6454–6465 *Nucleic Acids Research, 2009, Vol. 37, No. 19* *Published online 26 August 2009*
doi:10.1093/nar/gkp650

Yeast genome analysis identifies chromosomal translocation, gene conversion events and several sites of Ty element insertion

Yoshiyuki Shibata, Ankit Malhotra, Stefan Bekiranov and Anindya Dutta*

Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA 22908, USA

Received April 3, 2009; Revised July 20, 2009; Accepted July 20, 2009

ABSTRACT

Paired end mapping of chromosomal fragments has been used in human cells to identify numerous structural variations in chromosomes of individuals and of cancer cell lines; however, the molecular, biological and bioinformatics methods for this technology are still in development. Here, we present a parallel bioinformatics approach to analyze chromosomal paired-end tag (ChromPET) sequence data and demonstrate its application in identifying gene rearrangements in the model organism *Saccharomyces cerevisiae*. We detected several expected events, including a chromosomal rearrangement of the nonessential arm of chromosome V induced by selective pressure, rearrangements introduced during strain construction and gene conversion at the MAT locus. In addition, we discovered several unannotated Ty element insertions that are present in the reference yeast strain, but not in the reference genome sequence, suggesting a few revisions are necessary in the latter. These data demonstrate that application of the chromPET technique to a genetically tractable organism like yeast provides an easy screen for studying the mechanisms of chromosomal rearrangements during the propagation of a species.

INTRODUCTION

The precise identification of sites of chromosomal translocations, deletions and insertions holds the promise of cataloging recurrent gene rearrangements in diseases. This has been the primary impetus for the development of paired end mapping techniques in humans (1,2). The basic principle is to identify the short sequences at the ends of linear genomic DNA fragments of a specified size. In contrast to a complete sequencing strategy, mapping the 'paired-end-tag' (PET) sequences back to the genome allows one to identify structural variations such as insertions, inversions and translocations with far fewer sequence reads. Unlike array Comparative Genomic Hybridization (a-CGH), PET analysis provides quantitative digital information with no detectable signal saturation and ability to sample rare events by sequencing more DNA. In addition, a significant advantage to sequencing is the identification of the specific genomic location of these chromosomal rearrangements, and the emphasis till now has been on identifying such structural variations in unique chromosomal sequences. In clinical samples, such disease-specific junctional fragments can be useful molecular markers for the diagnosis of various diseases. Many oncogenes can be activated by recurrent chromosomal aberration and are frequently observed in hematological and solid tumors. For example, the t(9;22)(q34;q11) translocation results in the fusion of the BCR and ABL1 genes in chronic myelogenous leukemia, and the t(8;14)(q24;q32) translocation fuses *MYC* with the immunoglobulin heavy chain gene in Burkitt lymphoma (3,4). Recently, paired-end mapping studies have revealed extensive structural variation in the human genome (5,6) and identified somatically acquired rearrangements in cancer (7–9). This raises the question of the presence and extent of similar structural variation in other model organisms. In this study, we establish the experimental and computational methodology for paired-end mapping and use it to study structural variation in *Saccharomyces cerevisiae*.

Saccharomyces cerevisiae is a useful eukaryotic model organism for establishing PET applications for several reasons. Specifically, it is easy to produce uniform population of cells, since *S. cerevisiae* grow stably as either a haploid or diploid and are genetically tractable. The *S. cerevisiae* genome is 12 Mbp, which is ~2 orders of magnitude smaller than the human genome. In addition, *S. cerevisiae* is used as a model organism for system biology and has a well-annotated

*To whom correspondence should be addressed. Tel: +1 (434) 924 1227; Fax: +1 (434) 924 5069; Email: ad8q@virginia.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

genome database. A global approach for identifying and cataloging chromosomal structural variation in *S. cerevisiae* will facilitate the dissection of how these structural variations arise in the population. We developed the molecular and bioinformatics methods to use the chromosomal paired-end tag (ChromPET) technique to identify sites of chromosomal translocations and large insertions in *S. cerevisiae*. These methods were performed in parallel with the other paired end approaches reported in the literature and contain differences that provide our method with higher specificity. To ensure that the yeast cells contained some chromosomal rearrangements that could be detected, we utilized the powerful system developed by Chen and Kolodner (10) to select for gross chromosomal rearrangements (GCRs) in the nonessential portion of chromosome V in haploid cells.

In this article, we show that the ChromPET technology successfully identifies the junction of chromosomal rearrangement in chromosome V. In addition, we identified the expected disruption of *XRS2* by *HIS3* insertion and the replacement of the *HMRa* gene into the *MAT* locus. The last was expected since the experiments were done with the a-strain of yeast. Surprisingly, we also identified several sites of Ty element insertion not reported in the reference genome sequence but present in the reference yeast strain, suggesting (i) the need for revisions in the yeast genome sequence and (ii) that the ChromPET method is useful even for detecting rearrangements involving repeat elements.

MATERIALS AND METHODS

Selection of yeast cells with GCRs

Cells which have rearrangements involving chromosome V reporter region were isolated according to the protocol established by Chen and Kolodner as well as by Schmidt *et al*. (10,11) (Figure S1A). The clone used for this study was named RDKY3671GCR.

ChromPETs library construction

The ChromPET library was constructed according to (12,13) with modifications as summarized in Figure 1A. Genomic DNA of RDKY3671GCR was isolated with Piece Y-DER (Thermo Scientific) and sheared using a sonicator. DNA fragments of 1.5–2 kb were separated on 0.8% agarose gel and purified with DE81 paper. Following polishing with the End-It DNA End Repair Kit (EpiCentre) and cleaning up with QIAquick PCR Purification Kit (Qiagen), the fragments were A-tailed with *Taq* DNA polymerase (Roche). Furthermore, 0.25 pmol of A-tailed DNA fragments were ligated with 0.8 pmol of adaptor: 5′-end phosphorylated oligonucleotide (T30MmeI) with outward-facing MmeI sites on both ends and T-tailed. Linear DNA was removed with RingMaster Nuclease (Novagen). Following extraction with phenol/chloroform/isoamyl alcohol and ethanol precipitation, the circularized DNA was amplified by rolling circle amplification with REPLI-g-Mini Kit (Qiagen). Amplified DNA was digested with 90 U of MmeI (New England BioLabs) at 37°C for 2 h and the 70-bp ChromPET fragment excised from a 0.8% agarose gel after electrophoresis. The ChromPET DNA was extracted by the crush/soak method and ends polished (End-It DNA End Repair Kit, EpiCentre).

For deep sequencing, primer sequences (UA3A and UA3B, manufacturer's recommendation) were ligated to each end of ChromPET DNA fragments. Nicks at the 3′-junctions were removed by the large fragment of Bst DNA polymerase (New England BioLabs). Adaptor-ligated library was purified on M-270 Streptavidin beads (Dynal) because UA3B adaptor had biotin at its 5′-end, and single-stranded DNA recovered in Melt Solution (100 mM NaCl and 125 mM NaOH) and purified by QIAquick PCR Purification Kit (Qiagen). The ChromPETs were PCR amplified using MMP2A and MMP3B primer for 20 cycles, and the library was gel purified for sequencing. Four hundred and fifty four sequencing was performed according to the manufacturer's protocol (Roche) and 617 602 sequencing reads were obtained.

Southern blot hybridization

Two micrograms of genomic DNA were digested with restriction enzymes and electrophoresed in a 0.7% agarose gel. DNA was transferred to a Nitran SuPerCharge membrane (Schleicher & Schuell) using alkaline denaturing condition. The membrane was hybridized with a DNA probe labeled by Ladderman Labeling Kit (TAKARA) using [^{32}P]dCTP. The probe was amplified from S288C genomic DNA with primer 3E/3F.

