CORE
Metadata, citation and similar papers at core.ac.uk
Provided by DSpace@MIT

Ruler Arrays Detect Genomic Insertions and Deletions

by

Philip Alexander Rolfe

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2009

©2009 Massachusetts Institute of Technology. All rights reserved.

Author . . . .
Department of Electrical Engineering and Computer Science
May 11, 2009

Certified by . . . .
David K. Gifford
Professor
Thesis Supervisor

Accepted by . . . .
Terry P. Orlando
Chairman, Department Committee on Graduate Theses

Ruler Arrays Detect Genomic Insertions and Deletions

by

Philip Alexander Rolfe

Submitted to the Department of Electrical Engineering and Computer Science
on May 11, 2009, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science and Engineering

Abstract

A Ruler Array measures the distance between a set of microarray probes and a set of experimentally defined locations in a nucleic acid, offering new possibilities for locating and characterizing changes in the nucleic acid sequence. Despite the known relevance of genomic changes to pathogens, cancer, development, and evolution, many of these changes evade detection by existing high-throughput techniques. Since a microarray can interrogate thousands or millions of probes at once, Ruler Arrays can screen a small genome or part of a mammalian sized genome for insertions, deletions, and inversions in a single experiment.

Thesis Supervisor: David K. Gifford
Title: Professor

Acknowledgments

I thank Paula Grisafi and Douglas Bernstein, both from Dr. Fink's lab at the Whitehead Institute, for all of their work on this project. They performed all of the wet lab experiments (Ruler Array, PCR, etc) for this thesis; they are the "we" used throughout.

Robin Dowell provided the Σ1278b genome assemblies and performed most of the annotation and curation. She also provided the alignments and indel predictions between Σ1278b and S288C.

Contents

1 Introduction **23**

1.1 Genotyping Technologies 24

1.1.1 SNP Detection 24

1.1.2 TIP-Chip 25

1.1.3 CGH and FISH 25

1.1.4 Array-CGH 26

1.1.5 Sequencing 26

1.2 Our Contribution 27

2 Ruler Arrays **29**

2.1 Ruler Array Theory 29

2.2 Advantages of Ruler Arrays 31

2.3 Weaknesses of Ruler Arrays 35

3 Ruler Array Laboratory Protocol **37**

3.1 Overview 37

3.2 Digestion 38

3.3 Ligation 41

3.4 Extension and Labeling 41

3.5 Hybridization 43

4 Predicted Intensity Shape **45**

4.1 Extension Termination 45
4.2 The Shape 46

5 Predicted Ratios from Insertions 49

6 Log Likelihood of the Data 55

6.1 Model for Probe Intensities 55
6.2 Estimating the Variance of Intensity Observations 56
6.3 Penalizing Systematic Error 58
6.4 Limiting Outlier Influence 59

7 Theoretical Limits of Detection 61

7.1 Highly Simplified Indel Detection 61
7.2 Ratio Difference 62
7.3 Intensity Difference 62
7.4 Conclusion 67

8 Analyzing Ruler Array Data Channel Ratios 71

8.1 Hidden Markov Models 71
8.2 Our Model 72
8.3 HMM Learning 73
8.4 Evaluation 73

9 Analyzing Ruler Array Data with Segment Fitting 77

9.1 Segment Fitting 78
  9.1.1 Recursive Solution 79
  9.1.2 Correctness 79
  9.1.3 Priors on Segment Fitting 80
  9.1.4 Dynamic Programming Solution 80
9.2 Segment Fitting with Linear Regression 81
9.3 Runtime of Segment Fitting with Linear Regression 83

9.4 Handling Experimental Replicates 84

9.5 Independently Handling Two Channel Experiments 84

9.6 Jointly Handling Two Channel Experiments 85

9.7 A Non-Generative Model for Segment Fitting 86

9.8 Segment Fitting Efficiency Hacks 86

9.9 Calling Insertions and Deletions from Segment Fitting Output 87

9.10 Conclusion 88

10 Evaluating the Ruler Array **89**

10.1 Technical Evaluation Method 90

10.2 Analysis Parameters 91

10.3 The Test Set 92

10.4 Evaluating the False Negatives 94

10.4.1 1:182380-183125 95

10.4.2 2:643285-644061 and 2:644726-645126 96

10.4.3 4:434241-435318 98

10.4.4 4:1023149-1023496 99

10.4.5 11:388578-388978 100

10.4.6 Lessons from the False Negatives 101

10.5 Synthetic Diploid Experiment 101

10.6 Technical Evaluation of the Ruler Array 102

11 Biological Test Cases **103**

11.1 TY Elements 103

11.2 Repeat Length Changes 104

11.3 Gene Family Expansions and Contractions 105

11.3.1 Cup 105

11.3.2 Pho 108

11.3.3 Mal 109

11.4 Gross Rearrangements 110

12 Comparison to Other Indel Detection Techniques 113

12.1 Comparison to TIP-Chip 113

12.2 Comparison to aCGH 114

12.3 Comparison to Sequencing Methods 114

12.3.1 Long Read Assembly 119

12.3.2 Short Read Assembly 120

12.4 Evaluation Summary 122

13 Extensions of the Ruler Array Technique 123

13.1 Ruler Seq 123

13.2 Targeted Assembly 126

13.3 TIP-Seq 126

14 Future Work 127

14.1 Polymerase Characterization 127

14.2 Labeling Sites 128

14.3 Screening Closely Related Strains for Indels 128

14.4 Checking Genome Assemblies 128

14.5 Biological Significance of Repeat Length Changes 129

14.6 Ruler Arrays Expand Toolset for Discovering Genomic Differences 129

A Ruler Array Laboratory Protocol for *S. cerevisiae* 131

A.1 Growing Cells 131

A.2 DNA Extraction 131

A.3 Digestion 132

A.4 Ligation of Biotin Linker to Digested DNA 133

A.4.1 Making Biotin Linker 133

A.4.2 Ligation 133

A.4.3 Cleanup and Binding to Beads 134

A.5 Polymerase Extensions 134

A.5.1 Cy3/Cy5 and ExTaq 134

A.5.2 ULS 134

A.5.3 Aminoallyl-dUTP 135

A.5.4 Vent Exo- 135

A.5.5 Extensions 135

A.5.6 Isolating DNA 136

A.5.7 ULS Labeling 136

A.5.8 Aminoallyl Labeling 136

A.6 Hybridization 136

List of Figures

1-1 Sequencing and assembly methods have difficulty resolving repetitive sequences whose length is greater than the read length. In the top panel, two contigs A and B end in a repetitive element R while two other contigs X and Y begin with R. If the read length is less than the length of R, then unpaired reads will not be able to join A to X or Y as no read can contain unique sequence from A as well as unique sequence from X or Y. Consequently, an instance of R in the genome will result in a contig boundary. Paired-end reads will allow contigs to be joined if the insert size is greater than the size of R and if a pair spans R with both ends landing in unique sequence. The bottom panel shows how a tandem repeat that would confound single-end reads if R is longer than the read size; if the experimental sequence contains two adjacent repeat units then the assembler cannot determine the number of repeat units. Paired-end reads spanning R and with both ends in unique sequence provide probabilistic information, but not absolute information, about the copy count of R. Comparing the observed distance between ends than span the repeats to the expected distance indicates whether the assembly is correct. Similarly, comparing the observed to expected distance between a read in the unique sequence and a read in an instance of R provides information about whether the copy count of R in the assembly is correct. 28

