data. We propose a modification of binary segmentation (Sen and Srivastava, 1975) that we call *circular binary segmentation* (CBS) for this. Our method is novel in that it provides a natural way to segment a chromosome into contiguous regions and bypasses parametric modeling of the data with its use of a permutation reference distribution.

The rest of the manuscript is organized as follows. In Section 2 we show the relationship between the identification of aberrant genomic regions and change-point problems and introduce the CBS methodology. Results of using the approach of Section 2 to array CGH cell line data with known aberrations are covered in Section 3. In Section 4 results are shown from application to ROMA data from twenty three breast cancer cell lines. In Section 5 we study the accuracy of the CBS method via simulation. We summarize our results and discuss future directions in Section 6.

2 Change-point methods

We will now show the connection between estimating the locations of regions with aberrant DNA copy numbers and the change-point detection problem. This makes change-point methods a natural framework to approach the analysis of array DNA copy number data. Let $X_1, X_2, \dots$ be a sequence of random variables. An index ν is called a change-point if $X_1, \dots, X_\nu$ have a common distribution function F_0 and $X_{\nu+1}, \dots$ have a different common distribution function F_1 until the next change-point if one exists). Shaban (1980) and Basseville (1988) provide extensive reviews of change-point problems and methods.

In array copy number studies the data to be analyzed are naturally ordered by the marker location along the chromosome of interest. The data are the test and reference intensities for each marker (denoted I_{tm} and I_{rm} respectively for marker m). These intensities are related to the DNA copy number in the test and the reference samples, C_{tm} and C_{rm} respectively. This relationship is modeled as $I_{tm} = \beta_{tm}C_{tm}(1 + \epsilon)$ and $I_{rm} = \beta_{rm}C_{rm}(1 + \epsilon)$, where the parameters (β s) depend on factors such as sample amplification, probe affinity and labeling and ϵ s are random errors. Note that the normalization of data used to correct for the factors above, centers the log ratio of the intensities around zero. This makes the copy numbers unidentifiable without some additional modeling, since, for example, the normalized data from diploid and triploid test samples will appear similar. However, the location of the log ratio of the intensities changes whenever $\beta_{tm}C_{tm}/(\beta_{rm}C_{rm})$ changes and thus the corresponding marker indices are the change-points we want to detect. Since the reference sample is assumed to have no abnormalities and the log ratio of the β s is assumed to be constant, all the change-points correspond to changes in the copy numbers of the test sample.

The array data to be used for change-point detection are the log ratio of normalized intensities indexed by the marker locations. Observe that there may be multiple change-points in a given chromosome, each corresponding to a change in the copy number in the test sample. Our goal is to identify all the change-points which will then partition the chromosome into segments where copy numbers are constant. Once the chromosome is partitioned we can estimate the copy numbers of the segments with the help of additional information such as the ploidy of the

chromosome. This will provide the locations of copy number aberrations.

Let $X_1, \dots, X_n$ be the log ratios of the intensities, which are indexed by the locations of the n markers being studied and let $S_i = X_1 + \dots + X_i, 1 \leq i \leq n$, be the partial sums. When the data are normally distributed with a known variance (without loss of generality 1), the likelihood ratio statistic for testing the null hypothesis that there is no change against the alternative that there is exactly one change at an unknown location i (Sen and Srivastava, 1975) is given by $Z_B = \max_{1 \leq i \leq n} |Z_i|$, where

$$Z_i = \{1/i + 1/(n-i)\}^{-1/2} \{S_i/i - (S_n - S_i)/(n-i)\}.$$

The null hypothesis of no change is rejected if the statistic exceeds the upper α^{th} quantile of the null distribution of Z_B and the location of the change-point is estimated to be i such that $Z_B = |Z_i|$. Sen and Srivastava derived the critical value to be used for the test by Monte Carlo simulations. It can be computed quickly using the approximation for the tail probabilities of the test statistic given by Siegmund (1986). The *binary segmentation procedure* applies the test recursively until no more changes are detected in any of the segments obtained from the change-points detected thus far.

The binary segmentation procedure was shown to be consistent under suitable regularity conditions (Vostrikova, 1981). If the variance is unknown the procedure can be extended with a good estimate of it derived from the data. Note that in this case the statistics Z_i is replaced by the corresponding t -statistic and the overall statistic to test for a change is the maximum of these absolute t 's. Since the binary segmentation procedure is based on a test to detect a single change, a

potential problem with it is that it cannot detect a small changed segment buried in the middle of a large segment (Venkatraman, 1992). We propose the following modification of the binary segmentation procedure to address this problem.

This problem with the binary segmentation procedure is due to the fact that it looks for only one change-point at a time. Levin and Kline (1985) proposed a statistic to test for no change against the epidemic or square wave alternative with two change-points. (In the square wave alternative the mean up to the first change and after the second are assumed to be the same.) If we consider the segment to be spliced at the two ends to form a circle, the likelihood ratio test statistic for testing the hypothesis that the arc from $i + 1$ to j and its complement have different means is given by:

$$Z_{ij} = \{1/(j-i) + 1/(n-j+i)\}^{-1/2} \{(S_j - S_i)/(j-i) - (S_n - S_j + S_i)/(n-j+i)\}.$$

Our modification of the binary segmentation procedure, which we call *circular binary segmentation* (CBS), is based on the statistic $Z_C = \max_{1 \leq i < j \leq n} |Z_{ij}|$. Note that Z_C allows for both a single change ($j = n$) and the epidemic alternative ($j < n$). As before we declare a change if the statistic exceeds an appropriate threshold level based on the null distribution. This critical value when the X_i s are normal can again be computed using Monte Carlo simulations or the approximation given by Siegmund (1986) for the tail probability. Once the null hypothesis is rejected the change-point(s) is (are) estimated to be i (and j) such that $Z_C = |Z_{ij}|$ and the procedure is applied recursively to identify all the changes. Other change-point detection schemes such as one based on the Schwartz criterion (Yao, 1988) could also be used for the analysis of array copy number data.

An issue that can arise with the CBS procedure is the edge effect in the estimation of the change-points. That is, if the i and j that correspond to the maximal statistic are such that either i is “close” to 1 or j is “close” to n , then there might be only one true change instead of the two changes suggested by the data. We undo a change if the data does not support it as follows. First, we test whether the data supports i to be a viable change-point for the segment $X_1, \dots, X_j$ and undo the change at i if it is not a viable change-point. A similar test is performed for j . Note this is testing for a binary split. Since it is difficult to determine whether a change-point is “close” to the boundary based just on the values of i and j , we currently perform this test on all change-points derived from ternary splits, that is, splits that result in three different pieces.

The reference distributions used so far were derived using the normality of the data. We can generalize the procedure to non-normal data by generating a reference distribution using a permutation approach as follows. Under the null hypothesis of no change-point in the data, the X_i s are identically distributed. Let $X_1^*, \dots, X_n^*$ be a random permutation of the data and let $Z_C^* = \max |Z_{ij}^*|$ be the statistic derived as above from the permuted data. The threshold value can be chosen to be the upper α^{th} quantile of the permutation distribution given by the Z_C^* s. Since the significance level α used for the test is small we need a very large number of permutations (P) for the estimation of p -value (on the order of 10,000). Considerable computational efficiency can be achieved by stopping the permutation procedure once the number of $Z_C^* > Z_C$ exceeds αP . Note that α is the type I error in testing for a change in a single segment with no change-points.

Since the procedure tests for changes recursively on all resulting sub-segments the probability of finding spurious change-points is a function of the number true change-points and could be larger than α . Since the true number of change-points is unknown we do not correct for this multiple testing problem.

The permutation approach is computationally intensive. A modification is sometimes needed for large data sets. Our solution is to divide the data into K overlapping windows $W_k, k = 1, \dots, K$, of (approximately) equal size and search for change-points within each. The number of windows K depends on the window size and the overlap. The overall test statistic is defined as $Z_C = \max_k Z_k$ where Z_k is the maximum statistic for the data in the windows W_k . The permutation process is repeated as above, but with the new, faster maximization procedure.

There are two additional modifications to the basic procedure to make it more appropriate for array DNA copy number data. The first is to smooth outliers before segmenting. Outliers can be caused either by technical errors in an experiment or by aberrant copy number in a region covering only a single marker. The smoothing region for each i is given by $i - R, \dots, i, \dots, i + R$, where R is a small integer (say 2 to 5). Let m_i be the median of the data in the smoothing region and let $\hat{\sigma}$ be the standard deviation of the entire data. If the observation X_i is the maximum or the minimum of all the observations in the smoothing region we find j in the smoothing region closest to it. If the distance from X_i to X_j exceeds $L\hat{\sigma}$ we replace X_i with $m_i + \text{sign}(X_i - X_j)M\hat{\sigma}$. The values we use for L and M are 4 and 2, respectively.

The second modification is because, for reasons that are not totally under-

stood, there are local trends in the data that are not indicative of real copy number changes. This can lead to the identification of change-points that are not biologically meaningful. Therefore, we use a “pruning” procedure like in CART (Breiman et al., 1984) to eliminate some of them. Suppose there are C change-points after CBS. The sum of squared deviations of data points in segments around their segment average can be represented by $SS(C)$. (This is equivalent to the error sum of square in one way ANOVA.) We then compute $SS(1), \dots, SS(C-1)$, which are the sum of squares corresponding to the best set of change-points of sizes 1 to $C-1$, choosing only among the change-points previously identified. Then $c' = \min\{c : [SS(c)/SS(C) - 1] < \gamma\}$, where γ is some pre-specified constant (such as 0.05 or 0.10). The change-points are those that led to $SS(c')$.

3 Array CGH Example

We applied the CBS methodology with a permutation-based reference distribution to the aCGH data featured in Snijders et al. (2001). (These data are freely available for download at http://www.nature.com/ng/journal/v29/n3/supinfo/ng754_S1.html.) The data consisted of single experiments on 15 fibroblast cell lines. Each array contained 2276 mapped BACs spotted in triplicate. The variable used for analysis was the normalized average of the log base 2 test over reference ratio, as processed by the authors.

There were either one or two alterations in each cell line as identified by spectral karyotyping. Of these, all the alterations for six cell lines covered whole chromosomes and thus would not be identified by our methodology. Therefore,

we limited our analysis to the other nine cell lines. For those lines, we tested for change-points one chromosome at a time. As there is a multiple comparison issues from examining 23 chromosomes, we examined our procedure with the α values 0.01 and 0.001. Results can be found in Table 1. The data from a typical cell line experiment, specifically from cell line GM05296, can be seen in Figure 1.

Notably, both α levels lead to identification of the same regions for the chromosomes that were truly altered. Of the 15 altered regions, 12 were found. Of those not found, chromosome 9 on GM03563 had only two altered points among 139, so in the permutations it was not unlikely to find the two altered points together. For chromosome 12 on GM01535, the region of alteration is represented by only one point and single altered points cannot be found when using a permutation reference. Finally, for chromosome 15 on GM07081, our result is consistent with Snijders et al. (2001) that no evidence of an alteration is seen in the aCGH data. Therefore, our methodology found everything that it should have.

Our methodology also found a number of changes not detected by spectral karyotyping. We are calling these “false positives”, although some may be real and not detectable by spectral karyotyping. The number of false positive chromosomes ranged from 0 to 8, with averages of 4.1($SD = 2.6$) for $\alpha = 0.01$ and 1.8(2.1) for $\alpha = 0.001$. Most of the false positives were a result of what appeared to be local trends in the data, examples of which can be found in Figure 2, which shows the cell line GM03563. Note that these local trends were often in the same locations across cell lines suggesting that there may be a biological reason for them. The segmentation procedure detects change points as it approximates the local trend

by a step function leading to the “false positives.”

4 ROMA Example

Our methods were also applied to unpublished ROMA (Lucito et al., 2003) experiments on twenty three breast cancer cell lines. In this case, the labeled samples were hybridized to a slide containing 9820 unique probes that were each 70 bases long, with each probe spotted only once. Probes were mapped based on the draft human genome sequence. Each cell line was hybridized twice, once with the test sample labeled with Cy3 and the reference labeled with Cy5, and once with the two dyes swapped. This *dye-swapping* negates probe-specific bias in favor of Cy3 or Cy5. The arrays were imaged using the program Genepix. To show the robustness of our method, no spots were eliminated. The log base 2 test over reference intensities were normalized by subtracting off the log ratio that corresponded to a *lowess* (Cleveland, 1979) fit of the log ratio to the average of the test and reference log intensities, as suggested by Yang et al. (2002). This normalization was undertaken in each of the 16 sub-arrays of each array. The normalized log ratios from each dye-swap were averaged before segmentation. The α level for CBS was fixed at 0.01.

The results from applying CBS to these data are shown in Figure 3. Since we do not have external confirmation of these results, our interpretation is necessarily modest. We first focus on a region of chromosome 17 near 40 MB where the ERB-B2 (HER2NEU) gene resides. The ERB-B2 gene is important because it is amplified in 10-40% of breast cancers (Menard et al., 2000) and the drug Herceptin

can be used to treat ERB-B2-amplified cancers. CBS found change-points in the ERB-B2 region in 5 of the cell lines. The cell line in the seventh row and first column of Figure 3 is particularly interesting in this region. Note that the ratios for some of the probes in the region appear to be at the normal level, but CBS is still able to define a likely aberrant region. In addition, CBS helps to define the altered region in the cases where ERB-B2 is amplified.

Another noteworthy aspect of these data is that there were 8 cell lines with no ERB-B2 alteration where there appeared to be a copy number gain or loss in a whole arm of the chromosome. One would expect that the change-point would be found right at the centromere. Since, in most cases, the change-points were found 3 probes after the centromere, it is likely that those three probes are mis-mapped to the wrong side of the centromere. Thus the combination of the high-resolution ROMA data and the CBS algorithm was helpful in identifying these likely errors in the genome sequence.

Figure 4 shows results from the application of CBS to a whole breast cancer cell line. It can be seen that the sorted means of segments separated into plateaus. It is reasonable to assume that each plateau corresponds to a particular copy number, although what that copy number is remains unclear without additional information because the ploidy of the cell line is unknown. In addition, note the high degree of overlap of points that are part of segments in different plateaus. This overlap highlights the weakness of threshold-based methods.