Primers

Sequences of Primers used in this study are listed in Table S1.

Median absolute deviation. Instead of using the SD, which is the square root of the 'average' squared deviation from the mean value, we used the median absolute deviation (MAD), which is the median of the absolute deviations from the median. The MAD is a robust and quick estimator of variability in nonparametric data, and is less affected by outliers than SD.

Bioinformatic analysis of sequence reads as summarized

- Step 1· Identify unique ChromPETs.
  - Identify linker sequence in read.
  - Extract flanking sequences as 5′- and 3′-end of a ChromPET.
  - Remove any extra exact copies of a ChromPETs.
- Step 2· Map ChromPETs.
  - megaBLAST all tags of unique chromPETs to Yeast Reference Genome (Parameters: -W 12 -a 8 -p 100 -D 2 -e 10 -m 9).
  - Parse megaBLAST output file to get address(s) of each ChromPET tag (only perfect match to reference genome kept).

- Step 3· Identify and characterize aberrant ChromPETs.
  - Plot Inter-Tag distances, to identify cutoffs that will report on aberrant ChromPETs.
  - Apply cutoff to extract aberrant ChromPETs and normal ChromPETs.
  - Classify aberrant ChromPETs as either a direct inter-chromosomal (the two tags map to different chromosomes in the correct orientation), direct insertion (two tags map to within 258 bp of each other in the correct orientation), direct deletion (tags maps >1978 bp away from each other in the correct orientation) or their inverted (the orientation of the tags is not in the same direction, the other criteria is the same as above) analogs.
  - For normal ChromPETs, only keep addresses that report on normal genome architecture and discard all other address mappings.
- Step 4· Identify high-density windows (HDWs).
  - Run a sliding window across chromosomal profile of aberrant ChromPETs, calculating sum of aberrant ChromPET and sum of aberrant ChromPETs to normal ChromPETs ratio for each window.
  - Identify windows that have high sums of aberrant ChromPETs and the ratio (aberrant/normal).
  - Establish cutoff using the median and MAD of the distribution.
- Step 5· Identify aberrant linkages.
  - For each HDW identified, extract all tags that map to that region.
  - For each tag extracted, identify all windows in which its partner tag maps, keep a count of how many times a window is hit.
  - Identify the window with the maximum number of hits, call it the maximum linkage window (MLW).
  - Plot the distribution of number of chromPETs that identify maximal linkage windows and the percentage of total chromPETs that contribute to these calls.
  - Select cutoff and report any maximal linkage window clearing the cutoff as an 'Aberrant Linkage'.
  - Identify the class and directionality of aberrant chromPET to classify the aberrant linkage into Direct/Inverted Insertion, Deletion, Interchromosomal.

RESULTS

Induction of chromosomal rearrangements at the *URA3*/*CAN1* reporter region of chromosome V

Yeast cells harboring chromosome V rearrangements were generated by selective growth on L-canavanine and 5-FOA plates as described in (10). To increase the efficiency for the gross rearrangement formation, we used RDKY3615 derived strain, RDKY3671, which has a temperature-sensitive allele (*rfa1-t33*) of the *rfa1* gene and a deletion of *XRS2* by insertion of *HIS3* (10). Rfa1 is a subunit of single-strand DNA-binding protein, RFA which is required for several phases of DNA replication and repair and *XRS2* makes a heterotrimeric endo/exonuclease complex with Mre11, Rad50 required for both homology-dependent double-strand break repair and nonhomologous end-joining. Mutations in each of these genes increased the rate of GCRs. The nonessential arm of chromosome V was used as a reporter region for chromosomal rearrangements. In strain RDKY3615, the *HXT13* gene located distal to the *CAN1* gene on chromosome V was replaced by the *URA3* gene (Figure S1A). Since *CAN1* expression sensitizes cells to L-canavanine and *URA3* expression makes cells sensitive to 5-fluorotic acid (5-FOA), cells which have inactivated both *CAN1* and *URA3* can be selected for by growth on plates containing these drugs. The resulting colonies usually contain deletions in the *CAN1*/*URA3* region of chromosome V (10).

Chromosomal breakpoints induced by this method are expected to localize within the 12.1-kb nonessential region within *CAN1* and between *CAN1* and the first essential gene *PCM1* on the left arm of chromosome V (Figure S1A). PCR amplification using primer pairs directed to the essential (primers 5A/5B) and nonessential (primers 5C/5D) regions confirmed loss of the *CAN1* locus in our strain RDKY3671GCR (Figure S1B).

Generation of ChromPET library

MmeI is a class II restriction endonuclease that digests DNA 20/18 nt away from the recognition site. Genomic DNA fragments of 1.5–2 kb in length were circularized by ligation to an adaptor sequence that has two outward facing MmeI recognition sequences. Digestion of the circularized DNA with MmeI therefore leaves two tags from the ends of the genomic sequence Chromosomal Paired End Tags (ChromPET) linked to the adaptor sequence. A ChromPET library consisting of pairs of tags from either end of a genomic DNA fragment, separated by the T30MmeI adaptor was thus constructed from strain RDKY3671GCR. Since MmeI occasionally cuts DNA at even longer or shorter distances away from the recognition sequence, we get a distribution of tags, of 12–20-bp length (see Supplementary Figure S5A). Figure S5B shows that only tags in the range of 12–20 bp long could be mapped back to the genome reliably. This library was subjected to high-throughput sequencing using the Roche 454 sequencing platform and 617 602 sequencing reads were obtained. Using the protocol shown in Figure 1B and Supplementary Data, we first identified reads which had a perfect linker sequence and then took the flanking sequences as the PETs of a chromPET. Sequences that did not contain a perfect linker sequence (49 678 reads, 8.04% of total reads) were discarded. This yielded ~17 Mb of sequence data corresponding to 567 924 ChromPETs. Once we removed any additional copies of duplicate chromPETs, i.e. two chromPETs having identical 5′ and 3′ tags, we were left with 489 479 (86.2%) 'unique' ChromPETs, which were then mapped