2-1 A schematic of two ruler illumination sites and the resulting probe intensities observed by probes designed against one strand of a nucleic acid. At each labeling site, the intensities jump to some high level and then decrease as one moves farther from the illumination site. The locations of the labeling sites and the shape and length of the falloff in intensity depend on the details of the laboratory protocol. 30

2-2 Schematic Ruler Array data at an insertion. The top panel shows probes mapped to the reference sequence. The bottom panel shows that, when the reference sequence is modified to contain the insertion in the sample sequence, the observed intensities follow the expected shape. Note that this analysis requires no knowledge of the inserted sequence. 32

2-3 Schematic Ruler Array data at a deletion. As in aCGH, the probes corresponding to the deleted sequence produce little intensity since very little sample material hybridizes to them. Unlike aCGH, however, the probes beyond the change confirm the loss of material by producing higher intensities than one would expect in the absence of a deletion. 33

2-4 Schematic Ruler Array data at an inversion. As with insertions and deletions, inversions are recognizable because the observed relationship between probe intensity and distance from the labeling site does not match the expected relationship. 34

3-1 The Ruler Array laboratory protocol consists of four steps: digestion of the nucleic acid input material by a restriction enzyme, ligation of an adapter molecule onto the resulting sticky ends, generation of labeled material, and hybridization to a microarray 39

3-2 Sample Ruler Array data from yeast. The very bottom track shows EcoRI restriction sites as tick marks; these were the labeling sites for this experiment. Red marks in the main track indicate intensities from the S288C sample and green marks indicate intensities from the Σ1278b sample. The intensities are highest at the EcoRI site and fall off gradually over several kilobases (the intensity scale is on the right side). This array tiled only one strand, so the equivalent falloff in material on the opposite strand does not appear. The blue dots under the intensities show the ratio between S288C and Σ1278b (the ratio scale is on the left side). 40

3-3 The digestion and ligation steps of the Ruler Array protocol attach an adapter molecule to specific genomic loci. The adapter is biotinylated such that the ligated material can be separated from the unligated material. 42

4-1 Plot of predicted log-intensity vs distance for intervals of size 1000, 2000, 4000, and 8000. The probability of termination at any base is .001 in all four intervals. Note the relatively linear shape over most of the interval followed by a more rapid decrease as the end effects become dominant. 47

5-1 Predicted ratios in a 2000bp interval. Each plot shows an insertion of some size (200, 600, or 1000bp) at some distance from the restriction site (100,300,500,or 700bp) when the peak ratio with no insertion was one. 51

5-2 Predicted ratios in a 4000bp interval as in figure 5-1. 52

6-1 Plot of probe standard deviation vs intensity from the whole cell extract channel of ChIP-Chip experiments, measured in fluorescence units as reported by Agilent's scanner and feature extraction software. The experiments were normalized to have the same median intensity. While the noise in a ChIP-Chip experiment may not be identical to that in a Ruler Array experiment, we expect them to be similar and this plot does indicate that the relationship between the standard deviation of a probe's intensity and its intensity is close to linear. We have experimented both with a linear model and a piecewise linear model based on this data but found that the piecewise linear model offers relatively little improvement in the Ruler Array's overall performance. 57

6-2 The interpolation term in the variance estimate depends on the difference between the observed value for a probe (solid circle) and the value predicted by linear interpolation of the adjacent probes (hollow circle). When a set of probes fall closely along a line, this term of the variance will be small as the probe observations are consistent. When a probe falls far from the predicted value, this term of the variance will be large for that probe to reduce its weight. Computing the interpolated value using only adjacent probes will also downweight the probes on either side of the noisy probe. One might therefore use a larger set of probes near probe i, perform linear regression on those probes without i, and then compute the interpolated value. 58

6-3 Independently scoring probes may fail to distinguish between a good fit to noisy data and systematic error. On the left, the data clearly falls along a single line but seems noisy. On the right, the data is clearly drawn from a two different models; however, when fit with a single line, some probes fit well and others poorly. We use a scoring term that looks at the signs of the residuals to penalize models that seem to consistently over or under predict the true values over large, continuous regions. 59

8-1 Six state HMM to model the channel ratio in Ruler Array data. The states represent the underlying sequence around each probe: no change between the samples, an indel, or an added restriction site. The emissions from the HMM are the observed array intensities and the ratio between the two channels. The two "channels same" states represent the most common case in Ruler Array data- no difference between the two samples. The low intensity state permits a different distribution over the channel intensities when no restriction site illuminates a probe. The changed states represent either insertions or single strain restriction sites; both change the ratio between the two channels but will tend to have different intensities- indels can have high intensity in both channels whereas single strain restriction sites yield high intensities only in one channel. 72

8-2 Ruler Array extensions may result in exponential PCR amplification. In the standard or expected case of Ruler Array extensions, the amount of product increases linearly with the number of cycles- at most one product molecule comes from each genomic template during a cycle. If the products from opposite strands of an interval anneal during a later extension cycle, they may complete the extension to include the complement to the primer (i.e. the adapter) on both ends. In future cycles, this material may begin exponential amplification as in PCR, producing both full length product and partial products. The amount of product in the interval will therefore depend heavily on when the first product-product extension takes place Because the initial annealing of product material is probabilistic, the amount of product is therefore probabilistic and may differ between the two samples in the Ruler Array experiment, leading to an intensity ratio far from one. 75

9-1 Sample segment fitting results from the segment of chromosome seven shown in figure 3-2. The red and green marks indicate the datapoints and the purplish marks show the fitted line segments. 78

