AD ____________________

Award Number: DAMD17-01-1-0500

TITLE: A Genomics Approach to Tumor Genome Analysis

PRINCIPAL INVESTIGATOR: Colin C. Collins, Ph.D.

CONTRACTING ORGANIZATION: University of California, San Francisco
San Francisco, California 94143-0962

REPORT DATE: August 2002

TYPE OF REPORT: Final

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.

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	3. REPORT TYPE AND DATES COVERED
	August 2002	Final (16 Jul 01 - 15 Jul 02)

4. TITLE AND SUBTITLE
A Genomics Approach to Tumor Genome Analysis

5. FUNDING NUMBERS
DAMD17-01-1-0500

6. AUTHOR(S)
Colin C. Collins, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of California, San Francisco
San Francisco, California 94143-0962

E-Mail: collins@cc.ucsf.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

20030226 035

11. SUPPLEMENTARY NOTES
Original contains color plates: All DTIC reproductions will be in black and white.

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

Genomes of solid tumors are often highly rearranged and these rearrangements promote cancer progression through disruption of genes mediating immortality, survival, metastasis, and resistance to therapy. A growing number of these genes are targets for anti-tumor therapeutics. End Sequence Profiling (ESP) is a sequence-based method capable identifying all known structural rearrangements in a tumor genome with a resolution of ~ 5kb for approximately 1% the cost of whole genome sequencing. Because ESP is sequence-based it can readily integrate expression profiling and proteomic data. Analysis of the breast cancer cell line MCF7 using ESP revealed a far more different structure of rearranged genome than previously appreciated. Our observations reveal colocalization of amplified loci, inversions, translocation breakpoints and provide evidence for packaging of cancer related genes in complex tumor-specific structures. The results have important implications for understanding tumor evolution and identification of novel biomarkers and targets for development of therapeutics.

14. SUBJECT TERMS
breast cancer, ESP, MCF7

15. NUMBER OF PAGES
11

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Cover..1

SF 298..2

Table of Contents..3

Introduction..4

Body..4

Key Research Accomplishments...5

Reportable Outcomes...5

Conclusions...6

References..N/A

Appendices..N/A

Introduction

Genomes of solid tumors are often highly rearranged. Genes altered by rearrangements are known to mediate immortality, survival, metastasis, and resistance to therapy, and a growing number are targets for anti-tumor therapeutics. Powerful techniques now exist to detect and map changes in genome copy number, gene expression, and methylation pattern. However, these techniques do not allow high-resolution analysis of structural changes and they do not provide information at the DNA sequence level. Moreover, data from the many disparate techniques is not easily integrated. ESP is a sequence-based approach capable of precisely identifying and characterizing structural rearrangements and genome sequence variations in a tumor genome at very high resolution for approximately 1% the cost of whole genome sequencing. Because it is a sequence-based methodology, ESP is readily integrated with transcriptome and proteomic data. ESP begins with construction of a BAC library for the tumor of interest. BAC end sequences are then generated for individual BAC clones and mapped onto the normal "reference" genome sequence. This process reveals all classes of structural aberrations including copy number changes, translocations, and inversions, and identifies BAC clones carrying these structural aberrations (Fig. 1a). We have demonstrated the power of ESP in an analysis of the breast cancer cell line MCF7. Whole genome ESP confirmed most copy number changes revealed by array CGH, mapping them at much higher resolution, and revealed several recurrent structural rearrangements that are likely to have functional significance including translocations visible by spectral karyotyping. Interestingly, ESP clearly demonstrated that the gene ZNF217 in a region of amplification, at 20q13.2, is physically linked to DNA from 3p14 and 17q23. Fluorescence in situ hybridization (FISH), alignment to independent genome assemblies and shotgun sequencing confirmed these structural linkages. Information about specific breakpoints derived by sequencing BACs carrying these breakpoints will be presented as will an assessment of the mutation frequency in MCF7 derived from the BAC end sequences. ESP appears to be a powerful tool for precise identification and characterization of tumor genome structural and sequence level abnormalities, and because it is inherently integrative, it brings the power of genetic analysis to interpretation of transcriptome and proteomic data.