5 Simulations

In this section we will present the results from the Monte Carlo simulations we conducted to evaluate the performance of the CBS algorithm. The data to be segmented were generated from the model $x_i = \mu_i + \epsilon_i, 1 \leq i \leq n$, where n is the sample size, μ is the mean and ϵ the error term which is distributed as $N(0, \sigma^2)$. We used a permutation reference distribution to obtain the p -value of the segmentation procedure with values smaller than 0.01 ($= \alpha$) being considered significant.

In the first set of simulations the mean was set to be $\mu_i = c\sigma\mathcal{I}\{l < i \leq l + k\}$, where $\mathcal{I}$ is the indicator function and the parameters c , l and k control the change in the mean, the location of the change and the width of the changed segment, respectively. For these simulations we chose the value of c from $\{2, 3, 4\}$, l from $\{0, \lfloor (n - k)/2 \rfloor\}$ and k from $\{2, 3, 4, 5\}$. The two values for l correspond to the location of the changed segment being the edge and the center of the data with the correct number of change-points being 1 in the first case and 2 in the second. The CBS algorithm was designed to overcome a shortcoming of binary segmentation which is that it cannot detect a narrow changed segment buried in the middle of a wide segment. Hence in this set of simulations we ran both the procedures to compare their performance. The number of change-points detected from segmenting 1000 simulated data sets are summarized in Table 2.

The simulations show that the estimated number of changes exceeded the true number a maximum of 3% of the times (median: 1.75%; range: 0% – 3%). Even though it exceeds 1%, the excess is reasonable and is consistent with the

multiple testing issue discussed in Section 2. The segmentation procedures have low power to detect a change when the difference in means is small or if the width of the changed segment is small. The proportion of data sets in which the estimated number of change-points equals the true number increases as either c or k increases, except when the binary segmentation procedure was used to detect a changed segment in the middle. The simulation results are a clear demonstration of the inability of the binary segmentation to find a narrow aberrant region in the middle of a chromosome; no change was detected in over 99% of the data sets with the changed segment in the middle. The CBS procedure gains this ability by trading some of its power to detect a changed segment on the edge. Note that the power of both procedures increases more rapidly when the difference in means c increases than when the width k of the segment increases since change in c is equivalent to a change in the square root of k . Finally the exact changed segment is more easily identified when the difference in the means increases than when the width of the segment increases. The exact estimation requires that the test statistic (Z_{ij} for CBS and Z_i for binary segmentation) is maximized when the segment edges are the true change-points. This happens when the segments are clearly separated, which is likely only when the two means are far apart.

A second set of simulations were performed with data sets simulated based on the CBS fit to Chromosome 11 of a real ROMA breast cancer experiment. There were 497 markers in Chromosome 11 with six change-points estimated at 137, 224, 241, 298, 307, 331 and the average log-ratios of intensities within segments given

by:

i	1 – 137	138 – 224	225 – 241	242 – 298	299 – 307	308 – 331	332 – 497
$f(i)$	−0.18	0.08	1.07	−0.53	0.16	−0.69	−0.16

As earlier the data were generated using the model $x_i = \mu_i + \epsilon_i$ where μ is the mean and ϵ the error distributed as $N(0, \sigma^2)$. In these simulations a local trend component was incorporated into the mean in order to study its effect on segmentation giving the mean to be $\mu_i = f(i) + 0.25\sigma \sin(c\pi i)$. The noise parameter σ was set to be one of 0.1 or 0.2, and the trend parameter c was set to be one of 0, 0.01 or 0.025 corresponding to no trend and local trends with long and short periods respectively. Figure 5 shows typical data sets constructed by this model. The CBS algorithm was used to detect change-points in the simulated data using a permutation reference distribution with a p -value cutoff of 0.01. The change-point search was undertaken either over all data points or over overlapping windows of size 100 that had 75% overlap. The change-points detected thus were pruned using a sum of squares threshold γ of 0.05. In addition to the number of change-points detected the following distance measure was used to assess the accuracy of the procedure. Let $\nu_1 < \dots < \nu_k$ be the true change-points and $\hat{\nu}_1 < \dots < \hat{\nu}_{\hat{k}}$ be the estimated change-points where $\hat{k}$ is the number of change-points detected. The distance measure D is defined as $\max_{1 \leq i \leq k} |\nu_i - \hat{\nu}_i|$ for a data set with true number of changes (*i.e.* $\hat{k} = k$) and undefined otherwise. The number of change-points detected and the median and range of D are shown in Table 3 for the unpruned procedure and Table 4 for the pruned procedure. These results are based on 100 replicate simulations.

Table 3 shows that unpruned CBS found the correct number of change-points

in at least 52% of the simulations with false positive detection more likely. The number of false positives appear to be nominal when there is no trend in the mean function and the longer the period of the local trend the larger the number of false positives. This is the expected behaviour since the longer the period of local trend the easier it is to approximate it by a non-constant step function. Larger noise in the data makes it more difficult to detect a change-point and estimate its location correctly. This can be seen in the numbers of change-points and the median and range of D used to assess the accuracy of the procedure. The windowing scheme was devised to ease the computational burden which grows as a square of the sample size. Windowing appeared to have little impact when the noise was low, but it led to a change-point being missed occasionally when noise was high. Similarly there is a small drop in the accuracy of the estimated locations of the change-points as seen in the medians and ranges of D . Thus the computational gains from windowing appears to come at the cost of a small drop in the accuracy of the procedure.

Table 4 shows that pruning greatly improves the accuracy of the procedure by substantially reducing the false positives caused by the local trends in the data. This can be seen in the larger number of times the correct number of changes are estimated with minimal detrimental effect on D . Pruning also occasionally removes a false positive change with the correct number of change-points leading to the appearance of a missed change-point. Overall, these simulations show that the CBS procedure with pruning is a desirable method to analyze copy number data.

6 Discussion

We have developed a variant of binary segmentation that we call circular binary segmentation or CBS for identifying genomic alterations in array copy number experiments. We applied our procedure to copy number data from aCGH experiments on 15 fibroblast cell lines. The CBS algorithm identified all the expected alterations detected through spectral karyotyping of the cell lines. We also applied our procedure to ROMA data from 23 breast cancer cell lines. While there is no biological verification for all the changes detected, the procedure found alterations in the ERB-B2 region of Chromosome 17 in 5 of the 23 cases which is consistent with the known rate of abnormality in breast cancer in this region. Finally we showed through a series of simulations that the procedure performs well in identifying changes and estimating their locations, especially when detecting narrow regions of change of the square wave type.

Even though the step-function model is appropriate for copy number data, we have seen fluctuations in the log intensity ratio that are not due to copy number changes. We call these local trends since these fluctuations exhibit a similar pattern across cell line case and believe that they may have a biological reason. These local trends can lead to false positive detection of change-points. We developed a pruning component to our procedure to address this problem and showed through simulations that it achieves its goal of removing most false-positive change-points. We are currently exploring methods to estimate local trends from data across cell lines so that it would be possible to subtract out local trends before segmenting.

The number of computations needed to obtain the test statistic used in CBS is a function of the square of the sample size. Since our test procedure is based on a permutation reference distribution, these computations must be repeated thousands of times to accurately estimate the upper tail probability of the reference distribution. We have developed a windowing method that reduces this computational burden. We showed here through simulations that it has minimal effect on the accuracy of the procedure and have applied it to data from arrays that contained 85,000 markers (Lucito et al., 2003). We are devising additional methods to further speed up the computations in order to analyze even larger data sets.

Once the change-points have been estimated it is of interest to estimate the copy numbers of the test sample in every region. One possibility that operates on the output of the CBS procedure is presented in Lucito et al. (2003). As demonstrated in Figure 4, the CBS procedure segments array copy number data into regions whose means are consistent across chromosomes. These plateaus in the plot of the segment means reflect different copy number states in the test sample. It is not possible to know the true copy numbers in each of these states without additional data acquired using another technique.

The software used in this paper was written in R and Fortran and is freely available at <http://www.mskcc.org/biostat/~olshena/research/>.

7 References

1. AUTIO, R., HAUTANIEMI, S., KAURANIEMI, P., YLI-HARAJA, O., ASTOLA, J., WOLF, M. AND KALIONIEMI, A. (2003). CGH-Plotter: MATLAB toolbox fo CGH-data analysis. *Bioinformatics* **19** 1714-1715.
2. BASSEVILLE, M. (1988). Detecting changes in signals and systems - a survey. *Automatica* **24** 309-326.
3. BENJAMINI, Y. AND HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57** 289-300.
4. BREIMAN, L., FRIEDMAN, J, OLSHEN, R. AND STONE, C. (1984). Classification and Regression Trees. Wadsworth.
5. CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74** 829-836.
6. HODGSON, G., HAGER, J.H., VOLIK, S., HARIONO, S., WERNICK, M., MOORE, D., NOWAK, N., ALBERTSON, D.G., PINKEL D., COLLINS, C., HANAHAN, D. AND GRAY J.W. (2001). Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nature Genetics* **29** 459-464.
7. KALLIONIEMI, A., KALLIONIEMI, O-P, SUDAR, D., RUTOVITZ, D., GRAY, J.W., WALDMAN, F. AND PINKEL D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258** 818-821.
8. LEVIN, B. AND KLINE, J. (1985). The CUSUM test of homogeneity with an application in spontaneous abortion epidemiology. *Statistics in Medicine*

4 469-488.

9. LINN, S.C., WEST, R.B., POLLACK, J.R., ZHU, S., HERNANDEZ-BOUSSARD, T., NIELSEN, T.O., RUBIN, B.P., PATEL, R., GOLDBLUM, J.R., SIEGMUND, D., BOTSTEIN, D., BROWN, P.O., GILKS, C.B., AND VAN DE RIJN, M. (2003) Gene expression patterns and gene copy number changes in dermatofibrosarcoma protuberans. *American Journal of Pathology* **163** 2383-95.
10. LISITSYN, N., LISITSYN, N. AND WIGLER, M. (1993). Cloning the differences between two complex genomes. *Science* **259** 946-951.
11. LUCITO, R., HEALY, J., ALEXANDER, J., REINER, A., ESPOSITO D, CHI, M., RODGERS, L., BRADY, A., SEBAT, J., TROGE, J., WEST, J.A., ROSTAN, S., NGUYEN, K.C., POWERS, S., YE, K.Q., OLSHEN, A., VENKATRAMAN, E., NORTON, L. AND WIGLER, M. (2003). Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Research* **13** 2291-2305.
12. LUCITO, R., WEST, J., REINER, A., ALEXANDER, D., ESPOSITO, D., MISHRA, B., POWERS, S., NORTON, L. AND WIGLER, M. (2000). Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. *Genome Research* **10** 1726-36.
13. MENARD, S., TAGLIABUE, E., CAMPIGLIO, M. AND PUPA, S.M. (2002). Role of HER2 gene overexpression in breast carcinoma. *Journal of Cellular Physiology* **182** 150-162.

14. PINKEL, D., SEAGRAVES, R., SUDAR, D., CLARK, S., POOLE, I., KOWBEL, D., COLLINS, C., KUO, W-L, CHEN, C., ZHAI, Y., ZHAI, Y, DAIRKEE, S., LJUNG, B-M, GRAY, J.W. AND ALBERTSON, D. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics* **20** 207-211.
15. POLLACK, J.R., PEROU, C.M., ALIZADEH, A.A., EISEN, M.B., PERGAMENSCHIKOV, A., WILLIAMS, C.F., JEFFREY, S.S., BOTSTEIN, D. and BROWN, P.O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics* **23** 41-46.
16. POLLACK, J.R., SORLIE, T., PEROU, C.M., REES, C.A., JEFFREY, S.S., LONNING, P.E., TIBSHIRANI, R., BOTSTEIN, D., BORRESEN-DALE, A.L. AND BROWN P.O. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors *Proceedings of the National Academy of Sciences USA* **99** 12963-12968.
17. SEN, A. AND SRIVASTAVA, M. S. (1975). On tests for detecting a change in mean. *Annals of Statistics* **3** 98-108.
18. SHABAN, S. A. (1980). Change-point problem and two phase regression: an annotated bibliography. *International Statistical Review* **48** 83-93.
19. SIEGMUND, D. (1986). Boundary crossing probabilities and statistical applications. *Annals of Statistics* **14** 361-404.
20. SNIJDERS, A. M., FRIDLYAND, J., MANS, D. A., SEGRAVES, R., JAIN, A.N., PINKEL, D., AND ALBERSTON D.G. (2003). Shaping of tumor and drug-resistant genomes by instability and selection. *Oncogene* **22** 4370-4379.

21. SNIJDERS, A. M., NOWAK, N., SEGRAVES, R., BLACKWOOD, S., BROWN, N., CONROY, J., HAMILTON, G., HINDLE, A.K., HUEY, B., KIMURA, K., LAW, S., MYAMBO, K., PALMER, J., YLSTRA, B., YUE, J.P., GRAY, J.W., JAIN, A.N., PINKEL, D. AND ALBERSTON D.G. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics* **29** 263-264.
22. VENKATRAMAN, E. S. (1992). Consistency results in multiple change-point situations. Technical report, Dept. of Statistics, Stanford Univ.
23. VOSTRIKOVA, L. J. (1981). Detecting "disorder" in multidimensional random processes. *Soviet Mathematics Doklady* **24** 55-59.
24. YANG, Y.H., DUDOIT, S., LUU, P., LIN. D., PENG, V., NGAI, J. AND SPEED, T.P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **30**(4):e15.
25. WEISS, M.M., SNIJDERS, A.M., KUIPERS, E.J., YLSTRA, B., PINKEL, D., MEUWISSEN, S.G.M., VAN DIEST, P.J., ALBERTSON, D.G. AND MEIJER, G.A. (2003). Determination of amplicon boundaries at 20q13.2 in tissue samples of human gastric adenocarcinomas by high-resolution microarray comparative genomic hybridization. *The Journal of Pathology* **200** 320-326.
26. YAO, Y-C. (1988) Estimating the number of change-points via Schwarz' Criterion. *Statistics and Probability Letters*. **6** 181-189.