9-2 The four cases in which the Ruler Array analysis infers the presence of an indel from the segment fitting output. In (a), the segment fitting used one segment to fit the green channel but two segments to fit the red channel; consequently, the analysis makes a call at the split point in the red channel. In (b), the segment fitting used two segments in each channel. The green channel is greater to the right of the break but of lower magnitude to the left. If the change is large enough, the analysis calls this boundary an indel. This change is commonly observed at AT repeat length changes. Example (c) illustrates another change common at repeat length or repetitive element changes. There is a segment boundary in both channels, but the intensities drop much more in one channel than the other. A restriction site, or the insertion of an element that contains a restriction site such as a TY, generates the signature seen in (d). 88

11-1 The top panel shows the Ruler Array data (S288C in red, Σ1278b in green) over part of chromosome 15. The Σ1278b intensities fall suddenly over an AT repeat at the transcription stop of the THP1 gene whereas the S288C intensities continue a linear decline. The bottom panel shows the sequencing results for this locus; each strain was sequenced in both directions. The 14bp expansion of the AT repeat in Σ1278b seems the likely cause of the sudden intensity drop. 106

11-2 This example is similar to the previous AT repeat length change, though in this case the repeat expands by only 2 base pairs (one AT unit). Interestingly, the magnitude of the difference between the log-intensity drops across this repeat is greater than in the previous example. 107

11-3 The *Cup1-1*, *Cup1-2* locus. While these genes and the region around them are not tiled, the Ruler Array shows evidence of a change by the high ratio observed to the left of *Cup1-1*. 108

11-4 The *Pho3*, *Pho5* locus. The Ruler Array intensities over *Pho3* seem noisy and don't follow the expected falloff pattern. Since many genes in the *Pho* family exhibit high similarity, the intensities are not uniformly low in Σ1278b; however, enough probes detect the deletion to allow the analysis to identify the deletion of *Pho3* in Σ1278b. 109

11-5 The *Mal33*, *Mal31*, *Mal33* locus. Three copies of *Mal* genes have been inserted in Σ1278b between *Mal33* and *Mal31*. The sudden change in ratio over *Mal31* reveals the change. 110

11-6 The left arm of chromosome six has moved to the left arm of chromosome ten between S288C and Σ1278b. While the Ruler Array can't determine what moved where, it does make evident the sites at which some change occurred. The upper panel shows chromosome six; the break point is at 30kb. The lower panel shows the break around 24kb on chromosome ten. 111

12-1 While the Ruler Array data (top track) over the left arm of chromosome 6 clearly shows the location of the translocation between chromosomes 6 and ten in Σ1278b (at 30kb, marked with a black arrow), the aCGH data in the bottom track shows no difference. In the aCGH plot, the FY4 intensities are green and the Σ1278b intensities are red; the ratio is shown in blue. Both methods clearly show a deletion in Σ1278b at the left edge of the plot. . . . 115

12-2 The Ruler Array (data in top track) successfully detects the insertion of roughly 100bp on chromosome eight while the unique probes in the aCGH data show no difference. 116

12-3 The Ruler Array (data in top track) successfully detects the insertion of a TY element on chromosome eleven while the unique probes in the aCGH data show no difference. While the CGH data does show a difference in ratio over repetitive elements such as the TY family, it cannot localize the changes to particular insertion sites such as this one. 117

12-4 Both the Ruler Array and aCGH correctly detect the deletion of parts of the right arm of chromosome sixteen in Σ1278b (note that the channels are reversed between the two experiments). The low intensities and low ratio make the deleted regions obvious in both experiments. 118

13-1 The Ruler Seq protocol generates fragments in the same way as the Ruler Array Protocol. However, instead of labeling the fragments with fluorescent dyes, the Ruler Seq protocol ligates adapters to both ends of the fragment and then sequences from the 3' end. Mapping the read sequences to the genome produces the location and strand of the read. 124

13-2 Mapping the Ruler Seq reads to the genome produces stranded locations. After each read is extended back to the restriction site from which its fragment came, the number of reads crossing any point is the virtual intensity at that point. While one could generate an intensity at every base pair, those intensities would be repetitive. Instead, we generate a virtual intensity measurement at every position to which one or more reads align. 125

List of Tables

5.1 Difference in ratio across an insertion of the specified size at the given position in intervals of 1kb, 2kb, 4kb, and 8kb. The number shown is the channel ratio before the insertion (closer to the restriction site) divided by the ratio beyond the insertion. A value of 1.0 indicates no difference; a smaller value indicates more difference in ratios and therefore a more easily detectable insertion. . . 53

7.1 The p-value for detecting a difference in ratio given a single insertion of the specified size. These values were computing assuming a tiling density of one probe per 55bp and a standard deviation of the ratios of .66.. While the statistical test is performed under the null hypothesis that the mean ratio in each part of the interval is the same, our computational experiment is performed knowing that this is not the case. Hence a small p-value leads us to reject the null hypothesis and correctly detect the indel. In the case of no-indel, the p-value is the probability of a false positive call. These results indicate, for example, that a 300bp indel should be detectable in a 2kb interval but that a 70bp change will probably be missed. 63

7.2 The p-value for detecting a difference in ratio given a single insertion of the specified size with a probe spacing of 10bp. 64

7.3 The p-value for detecting a difference in ratio given a single insertion of the specified size with a probe spacing of 1bp. 65

7.4 Maximum difference in log-likelihood for various insertions with a probe spacing of 55bp and $\sigma = .3 \cdot \mu$. 68

7.5 same thing again for probe spacing 10bp 69

10.1 The 35 indels that must be found by the Ruler Array analysis. Alignments of the curated Σ1278b assembly to the S288C reference sequence predicted each indel, which we then confirmed with PCR, CGH, or chromoblot. 93

12.1 Indels not found by comparing the short read assembly to the S288C reference sequence. 121

Chapter 1

Introduction

We aim to improve the available techniques for studying insertions, deletions, and inversions in genomic samples. While much previous work has focused on SNPs and on certain types of structural changes, no single current technique can detect the full range of changes and changes of moderate size (tens to thousands of nucleotides) have received relatively little attention despite their obvious importance. For example, numerous studies have employed variable length repeats as genetic markers in humans and other species[13, 52, 16]. Repeat sequences have been implicated in several neural diseases such as spinal and bulbar muscular atrophy and Huntington's disease[8]. A recent study of the *Gallus gallus* genome indicated that length polymorphisms account for 10% of the polymorphism events and 20% of the polymorphic bases between three strains of chicken[6]. A recent resequencing in human found hundreds of thousands of indels between the sample and NCBI reference sequence[15].