Body

Aim 1 was as follows: Construct and array a 30-fold redundant BAC library from the breast cancer cell line MCF7 and select cloned carrying genes known to be amplified in this cell line such as AIB1, ERBB2, and ZNF217. Computer modeling suggested that a 3-fold redundant BAC library would function equally well as a 30-fold library for analysis of amplicon structures and for validation of ESP. Thus, a 3-fold redundant BAC library was constructed. This library is comprised of 68,000 clones and is arrayed in microtiter plates and on nylon filters for screening by hybridization. The average insert size of the library is 130 kb as determined by pulsed field gel electrophoresis. This was determined by pulsed field gel electrophoresis. Eleven hybridization probes were designed spanning the ZNF217 breast cancer amplicon at 20q13.2 and used to screen the MCF7 library. This process yielded ~ 150 BAC clones. This is significant because it confirmed our computer modeling suggesting a smaller library would function for validation of ESP.

Aim 2 was as follows: End sequence several hundred BAC clones carrying these genes in collaboration with the Department of Energy's Joint Genome Institute (JGI). To facilitate and significantly expand this aim we initiated a collaboration with The Institute for Genome Research (TIGR) rather than the JGI. All 150 BAC clones from the 20q13.2 amplicon were sequenced and ~8000 random BAC clones from the MCF7 library were also end sequenced. This process enabled us to determine the structure of the ZNF217 amplicon at molecular resolution and to establish that ESP works as modeled.

Aim 3 was as follows: Apply custom genome analysis software to identify: (a) structural

rearrangements in the MCF7 genome within BAC carrying oncogenes such as AIB1, ERRB2, and ZNF217; (b) mutations in the BAC end sequences; (c) genes that are brought into close proximity to ERRB2, MYC, and ZNF217 as a result of structural rearrangement. Figure 1b illustrates the result of running our software on the BAC end sequence data. This software maps BAC end sequence onto the normal reference genome and then generates a graphical display of the data making amplifications, deletions, inversions, complex rearrangements, and translocations immediately apparent. *Figure 1 represents the first structural genomics map of any tumor genome.* Analysis of BAC end sequences for mutations was not performed. Detailed analysis of the genome wide and targeted 20q13.2 ESP data revealed a number of genes being brought together as a result of complex rearrangements. For example the ZNF217 locus at 20q13.2 becomes fused to DNA from chromosomes 1p21, 3p14, and 17q23 (fig.1a, c). In addition, within 20q, DNA from the AIB1 locus at 20q12 is contained on the same BAC clones as ZNF217 and bone morphogenic factor seven (BMP7). This is remarkable given the fact that ZNF217 and AIB1 are separated by ~ 15 Mb and ZNF217 and BMP7 by ~ 5 Mb. These structural rearrangements were confirmed using fluorescent in situ hybridization (FISH) and sequencing (figs. 2a,b, and 3). Evidence has been obtained that high level amplicons at 1p21, 3p14, 17q23, and 20q13 are packaged together in the MCF7 genome (fig.1b). Significantly we also identified, cloned *en masse,* and confirmed both inversions and translocations from the MCF7 genome (figure 2c,d,e, and f)).

Aim 4 was as follows: Completely sequence BAC clones spanning structural rearrangements to identify the sequences involved in the rearrangement. One BAC clone from the ZNF217 amplicon at 20q13.2 was sequenced to completion (fig. 3). This clone contains the ZNF217 gene, BMP7 gene and other 20q DNA fused to sequence from chromosome 3p14. In this 109 kb BAC clone ZNF217 is the only intact gene, and all rearrangement breakpoints were identified. Three of the four breakpoint are associated with a high density of repetitive elements and one occurs in single copy DNA.

In conclusion the aims of this grant were met and greatly exceeded.

Key Research Accomplishments:

- Constructed a BAC library from MCF7 breast cancer cell line.
  End sequenced ~ 8000 of the BAC clones.
- Developed software for graphical representation of ESP data.
- Established ESP as a rational approach to determining the structural organization of tumor genomes.
- Determined the molecular structure of the ZNF217 amplicon.
- Determined the first structural genomic organization of any tumor genome and did so ~ 370 kb resolution.
- Determined the sequence and fine structure organization of 1 BAC clone from within the ZNF217 amplicon.