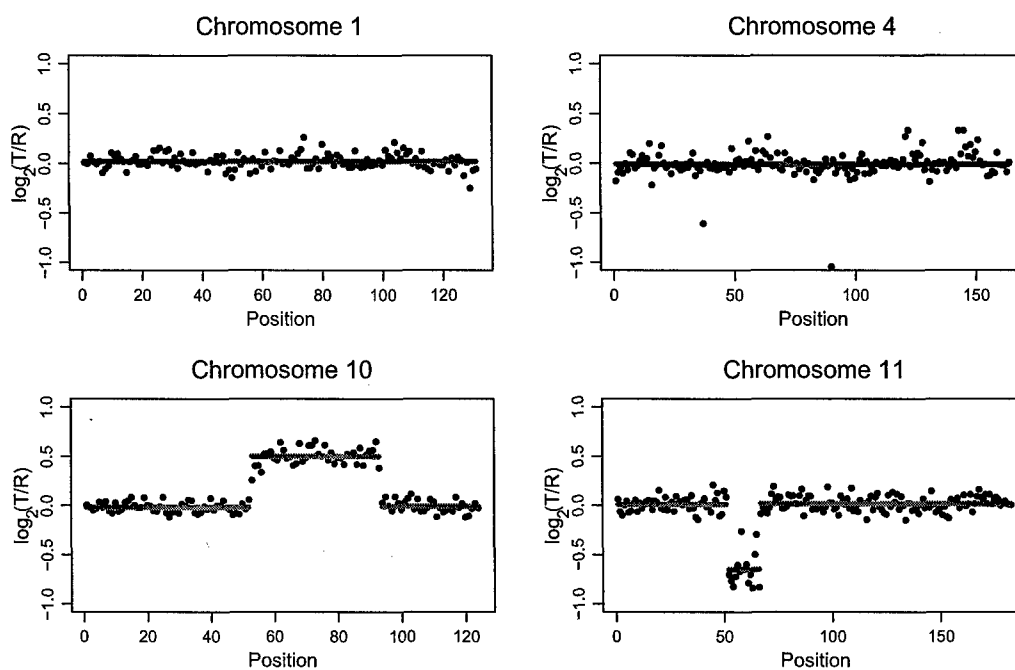


Figure 1: A CBS analysis of the fibroblast cell line GM05296 which has known alterations on chromosomes 10 and 11. The points are normalized log ratios, and the lines are the mean values among points in a segment. The red and dark green lines represent regions defined by CBS for chromosomes with and without known alterations, respectively.

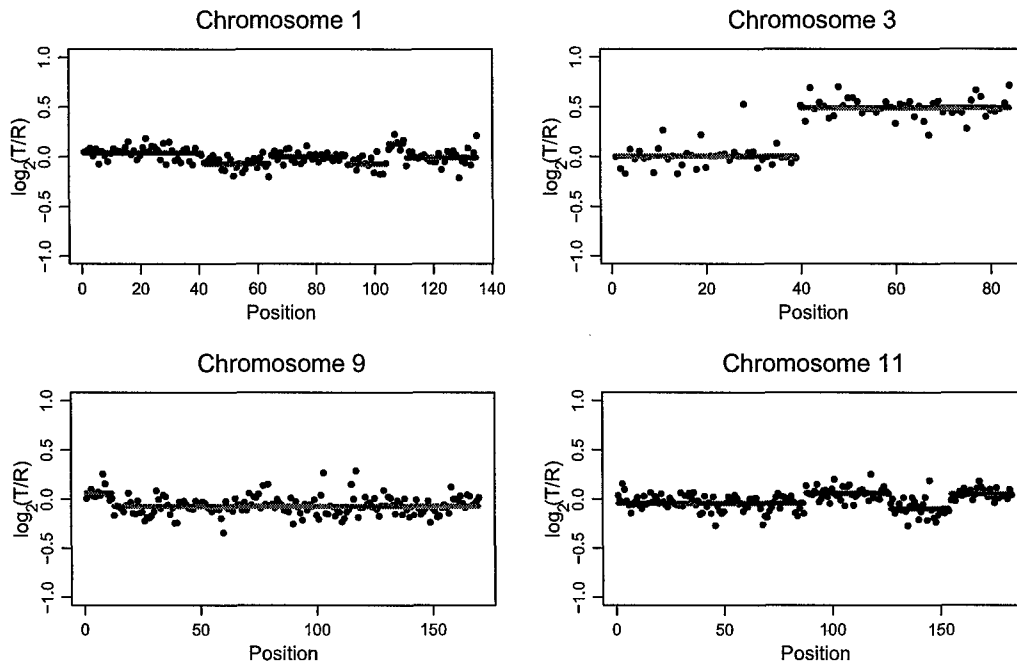


Figure 2: A CBS analysis of the fibroblast cell line GM03563 which has known alterations on chromosomes 3 and 9. The red and dark green lines represent regions defined by CBS with and without known alterations, respectively. Note that change-points found on chromosome 1 and chromosome 11 appear to be because of local trends in the data.

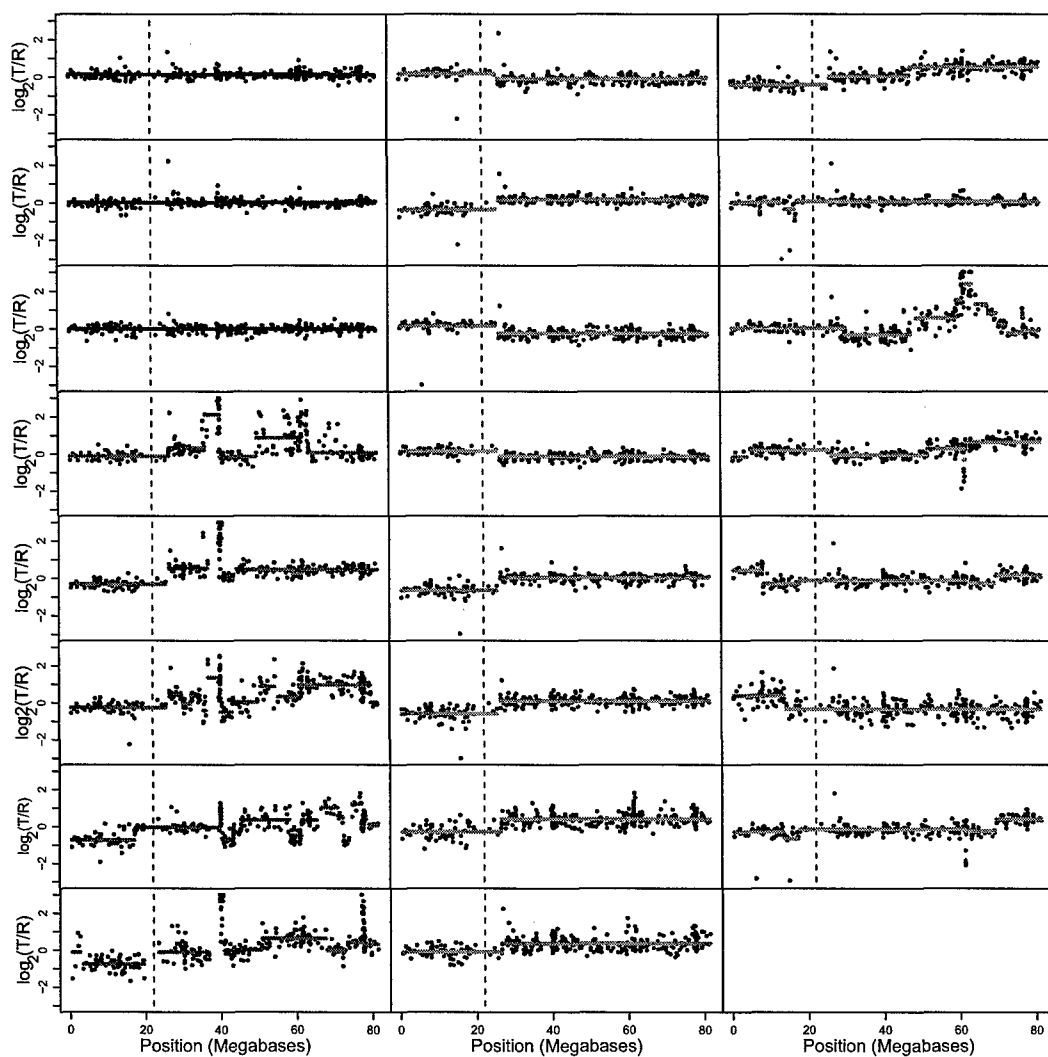


Figure 3: CBS applied to 23 breast cancer cell lines using the ROMA technology. Cell lines with no changes are dark green, those that appear to have ERB-B2 amplifications (near 40 MB) are red, those that appear to have whole-arm alteration are blue, and those with other changes are purple. The dashed line is at the centromere.

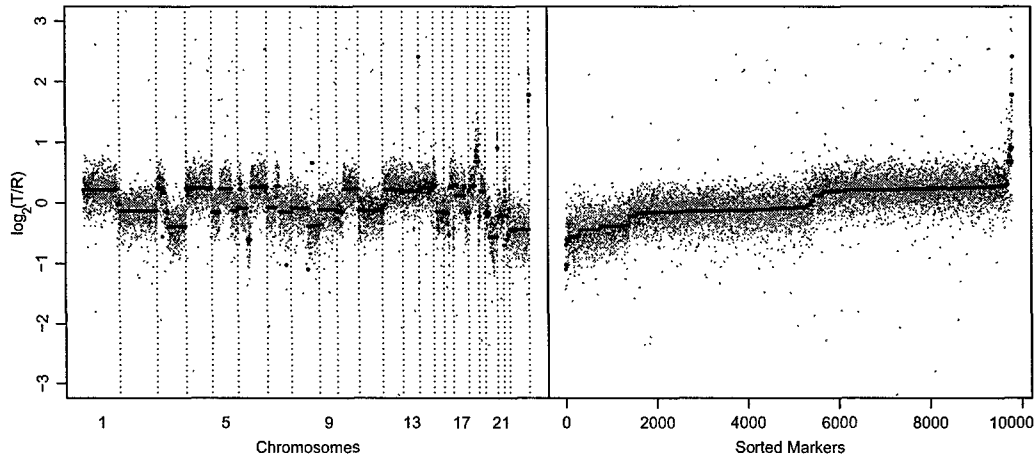


Figure 4: A CBS analysis of a whole breast cancer cell line. The left panel shows markers arranged by chromosome, while the right shows the markers sorted by the mean of the corresponding segment. The red points are the actual normalized log ratios, while the black lines are the segment means. Each plateau in the segment means implies a different copy number in the test sample.

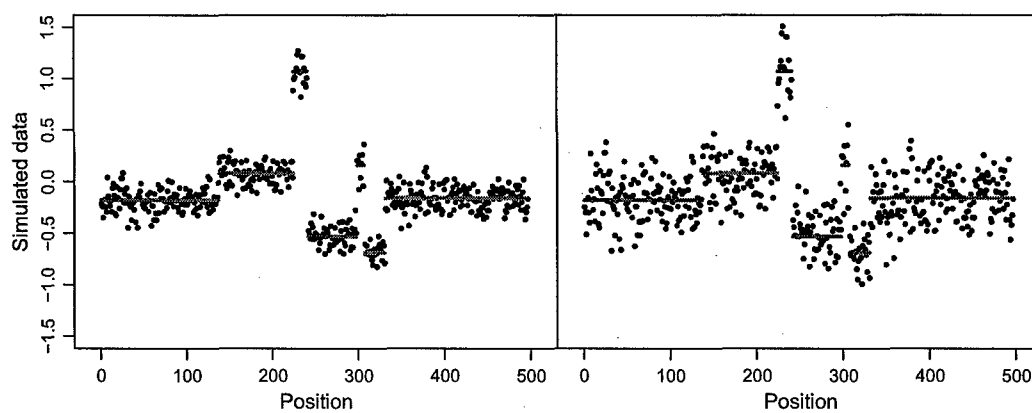


Figure 5: Example simulation data sets. The lines are the step function. The picture on the left is for the data set when $\sigma = 0.1$ and no local trend. The picture on the right is for the same data set except $\sigma = 0.2$ and the local trend has a long period.

Cell Line/Chrom.	$\alpha = 0.01$	$\alpha = 0.001$
GM03563/3	Yes	Yes
GM03563/9	No	No
GM03563/False	8	5
GM05296/10	Yes	Yes
GM05296/11	Yes	Yes
GM05296/False	3	0
GM01750/9	Yes	Yes
GM01750/14	Yes	Yes
GM01750/False	1	0
GM03134/8	Yes	Yes
GM03134/False	3	1
GM13330/1	Yes	Yes
GM13330/4	Yes	Yes
GM13330/False	8	5
GM01535/5	Yes	Yes
GM01535/12	No	No
GM01535/False	2	0
GM07081/7	Yes	Yes
GM07081/15	No*	No*
GM07081/False	1	0
GM13031/17	Yes	Yes
GM13031/False	5	3
GM01524/6	Yes	Yes
GM01524/False	6	2

Table 1: Results from applying CBS to nine cell lines with known copy number alterations. “Yes” means the alteration was found for the particular cell line and chromosome at the given α level, while “No” means that it was not. For GM07081/15, the asterisk is because there was no evidence in the array data of an alteration. “False” is the number of chromosomes for the cell line where change-points were found that do not have known alterations.