In a more dynamic context, genomic changes play a key role in evolution and may influence vertebrate development as well[40]. Pathogens modify their genome to evade detection by a host immune system[37, 33]. Numerous cancers correlate with genomic changes[43, 36, 39]. Past transposon activity may influence gene regulation[31]. Most interestingly, some recent studies have linked genomic changes of repetitive elements with differentiation during vertebrate development[34, 22, 50, 38].

1.1 Genotyping Technologies

Traditional genotyping techniques use karyotypes, chromoblots, and fluorescent in-situ hybridization (FISH) to detect gross chromosomal changes and long-read sequencing to detect small indels and single nucleotide polymorphisms (SNPs). Microarray technologies and high throughput sequencing have enabled the detection of small, even single base pair, changes across the genome and at millions of loci simultaneously.

FISH and Comparative Genomic Hybridization (CGH) can detect insertions and deletions in nucleic acid samples. However, these techniques typically targeted either a few genomic loci or detected only gross changes (megabases or more) across the entire genome. More recently, microarray technology has enabled high resolution CGH that can interrogate potentially thousands of genomic loci at a resolution of hundreds of bases. Microarray technologies have also been employed to detect SNPs and transposon insertion sites. Sequencing genomic material from the sample of interest can also locate certain insertions and deletions but is not yet a practical approach to quickly and cheaply locate changes in mammalian-sized genomes. For example, 5X coverage of human would require fifteen million kilobase reads, making sequencing many individuals too expensive even at ten cents per read. Recently, 454 Life Sciences claimed to have sequenced a full human genome for only two million dollars, which is still impractical if one wishes to study a limited set of genomic regions in a large number of individuals.

1.1.1 SNP Detection

SNPs are changes in a single base pair of a genomic sequence[46, 29]. A microarray experiment can detect such a change by including a set of probes that differ only at one base; the probe yielding the strongest signal corresponds to the sequence of the sample material[20]. This type of array can determine the genotype at hundreds of thousands of bases simultaneously, though it typically requires knowledge of the polymorphic positions in advance. Another SNP array design uses single base extension (SBE) on a microarray; here, an oligo is designed for the bases immediately 5' of the polymorphic location[26]. After the nucleic acid is hybridized to the array, fluorescently labeled ddNTPs are incorporated; the array probe is extended by a single base using the hybridized material as a template. By scanning at several wavelengths corresponding to the different fluorophores (each ddNTP carries a different color fluorophore), the identity of the incorporated base and hence the genotype become apparent.

1.1.2 TIP-Chip

Transposon screens detect the location of a transposon insertion. All previously described transposon screens work only on a single transposon or transposon family; they do not find insertions of any other material[55, 53]. The simplest transposon screen involves a primer complementary to the transposon sequence (typically near one or both ends of the transposable element). This primer initiates an extension using the genomic DNA as a template; further manipulations result in PCR around the transposon to produce labeled DNA. The resulting material is hybridized to a microarray illuminating probes near a transposon site.

1.1.3 CGH and FISH

FISH encompasses a set of techniques involving the hybridization of labeled DNA to potentially complementary DNA to detect the presence of a single, specific sequence[11, 21, 10]. One protocol requires a labeled probe complementary to some region of the sample DNA. The probe hybridizes when the complementary sequence is present and fails to hybridize if the target sequence is missing. Alternatively, the sample DNA might be labeled and hybridized to known reference DNA. If two samples, labeled with different fluorophores, compete for the reference sequence, then the technique is known as Comparative Genomic Hybridization (CGH) [24]. One approach hybridizes labeled sample material to metaphase chromosomes, permitting the detection of gross deletions. Appropriately designed FISH probes can detect insertions or deletions at specific loci, but this approach is not suitable for high-throughput screens[30].

1.1.4 Array-CGH

Array-based comparative genome hybridization (aCGH) requires two genomic samples labeled with different fluorophores and can detect copy number changes but not necessarily the site of the change[44, 43, 45, 2, 54, 42, 9, 28]. The two samples are hybridized to a single microarray and scanned. Comparing the intensities in the two channels at each probe or set of genomicaly proximal probes determines the presence of duplications (higher intensity than expected compared to other probes in the same sample or compared to the other sample) and deletions (lower than expected intensities). While the location of a deletion in the genome is apparent if one knows the genomic location of the relevant probes and if the deletion removes enough probes from the sample sequence, aCGH does not provide the genomic location of duplications. Furthermore, aCGH cannot necessarily detect rearrangements (very high density arrays may be able to detect candidate rearrangements when low intensity probes, those spanning the relocation boundary, surround probes of the expected intensity).

1.1.5 Sequencing

Sequencing has the potential to detect all sequence changes, but its practical limitations depend on the technology used (which determines the read length, the availability of paired reads, and the mean and standard deviation of the distance between pairs of reads) and the coverage depth. In general, all sequencing approaches can detect SNPs since single nucleotide changes are small compared even to the short reads produced by current high throughput techniques[35, 4].

Detecting insertions and deletions by sequencing and assembly presents more challenges. Reads spanning a change potentially contain only a small amount of overlap with adjacent reads on either side making it difficult to map the reads to a reference genome. Repetitive sequences cause difficulties for sequencing when the length of the repetitive element is comparable to the read length; repeats longer than a read will be particularly difficult to resolve since no read will span the repeat and include unique sequence on either side[6, 15]. Figure 1-1 shows examples of genomic sequences that pose particular challenges to assemblers.

Paired-end reads can permit assembly across difficult regions. Both reads in the pair come from the same underlying template molecule, one from each end. By controlling the distribution of template molecule lengths, the technique provides the location of an ambiguous read (e.g., one that contains primarily repetitive sequence) if its partner can be mapped unambiguously. If only one read of a pair can be mapped, then the other read is presumably from an insertion and the location of the insertion can be estimated based on the expected distance between reads. While paired-end reads can enable assembly across repetitive regions as large as the underlying template molecules (rather than across repetitive regions as large as the read), there may still be regions that do not assemble or assemble incorrectly, depending on the genome and the technology used to generate the templates.

Sequencing approaches to indel detection (as opposed to sequencing approaches to genome assembly) such as Paired-End Mapping (PEM) use libraries of paired-end reads[49, 27] . After the resulting reads are mapped to the genome, the distance between each pair can be compared to the expected distribution of the library generation technology. If the reads seem to be too close or too far apart, then one concludes that an insertion (too close) or deletion (too far) has occurred between the reads. Given current techniques for generating the template molecules, these methods can detect changes in the kilobase and larger range. Methods that generate DNA fragments with a smaller variance would permit detection of smaller changes.