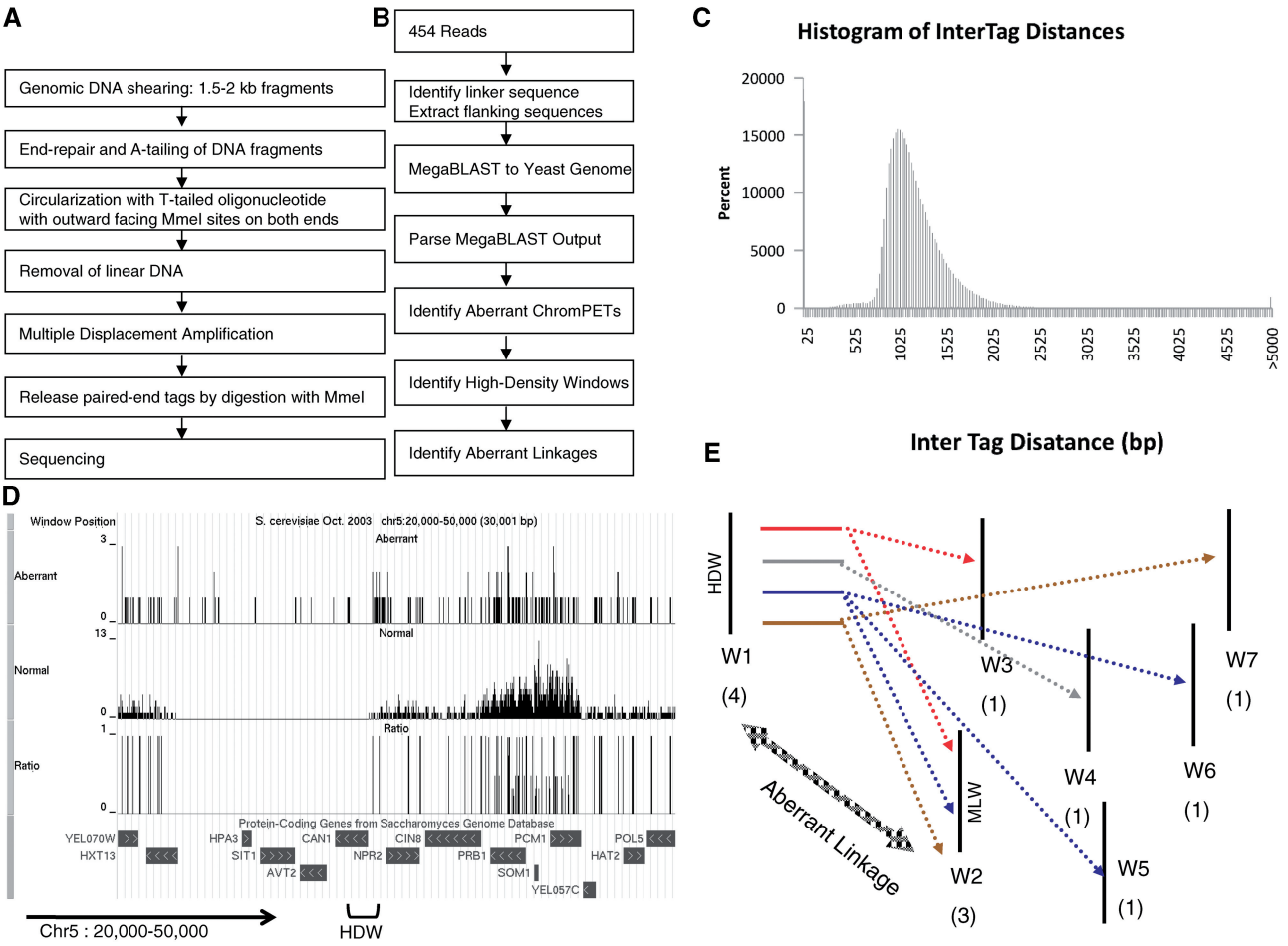

Figure 1. (**A**) Flowchart illustrating the ChromPET library preparation and (**B**) bioinformatics analysis. (**C**) Histogram showing the minimum inter-tag distances for all the ChromPETs that were mapped to the yeast genome. Any distance >10 kb was changed to 10 kb. (**D**) A UCSC genome browser snapshot showing the data from the *CAN1-PCM1* locus on chromosome V. From top to bottom, the tracks shows the number of recombinant aberrant ChromPET tags, the number of normal ChromPET tags, and the ratio of recombinant aberrant to normal ChromPET tags mapping to each base pair, and the positions of protein coding genes. (**E**) Schematic of aberrant linkage analysis. ChromPETs linking the high-density window, W1 with its partner windows (W2, W3, . . . , representing different genomic locations) are indicated by dotted lines. The number of unique ChromPETs linking each window is denoted in brackets. In this example, W2 is defined as the MLW for the HDW W1 with 3 chromPETs (and 3/4 = 75% of total chromPETs anchored in W1).

back to the yeast reference genome using MegaBLAST (details in 'Materials and Methods' section). Of the unique ChromPETs, 380 987 (77.8%) had both ends mapping back to the yeast genome, 84 256 (17.2%) had only one end mapping back to the yeast genome and 24 236 (5%) had neither of the ends mapping back to the yeast genome (Table 1).

Identification of aberrant ChromPETs

Aberrant ChromPETs are those that (i) link two different chromosomes (interchromosomal), (ii) are too close or too far from each other on the same chromosome (intra-chromosomal deletions or insertions) or (iii) inverted in orientation relative to each other on the same chromosome. While aberrant ChromPETs of types (i) and (iii) are easy to call, we had to use a statistical cutoff to call those of type (ii).

Since the ChromPET library was generated using DNA fragments of ~1.5–2 kb in size, any intra-chromosomal ChromPET whose inter-tag distance was sufficiently far from this range should be classified as an aberrant ChromPET. To examine this, we plotted the minimum inter-tag distance for all intra-chromosomal chromPETs as a histogram (Figure 1C). The distribution of inter-tag

Table 1. Number of reads and identified ChromPETs for each category

	ChromPET numbers	
	Count	Percentage
Total reads	617 602	
(i) ChromPETs	567 924	92
(ii) Unique ChromPETs	489 479	86.2
(iii) Mapped ChromPETs	380 987	77.84
One-sided ChromPETs	84 256	17.21
Unknown ChromPETs	24 236	4.95

Percentages are calculated for: (i) reads that are ChromPETs; (ii) ChromPETs that are unique; (iii) unique ChromPETs in each category based upon mapping of the two tags and the inter-tag distance.

Table 2. Number of uniquely mapped ChromPETs and their distribution into the indicated categories

	Aberrant ChromPETs	
	Count	Percentage
(i) Mapped unique ChromPETs	380 987	
Direct		
All	26.373	6.92
Deletions	7 763	29.44
Insertions	123	0.47
Inter-chromosomal	6 470	24.53
(ii) Aberrant ChromPETs		
Inverse		
Deletions	605	2.29
Insertions	137	0.52
Inter-chromosomal	6 491	24.61
Normal	123	0.47

(i) normal ChromPETs, (ii) aberrant ChromPETs and their further classification.

distances appears Poisson-like (Figure 1C) with a median of 1118 bp and MAD (Supplementary Data) of 172 bp. ChromPETs with inter-tag distances 5 MAD away from the median ($\leq$258 bp or $\geq$1978 bp) were classified as aberrant chromPETs.

Because of the short size of the tags (12–20 nt) and the presence of repeat sequences in the genome, ~30% of the tags map to multiple sites in the genome. Instead of using a heuristic to guess the most probable alignment of a tag, we examined all possible combinations of the 5′ and 3′ tag addresses of a ChromPET. If even one combination reported on a normal linkage (i.e. the two tags map with an inter-tag distance between 258 bp and 1978 bp), the ChromPET was classified as a normal ChromPET. However, 6.92% of the chromPETs could not be explained by a normal linkage (Table 2).

Of the aberrant chromPETs, ~27.89% mapped in the reverse orientation (both inter- and intra-chromosomal), and 24.53% mapped to different chromosomes in the correct orientation. Of the aberrant chromPETs, 29.44% represented direct, intra-chromosomal deletions with tags mapping to the same chromosome but separated by >1978 bp in the reference genome, while 0.47% of aberrant ChromPETs represented intra-chromosomal insertion events with tags separated by <258 bp on the same chromosome in the reference genome. The remaining chromPETs had tags, which could not be classified into one category as defined above, because they showed combinations of more than one category. Hence, these were determined as 'ambiguous' chromPETs.