<i>k</i>	<i>c</i>	Change-points (edge)					Change-points (center)				
		0	1	2	3-4	# Exact	0	1	2	3-4	# Exact
2	2	980	11	8	1	5	968	0	32	0	9
		762	238	0	0	170	990	10	0	0	0
	3	832	159	4	5	128	821	0	175	4	174
		297	699	4	0	607	992	8	0	0	0
	4	430	556	5	9	518	405	0	583	12	516
		34	957	9	0	900	995	5	0	0	0
3	2	874	115	7	4	81	857	0	141	2	79
		539	458	3	0	294	992	8	0	0	0
	3	348	635	8	9	538	330	0	654	16	496
		76	914	10	0	754	994	6	0	0	0
	4	35	947	2	16	891	23	0	954	23	847
		1	989	10	0	925	995	5	0	0	0
4	2	720	261	15	4	192	689	0	307	4	159
		334	662	4	0	439	992	8	0	0	0
	3	115	863	10	12	716	97	0	883	20	648
		12	979	9	0	802	994	5	1	0	0
	4	3	977	5	15	918	0	0	978	22	867
		0	990	10	0	931	996	3	1	0	0
5	2	531	439	23	7	297	511	0	481	8	232
		192	801	7	0	516	991	7	2	0	0
	3	24	954	6	15	818	19	0	961	20	692
		1	988	11	0	842	994	5	1	0	0
	4	0	982	4	14	937	0	0	981	19	877
		0	989	11	0	943	997	1	1	1	1

Table 2: Counts of the number of change-points observed when applying CBS and binary segmentation to 1000 data sets of 250 points simulated from the Gaussian distribution. The columns under the heading “Exact” provide the number of cases in which the exact number (1 for edge and 2 for center) and locations of the change-points are observed. Here k is the width of the changed segment and c is the number of standard deviations between the two means. The large font corresponds to the CBS results and the small font corresponds to the binary segmentation result. Each data set had one elevated region ranging from 2 – 5 points, and the elevated region varied from 2 – 4 SDs above the mean. The elevated region was all the way to one edge of the data set or at the exact center of the data set. The α for the simulation was 0.01.

σ	Window	Trend	Number of Change-points							Max. Distance	
			5	6	7	8	9	10	11	Median	Range
0.1	No	None	0	96	1	2	1	0	0	0	0-4
0.1	Yes	None	0	92	5	3	0	0	0	0	0-5
0.2	No	None	0	91	6	2	1	0	0	1	0-11
0.2	Yes	None	9	78	10	3	0	0	0	1	0-28
0.1	No	Short	0	82	5	10	2	1	0	0	0-6
0.1	Yes	Short	0	73	8	15	2	2	0	0	0-6
0.2	No	Short	0	66	23	8	3	0	0	1	0-23
0.2	Yes	Short	7	58	17	11	4	3	0	2	0-28
0.1	No	Long	0	71	7	18	4	0	0	0	0-5
0.1	Yes	Long	0	75	12	9	2	2	0	0	0-6
0.2	No	Long	0	68	11	17	2	1	1	1	0-28
0.2	Yes	Long	18	52	20	6	2	2	0	2	0-46

Table 3: Counts of the number of change-points observed when applying CBS to data simulated from the step-function f from Section 5. If “Window” is “No”, a search over all points was undertaken to find the best change-point. If “Window” is “Yes”, the search was only over overlapping window (window size=100; overlap=75%). If “Trend” is “Long”, $c = 0.01$. If Trend is “Short”, $c = 0.025$. The true number of change-points was six. Distances are the maximums of the minimum distances from the i th observed change-point to the i th true change-point in every simulation. The median and range was then computed over 100 simulations.

σ	Window	Trend	Number of Change-points							Max. Distance	
			5	6	7	8	9	10	11	Median	Range
0.1	No	None	0	99	1	0	0	0	0	0	0-4
0.1	Yes	None	0	100	0	0	0	0	0	0	0-5
0.2	No	None	0	99	1	0	0	0	0	1	0-11
0.2	Yes	None	9	91	0	0	0	0	0	1	0-28
0.1	No	Short	0	94	5	1	0	0	0	0	0-6
0.1	Yes	Short	0	94	4	2	0	0	0	0	0-6
0.2	No	Short	0	95	5	0	0	0	0	1	0-23
0.2	Yes	Short	8	84	6	2	0	0	0	1	0-31
0.1	No	Long	0	91	9	0	0	0	0	0	0-5
0.1	Yes	Long	0	94	4	1	1	0	0	0	0-6
0.2	No	Long	0	90	8	1	1	0	0	2	0-40
0.2	Yes	Long	20	75	4	1	0	0	0	1	0-46

Table 4: Same as Table 3, except that the change-points have been pruned.

vaccine was both as safe and as immunogenic as the commercial control vaccine (Fig. 3).

The above results set the stage for more detailed clinical assessment of the vaccine in a targeted infant population. Thus, a phase I trial was initiated with 139 2-month-old infants who received three vaccine doses scheduled at 2, 4, and 6 months, as recommended for other conjugate anti-Hib vaccines. The test vaccine induced a strong and bactericidal antibody response against Hib in infants (Fig. 4) that fell to values ranging from 5 to 7 $\mu\text{g/mL}$ at 18 months of age but remained at least five times that required for long-term protection (Fig. 4A). A booster dose with sPRP-TT applied to all groups increased the antibody against Hib titers by 10-fold. Thus, the capacity of sPRP-TT to prime an effective immune response against Hib was demonstrated.

In a second phase II trial, a total of 1141 infants distributed in three groups received three doses of either sPRP-TT conjugate, sPRP-TT mixed with aluminum phosphate, or the control vaccine (Vaxem-Hib). Of the test infants, 99.7% reached antibody titers above 1 $\mu\text{g/mL}$, which is considered appropriate for long-lived protection against Hib (28, 29). The mean IgG anti-PRP titer was 27.4 $\mu\text{g/mL}$ for all infants vaccinated with the sPRP-TT, which is consistent with previously reported clinical trials (between 7.67 and 35 $\mu\text{g/mL}$) for anti-Hib vaccines without adjuvant (30, 31).

The present study demonstrates that a synthetic capsular polysaccharide antigen can be produced on a large scale under GMP conditions and used to manufacture an effective vaccine for human use. The resulting conjugate vaccine incorporating a synthetic bacterial carbohydrate antigen was demonstrated to be as safe and immunogenic in humans as already-licensed vaccines incorporating the native polysaccharide (32–34). Access to synthetic complex carbohydrate-based vaccines is therefore feasible and provides an alternative strategy in the fight against Hib infections. It also sets the stage for further development of similar approaches against other human pathogens.

References and Notes

1. J. B. Robbins, R. Schneerson, P. Anderson, D. H. Schmidt, *JAMA* **276**, 1181 (1996).
2. H. Peltola, *Clin. Microbiol. Rev.* **13**, 302 (2000).
3. P. Anderson, G. Peter, R. B. Johnston Jr., L. H. Wetterlow, D. H. Smith, *J. Clin. Invest.* **5**, 39 (1972).
4. L. P. Rodrigues, R. Schneerson, J. B. Robbins, *J. Immunol.* **107**, 1071 (1971).
5. H. Peltola, H. Kayty, M. Virtanen, P. H. Makela, *N. Engl. J. Med.* **310**, 1561 (1984).
6. R. Schneerson, O. Barrera, A. Sutton, J. B. Robbins, *J. Exp. Med.* **152**, 361 (1980).
7. S. L. Cochi, D. O'Mara, S. R. Preblud, *Pediatrics* **81**, 166 (1988).
8. M. E. Ramsey, N. Andrews, E. B. Kaczmarski, E. Miller, *Lancet* **357**, 195 (2001).
9. S. Black et al., *Pediatr. Infect. Dis. J.* **19**, 187 (2000).
10. H. J. Jennings, R. A. Pon, in *Polysaccharides in Medicinal Applications*, S. Dimitriu, Ed. (Marcel Dekker, New York, 1996), p. 443.
11. J. B. Robbins, R. Schneerson, S. C. Szu, V. Pozsgay, in *Vaccinia, Vaccination and Vaccinology: Jenner, Pas-*

12. P. Constantino et al., *Vaccine* **17**, 1251 (1999).
13. P. Hoogerhout et al., *Tetrahedron Lett.* **28**, 1553 (1987).
14. A. A. Kandil, N. Chan, P. Chong, M. Klein, *Synlett* **7**, 555 (1992).
15. S. Nilsson, M. Bengtsson, T. Norberg, *J. Carbohydr. Chem.* **11**, 265 (1992).
16. C. C. Peeters et al., *Infect. Immun.* **60**, 1826 (1992).
17. P. Chong et al., *Infect. Immun.* **65**, 4918 (1997).
18. I. Chiu-Machado, J. C. Castro-Palomino, O. Madrazo-Alonso, C. Lopetegui-Palacios, V. Verez-Bencomo, *J. Carbohydr. Chem.* **14**, 551 (1995).
19. A. V. Nikolaev, J. A. Chudek, M. A. J. Fergusson, *Carbohydr. Res.* **272**, 179 (1995).
20. V. Fernandez Santana, R. Gonzalez Lio, J. Sarracent Perez, V. Verez-Bencomo, *Glycoconjugate J.* **15**, 549 (1998).
21. D. C. Phipps et al., *J. Immunol. Methods* **135**, 121 (1990).
22. V. Fernandez-Santana et al., unpublished observations.
23. ELISA was performed according to (21) with the use of Hib reference serum pool for calibration, Center for Biologics Evaluator and Research, Food and Drug Administration.
24. P. Farrington, E. Miller, in *Vaccine Protocols*, vol. 87 of *Methods in Molecular Medicine*, A. Robinson, M. J. Hudson, M. P. Cranage, Eds. (Humana, Totowa, NJ, 2003), p. 335.
25. Materials and methods are available on Science Online.
26. Y. Schlesinger, D. M. Granoff, *JAMA* **267**, 1489 (1992).
27. S. Romero-Steiner et al., *Clin. Diagn. Lab. Immunol.* **8**, 1115 (2001).
28. H. Kayty, H. Peltola, V. Karamko, P. H. Makela, *J. Infect. Dis.* **147**, 1100 (1983).
29. J. Amir, X. Liang, D. M. Granoff, *Pediatr. Res.* **27**, 358 (1990).
30. S. Holmes et al., *Am. J. Dis. Child.* **147**, 832 (1993).
31. D. M. Granoff, S. J. Holmes, *Vaccine* **9** (suppl.), S30 (1991).
32. P. Anderson, *Infect. Immun.* **39**, 233 (1983).
33. C. Chu, R. Schneerson, J. B. Robbins, S. C. Rastorgi, *Infect. Immun.* **40**, 245 (1983).
34. S. Marburg et al., *J. Am. Chem. Soc.* **108**, 5282 (1986).
35. We would like to thank the World Health Organization, the Pan-American Health Organization, and the following Cuban institutions: State Council, Ministers of Science Technology and Environment, and Ministry of Health. We are particularly thankful to J. M. Miyar, M. C. Santana, L. Yañez, J. L. DiFabio, M. Beurret, and C. Jones for their fundamental contributions and the many laboratory assistants, medical doctors, and nurses that were involved in the project over the years. C. H. Fox critically reviewed the manuscript.

Supporting Online Material

www.sciencemag.org/cgi/content/full/305/5683/522/DC1

Materials and Methods

Fig. S1

Table S1

31 December 2003; accepted 23 June 2004

Large-Scale Copy Number Polymorphism in the Human Genome

Jonathan Sebat,¹ B. Lakshmi,¹ Jennifer Troge,¹ Joan Alexander,¹ Janet Young,² Pär Lundin,³ Susanne Månér,³ Hillary Massa,² Megan Walker,² Maoyen Chi,¹ Nicholas Navin,¹ Robert Lucito,¹ John Healy,¹ James Hicks,¹ Kenny Ye,⁴ Andrew Reiner,¹ T. Conrad Gilliam,⁵ Barbara Trask,² Nick Patterson,⁶ Anders Zetterberg,³ Michael Wigler^{1*}

The extent to which large duplications and deletions contribute to human genetic variation and diversity is unknown. Here, we show that large-scale copy number polymorphisms (CNPs) (about 100 kilobases and greater) contribute substantially to genomic variation between normal humans. Representational oligonucleotide microarray analysis of 20 individuals revealed a total of 221 copy number differences representing 76 unique CNPs. On average, individuals differed by 11 CNPs, and the average length of a CNP interval was 465 kilobases. We observed copy number variation of 70 different genes within CNP intervals, including genes involved in neurological function, regulation of cell growth, regulation of metabolism, and several genes known to be associated with disease.

Many of the genetic differences between humans and other primates are a result of large duplications and deletions (1–3). From these observations, it is reasonable to expect that differences in gene copy number could be a significant source of genetic variation between humans. A few examples of large duplication polymorphisms have been reported (4). However, because of previous limitations in the power to determine DNA copy number at high resolution throughout the genome, the extent to which copy number polymorphisms (CNPs) contribute to human genetic diversity is unknown.

In our previous studies of human cancer with the use of representational oligonucleotide microarray analysis (ROMA), we have detected many genomic amplifications and deletions in tumor genomes when analyzed in comparison to an unrelated normal genome (5), but some of these genetic differences could be due to germline CNPs. To correctly interpret genomic data relating to cancer and other diseases, we must distinguish abnormal genetic lesions from normal CNPs.

We used ROMA to investigate the extent of copy number variation between normal

REPORTS

individuals. ROMA measures the relative concentration of DNA in two samples by hybridizing differentially labeled samples to a set of probes. Briefly, the complexity of the samples is reduced by making Bgl II genomic representations, consisting of small (200 to 1200 base pair) Bgl II restriction fragments amplified by adaptor-mediated polymerase chain reaction of genomic DNA (6). Oligonucleotide microarray probes are designed in silico from the human genome sequence assembly to be complementary with these fragments and are further optimized by performance (7). Microarrays are used to analyze genomic representations of unrelated individuals. Hybridization data are analyzed with a hidden Markov model (HMM) that is designed to distinguish differences between the DNA copy number and other variation in probe ratios, which can result from experimental noise or sequence polymorphisms at the restriction endonuclease sites used to make the representations (8).

Observed differences in the copy number of genome segments between samples from two individuals could reflect germline differences or somatic variation. Therefore, we sampled multiple tissues and Epstein-Barr virus-immortalized lymphoblastoid cell lines (LCLs) from a subset of the donors in this study (8), and by comparing the variants detected in the same donor, we determined that somatic mutations occurring in whole blood and LCLs were located exclusively within gene clusters encoding T cell receptors or immunoglobulins (fig. S1 and table S2), which most likely reflects normal V(D)J-type recombination of T cells and B cells, respectively. Therefore, the use of blood and LCLs as sources of genetic material for this study was not problematic.

In experiments with Bgl II representations, we identified 210 differences in 20 donors (excluding somatic differences, Fig. 1). For the sake of simplicity, overlapping CNPs from different experiments were assumed to represent the same polymorphism even if they did not overlap perfectly. Based on these criteria, we identified a nonredundant set of 71 CNPs (table S1).