1.2 Our Contribution

Since no existing genotyping technology could inexpensively detect sub-kilobase length polymorphisms across an entire genome, we sought to develop a new technique that would detect insertions, deletions, and repeat-length changes in a genome similar to a previously characterized genome but without requiring prior knowledge of the inserted sequence.

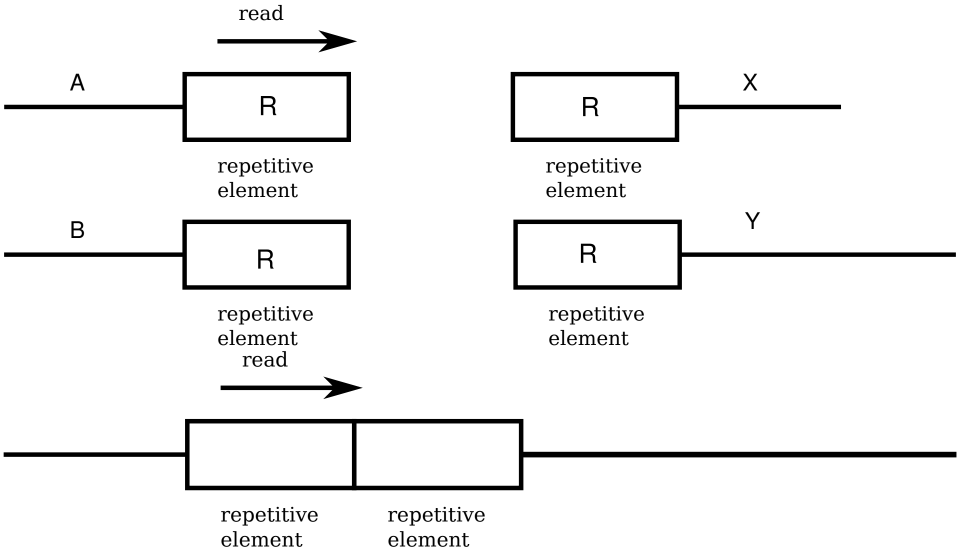


Figure 1-1: Sequencing and assembly methods have difficulty resolving repetitive sequences whose length is greater than the read length. In the top panel, two contigs A and B end in a repetitive element R while two other contigs X and Y begin with R. If the read length is less than the length of R, then unpaired reads will not be able to join A to X or Y as no read can contain unique sequence from A as well as unique sequence from X or Y. Consequently, an instance of R in the genome will result in a contig boundary. Paired-end reads will allow contigs to be joined if the insert size is greater than the size of R and if a pair spans R with both ends landing in unique sequence. The bottom panel shows how a tandem repeat that would confound single-end reads if R is longer than the read size; if the experimental sequence contains two adjacent repeat units then the assembler cannot determine the number of repeat units. Paired-end reads spanning R and with both ends in unique sequence provide probabilistic information, but not absolute information, about the copy count of R. Comparing the observed distance between ends than span the repeats to the expected distance indicates whether the assembly is correct. Similarly, comparing the observed to expected distance between a read in the unique sequence and a read in an instance of R provides information about whether the copy count of R in the assembly is correct.

Chapter 2

Ruler Arrays

Ruler arrays detect changes in the distance between pairs of defined genomic sequences. They can also detect differences between a sample sequence and an assumed reference sequence, allowing genome assemblies to be checked for certain errors. Unlike aCGH, a Ruler Array does not directly detect the presence or absence of a DNA sequence; instead, it detects the sequence's effect on the surrounding sequences.

2.1 Ruler Array Theory

Ruler arrays define arbitrary sites of illumination throughout the nucleic acid sample such that probes interrogating nearby sequences yield high intensities while more distant probes yield lower intensities. Since each probe detects material from only one strand of the genome, the observed intensity provides information only about the distance to one of the two surrounding illumination sites. The computational analysis evaluates the fit of the observed intensities to some model to determine the presence and location of indels from the observed intensities.

Figure 2-1 shows schematic of a segment of nucleic acid with several labeling sites and the resulting probe intensities on one strand. Analyzing the resulting intensities indicates the distance between the probe and the closest illumination site. By comparing these distances to

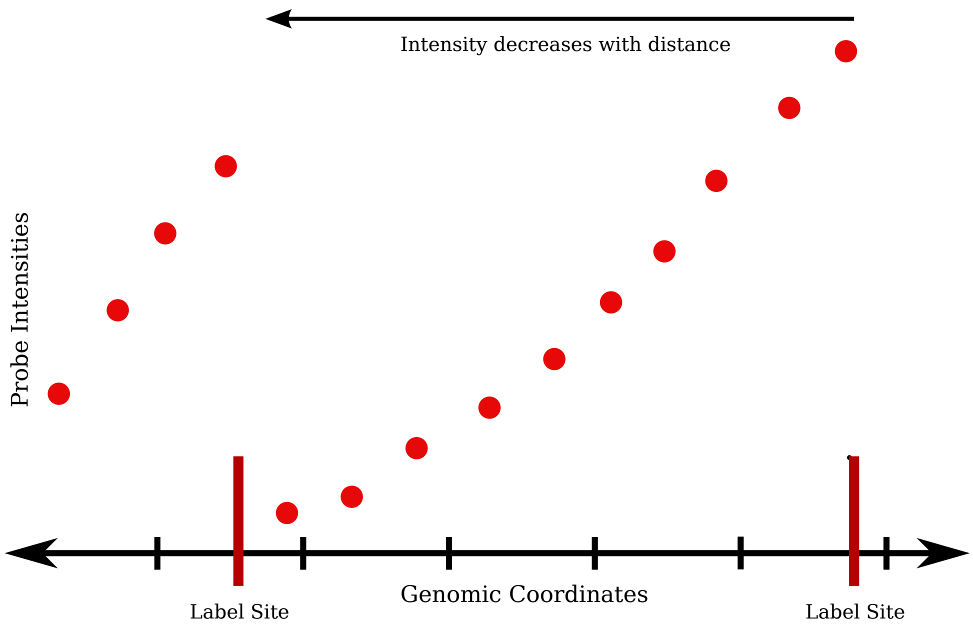


Figure 2-1: A schematic of two ruler illumination sites and the resulting probe intensities observed by probes designed against one strand of a nucleic acid. At each labeling site, the intensities jump to some high level and then decrease as one moves farther from the illumination site. The locations of the labeling sites and the shape and length of the falloff in intensity depend on the details of the laboratory protocol.

another sample or to a reference sequence, Ruler Arrays detect the presence of an insertion (schematic shown in figure 2-2), deletion (figure 2-3), or inversion (figure 2-4). Since an indel only effects the probes beyond the change, and not those between the change and the illumination site, Ruler Arrays determine the location of the change to within the resolution of the probes on the array. Furthermore, careful analysis of the intensities can estimate the size of the change. The expected relationship between intensity and distance from the illumination site can be found either from a separate control experiment, theoretical predictions, or from the average observed relationship.