Prediction of aberrant linkages representing structural variation of the chromosomes

Some aberrant chromPETs may be the result of artifactual intermolecular ligation between genomic fragments during library construction and/or mis-mapping back to the genome. To reduce such false-positive calls, we required that multiple independent chromPETs report on an aberrant linkage, and rather than using a arbitrary number (e.g. two) we chose to find this number (of multiple independent chromPETs) using a statistically rigorous approach. First, we calculated the number of normal and aberrant chromPETs that cover each base pair in the genome and then determined the ratio of aberrant to normal ChromPETs at each base pair. This removes the bias toward repeat sequences in the genome since, such regions would have a high coverage by normal ChromPETs as well. Figure 1D shows the coverage per base pair for aberrant and normal tags and the fold enrichment of aberrant ChromPET tags compared to normal tags for a 30-kb region of chromosome V. Next, we used a sliding window analysis (with a window size of 2000 bp and step size of 200 bp) to survey the number of aberrant chromPET covering each base pair and the fold enrichment of aberrant ChromPETs over normal chromPETs per base pair for the whole yeast genome. The distribution of these two variables across the entire genome is shown in Figure S2. We then selected a cutoff of one MAD higher than the median for both aberrant chromPET coverage and fold enrichment over normal to determine areas of the genome with a high density of aberrant tags (the cutoffs are indicated in Figure S2). This yielded 14 423 HDWs of 2 kb each for further analysis. Since this is an intermediate step in the analysis pipeline, we chose the most permissive cutoffs so that even windows that are moderately enriched in aberrant ChromPETs pass to the next stage of the pipeline.

For each HDW, we identified all the aberrant ChromPETs that were mapped to that window and the genomic locations to which the corresponding paired tags mapped. For example, in Figure 1E, an HDW (W1) contains multiple tags belonging to aberrant ChromPETs whose paired tag resides in various different 'partner' windows (W2, W3, . . . , W7). The partner window with the most linkages was identified as MLW.

The distribution of ChromPETs linking a HDW with a MLW for the whole genome is shown in Figure S3. To assess the statistical significance of our predicted aberrant linkages (between a HDW and the corresponding MLW), we developed a null model where all the addresses for the aberrant chromPETs were randomized, and aberrant linkages in this random population were identified (Figure S3). The linkages between an HDW and a MLW were clearly more frequent (Figure S3A) and more specific for a single MLW (Figure S3B) in the experimental (or observed) data set than in the random control. To call a significant aberrant linkage in the experimental data set, we required the number of aberrant chromPETs linking a HDW to its partner MLW to be at least 3 MAD away from the median of the experimental distribution. At least 11 chromPETs were required to link a HDW with its MLW (Figure S3A). In addition, the number of ChromPETs linking these two windows was required to represent at least 35% of the total aberrant ChromPETs present in the HDW being interrogated (35% is again 3 MAD away from the median percentage for the experimental distribution shown in Figure S3B). These cutoffs are >9 SDs above the mean of the random model, >12 SDs for the mean percentage of the random model.

Table 3. Summary of all tested aberrant linkages and their experimental validation results

Name	Type	Region 1			Region 2			Validated
		Chrom	Start	End	Chrom	Start	End	
CAN1 Locus	Inter-chromosomal	chr5	33 691	34 714	chr12	460 414	462 215	Yes
TY-SRD 1	Inter-chromosomal	chr3	146 785	149 774		TY element		Yes
TY-TY 1	Inter-chromosomal	chr3	82 706	84 962		TY element		Yes
TY-TY 2	Inter-chromosomal	chr3	166 948	170 380		TY element		Yes
TY-TY 3	Inter-chromosomal	chr12	818 104	820 010	chr7	54 0310	541 605	Yes
Ty-URA3	Inter-chromosomal	chr5	115 157	117 659		TY elementt		Yes
XRS2-HIS3	Inter-chromosomal	chr15	721 780	722 754	chr4	12 13436	121 4975	Yes
MSG5-REV3	Negative region	chr14	530 168	530 185	chr16	23 6749	236 766	No
MATALPHA-HMRA	Deletion	chr3	198 722	199 993	chr3	29 3069	295 437	Yes[a]

[a]MATALPHA deletion candidate was not validated experimentally as such a linkage is expected given the strain is of mating type a (see text).

Streamlining of predicted aberrant linkages and summary of predictions

This generated 184 aberrant linkages. For a given structural alteration, it was common to find multiple contiguous windows on each side of the alteration to be aberrantly linked to each other. Once we merged such overlapping predictions we were left with 37 aberrant linkages (shown in Table S2). In addition, when we had a unique genomic locus linked to a repeat element (like a Ty element), the unique locus appears to be linked to multiple sites, one for each site where the repeat element maps. In addition, such events are reported in both directions, doubling the number of reported linkages. After merging such unique locus-repeat element linkages, we were left with a total of 21 aberrant linkages. These linkages might represent any one of three types of structural variations—inter-chromosomal recombinations, intra-chromosomal insertions and intra-chromosomal deletions.

Of the 11 linkages reporting deletions, 10 pairs of linked sites were <3 kb apart from each other. Because this inter-tag distance was so close to the upper limit of normal inter-tag spacing (1978 bp) they either represent small deletions in the genome or arise from normal genomic architecture. The remaining one region with larger deletion, has been shown in Table 3. (All the predictions are reported in Table S2).

The inter-tag distances of all the insertion chromPETs were found to be very close to the 258 bp inter-tag distance cutoff for defining normal chromPETs (Figure 1C) and could therefore represent normal genomic fragments. PCR analysis of one of the candidate 'insertions' confirmed the reference genomic architecture (data not shown). Since several normal ChomPETs were obtained which spanned the aberrant linkages, these apparent insertions were considered to be false positives and not followed up further.

The most interesting aberrant linkages are the inter-chromosomal rearrangements. However, even in this group it quickly became apparent that several were anchored on one side by unique sequence, but were linked on the other side to a repeat element (e.g. a Ty element or telomeric repeat) so that instead of true inter-chromosomal rearrangements, they represent the insertion of a repeat element at the unique sequence site. We decided to validate by PCR both true inter-chromosomal rearrangements and a subset of the ones where a Ty element appeared to be inserted in an unique sequence (Table 3). Rearrangements where repeat elements anchored both ends of the aberrant linkage were not validated because of a difficulty in picking unique PCR primers.

Detection of expected chromosomal rearrangements

As mentioned above, RDKY3671GCR lacks a nonessential portion of the left arm of chromosome V. The majority of tags that mapped to region 33 500–35 000 of chromosome V had paired-tags, which mapped to the ribosomal DNA (rDNA) region of chromosome XII (Figure 2A). PCR amplification using a unique primer sequence from chromosome V and a primer from the rDNA repeat sequence specifically yielded a product from RDKY3671GCR genomic DNA but not genomic DNA from other strains (Figure 2B). A second PCR reaction on the amplified fragment using internal primers successfully amplified DNA (5F/12B. Figure 2C) and sequencing of this amplicon identified the breakpoint. This breakpoint was flanked by a few base pairs of homology (microhomology) (Figure 2D), consistent with previous reports that nonhomologous-end-joining using sites of microhomology are responsible for most of the translocations obtained in this system (10).

In the parental strain RDKY3671, *XRS2* on chromosome IV in RDKY3671 was disrupted by insertion of the *HIS3* gene and thus we expected to detect this aberrant linkage in our study. Indeed, many aberrant ChromPETs linked *XRS2* on chromosome IV (region 1 212 600–1 219 000) to the *HIS3* locus located on chromosome XV (Figure 3A) and we did not detect any normal tags that contained *XRS2* gene sequence. PCR primers based on the paired-tag sequences confirmed the *HIS3* insertion in the *XRS2* locus (Figure 3B).