Reportable Outcomes

Established and archived a BAC library of MCF7

Obtained funding to continue the work from the California Breast Cancer Research Program (BCRP0

A manuscript is in preparations.

The data was presented at the 2002 Oncogenomics Meeting in Dublin Ireland as an oral presentation. "End sequence profiling (ESP): a sequence-based approach to structural

analysis of tumor genomes"

Conclusions.

ESP is an important genomics tool for determining the structural organization of tumors. Its power is derived from the fact that it is a sequenced based method and can thus be integrated with expression microarray and proteomic data. Further, because it uses BAC libraries of the tumors to be analyzed, aberrations such as translocations, complex rearrangements, and inversions are not only detected but also cloned, making their validation and sequence level analysis of breakpoints and involved genes straightforward. Because ESP provides a rational framework for sequencing tumor genomes it may revolutionize cancer genomics.

Figure Legends:

Fig 1. Depiction of ESP. (**A**) An end-sequenced tumor BAC library is mapped onto normal human sequence. Amplified loci will be over-represented in the library resulting in BES stacking (1) above the expected average. BACs spanning translocations, inversions and other structural rearrangements connect non-adjacent loci (2,4). Deletions are detected by a deficit of BES (3). (**B**) Structural view of MCF7 genome. 8320 MCF7 BES were mapped onto the normal reference sequence. Reference sequence is represented as a horizontal line across the top of the figure (see panel **D** for enlargement). Dark green plot represents number of BES mapped per analysis interval (1 MB as shown). Red dotted line across the BAC number plot represents the average number of BES mapped to an analysis interval. BAC ends with ends mapping to different chromosomes (thus possibly harboring translocation breakpoints) are shown as red lines. BAC clones with ends in the wrong orientation (with ends not pointing towards each other), possibly harboring inversion breakpoints, are shown in blue. BAC clones with apparent insert size too big or too small for expected library size distribution are shown as green lines (these BACs may span deletions in the tumor genome). BES that were mapped ambiguously are shown in purple. Mouse click on the chromosome name brings up a detailed representation of this chromosome (**C**). Chromosome-specific view follows the same conventions as whole-genome view. Blue arrows indicate BAC clones, detecting inversions and translocations validated by FISH, and red arrows indicate BAC clones, detecting complex structural rearrangements, associated with gene amplification, that were confirmed by FISH and sequencing (see fig. X & Y).

Fig 2. FISH-based validation of genome rearrangements identified by ESP. Complete metaphase images can be viewed at http://shark.ucsf.edu/~stas/ESP_1.10.02/fish.html.
A. Multiple independent BAC clones have BES connecting amplicons on 20q13.2 and 17q23. One such BAC 1A11 detects FISH signals on chromosomes 17q23 and 20q13.2 as predicted using ESP. **B.** Hybridization of BAC 1A11 to MCF7 metaphase chromosomes reveals multiple loci of amplification. **C.** Confirmation of translocations identified by ESP. ESP places BES of 5H15 on chromosomes 15q11.2 and 16q22.2. This putative translocation was confirmed using FISH on normal metaphase chromosomes meaning the translocation breakpoint is within clone 5H15. Signal on chromosome 1 may suggest a more complex rearrangement (data not shown). **D.** Detection of an inversion involving the ABL1 oncogene. FISH using BAC 5K16 to normal metaphase chromosomes detects two hybridization domains at 9q22.3 and 9q34.1. The distal breakpoint is located within a first intron of the ABL1 oncogene. **E.** Detection of a pericentric inversion on chromosome 11 using BAC 9I10 as a FISH probe BAC 9I10 detects hybridization domains at 11p11.2 and 11q14.3 as predicted by ESP mapping. **F.** FISH using BAC 9F10 to MCF7 metaphase chromosomes detects 5 chromosomes. The two on the left contain double hybridization domains and the three on the right only one. These data are consistent with a model whereby the inversion occurred on one of the two homologous chromosomes in a hyperdiploid cell line. **G.** Detection of complex rearrangements associated with gene amplification. ESP mapping predicts BAC clone 3F5 has one BES at the ZNF217 locus at 20q13.2 and another at 3p14. FISH with BAC 3F5 on normal metaphase chromosomes confirms the ESP mapping. PCR-based mapping and sequencing located the BMP7 gene in this BAC as well. **H.** Dual color FISH using <u>normal</u> BACs spanning the BMP7 locus (red) and the ZNF217

locus (green) to MCF7 metaphase to metaphase chromosomes. Yellow FISH signals show co-amplification and co-localization of these BACs in MCF7 genome.