Nine of twelve CNPs were unambiguously confirmed by cytogenetic analysis (Fig. 2 and fig. S2). Five CNPs were found to be hemizygous deletions, and four were dupli-

cations. Figure 2 presents array data and fluorescence in situ hybridization (FISH) confirmation for CNPs 15, 21, 32, and 56, which encompass the full length of genes *RAB6C*, *NT_016297.17*, *DUSP22*, and *PPYR1*, respectively. By interphase FISH, we confirmed a deletion of *RAB6C* (Fig. 2B), a duplication of *PPYR1* (Fig. 2D), and a deletion of *NT_016297.17* (Fig. 2F). By metaphase FISH, CNP32 was determined to involve an interchromosomal duplication of a region containing the *DUSP22* gene on 6p25 and 16p11.2 (Fig. 2, G, H, and I). FISH results were inconclusive for CNPs 68, 69, and 73. In these cases, FISH signals were too numerous, and a consensus copy number could not be reached. CNPs 68 and 69 were validated by other means (table S2); thus, 11 of 12 CNPs were validated by one of two methods, which is consistent with a false positive rate of about 10%.

Additional validation of CNPs was obtained by microarray analysis of genomic representations made with a different restriction enzyme. A pair of individuals analyzed by Bgl II-ROMA (experiment JA437, table S1) was also analyzed with Hind III representations and arrays of Hind III probes (JT393). The results of Bgl II-ROMA and Hind III-ROMA were generally in agreement (8). In addition, because of differences in the genomic distribution of Hind III probes,

some unique CNPs were identified, bringing the total of copy number differences identified in this study to 221 and the total of unique CNPs to 76.

Our study population consisted of 20 individuals from a variety of geographic backgrounds. These results provide an indication of the extent of human copy number variation and the frequency of the most common alleles. In all experiments, there were a total of 221 observed copy number differences (not including somatic differences) comprising a nonredundant set of at least 76 CNPs (Fig. 1 and table S2). There was an average of 11 CNPs between two individuals, with an average length of 465 kb and a median length of 222 kb. At least five of these polymorphisms have been described previously (9–13). The overwhelming majority of CNPs were previously unidentified. About half of the above CNPs were recurrent in multiple individuals.

The CNPs observed here represent only a subset of the total CNPs in the population. For example, some CNPs that have previously been reported were not observed in this study (14, 15). Undoubtedly, an increase in the size of our study population would reveal additional CNPs, as would an increase in the density of probe coverage. By comparing Hind III and Bgl II results and analyzing Bgl II results with replicate samples, we estimate that in any given experiment we may miss up

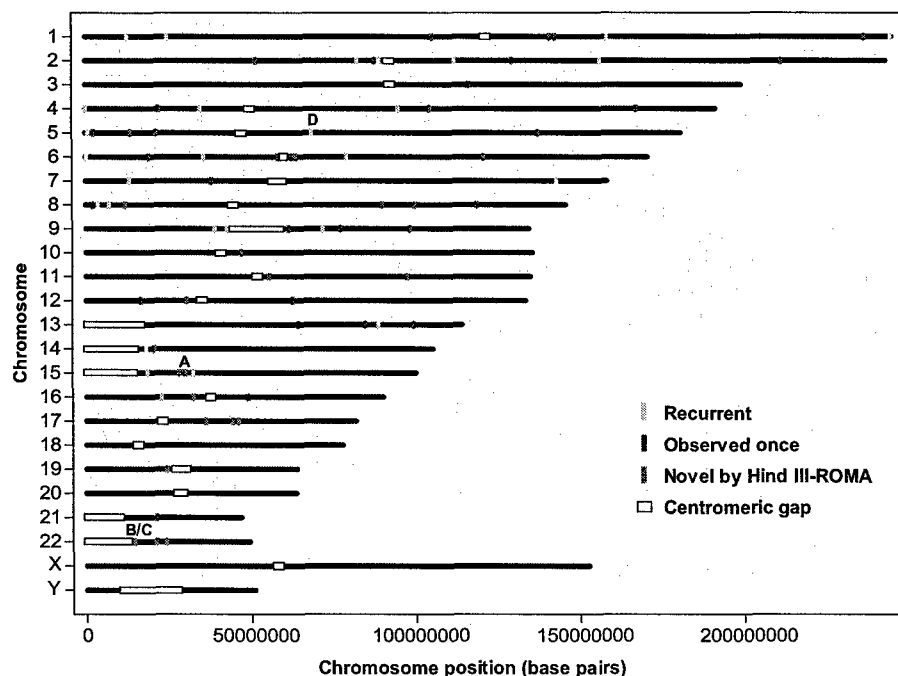


Fig. 1. Genome-wide map of CNPs identified by ROMA. The position of all CNPs (excluding somatic differences) is shown. CNPs identified in multiple individuals (by Bgl II-ROMA) are indicated in yellow, and CNPs observed in only one individual are indicated in red. Additional CNPs identified by one Hind III-ROMA experiment are indicated in blue. Symbols denoting CNPs are not drawn to scale. Genome assembly gaps in pericentromeric and satellite regions are indicated by gray boxes. Genomic regions where recurring de novo rearrangements cause the developmental disorders Prader-Willi and Angelman syndromes, cat eye syndrome, DiGeorge/velocardiofacial syndrome, and spinal muscular atrophy are labeled A, B, C, and D, respectively.

¹Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA. ²Division of Human Biology, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA. ³Karolinska Institute, Stockholm SE-17176, Sweden. ⁴Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794, USA. ⁵Columbia Genome Center, Columbia University, New York, NY 10032, USA. ⁶Broad Institute, Cambridge, MA 02139, USA.

*To whom correspondence should be addressed. E-mail: wigler@cshl.edu

to 30% of the large-scale copy number changes that we ought to find (table S3). In addition, there are theoretical limits to the detection of CNPs with only 85,000 probes. Based on Poisson distributions of probes and the probabilities of detecting CNPs of given lengths, we estimate that there are 226 non-redundant CNPs in our study population covering 44 Mb of the genome (table S4).

CNPs were widely distributed throughout the genome. Some locations such as 6cen, 8pter, and 15q13-14 contained clusters of three to four CNPs, which may be evidence that these regions are "hotspots" of copy number variation. We observed no CNPs on the X chromosome. This may be due to the underrepresentation of females in our study population (16 donors and SKN1 were male). A larger study would be necessary to determine if selective pressure against copy number variation is greater on the X chromosome than on autosomes, or if it is especially apparent in the X chromosomes present in males.

CNPs were frequently located near other

types of chromosomal rearrangements. Some CNPs occurred within genomic regions where recurring de novo rearrangements are causes of developmental disorders, specifically, Prader-Willi and Angelman syndromes, cat eye syndrome, DiGeorge/velocardiofacial syndrome, and spinal muscular atrophy (labeled A, B, C, and D, respectively, in Fig. 1). These CNPs are not directly implicated in the above diseases, but they may reflect the instability of these genomic regions. A preliminary analysis of the duplication content of CNPs determined that 30% of the sequence within intervals of polymorphic deletions consists of segmental duplications, a sixfold enrichment relative to the genome average. As would be expected, a greater enrichment (12-fold) was observed for polymorphic duplications (16). The former is consistent with previous observations of a positive correlation between segmental duplications and microdeletions (17, 18). A more thorough characterization of CNP junctions at the sequence level is necessary to determine a causal relationship between the two. Fixed segmental duplications, unstable regions, and

CNPs are probably manifestations of the same underlying process. Just as chromosomal rearrangements have played a significant role in primate evolution and human disease, structural polymorphisms may play an analogous role in determining genetic diversity within the human population.

We observed copy number variation of 70 genes (table S5). Variation in the dosage of individual genes can lead to a profound phenotype; for instance, the familial inheritance of gene copy number variants is a cause of some neurological disorders (19, 20). Notably, one of the donors in this study was determined to carry a deletion of *COH1* (CNP48), a gene whose inactivation causes the autosomal recessive disease Cohen syndrome (21). Several additional CNPs contained genes involved in neurodevelopment, such as *GTF2H2*, *ATOH1*, *CASPR3*, *CHRFAM7A*, and *NCAM2*. Other compelling examples from table S5 include the Enhancer of Split (*TLE1*) and *RAB6C*, which are implicated in leukemia and drug resistance in breast cancer, respectively (22, 23). Lastly, some CNPs identified in this study involve genes with a known influence on "normal" human phenotypes. For example, we observed triplication of the neuropeptide-Y4 receptor (*PPYR1*, Fig. 2, C and D), a gene that is directly involved in the regulation of food intake and body weight (24). Thus, a relationship between CNPs and susceptibility to health problems such as neurological disease, cancer, and obesity is an intriguing possibility.

Owing to their size and gene content, CNPs are unlikely to be selectively neutral. Indeed, a large proportion of CNPs observed in this study are rare (i.e., they occur once in 20 donors). A preliminary analysis of the comparative frequency of variants (25) suggests that CNP as a class is under negative selection. However, more data are required to reach this conclusion with confidence.

As evident by ROMA, there is considerable structural variation in the human genome, most of which was not previously apparent by other methods of genomic analysis. Previous studies using array comparative genomic hybridization have identified a handful of large-scale polymorphisms (26, 27). For example, by using a 1-Mb-resolution bacterial artificial chromosome (BAC) array, Shaw-Smith *et al.* detected five inherited CNPs from a set of 50 patients with developmental disabilities (27). The ROMA chips used here have a resolution of approximately one probe every 35 kb, which accounts for much of the enhanced sensitivity of our method. Furthermore, by designing oligonucleotide probes that are free of repetitive sequence, by empirically selecting 85,000 probes that yield maximum signal, and by reducing the complexity of the genome, ROMA achieves a ratio of signal-to-background superior to that which can be

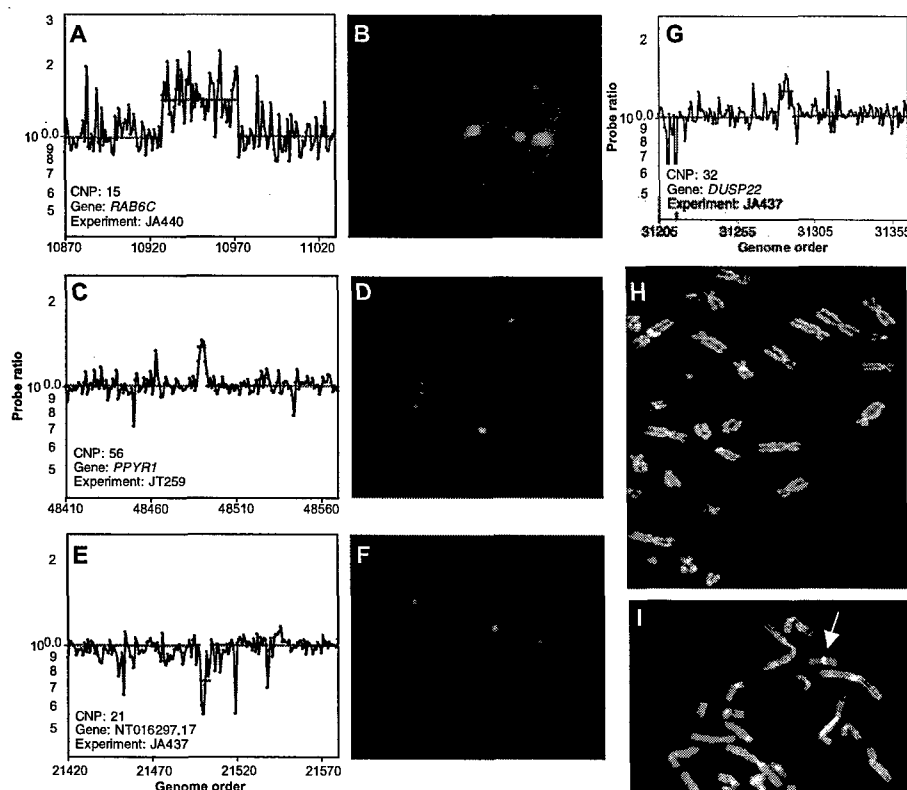


Fig. 2. Validation of ROMA results by FISH. (A), (C), (E), and (G) show CNPs identified by ROMA and include the CNP identification number, the name of one gene located entirely within the interval, and the experiment name. (B), (D), (F), (H), and (I) show cytogenetic analyses of one or both individuals with probes that target the same CNP intervals. In all panels, the polymorphic probe is labeled red. In interphase cells [(B), (D), and (F)], a control probe (labeled green) was also included to confirm that cells were diploid. (B) CNP15 probe in GM11322 cells; (D) CNP56 probe in GM10470 cells; (F) CNP21 probe in GM10470 cells; (H) CNP32 probe in SKN1 cells; (I) CNP32 probe in SKN1 cells. In (I), one parental copy of chromosome 16 in SKN1 lacks the duplication (arrow).

attained by hybridization of total genomic DNA to an array of BACs. Thus, ROMA has additional advantages even compared with arrays with "complete" coverage of the genome, such as the 32,000-probe tiling-path BAC array (28). Further developments of ROMA are under way, including a 380,000-probe microarray, which promise to reveal a great deal more about large-scale polymorphism in the human genome.