Detection of Ty element insertions in the *URA3* gene and at several sites in chromosome III

Several of the aberrant chromosomal linkages determined computationally were anchored on one side at a unique map position but were computationally linked

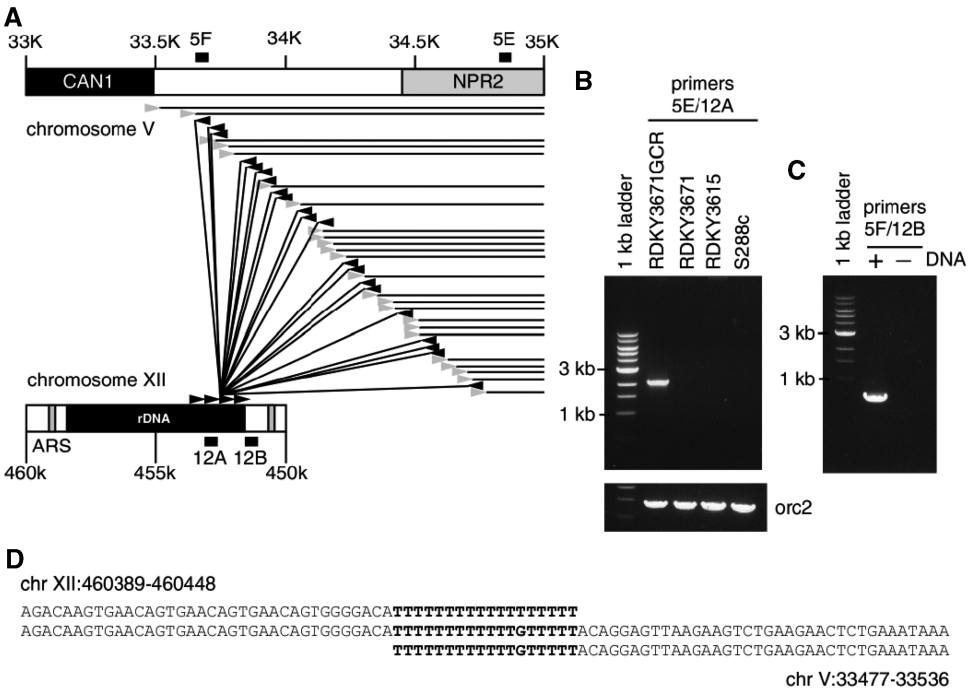

Figure 2. Aberrant PETs on chromosome V reporter region. (**A**) Many aberrant paired tags linked region 33 500–34 500 of chromosome V and the ribosomal DNA locus on chromosome XII. Chromosome V is represented at the top, and the rDNA locus on chromosome XII at the bottom. Tags belonging to aberrant ChromPETs are shown as black arrowheads. Tags belonging to normal ChromPETs are shown as gray arrowheads. The solid lines indicate linkage between the two chromosomes. Location of PCR primer pairs used for validation (5E, 5F, 12A and 12B) are shown. (**B**) Confirmation of chromosomal rearrangements by PCR analysis. Primer pair 5E/12A yielded DNA fragment using genomic DNA from RDKY3671GCR but not from the parental strain RDKY3671, the grand-parental strain RDKY3615 or the reference strain S288c. (**C**) To examine the specificity of initial PCR, a second round of PCR was performed with internal primers (5F and 12B) using initially amplified DNA as a template. (**D**) Sequence of the nested PCR product (middle) identified the break point. The corresponding sequences from chromosome XII (rDNA locus) and chromosome V are shown above and below the PCR product sequence. The site of microhomology between the two sequences in shown in bold (10).

on the other side to multiple sites in the genome. An examination of these multiple linkage sites revealed that they mapped within Ty elements, raising the possibility that these rearrangements were pointing to Ty element insertions.

The first of these types of anomalous linkages mapped on one side to chromosome V (region 115 400–117 000) near the *URA3* locus and on the other side to Ty element sequences (Figure 3C). The parental strains, RDKY3671 and RDKY3615, carry the mutant *ura3-53* allele, which is caused by a Ty element insertion within the coding region of the *URA3* gene (14). PCR amplification across this region of chromosome V from RDKY3671GCR genomic DNA yielded a DNA fragment ~6 kb larger than the predicted size fragment obtained using S288C genomic DNA (Figure 3D), consistent with the presence of a full-length Ty element insertion in the *URA3* gene.

ChromPETs were identified that linked the chromosome III region 147 000–153 000 (Figure 4A) to a full-length Ty element sequence. As a Ty element was not reported in the reference genome at this locus, we confirmed this chromosomal rearrangement using PCR. According to the reference genome, the primer pair 3E/3F should generate a PCR product of 3.8 kb; however, we obtained fragments >10 kb (Figure 4B), supporting the unexpected insertion. In order to confirm that the PCR product was derived from the correct region, we sequenced both ends of the product. Both ends mapped to the expected sites in the genome, but the internal region of the PCR amplified fragment was not present in the reference genome and was perfectly matched to Ty1 element sequence (data not shown). Considering the PCR-amplified fragment size, it is possible that there are two copies of Ty element in this region and this is supported by a previous report (15). This Ty element is present in the reference strain S288c, but as will be discussed, there is a reason why it was absent in the reference sequence.

The area of chromosome III around 83 000 nt was aberrantly linked to a Ty element sequence (Table 3). This observation was also confirmed by PCR (Figure 5A and B). DNA fragments were amplified with primer pair 3A/T1 using RDKY3671GCR genomic DNA as a template. Interestingly, PCR with S288C genomic DNA also yielded similar DNA fragments (Figure 5B), suggesting the presence of a Ty element at this locus even in S288c. In order to confirm this observation further, we designed additional primers based on Ty1 sequence and the region upstream of the predicted Ty element insertion site. DNA fragments consistent with a Ty element insertion were obtained with these primer pairs. Primer pairs that flanked the predicted insertion site amplified a fragment of 6 kb, consistent with insertion of a full-length Ty1 element (Figure 5B) and sequencing confirmed that this is the case.