A two sample Ruler Array experiment produces a ratio of intensities at each probe in which probes equidistant in the two samples yield ratios near one. A probe that is farther from the illumination site in one channel than the other yields a ratio either above or below one depending on the direction of the change. All probes between the change and the next restriction site will show a change of intensity and hence a change in ratio. An analysis method can detect this series of probes with either elevated or depressed ratios and detect the changed site by determining where the series of non-unity ratios begins. Thus a two channel experiment depends less on knowing the expected intensity shape and relies instead on identifying the differences between the two samples.

2.2 Advantages of Ruler Arrays

Ruler arrays resemble aCGH in that, at the highest level, both attempt to detect insertions and deletions. However, Ruler Arrays offer several key advantages when the changed sequence is unknown or difficult to handle.

Ruler arrays can detect insertions of any material, not just sequence represented by probes on the array. aCGH will not detect an insertion of foreign or untiled material because no probe interrogates that sequence. Ruler arrays detect insertions by their effect on the probes surrounding the insertion and therefore work even on insertions of unknown material.

Ruler arrays can also detect insertions of repetitive sequence (e.g. the 6kb TY1 transposon or 300bp Σ element in yeast) that aCGH may miss because the relative change in copy number may be small or that sequencing may miss because the repetitive sequence is hard to assemble. Ruler arrays can still detect the insertion of, for example, the thirty eighth copy of the TY1 element because they will notice the change in distance between an illumination site and some set of probes. An aCGH experiment may miss the change because it cannot detect that a ratio of $\frac{38}{37} = 1.03$ as different from one.

Changes in the length of simple repeats (e.g. repeats of A, AT, GC, ATT, etc) cannot be easily detected by any existing high throughput technology but represent a primary use for Ruler Arrays. While one cannot design probes against a simple repeat (since the probe may have complicated hybridization dynamics and may also match multiple genomic loci), the Ruler Array only needs an illumination site on one side of the repeat and unique probes on the other side. The intervening sequence is irrelevant; the Ruler Array merely detects its length.