References and Notes

1. D. P. Locke et al., *Genome Res.* **13**, 347 (2003).
2. K. A. Frazer et al., *Genome Res.* **13**, 341 (2003).
3. G. Liu et al., *Genome Res.* **13**, 358 (2003).
4. P. R. Buckland, *Ann. Med.* **35**, 308 (2003).
5. R. Lucito et al., *Genome Res.* **13**, 2291 (2003).
6. N. Lisitsyn, M. Wigler, *Science* **259**, 946 (1993).
7. J. Healy, E. E. Thomas, J. T. Schwartz, M. Wigler, *Genome Res.* **13**, 2306 (2003).
8. Materials and methods are available as supporting material on *Science* Online.
9. T. Blunt, F. Steers, G. Daniels, B. Carritt, *Ann. Hum. Genet.* **58**, 19 (1994).
10. F. Gilles, A. Goy, Y. Remache, K. Manova, A. D. Zeleznitz, *Genomics* **70**, 364 (2000).
11. E. J. Hollox, J. A. Armour, J. C. Barber, *Am. J. Hum. Genet.* **73**, 591 (2003).
12. J. K. Kulski, T. Shiina, T. Anzai, S. Kohara, H. Inoko, *Immunol. Rev.* **190**, 95 (2002).
13. B. Riley, M. Williamson, D. Collier, H. Wilkie, A. Makoff, *Genomics* **79**, 197 (2002).
14. J. R. Townson, L. F. Barcellos, R. J. Nibbs, *Eur. J. Immunol.* **32**, 3016 (2002).
15. R. A. McLellan, M. Oscarson, J. Seidegard, D. A. Evans, M. Ingelman-Sundberg, *Pharmacogenetics* **7**, 187 (1997).
16. Copy number gains are, by definition, recent segmental duplications. Not surprisingly, many of the duplications observed in this study overlap with regions that have been previously annotated as segmental duplications.
17. J. A. Bailey et al., *Science* **297**, 1003 (2002).
18. C. J. Shaw, J. R. Lupski, *Hum. Mol. Genet.* **13** Spec, R57 (2004).
19. J. R. Lupski et al., *Cell* **66**, 219 (1991).
20. P. F. Chance et al., *Cell* **72**, 143 (1993).
21. J. Kolehmainen et al., *Am. J. Hum. Genet.* **72**, 1359 (2003).
22. J. Shan et al., *Gene* **257**, 67 (2000).
23. Y. Imai et al., *Biochem. Biophys. Res. Commun.* **252**, 582 (1998).
24. A. Sainsbury et al., *Genes Dev.* **16**, 1077 (2002).
25. N. Patterson, data not shown.
26. D. P. Locke et al., *J. Med. Genet.* **41**, 175 (2004).
27. C. Shaw-Smith et al., *J. Med. Genet.* **41**, 241 (2004).
28. A. S. Ishkanian et al., *Nature Genet.* **36**, 299 (2004).
29. We thank E. Nuwaysir and T. Richmond (of Nimble-Gen Systems Inc.) and E. Thorolfsson and K. Olafsdottir (of LindGen Einkahlutafélag) for providing technical support; and E. Linardopoulou, L. Iakoucheva, and J. Simons for helpful discussions. Supported by grants to M.W. from NIH and the National Cancer Institute (CA81674; CA078544; HG02606), New York University/Defense Advanced Research Projects Agency (DARPA) F5239, Tularik Inc., 1 in 9: The Long Island Breast Cancer Action Coalition, Lillian Goldman and the Breast Cancer Research Foundation, Starr Foundation, Marks Family Foundation, The Miracle Foundation, and The Simons Foundation. M.W. is an American Cancer Society Research Professor. B.T. was supported by NIH grants (GM057070; DC004209). J.S. was supported by the NIH Training Grant (5T32; CA069311-25). A.Z. was supported by grants from the Swedish Cancer Society and Cancerföreningen, Stockholm.

Supporting Online Material

www.sciencemag.org/cgi/content/full/305/5683/525/DC1
Materials and Methods
SOM Text
Figs. S1 and S2
Tables S1 to S5

8 April 2004; accepted 24 June 2004

Integrase Inhibitors and Cellular Immunity Suppress Retroviral Replication in Rhesus Macaques

Daria J. Hazuda,^{1*} Steven D. Young,^{2*} James P. Guare,² Neville J. Anthony,² Robert P. Gomez,² John S. Wai,² Joseph P. Vacca,² Larry Handt,³ Sherri L. Motzel,³ Hilton J. Klein,³ Geethanjali Dornadula,¹ Robert M. Danovich,¹ Marc V. Witmer,¹ Keith A. A. Wilson,⁴ Lynda Tussey,⁴ William A. Schleif,⁴ Lori S. Gabryelski,⁴ Lixia Jin,⁵ Michael D. Miller,¹ Danilo R. Casimiro,⁴ Emilio A. Emmini,⁴ John W. Shiver⁴

We describe the efficacy of L-870812, an inhibitor of HIV-1 and SIV integrase, in rhesus macaques infected with the simian-human immunodeficiency virus (SHIV) 89.6P. When initiated before CD4 cell depletion, L-870812 therapy mediated a sustained suppression of viremia, preserving CD4 levels and permitting the induction of virus-specific cellular immunity. L-870812 was also active in chronic infection; however, the magnitude and durability of the effect varied in conjunction with the pretreatment immune response and viral load. These studies demonstrate integrase inhibitor activity in vivo and suggest that cellular immunity facilitates chemotherapeutic efficacy in retroviral infections.

The substantial incidence of resistance observed in therapy-experienced patients and newly acquired HIV-1 infections (1–5) underscores the need for new antiretroviral agents, as well as the importance of maximizing the du-

rability of available therapies. All oral agents licensed to treat HIV-1 disease target two of the three essential, virally encoded enzymes, reverse transcriptase and protease (6–8). The third HIV-1 enzyme, integrase, inserts the viral DNA into the cellular genome through a multistep process that includes two catalytic reactions: 3' endonucleolytic processing of the viral DNA ends and strand transfer or joining of the viral and cellular DNAs (9, 10). Compounds that selectively inhibit strand transfer have provided proof of concept for integrase as a chemotherapeutic target for HIV-1 infection in vitro (11).

In this investigation we used a novel strand-transfer inhibitor, L-870812 (12) (Fig. 1), which exhibits potent antiviral activity in vitro against both HIV-1 and the simian lentivirus, SIV [95% inhibition concentration (IC₉₅) of 250 and 350 nM, respectively, in 50% human and rhesus serum] and favorable pharmacokinetics in rhesus macaques [oral bioavailability = 64% and half-time (t_{1/2}) = 5 hours] to assess the efficacy of such inhibitors in vivo. The studies were designed to evaluate integrase inhibitors as a new class of antiretroviral agents and to examine the role of viral-specific cellular immunity in chemotherapeutic intervention using SHIV 89.6P-infected rhesus macaques as an experimental model of early- and late-stage retroviral infection.

Rhesus macaques infected with SHIV 89.6P exhibit an atypical, accelerated disease marked by a profound depletion of CD4 cells concomitant with progression from acute viremia to a chronic phase at about 2 weeks after infection

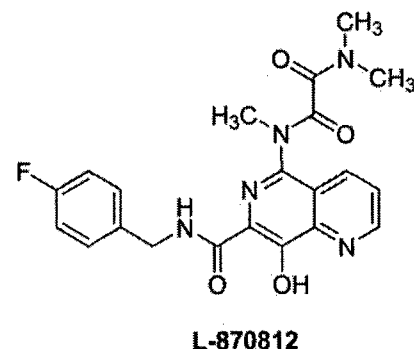


Fig. 1. The structure of L-870812, a naphthyridine carboxamide that inhibits the strand-transfer activity of recombinant HIV and SIV integrase in vitro (IC₅₀ = 40 nM).

¹Department of Biological Chemistry, ²Department of Medicinal Chemistry, ³Department of Laboratory Animal Research, ⁴Department of Vaccine Research, ⁵Drug Metabolism and Pharmaceutical Research, Merck Research Laboratories, Post Office Box 4, West Point, PA 19486, USA.

*To whom correspondence should be addressed. E-mail: steve_young@merck.com, daria_hazuda@merck.com

A versatile statistical analysis algorithm to detect genome copy number variation

Raoul-Sam Daruwala^{†*}, Archisman Rudra^{†*}, Harry Ostrer[‡], Robert Lucito[¶], Michael Wigler[¶], and Bud Mishra^{†¶||}

[†]Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012; [¶]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724; and [‡]Human Genetics Program, New York University School of Medicine, New York, NY 10012

Communicated by Jacob T. Schwartz, New York University, New York, NY, September 30, 2004 (received for review April 10, 2004)

We have developed a versatile statistical analysis algorithm for the detection of genomic aberrations in human cancer cell lines. The algorithm analyzes genomic data obtained from a variety of array technologies, such as oligonucleotide array, bacterial artificial chromosome array, or array-based comparative genomic hybridization, that operate by hybridizing with genomic material obtained from cancer and normal cells and allow detection of regions of the genome with altered copy number. The number of probes (i.e., resolution), the amount of uncharacterized noise per probe, and the severity of chromosomal aberrations per chromosomal region may vary with the underlying technology, biological sample, and sample preparation. Constrained by these uncertainties, our algorithm aims at robustness by using a priorless maximum *a posteriori* estimator and at efficiency by a dynamic programming implementation. We illustrate these characteristics of our algorithm by applying it to data obtained from representational oligonucleotide microarray analysis and array-based comparative genomic hybridization technology as well as to synthetic data obtained from an artificial model whose properties can be varied computationally. The algorithm can combine data from multiple sources and thus facilitate the discovery of genes and markers important in cancer, as well as the discovery of loci important in inherited genetic disease.

array-based comparative genomic hybridization | copy-number fluctuations | maximum *a posteriori* estimator

Genomes in a population are polymorphic, giving rise to diversity and variation. In cancer, even somatic cell genomes can rearrange themselves, often resulting in genomic deletion (hemi- or homozygous) and amplifications. Means for assessing these chromosomal aberrations quickly, inexpensively, and accurately have many potential scientific, clinical, and therapeutic implications (1, 2), particularly in the genomics of cancer and inherited diseases. Genome-based methods for studying cancer, in contrast to the gene expression-based methods, can exploit the stability of DNA (as a component of the cancerous cell, which does not vary as a function of the cell's physiological state). Karyotyping, determination of ploidy, and comparative genomic hybridization have been useful tools for this purpose even though they are crude and produce data that must be processed by sophisticated statistical algorithms to serve as useful guides to diagnosis and treatment.

Microarray methods are an important new technology that can be used to study variations between regular and cancer genomes. Imagine that one can sample the genome uniformly (independently and identically distributed) and reproducibly to create a large number of oligonucleotides (on the order of 100,000 probes) located every 30 kb or so. These oligonucleotides almost always come from regions of the genome that do not share homologous sequences elsewhere in the genome. These sequences (typically less than a few hundred base pairs long) occupy unique positions in the normal genome and have exactly two copies.

If one such oligonucleotide belongs to a region in a cancer genome that has an altered copy number, say, c ($0 \leq c \neq 2$), then

when the cancer genome is sampled, this oligonucleotide will occur with a probability that is $c/2$ times that in the regular genome. The copy number can be computed by a ratiometric measurement of the abundance of an oligonucleotide in a cancer sample measured against that in the regular genome. This technique can be generalized to measure the copy number variations for many probes simultaneously with high-throughput microarray experiments. Even though the ratiometric measurements used and the associated regularizations tame the multiplicative noises in the system to some extent, there remains a large amount of uncharacterized noise (generally additive) that can render the data worthless unless a proper data-analysis algorithm is applied. Because the data may come from multiple sources collected with varying protocols, such an algorithm must be general and be based on a minimal set of prior assumptions about the methods. The algorithm we describe below reflects these desiderata.

Our Bayesian approach constructs a most plausible hypothesis concerning regional changes and the corresponding associated copy number. It can be viewed as an optimization process minimizing a score function that assigns penalties of different type for each kind of deviation from genomic normality (break-points, unexplainable probe values, noise, etc.); we discuss how these penalties are derived. We describe various algorithmic alternatives, their implementations, and the empirical results derived using real data (where the underlying facts are not directly verifiable) and simulated data (where the true facts are known).

Statistical Model

We start by describing a probabilistic generative model for observed copy number data. The model is Bayesian in spirit, in that we use parameterized prior distributions and use the posterior distribution function to estimate the underlying model. We use a maximum *a posteriori* (MAP) technique to estimate the underlying model. This idealized statistical model takes into account some major sources of copy number variation in an irregular genome and is described by two scalar parameters $0 \leq p_r, p_b \leq 1$.

We assume that there is a copy-number distribution for probes at locations that have not been affected by the chromosomal aberrations associated with cancer. We call these probes regular probes. We also assume that the probability for a particular probe being regular is p_r and that the associated regular copy-number distribution, after log transformation, is Gaussian, with mean μ_r and standard deviation σ_r . For the other probes, which we call deviated, the log-transformed copy-number distributions also are assumed to be Gaussian, with unknown mean and standard deviation, distinct from the regular distribution. There

Abbreviations: MAP, maximum *a posteriori*; ROMA, representational oligonucleotide microarray analysis; CGH, comparative genomic hybridization; arrayCGH, array-based CGH; HMM, hidden Markov model.

*R.-S.D. and A.R. contributed equally to this work.

||To whom correspondence should be addressed. E-mail: mishra@nyu.edu.

© 2004 by The National Academy of Sciences of the USA

are usually many sets of probes drawn from different deviated distributions.

We also assume that there are locations in the genome that are particularly susceptible to amplification (also known as duplication) and deletion events. These aberrations change the copy numbers of probes locally. We model the number of such mutations as a Poisson process with parameter $p_b N$, where N is the length of the genome (i.e., total number of probes).

We subdivide the probes along the genome into k nonoverlapping intervals. Probes belonging to a particular interval are assumed to have a similar evolutionary history of duplication and deletion events, and therefore have similar copy-number distributions. The number of intervals into which the probes can be separated represents the progressive degeneration of a cancer cell line. We do not model single nucleotide polymorphisms and other point-mutation events, and this undermodeling reappears as localized noise in our analyzed data.

In our picture, each interval in this subdivision has a “true” copy number. Our goal is to estimate the correct subdivision and the copy numbers associated with each subinterval. Despite its simplicity, our model can serve as the basis of a statistical algorithm to infer the aberrations without overfitting the data.

More formally, given a set of N probe copy-number values arranged on the genome, we assume that there is an unknown partition of this set into nonoverlapping subintervals. The probe copy number values in the j th interval are assumed to arise as independent samples from a Gaussian distribution $\mathcal{N}(\mu_j, \sigma_j)$. The parameters relating to the j th interval can be represented as the tuple $I_j = (\mu_j, \sigma_j)$, where μ_j and σ_j are the mean and standard deviation of the appropriate Gaussian distribution and i_j is the position of the last probe in the interval. We call such a set of intervals $I = \{I_j | j = 1, \dots, k\}$ an interval structure. When a particular interval in I is regular, its mean is the regular mean μ_r . If an interval I_j is deviated, then its population mean μ_j is unknown and is estimated by using the sample mean over the interval. In this work, we assume that all of the σ_j terms are equal to some common value σ , and we therefore omit them from the notation. We denote an interval structure I_N with k intervals and whose intervals have associated means $\mu_1, \dots, \mu_k$ and endpoints $i_1, \dots, i_k$ (necessarily $i_k = N$) as $\langle i_1, \mu_1, i_2, \mu_2, \dots, i_k, \mu_k \rangle$.