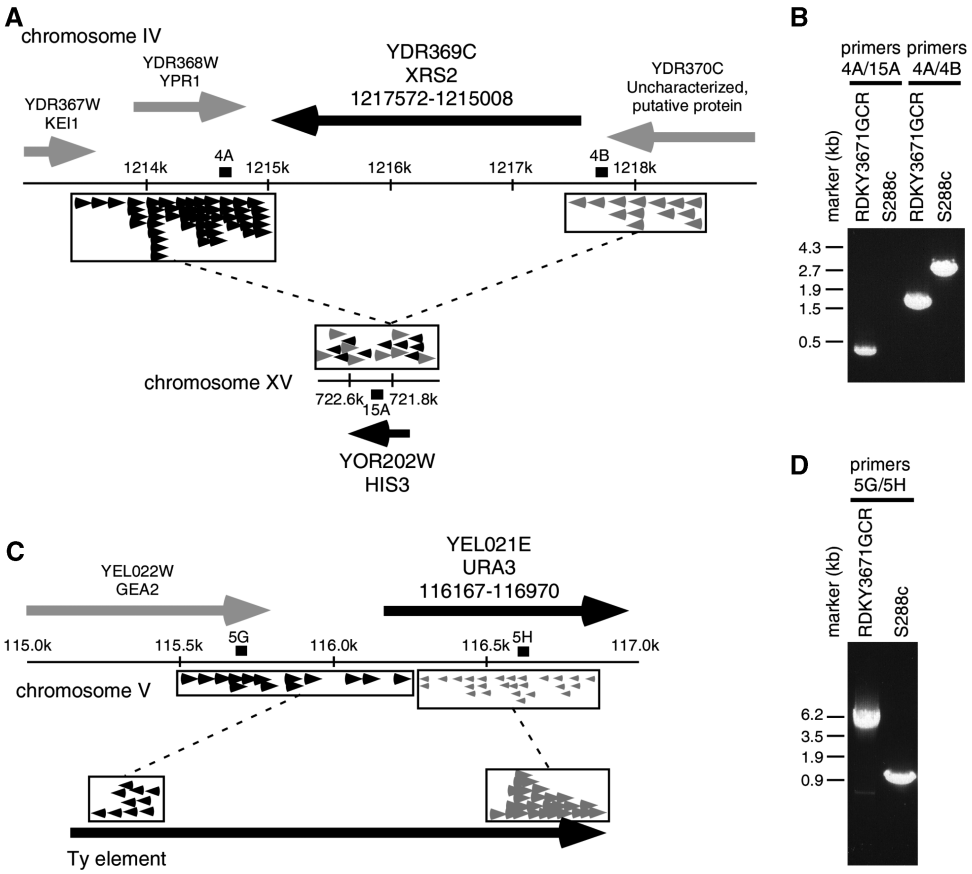

Figure 3. Aberrant PETs on *XRS2* and *URA3* locus. (**A**) Schematics of aberrant tags on *XRS2* locus. Aberrant tags which link the *XRS2* locus on chromosome IV (top) with the*HIS3* locus on chromosome XV (bottom) are shown by arrowheads. Black arrowheads designate ChromPETs that span the left recombination junction and gray arrowheads designate ChomPETs that span the right recombination junction. PCR primer pairs used for experimental validation are indicated. (**B**) Amplified DNA fragments using RDKY3671GCR and S288C genomic DNA as templates. PCR with primer pair 4A/15B (on chromosomes IV and XV) successfully amplified DNA fragment from RDKY3671GCR genomic DNA but not from genomic DNA of the reference strain S288C, confirming the presence of the identified rearrangement in the former. Additionally, PCR with primer pair 4A/4B directed to the endogenous *XRS2* locus yielded a different sized product from RDKY3671GCR genomic DNA as compared to the predicted fragment obtained with amplification from S288C genomic DNA. (**C**) Schematics of aberrant tags on *URA3* locus. Aberrant paired tags which link *URA3* locus on chromosome V and Ty element are shown by arrowheads. (**D**) Confirmation of ChromPET analysis by PCR. PCR with primer pair 5G/5H yielded different size products from RDKY3671GCR and S288c genomic DNA.

Because the Ty element at chromosome III, 83 000 nt, is missing in the reference sequence obtained from S288c, we wanted to show that the Ty element is indeed present in the S288c genome by Southern blotting independent of PCR. As shown in Figure 5C, AseI digestion should yield 2210 bp and 972 bp fragments based on the reference sequence; instead, Southern blots showed a closely migrating doublet of around 2 kb (Figure 5D). In addition, ClaI digestion yielded a 5-kb fragment instead of the expected 3-kb fragment based on the reference sequence (Figure 5D). Both of these results are consistent with the sequencing results and indicate the presence of an unannotated Ty element in this region.

Similarly, we found two other ChromPETs that did not correspond to annotated Ty elements on chromosome XII (818 200–820 400) and on chromosome III (region 169 800–171 900). Primer pairs 3G/T4 and 12C/T5 yielded amplified products (Figure 4C and D), consistent with the insertion of a full length Ty element in these regions, even in the S288c reference strain.

The aberrant linkage reporting a deletion event turned out to link the *MAT* locus to the *HMRa* locus (Figure S4), bringing them closer together than expected in the reference sequence. This is consistent with the *HMRa1* cassette being copied into the *MAT* locus, as expected in our yeast strain of mating type a.

DISCUSSION

In this article, we report a bioinformatics approach applied to PET analysis of structural changes in *S. cerevisiae* chromosomes. Our analysis effectively reported the disruption of *XRS2* with *HIS3*, the inactivation of the endogenous *URA3* with insertion of a Ty element and the copying of the *HMRa* cassette into the *MAT* locus. More importantly, we could detect a translocation in chromosome V reporter region induced by selective pressure and several unannotated transposon insertions within the genome. This may be the cheapest and most powerful approach for detecting insertions of new repeat elements.

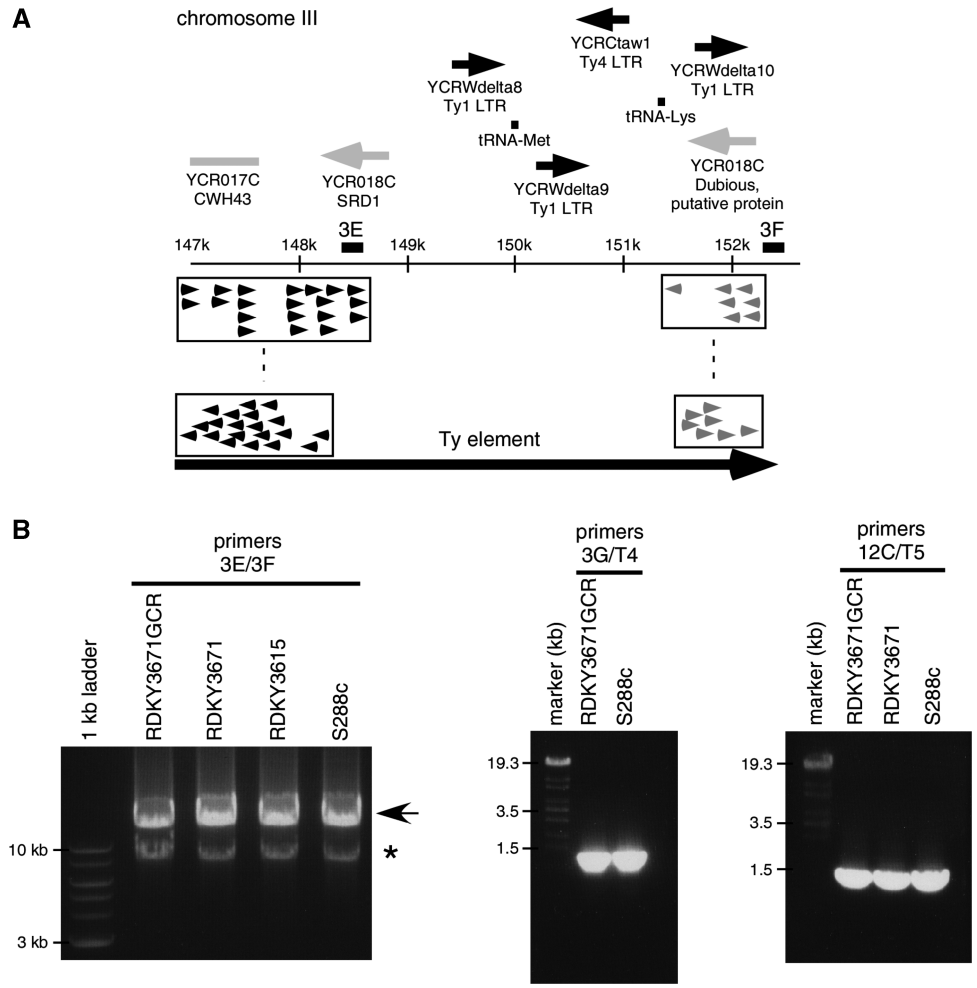

Figure 4. Aberrant PETs in RAHS locus, chromosome III (region 169 800–171 900) and chromosome XII (region 818 200–820 400). (**A**) Aberrant tags linking chromosome III position 147 000–152 000 with a Ty element are shown as arrowheads. Black arrowheads designate ChromPETs that span the left recombination junction and gray arrowheads designate ChomPETs that span the right recombination junction. PCR primer pairs used for experimental validation are indicated. (**B**) Confirmation of ChromPET analysis results by PCR. PCR amplified products with primer pair 3E/3F are larger than the size expected from the reference genome sequence. Arrow indicates PCR amplified fragments (>10 kb). Asterisk indicates a fragment synthesized as an artifact by the PCR reaction. (**C**) Confirmation of Ty element insertion in chromosome III at 169 800 by PCR. PCR primer pair 3G [chromosome III: 170 070–170 046)/T4 (chromosome XVI: 849 610–849 634 (Ty element)] yielded amplified DNA fragments using genomic DNA prepared from RDKY3671GCR and S288c as a template. (**D**) Confirmation of Ty element insertion in chromosome XII at 818 200 by PCR. PCR primer pair 12C [chromosome XII: 819 157–819 132)/T5 (chromosome VII: 540 890–540 914 (Ty element)] amplified DNA fragments using genomic DNA prepared from RDKY3671GCR, RDKY3671 and S288c as a template.