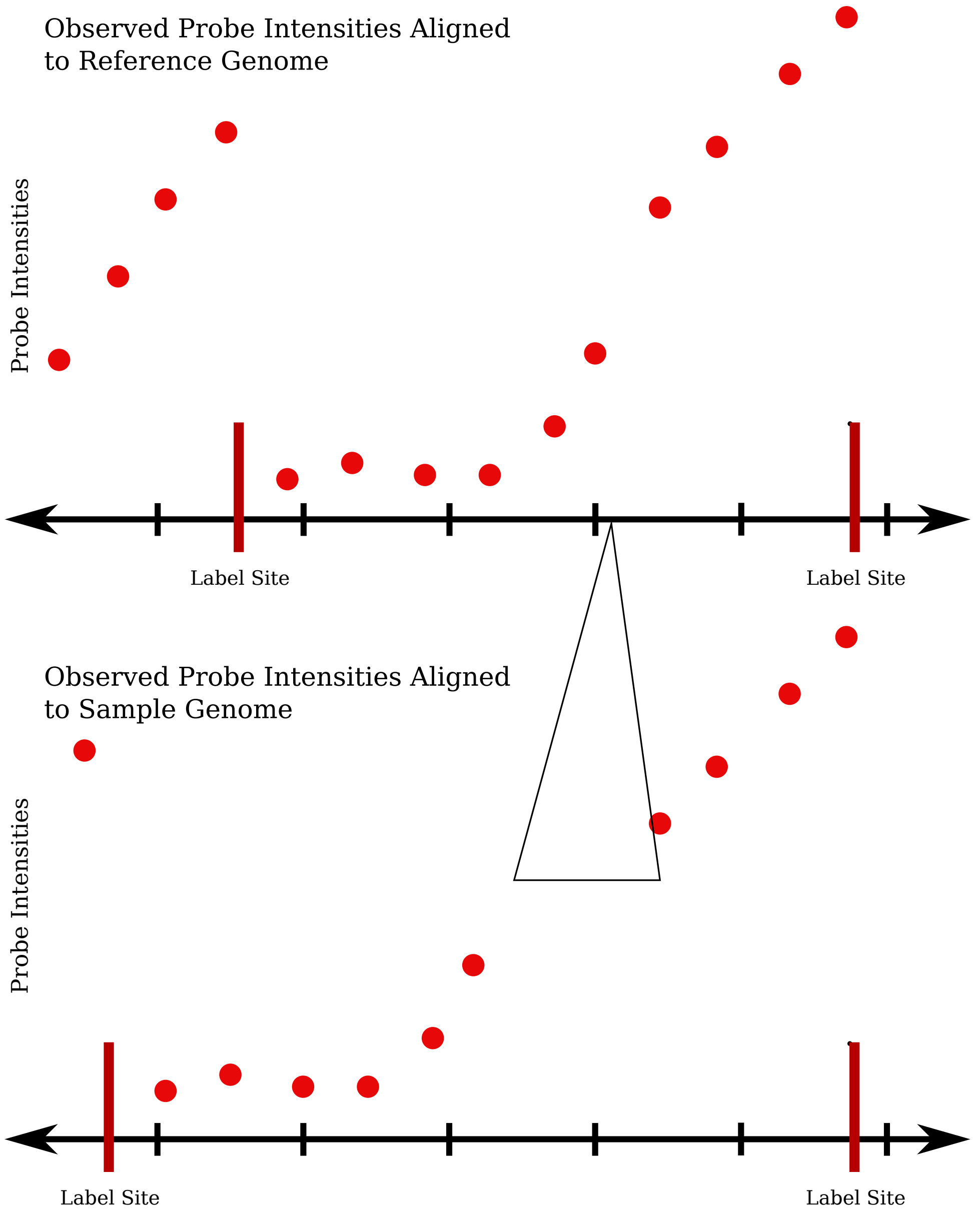
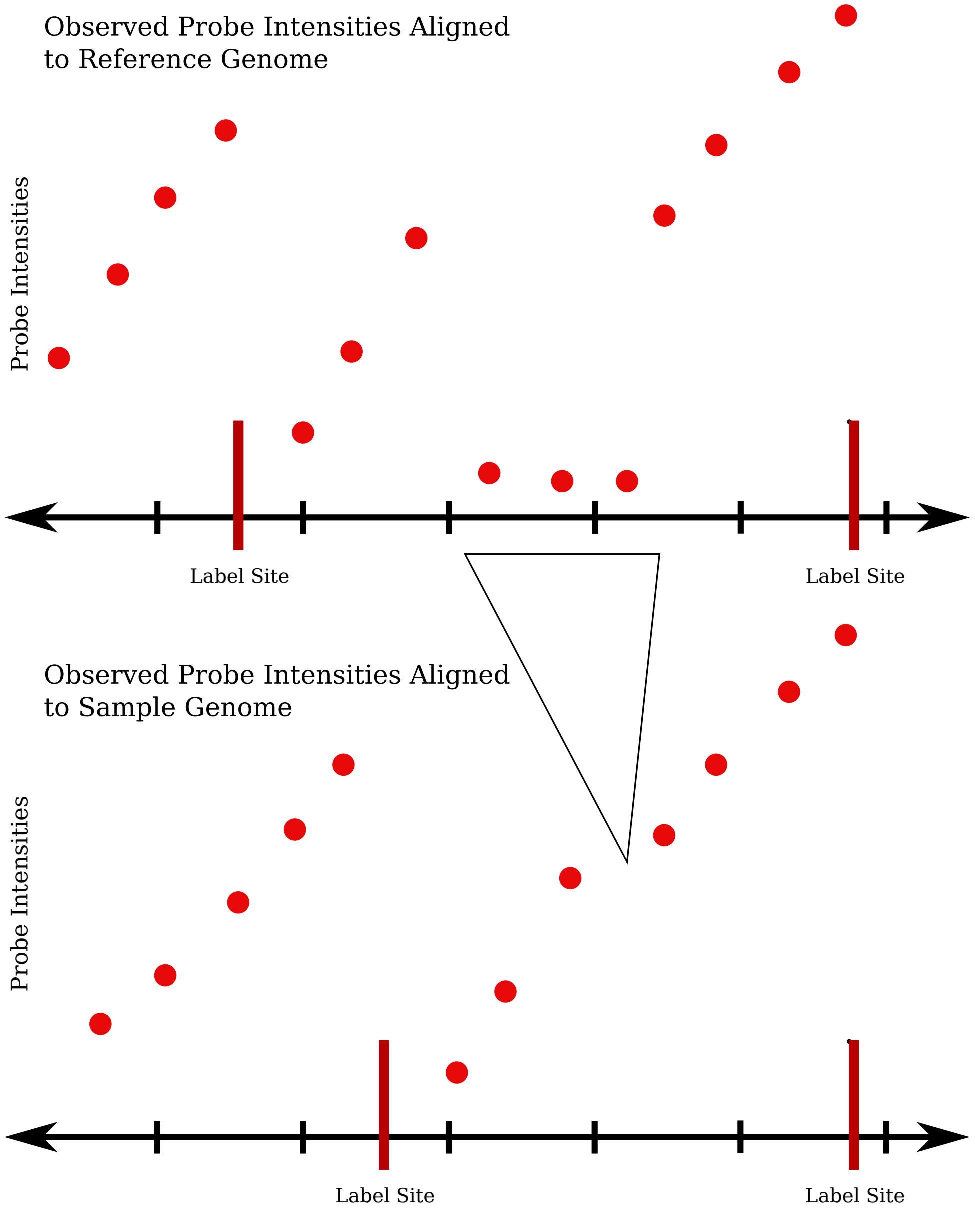

Observed Probe Intensities Aligned to Reference Genome
Probe Intensities
Label Site
Label Site
Observed Probe Intensities Aligned to Sample Genome
Probe Intensities
Label Site
Label Site

Figure 2-2: Schematic Ruler Array data at an insertion. The top panel shows probes mapped to the reference sequence. The bottom panel shows that, when the reference sequence is modified to contain the insertion in the sample sequence, the observed intensities follow the expected shape. Note that this analysis requires no knowledge of the inserted sequence.


Observed Probe Intensities Aligned to Reference Genome

Observed Probe Intensities Aligned to Sample Genome

Probe Intensities
Label Site
Label Site
Probe Intensities
Label Site
Label Site

Figure 2-3: Schematic Ruler Array data at a deletion. As in aCGH, the probes corresponding to the deleted sequence produce little intensity since very little sample material hybridizes to them. Unlike aCGH, however, the probes beyond the change confirm the loss of material by producing higher intensities than one would expect in the absence of a deletion.