Our goal is to estimate the unknown interval structure I_N from an input sequence $\{v_i, i = 1 \dots N\}$ of copy numbers of N successive probes.

The statistical model described thus far fits naturally into a Bayesian setting. We can start with a prior distribution on the set of interval structures depending only on the number of intervals and the number of regular probes with two scalar parameters p_r and p_b whose significance is described above.

This prior has two components, the first a Poisson distribution to model the number of intervals with Poisson parameter $p_b N$. The second component is a sequence of Bernoulli trials, one for each probe with probability p_r that a given probe is regular. Combining these factors, the prior distribution becomes

$$Pr(I_N) = e^{-p_b N} \frac{(p_b N)^k}{k!} p_r^{\#regular} (1 - p_r)^{\#deviated} \quad [1]$$

where $\#regular$ is the number of regular probes with the “regular” copy-number distribution and $\#deviated$ is the number of remaining probes in the interval structure I_N . In each interval I_j , the data points are modeled by adding independent Gaussian noise to this prior structure and are drawn from the Gaussian distribution $\mathcal{N}(\mu_j, \sigma)$.

The data likelihood function for the first n probes is given by the product of Gaussians:

$$Pr(\mathbf{x} | I_N) = \prod_{i=1}^n \phi(x_i, \mu_j, \sigma^2) \quad [2]$$

where the i th probe is covered by the j th interval of the interval structure I_N and μ_j is the mean of the corresponding Gaussian distribution. ϕ denotes the density function of the Gaussian distribution. By multiplication, we obtain the posterior likelihood function:

$$L(I_N | \mathbf{x}) = e^{-p_b N} \frac{(p_b N)^k}{k!} p_r^{\#regular} (1 - p_r)^{\#deviated} \prod_{i=1}^n \phi(x_i, \mu_j, \sigma^2). \quad [3]$$

In the above expression for L , only the μ values of nonregular processes are unknown, and we estimate these values by using the sample mean for the interval. The MAP solution to the segmentation problem is obtained by finding the interval structure I^* that maximizes this likelihood function or, equivalently, minimizes the negative log likelihood of L .

Algorithm and Implementation

A dynamic programming algorithm efficiently minimizes the negative posterior log likelihood function obtained above. Starting with an interval structure $I = \langle i_1, \mu_1, \dots, i_k, \mu_k \rangle$, we can extend it to the interval structure $I' = \langle i_1, \mu_1, \dots, i_{k+1}, \mu_{k+1} \rangle$, where $i_{k+1} > i_k$. The following formula computes the log likelihood for such an extension

$$\begin{aligned} -\log L(I') = & -\log L(I) + \frac{1}{2\sigma^2} \sum_{j=i_k+1}^{i_{k+1}} (x_j - \mu_{k+1})^2 \\ & - \log(p_b N) + \log(k+1) \\ & + \frac{i_{k+1} - i_k}{2} \log(2\pi\sigma^2) - (i_{k+1} - i_k) \\ & \cdot [\mathbb{1}_{k \in regular} \log p_r + \mathbb{1}_{k \in deviated} \log(1 - p_r)] \end{aligned} \quad [4]$$

where the last term on the right side is chosen according to whether the last added interval (i.e., the one extending from $i_k + 1$ to i_{k+1}) is regular or not. $\mathbb{1} = 1$ if the Boolean formula e is true, and 0 otherwise. We also point out that the MAP approach permits the estimation of the μ terms in a uniform manner. When the last added interval is regular, the value μ_{k+1} is fixed at the global mean μ . When the last added interval is deviated, however, the MAP criterion automatically forces the choice of the sample mean of the data points covered by the last interval as the value for μ_{k+1} . One could build hierarchical models for the mean and use these “shrinkage-like” estimators as well (3), although we do not explore that approach here.

The negative log likelihood function satisfies an optimality condition that allows one to use a standard dynamic programming algorithm [of time-complexity $O(N^2)$] in this setting.

Results

We evaluate the performance of this simple Bayesian scheme on three kinds of data. For each of these data sets, we will see that proper choice of the parameter values p_r and p_b leads to good segmentation. Indeed, coefficients chosen from within a fairly large region of the “ p_r - p_b space” lead to a good segmentation because our procedure is stable over a large domain. The

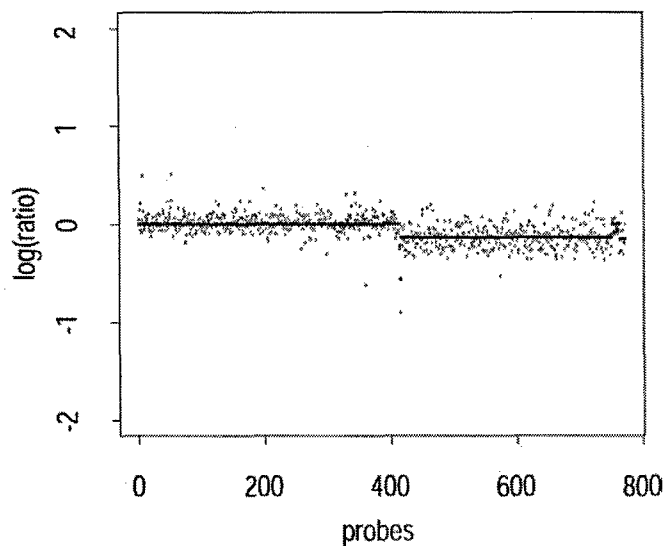


Fig. 1. Segmented probes on chromosome 2, $p_r = 0.55$, $p_b = 0.005$, sampling rate 1 in 10.

parameters p_b , p_r , μ , and σ play different roles: an increase in p_b yields more intervals in the segmentation, and, as p_r is increased, more probes come to be classified as regular, and therefore the number of different segments diminish.

The choice of μ is critical because it controls the bias in the resulting segmentation. The choice of σ is also important because increasing σ weakens the influence of the data on the segmentation obtained.

Representational Oligonucleotide Microarray Analysis (ROMA) Data from Breast Cancer Cell Lines. ROMA is a comparative genomic hybridization (CGH) technique developed by Wigler and colleagues (1) at Cold Spring Harbor Laboratory. It evolved from an earlier method, representational differential analysis, which was adapted for greatly increased volumes of data obtained by using an oligonucleotide microarray. ROMA uses a comparative "two-color" scheme to compare multiple genomes, each represented with reduced complexity by using a PCR-based method (4, 5). As in other array-based methods, ROMA performs simultaneous array hybridization to compare a normal genome at one fluorescent wavelength and a tumor genome at another. The DNA representations used by ROMA are based on amplification of short restriction endonuclease fragments and hence are predictable from the nucleotide sequence of the genome. We have tested our algorithm on the data sets from the Wigler laboratory obtained by ROMA from the genomes of breast cancer cell lines. The data set is based on 85,000 well characterized probes, each of length 70 bp, providing a resolution of a probe every 15–30 kb.

Figs. 1 and 2 show subsampled ROMA breast cancer data from chromosomes 2 and 8, respectively, overlaid with the segmentation found by our algorithm. The low-complexity DNA representation used in ROMA, together with a careful choice of probes, provides low-noise data that can be characterized accurately by the algorithm.

Array-Based CGH (arrayCGH) Data from Prostate Cancer Cell Lines. arrayCGH is a recently developed technique that maps duplicated or deleted chromosomal segments onto high-density arrays of well characterized bacterial artificial chromosomes (BACs), rather than onto metaphase chromosomes. This method has been used for precise mapping of duplications and deletions occurring in cancers and other human diseases, including birth

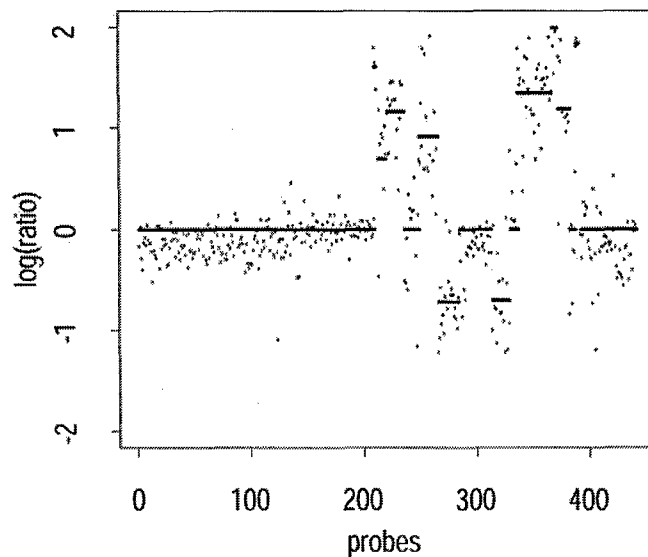


Fig. 2. Segmented probes on chromosome 8, $p_r = 0.55$, $p_b = 0.01$, sampling rate 1 in 10.

defects and mental retardation (see ref. 6 for review and applications of this technique). Tumors that have been studied by using this method are breast, head and neck, Wilms, esophageal, pulmonary artery intimal, adrenocortical, renal, and prostate cancers and lymphomas. We have tested our algorithm on a data set obtained by high-resolution arrayCGH analysis of prostate cancer tissue. The data were supplied by a group at Nijmegen University Medical Center and obtained by hybridization on their custom array composed of $\approx 3,500$ fluorescence *in situ* hybridization-verified clones selected to cover the genome with an average of one clone per megabase (7).

Fig. 3 shows the performance of our segmentation algorithm on data from prostate cancer cell lines obtained through arrayCGH experiments. We note that these data are noisier than the ROMA data considered previously. But, despite the increased noise, the segmentation algorithm is robust and yields reasonable segmentations.

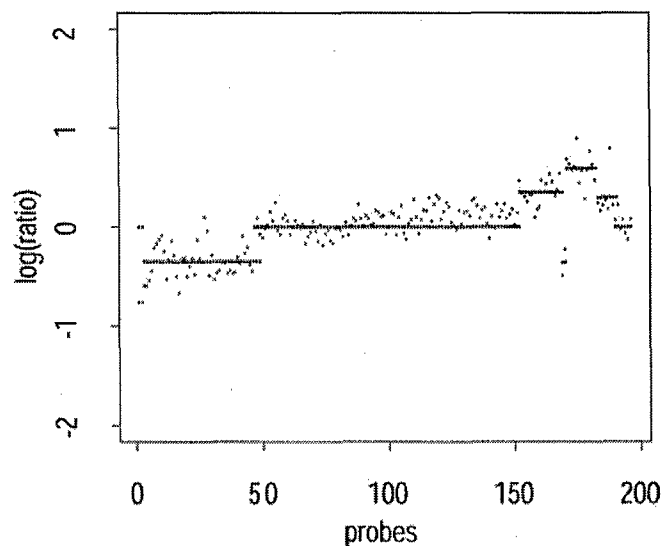


Fig. 3. Segmented probes on array-based CGH data. Chromosome 8, $p_r = 0.55$, $p_b = 0.01$, sampling rate 1 in 10.

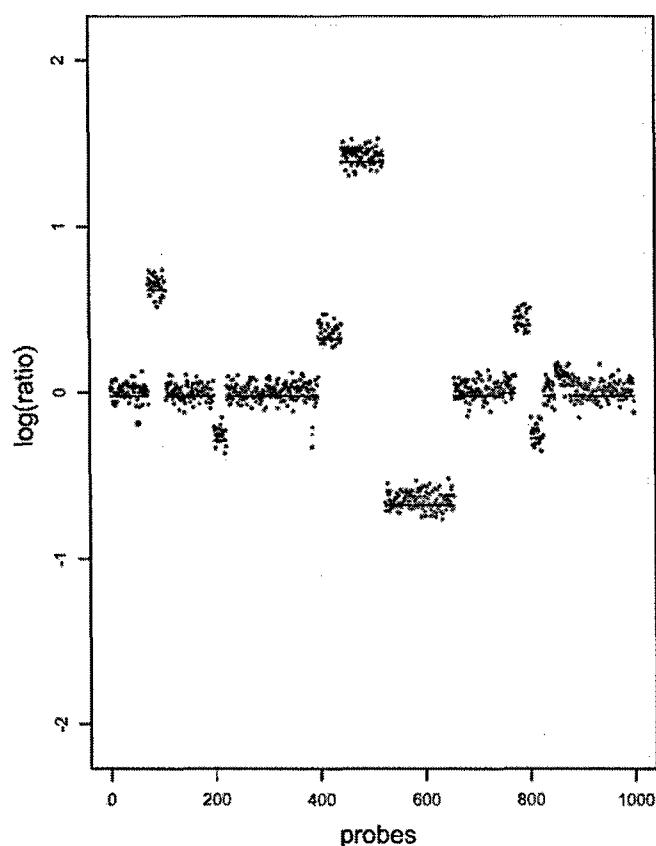


Fig. 4. A simulated genome with $\mu = 0.0$ and $\sigma = 0.15$ and the corresponding segmentation.

Simulated Data. To further test our algorithm, we can use an artificial but biologically inspired model to generate synthetic data. To generate simulated copy-number data, we choose loci uniformly over a genome such that the probability of a duplication or deletion event taking place at that location on the genome is given by p_b . At each of these points, we assign a new copy-number value that represents the mean for the new interval. The mean values are drawn from a power-transformed γ distribution to mimic the observed distribution of means in experimental data. The lengths of the intervals follow a geometric distribution such that the ratio of expected fragment length and the expected distance between the beginning of each interval is p_r . Once the segmentation and the mean values are chosen, we generate the simulated data by adding random Gaussian noise. A typical simulated genome is shown in Fig. 4.

Effect of Noise on Performance. We investigate the effect of increasing the σ of the underlying model on the performance of the segmenter. Assuming the parameters of the model are correctly estimated, the segmenter can output the estimated mean value at every probe position. Using the known mean values, we can compare these two sequences of means. In our setting, a good measure of error is the number of misclassified probes, i.e., the number of probes that are known to be regular but were classified as amplified or deleted and vice versa. Fig. 5 shows the increase in the rate of misclassification as σ increases.