Chromosome III, Ty element insertion at 148 000

In the reference sequence, four partial Ty elements were mapped in this region, which is known to be a right-arm transposition hot-spot (RAHS) yet we detected a full-length Ty element insertion at this locus in the strain S288C, from which the *S. cerevisiae* reference sequence was mostly derived. Several, Ty element polymorphisms are, in fact, known to be present at this locus (16,17). It turns out that the reference sequence of chromosome III is a compilation of sequence from four strains related to S288c: XJ24-24a, A364A, AB972 and DC5 (16,17). Furthermore, the reference sequence contains sequence data of strain CN31c at this locus because this strain does not contain a transposon at the RAHS and instead includes an ORF and a tRNA gene in this region (18). This explains why the full-length Ty element insertion was missing in the reference sequence even though it was present in the reference yeast strain, S288c.

Chromosome III, Ty element insertion at 83 000

This insertion is also present in S288c, but not in the reference sequence. This region is a transposition hot-spot in the left-arm (LAHS) of chromosome III. The extra Ty element is located between tRNAglu and a Ty2 element. Previous DNA sequence analysis has revealed a number of sequences with homology to delta in this region and transposon insertion site profiling chip (TIP-chip) data (19) suggested the presence of Ty element here. Our ChromPET analysis and experimental validation clearly detected a Ty element insertion at this locus. In fact. PCR primer pairs designed to amplify this region yielded a 6-kb fragment and

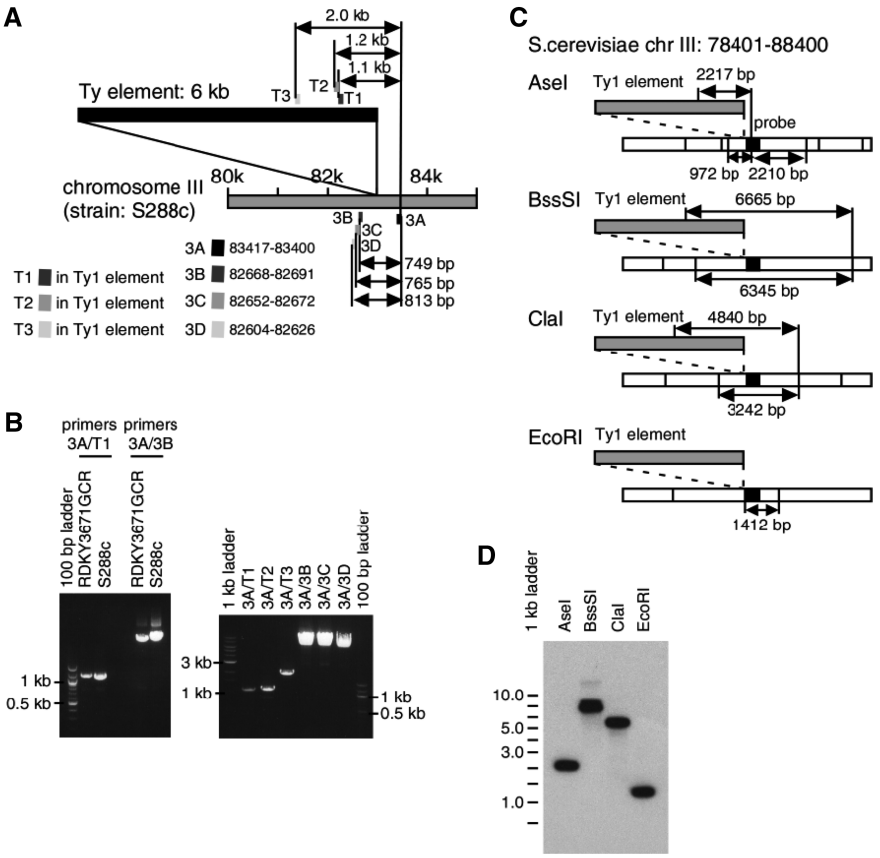

Figure 5. Aberrant PETs at LAHS locus. (**A**) Schematics of Ty insertion on chromosome III around region 83 000. PCR primer pairs used for experimental validation are indicated. Expected sizes of PCR amplified DNA fragments for confirmation of Ty element insertion are shown by double-headed arrows. (Top) Sizes with Ty element insertion. (Bottom) Sizes in the absence of a Ty element. (**B**) Confirmation of Ty element insertion by PCR. Primer pairs 3A/T1, 2 or 3 and 3A/3B, C or D yield DNA fragments of size in agreement with Ty element insertion. (**C**) Schematics of expected size of DNA fragments by restriction enzyme digestion. White bars show chromosome III region 78 401–88 400 and gray bars represent Ty element. Vertical lines show restriction enzyme sites. Black square shows the probe used for Southern blot hybridization in (**D**). (D) Confirmation of Ty element insertion by Southern blot hybridization. Sizes of DNA fragments that hybridize with the probe are consistent with Ty element insertion.

the sequence of the PCR amplified fragments is very similar to Ty1-1. Additionally, the result of the Southern blot analysis (Figure 5D) also supports the presence of an extra Ty element. We propose that there are two copies of Ty element at the LAHS and the sequence data in the reference genome may have assembly errors.

Ty element insertion at 169 800 of chromosome III and at 818 200 of chromosome XII

These insertions are not reported in the reference sequence, but again are present in the reference strain S288c. In fact, we can detect four of six unannotated Ty 'elements' which were reported by Wheelan *et al.* (19). One of the previously reported insertions is localized to the rDNA locus on chromosome XII. The fact we did not detect this insertion can be explained by the difficulty in mapping aberrant ChromPETs when both tags mapped to repetitive sequences. Another Ty element insertion reported by Wheelan *et al.* (19), which maps to chromosome X, was not detected by our analysis and therefore could be specific to the FY2 strain used in their study.

Chromosome III, MAT locus

Haploid yeast cells switch mating type by replacing the information present at the *MAT* locus. This depends on the presence of the two silent mating-type cassettes, *HMLα* on the left arm and *HMRa* on the right arm on chromosome III, respectively. A gene conversion event following the cut by the HO endonuclease at the MAT locus copies either the *HMLα* or *HMRa* gene to the *MAT* locus. As shown in Figure S4, we found 12 aberrant chromPETs linking the *MAT* locus with the *HMRa*, sites that are ~100 kb apart in the reference sequence. The mating type of our strain is *MATa* so that *HMRa* is expected to be copied into the *MAT* locus. The ability to detect even a physiological gene conversion event confirms the comprehensiveness of this method in detecting structural rearrangements in chromosomes.