Fig 3. **(A)** A graphical representation of the structural organization of BAC 3F5 from the ZNF217 amplicon at 20q13.2. Red arrows demarcate BES. BAC 3F5 was determined to have BES on chromosomes 3p14 and 20q13.2 at the ZNF217 locus. STS content mapping localized the 5' region of the BMP7 gene within the BAC. BAC 3F5 is one of 26 independent clones in the library juxtaposing ZNF217 and BMP7, and one of four that also contain BES in the 3p14 amplicon. Sequencing BAC 3F5 identified five widely separated chromosomal regions fused together in the orientations shown. Only the ZNF217 locus is structurally intact. The PTPRT gene, BMP7, and L39 genes are all truncated. The PTPRT intron 6 is fused to BMP7 intron1 in opposing polarity. L39 intron is fused to nontranscribed DNA 3'of ZNF217. A large CpG island shared by BMP7 and L39 is structurally intact. GenScan and FGENES predict at least two novel genes created by these genome rearrangements (blue arrows). **(B)** Sequences spanning each genome breakpoint are presented with the fusion site in red. **(C)** Genome cryptographer plot of the density an classification of repetitive elements (Alu elements red and L1 elements green) at each breakpoint. Breakpoints 1, 2, and 4 occur in regions of very high repetitive element density whereas breakpoint 3 occurs in single copy DNA.

Figure 1

A.

Aberration Detection

Copy Number

Distance Along Genome

1

2

3

4

D.

chr11

20000000 40000000 60000000 80000000 100000000 120000000

11P17

MCF7-7P24

MCF7-7N24

MCF7-24L21

MCF7-9I10

MCF7-23A9

MCF7-3B8

C.

chr20

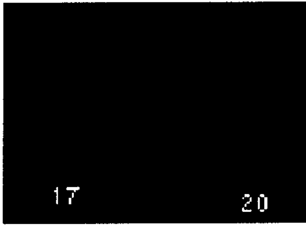
B.

Translocations Wrong end orientation Wrong insert size Ambiguous mapping Unremarkable clones

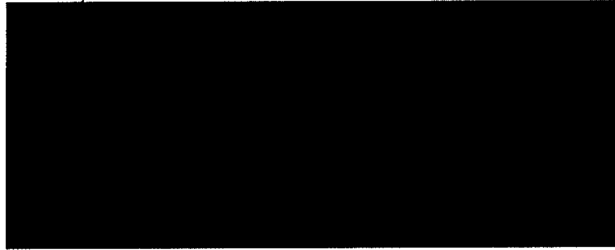
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22 chrX

Figure 2.

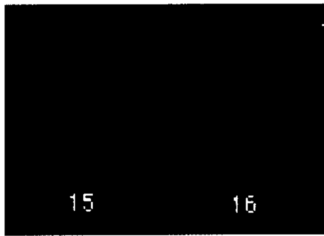
A



B



C

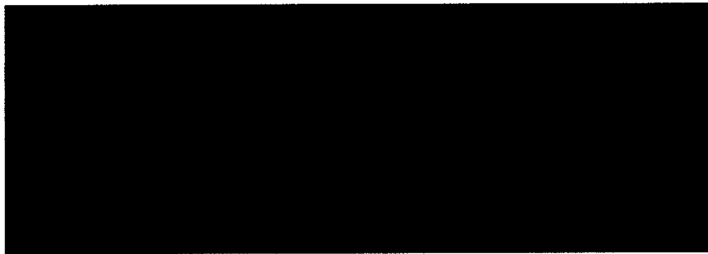


D

E

F

G



H

Figure 3