Prior Selection

Proper selection of a prior distribution has received extensive attention in the literature. Approaches include noninformative priors [Jeffreys (8)], reference priors [Bernardo (9); see also

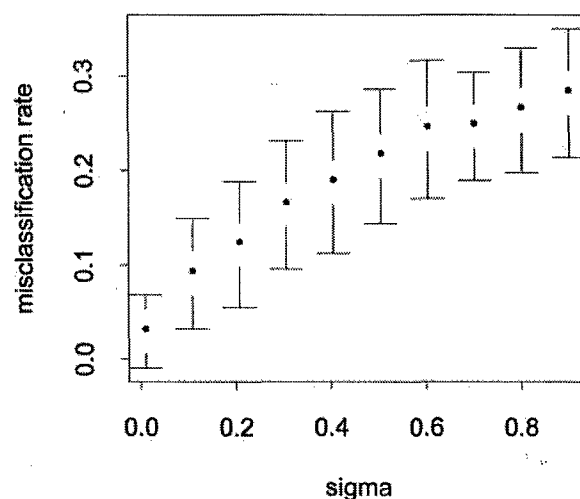


Fig. 5. Average number of misclassified probes plotted against increasing σ on synthetic data. The average number of misclassified probes in >100 trials is normalized against the length of the simulated genome.

Berger and Bernardo (10) and Kass and Wasserman (11)], and conjugate priors [Raiffa and Schlaifer (12)] among others. Conjugate prior methods frequently arise in connection with exponential families of distributions [see Brown (13)]. Other approaches include using invariance properties to posit prior distributions with good performance. More recent and somewhat more data-dependent techniques include hierarchical and empirical Bayes techniques. Textbooks such as those by Bernardo and Smith (14), Berger (15), Carlin and Louis (16), Gelman *et al.* (17), and Robert (18) cover model selection as a part of Bayesian learning.

For the problem of estimating probe copy numbers, the prior distribution is specified by the two probability parameters, p_r and p_b . The other parameters (μ , the regular mean, and σ^2 , the regular variance) can be estimated by experiment. The problem of prior selection reduces to the problem of optimally selecting the values of p_r and p_b to prevent overfitting of the data.

Minimax approaches choose prior distributions that minimize the maximum value of the likelihood function (Eq. 4). This criterion is pessimistic, in that it chooses the prior that generates the worst likelihood value. See, for example, Berger (15), Brown (19–21), and Strawderman (22–24). In the non-parametric setting of function estimation, multiscale methods have been proven to be asymptotically minimax by Donoho and Johnstone (25–27).

We adapt an approach, based on statistical decision theory, that directly controls the level of overfitting without explicitly depending on the asymptotic performance guarantees of minimax approaches. We rely on the fact that, in any segmentation, each jump separates the probes in the two adjoining intervals. If a segmentation is overfitted, at least one of its jumps must be overfitted, too. We use Hotelling's t^2 statistic [see Anderson (28) or Wilks (29)] at each jump to compute a measure of this overfitting.

We apply an F test to Hotelling's t^2 statistic to test whether two sets of independent samples come from populations with the same mean. This F test is possible because we assume that the two sets of samples have the same (but still unknown) variance. Let $x_1, x_2, \dots, x_{N_1}$ and $y_1, y_2, \dots, y_{N_2}$ be the two sets of independent samples taken from successive intervals of size N_1 and N_2 , respectively. Then, we define the statistic:

$$t^2 = \frac{\frac{N_1 N_2}{(N_1 + N_2)} (\bar{x} - \bar{y})^2}{\frac{1}{df_1 + df_2} (\sum_i (x_i - \bar{x})^2 + \sum_j (y_j - \bar{y})^2)} \quad [5]$$

where $\bar{x}$ and $\bar{y}$ are the respective sample means, and df_1 and df_2 refer to the respective degrees of freedom of the two samples. Under the null hypothesis (of equal means), t^2 follows an F distribution with 1, $(df_1 + df_2)$ degrees of freedom. This leads to a one-tailed F test.

Intuitively, t^2 needs to be large to avoid overfitting. The cumulative probability for the appropriate F distribution yields a score that determines the quality of the break. We also compute a score for the goodness of fit for the whole segmentation. This procedure yields a set of scores: one for each break and one for the goodness of fit. The minimum of these scores is used to evaluate the whole segmentation. We select the parameters p_r and p_b to maximize this score by searching at regular intervals over the parameter space. We can continue to refine the search in the neighborhood of the optimal values obtained. However, the algorithm is already extremely stable in a large region of the p_r - p_b space and yields, in practice, very good segmentations.

Discussion

The problem of detecting copy-number variations has assumed biological importance in recent years. Most extant algorithms use a global thresholding approach for this problem. This is the case, for example, in Vissers *et al.* (7) as well as many commercially available packages. These algorithms have the advantage of simplicity but perform poorly in the presence of noise and correlations. Other published approaches have used smoothing (30), hidden Markov models (HMMs) (31, 32), and mixtures of Gaussians, as well as approaches that try to estimate the correlations between probes (33, 34). Although smoothing certainly improves the performance of threshold-based approaches, the specifics remain somewhat ad hoc, and the method requires tuning dependent on the source and resolution of the data.

HMMs have the advantage of having a general (although slow) learning algorithm; however, their performance is very sensitive to the topology of the HMM. For this reason, researchers tend to analyze very narrow classes of data with a particular HMM, e.g., a prostate cancer cell line. Very rarely are normal-normal data so analyzed, because this analysis usually necessitates the construction of an HMM with a different topology, leading to questions about the comparative power of such analyses. The main problem with both this approach and others based on assuming a distributional form for cancerous data is that the cancerous insertion-deletion polymorphisms are characterized by being nonnormal, rather than belonging to a specific distributional form. Therefore, fitting cancer data leads to the construction of a specific HMM topology that might depend on the specific cancer as well as the goodness of fit desired by the statistician.

Olshen and Venkatraman (35) have advocated another approach based on recursive change-point detection in the copy number data. The existence of a large literature on change-point analysis makes this approach attractive. Conversely, an efficient implementation of this algorithm is difficult. The specific statistic chosen for change-point detection in this and other work of

the group perform poorly on normal-normal data due to overly pessimistic criteria. In some sense, our approach of putting a Bayesian prior on the number of change points enables us to be aggressive about detecting change points.

We have devised a versatile MAP estimator algorithm to analyze arrayCGH data. This algorithm uses a model that captures the genomic amplification-deletion processes but is relatively insensitive to additive noise in the data. When the algorithm was tested on a wide variety of data from ROMA- and arrayCGH-based methods, this particular feature of the algorithm provided strength and robustness. We note that the correct choice of p_r and p_b is critical in the segmentation algorithm. High values of p_b tend to yield overfitted solutions, whereas high values of p_r drive us toward biased solutions that mark all segments as regular. The advantage of having an algorithm with only two numerical parameters is that a simple and natural statistical criterion enables the proper choice of these parameters in all cases.

We parenthetically note that our approach extends to multi-dimensional data *mutatis mutandis*. The relevant likelihood function needs to be changed to the following

$$L((i_1, \mu_1, i_2, \mu_2, \dots, i_k, \mu_k)) \\ = e^{-p_b N} \frac{(p_b N)^k}{k!} \frac{1}{(2\pi|\Sigma|)^{n/2}} \\ \cdot \prod_{i=1}^n e^{-(x_i - \mu_i)' \Sigma^{-1} (x_i - \mu_i)/2} \cdot p_r^{\# \text{regular}} (1 - p_r)^{\# \text{deviated}}. \quad [6]$$

The t^2 statistic can be modified similarly.

Prior work by Donoho and colleagues (36–38) on detecting geometrical features in point clouds by using multiresolution methods relates to the ideas presented here. These papers focus on the use of multiresolution approaches for efficiency and statistical stability. There is also prior work by Kolaczyk (see ref. 39, for example) that gives a unified Bayesian treatment to multiresolution analysis and covers large classes of both continuous as well as discrete processes. Our approach leads to an efficient algorithm for sequence-like data, which can be used in a multiscale setting if desired. Furthermore, in our approach, the probabilistic generative model directly leads to the cost function; thus, other generative models, e.g., poisson models, can be easily considered in this setting. It should be noted that the Hotelling's t^2 statistic cannot be easily generalized to this setting.

We thank Yi Zhou (New York University Bioinformatics Group) and two anonymous reviewers for many helpful discussions, suggestions, and relevant references to statistical literature. We also thank Lakshmi Muthuswami (Cold Spring Harbor Laboratory), and Eric Schoenmakers and Joris Veltman (Nijmegen University Medical Center, Nijmegen, The Netherlands) for providing the data used here and for explaining their biological significance. The work reported in this paper was supported by grants from the National Science Foundation (NSF) Qubic Program, the NSF Information Technology Research Program, the Defense Advanced Research Projects Agency, a Howard Hughes Medical Institute Biomedical Support Research Grant, the U.S. Department of Energy, the U.S. Air Force (Air Force Research Laboratory), the National Institutes of Health, and the New York State Office of Science, Technology and Academic Research.

- Lucito, R., West, J., Reiner, A., Alexander, J., Esposito, D., Mishra, B., Powers, S., Norton, L. & Wigler, M. (2000) *Genome Res.* **10**, 1726–1736.
- Mishra, B. (2002) *Comput. Sci. Eng.* **4**, 42–49.
- Daniels, M. J. & Kass, R. E. (1999) *J. Am. Stat. Assoc.* **94**, 1254–1263.
- Lisitsyn, N., Lisitsyn, N. & Wigler, M. (1993) *Science*, **258**, 946–951.
- Lucito, R., Nakimura, M., West, J. A., Han, Y., Chin, K., Jensen, K., McCombie, R., Gray, J. W. & Wigler, M. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 4487–4492.

- Albertson, D. G. & Pinkel, D. (2003) *Hum. Mol. Genet.* **12**, Suppl. 2, R145–R152.
- Vissers, L. E. L. M., de Vries, B. B. A., Osoegawa, K., Janssen, I. M., Feuth, T., Choy, C. O., Straatman, H., van der Vliet, W., Huys, E. H. L. P. G., van Rijk, A., *et al.* (2003) *Am. J. Hum. Genet.* **73**, 1261–1270.
- Jeffreys, H. (1946) *Proc. R. Soc. London Ser. A* **186**, 453–461.
- Bernardo, J. M. (1979) *J. R. Stat. Soc. Ser. B* **41**, 113–147.

10. Berger, J. O. & Bernardo, J. M. (1992) in *Bayesian Statistics 4*, eds. Berger, J. O., Bernardo, J. M., Dawid, A. P. & Smith, A. F. M. (Oxford Univ. Press, Oxford), pp. 35–60.
11. Kass, R. E. & Wasserman, L. A. (1996) *J. Am. Stat. Assoc.* **91**, 1343–1370.
12. Raiffa, H. & Schlaifer, R. (1961) *Applied Statistical Decision Theory* (Wiley, New York).
13. Brown, L. D. (1986) *Foundations of Exponential Families* (Institute of Mathematical Statistics, Hayward, CA), Monograph Series 6.
14. Bernardo, J. M. & Smith, A. F. M. (1994) *Bayesian Theory* (Wiley, New York).
15. Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis* (Springer, New York), 2nd Ed.
16. Carlin, B. P. & Louis, T. A. (1996) *Bayes and Empirical Bayes Methods for Data Analysis* (Chapman & Hall, London).
17. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. D. (1995) *Bayesian Data Analysis* (Chapman & Hall, London).
18. Robert, C. P. (2001) *The Bayesian Choice* (Springer, New York).
19. Brown, L. D. (1971) *Ann. Math. Stat.* **42**, 855–903.
20. Brown, L. D. (1993) in *Statistical Decision Theory and Related Topics 5*, eds. Gupta, S. S. & Berger, J. O. (Springer, New York), pp. 1–18.
21. Brown, L. D. (2000) *J. Am. Stat. Assoc.* **95**, 1277–1282.
22. Strawderman, W. E. (1971) *Ann. Math. Stat.* **42**, 385–388.
23. Strawderman, W. E. (1974) *J. Multivariate Anal.* **4**, 255–263.
24. Strawderman, W. E. (2000) *J. Am. Stat. Assoc.* **95**, 1364–1368.
25. Donoho, D. & Johnstone, I. M. (1998) *Ann. Stat.* **26**, 879–921.
26. Donoho, D. & Johnstone, I. M. (1999) *Stat. Sinica*, **9**, 1–32.
27. Donoho, D. L. (1999) *Ann. Stat.* **27**, 859–897.
28. Anderson, T. W. (1958) *An Introduction to Multivariate Statistical Analysis* (Wiley, New York).
29. Wilks, S. S. (1962) *Mathematical Statistics* (Wiley, New York).
30. Jong, K., Marchiori, E., Meijer, G., van der Vaart, A. & Ylstra, B. (June 16, 2004) *Bioinformatics*, 10.1093/bioinformatics/bth355.
31. Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., et al. (2004) *Science* **305**, 525–528.
32. Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G. & Jain, A. N. (2004) *J. Multivariate Anal.* **90**, 132–153.
33. Wang, J., Meza-Zepeda, L. A., Kresse, S. H. & Myklebost, O. (2004) *BMC Bioinformatics* **5**, 74.
34. Wang, Y. & Guo, S. W. (2004) *Front. Biosci.* **9**, 540–549.
35. Olshen, A. B. & Venkatraman, E. S. (2002) in *American Statistical Association Proceedings of the Joint Statistical Meetings* (American Statistical Association, Alexandria, VA) pp. 2530–2535.
36. Arias-Castro, E., Donoho, D. & Huo, X. (2003) *Technical Report 2003-22* (Department of Statistics, Stanford University, Stanford, CA).
37. Arias-Castro, E., Donoho, D. & Huo, X. (2003) *Technical Report 2003-17* (Department of Statistics, Stanford University, Stanford, CA).
38. Donoho, D. & Huo, X. (2001) in *Multiscale and Multiresolution Methods*, Springer Lecture Notes in Computational Science and Engineering, eds. Barth, T. J., Chan, T. & Haimes, R. (Springer, New York), Vol. 20, pp. 149–196.
39. Kolaczyk, E. D. (1999) *J. Am. Stat. Assoc.* **94**, 920–933.