Bioinformatics analysis

By pooling the tags into windows, we hoped to bypass complications stemming from sequencing errors and mis-mapping. We chose to keep all possible mappings of a tag

in our pipeline to improve the sensitivity of the method. Because of our reliance on windows that have the possibility of including multiple aberrant chromPETs, we can call aberrant linkages in a statistically rigorous way providing more specificity in our results. The ChromPET technology in higher eukaryotes is revealing hundreds of 'abnormal' linkages in 'normal' cell lines, much higher than expected (unpublished data). This is why we biased our bioinformatics analysis toward finding true positives. The null model tested in the current approach gives the lower bound on the permissiveness of the cutoffs and sets the threshold beyond which our predictions would not happen by chance. This turns out to be correct: our true positive rate is very respectable. Obtaining an accurate estimate of the false-negative rate will require more extensive sampling of the genome by PCR and by additional ChromPET screens and so we do not intend to be too categorical about the sensitivity of our method.

Although we have not validated the intra-chromosomal deletions or insertions because most of them could have arisen from normal variation in fragment size distribution, the statistics for determining the cutoffs for calling insertions or deletions was interesting. Since the inter-tag distance distribution looks like a Poisson distribution, we elected to use Median and MAD to establish the cutoffs. An alternate approach could be to estimate the size and distribution of the input DNA fragments that are from the normal population (excluding fragments reporting abnormal linkage and virtual fragments arising from mis-mapping). The mean and SD from this normal population could be used to establish the thresholds for calling insertions or deletions.

The identification of cancer-associated chromosomal aberrations using ChromPET technology will be a powerful tool for diagnosis of cancer and for the elucidation of how translocations contribute to cancer progression. The ChromPET technology is more cost efficient than whole genome sequencing for screening for such translocations. For example, in the current study, with ~380 000 usable reads (without linker sequencing errors and without redundancy) we obtained ~33× coverage of the genome by fragments whose ends are sampled by the PETs. To obtain similar coverage of the genome by complete genome sequencing we need ~2 100 000 reads of 300 bases each. This technology also has the advantage that it can identify aberrations that are currently unidentified because of the difficulty of establishment of cultured cell lines. This technology was highly successful in identifying new sites of insertion of repeat sequences in the yeast genome and so may reveal transposition of repeat elements when applied to clinical specimens. However, as the data from the yeast genome reveals, studies on human cancer will need to take into consideration the many naturally occurring variations in human genome sequences for detecting cancer-specific rearrangements. Thus, it is important to collect a large amount of sequence data in order to identify chromosomal aberrations that are specific to a cancer and, where possible, compare the translocation site with the normal genomic DNA from the same patient.

In conclusion, the ChromPET technology is a powerful and cost-effective method to identify chromosomal aberrations. The accuracy of the data depends on the number of paired-tags recovered for a given rearrangement. The more independent paired-tags report an abnormal linkage, the higher the reliability of the screen. A PCR-based post-sequence analysis step is another tool that will help exclude chimera products that are expected to contaminate such ChromPET libraries. In addition, the application of this technology to simpler genomes like yeast has already revealed sites where the reference sequence needs to be revised and will allow us to understand the mechanisms of genomic instability in greater detail.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Mignon Keaton, Adam Mueller, Neerja Karnani and the other members of the Dutta laboratory for critical help, insight and advice.

FUNDING

National Institutes of Health (R01 CA60499 and CA89406 to A.D.). Funding for open access charge: CA89406.

Conflict of interest statement. None declared.

REFERENCES

1. Wei,C.L., Wu,Q., Vega,V.B., Chiu,K.P., Ng,P., Zhang,T., Shahab,A., Yong,H.C., Fu,Y., Weng,Z. *et al.* (2006) A global map of p53 transcription-factor binding sites in the human genome. *Cell*, **124**, 207–219.
2. Ng,P., Tan,J.J., Ooi,H.S., Lee,Y.L., Chiu,K.P., Fullwood,M.J., Srinivasan,K.G., Perbost,C., Du,L., Sung,W.K. *et al.* (2006) Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. *Nucleic Acids Res.*, **34**, e84.
3. Albertson,D.G., Collins,C., McCormick,F. and Gray,J.W. (2003) Chromosome aberrations in solid tumors. *Nat. Genet.*, **34**, 369–376.
4. Mitelman,F., Johansson,B. and Mertens,F. (2007) The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer*, **7**, 233–245.
5. Korbel,J.O., Urban,A.E., Affourtit,J.P., Godwin,B., Grubert,F., Simons,J.F., Kim,P.M., Palejev,D., Carriero,N.J., Du,L. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.
6. Kidd,J.M., Cooper,G.M., Donahue,W.F., Hayden,H.S., Sampas,N., Graves,T., Hansen,N., Teague,B., Alkan,C., Antonacci,F. *et al.* (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**, 56–64.
7. Campbell,P.J., Stephens,P.J., Pleasance,E.D., O'Meara,S., Li,H., Santarius,T., Stebbings,L.A., Leroy,C., Edkins,S., Hardy,C. *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.*, **40**, 722–729.
8. Bignell,G.R., Santarius,T., Pole,J.C., Butler,A.P., Perry,J., Pleasance,E., Greenman,C., Menzies,A., Taylor,S., Edkins,S. *et al.* (2007) Architectures of somatic genomic rearrangement in human

cancer amplicons at sequence-level resolution. *Genome Res*., **17**, 1296–1303.
9. Bashir,A., Volik,S., Collins,C., Bafna,V. and Raphael,B.J. (2008) Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS Comput. Biol*., **4**, e1000051.
10. Chen,C. and Kolodner,R.D. (1999) Gross chromosomal rearrangements in Saccharomyces cerevisiae replication and recombination defective mutants. *Nat. Genet*., **23**, 81–85.
11. Schmidt,K.H., Pennaneach,V., Putnam,C.D. and Kolodner,R.D. (2006) Analysis of gross-chromosomal rearrangements in Saccharomyces cerevisiae. *Methods Enzymol*., **409**, 462–476.
12. Shendure,J., Porreca,G.J., Reppas,N.B., Lin,X., McCutcheon,J.P., Rosenbaum,A.M., Wang,M.D., Zhang,K., Mitra,R.D. and Church,G.M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**, 1728–1732.
13. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z. *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
14. Rose,M. and Winston,F. (1984) . Identification of a Ty insertion within the coding sequence of the S. cerevisiae URA3 gene. *Mol. Gen. Genet*., **193**, 557–560.
15. Umezu,K., Hiraoka,M., Mori,M. and Maki,H. (2002) Structural analysis of aberrant chromosomes that occur spontaneously in diploid Saccharomyces cerevisiae: retrotransposon Ty1 plays a crucial role in chromosomal rearrangements. *Genetics*, **160**, 97–110.
16. Warmington,J.R., Green,R.P., Newlon,C.S. and Oliver,S.G. (1987) Polymorphisms on the right arm of yeast chromosome III associated with Ty transposition and recombination events. *Nucleic Acids Res*., **15**, 8963–8982.
17. Newlon,C.S., Lipchitz,L.R., Collins,I., Deshpande,A., Devenish,R.J., Green,R.P., Klein,H.L., Palzkill,T.G., Ren,R.B., Synn,S. *et al*. (1991) Analysis of a circular derivative of Saccharomyces cerevisiae chromosome III: a physical map and identification and location of ARS elements. *Genetics*, **129**, 343–357.
18. Oliver,S.G., van der Aart,Q.J., Agostoni-Carbone,M.L., Aigle,M., Alberghina,L., Alexandraki,D., Antoine,G., Anwar,R., Ballesta,J.P., Benit,P. *et al*. (1992) The complete DNA sequence of yeast chromosome III. *Nature*, **357**, 38–46.
19. Wheelan,S.J., Scheifele,L.Z., Martinez-Murillo,F., Irizarry,R.A. and Boeke,J.D. (2006) Transposon insertion site profiling chip (TIP-chip). *Proc. Natl Acad. Sci. USA*, **103**, 17632–17637.