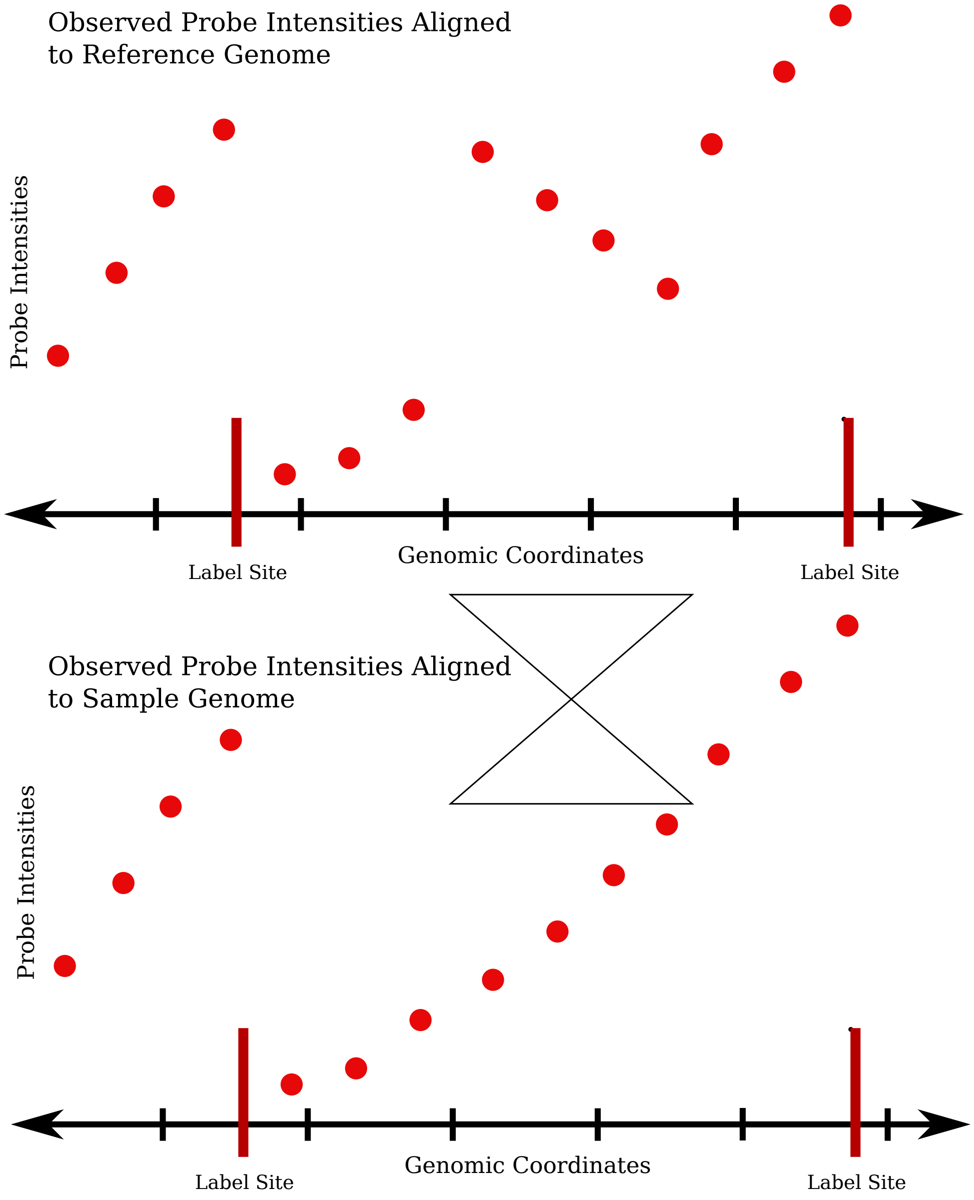

Observed Probe Intensities Aligned to Reference Genome
Probe Intensities
Label Site
Genomic Coordinates
Label Site
Observed Probe Intensities Aligned to Sample Genome
Probe Intensities
Label Site
Genomic Coordinates
Label Site

Figure 2-4: Schematic Ruler Array data at an inversion. As with insertions and deletions, inversions are recognizable because the observed relationship between probe intensity and distance from the labeling site does not match the expected relationship.

Finally, Ruler Arrays provide the location of a change. Even when aCGH detects a given change in copy number, it often cannot determine the location of the change (though the case of one to zero copies is easy). While the Ruler Array may not reveal what material has been inserted at a genomic location, it does determine the location. Given the location, it is then easy to amplify and sequence the insertion.

2.3 Weaknesses of Ruler Arrays

While Ruler Arrays offer a high throughput method to detect and localize insertions, deletions, and inversions, they do not fully solve the problem of detecting genomic changes.

- Ruler arrays do not provide the sequence of an insertion; this must be obtained through traditional methods such as primer-based sequencing.
- Ruler arrays rely on a sufficient prior knowledge of the expected sequence to choose microarray probes.
- The presence and distribution of restriction sites influences how much of a given genome or sequence can be interrogated at once by a single experiment
- The protocol assumes that localized sequence peculiarities do not substantially alter the intensity versus distance relationship. This may be problematic, for example, when trying to determine the length of a tandem repeat as many polymerases have difficulty working through repetitive sequence.

Chapter 3

Ruler Array Laboratory Protocol

The Ruler Array requires a population of labeled DNA fragments to exhibit a known relationship between fragment length and frequency in the population. In our experiments, the population contains many short fragments but few long fragments. Furthermore, one end of each fragment is taken from a small set of known genomic loci. The laboratory protocol to produce this population of fragments consists of four basic steps: digestion, ligation, labeling, and hybridization. Figure 3-1 shows the process through labeling.

3.1 Overview

The Ruler Array protocol uses restriction enzymes to define the sites of illumination. The large variety of restriction enzymes allows the researcher to choose an enzyme that yields a desirable set of illumination events for any particular sample. Furthermore, the sticky ends left by many restriction enzymes provide a substrate for efficient ligation.

An adapter molecule is ligated onto the sticky end left by the restriction enzyme. The adapter's 5' end is complementary to the restriction site while the remainder of the adapter consists of arbitrary sequence that we have chosen to minimize its hybridization to the genome.

A primer complementary to the adapter oligo initiates a labeling reaction at the illumination site. This reaction extends 5' to 3' along each strand, incorporating labeled nucleotides. Either the polymerase's natural processivity, the inclusion of ddNTPs in the reaction mix, or the length of the template material terminates the labeling reaction. If the polymerase terminates with sufficiently high probability at each base then most fragments will not be full length (i.e. shorter than the template). If the termination probability is relatively uniform, then short fragments will be more common than long fragments.

The relative abundance of short products and relative scarcity of long products from the extension reaction yields the desired falloff in intensity when the material is hybridized to the microarray. Furthermore, if the restriction enzyme completely digested the input material, each interval between restriction sites will be completely independent of the neighboring intervals since no template material (and hence no labeled product) crosses a restriction site. Figure 3-2 shows the results from an experiment between two yeast strains.

3.2 Digestion

Any restriction enzyme that leaves sticky ends to which an adapter can be ligated is suitable for use with Ruler Arrays. In most situations, an enzyme that recognizes a six nucleotide sequence and leaves a four nucleotide overhang will generate a good distribution of restriction sites and provide reasonable ligation efficiency. For example, EcoRI leaves four nucleotide overhands and produces an average interval size of about 3kb in yeast.

The variety of restriction enzymes allows a range of interval sizes. Short intervals between restriction sites will contain few probes but produce a high slope- the falloff from high intensities near the illumination site to low intensities near the neighboring site happens over a short distance. Few probes makes for more difficult analysis since there are fewer observations to provide statistical power and robustness against noise. On the other hand, the high slope means that small insertions or deletions will yield large changes in intensity ($\Delta x \cdot \text{slope} \rightarrow \text{large}\Delta y$). Very large intervals yield template fragments that cannot be labeled to the end and thus produce holes in the genome that the Ruler Array cannot interrogate.

Our protocol calls for phosphatasing the input material after digestion. This prevents

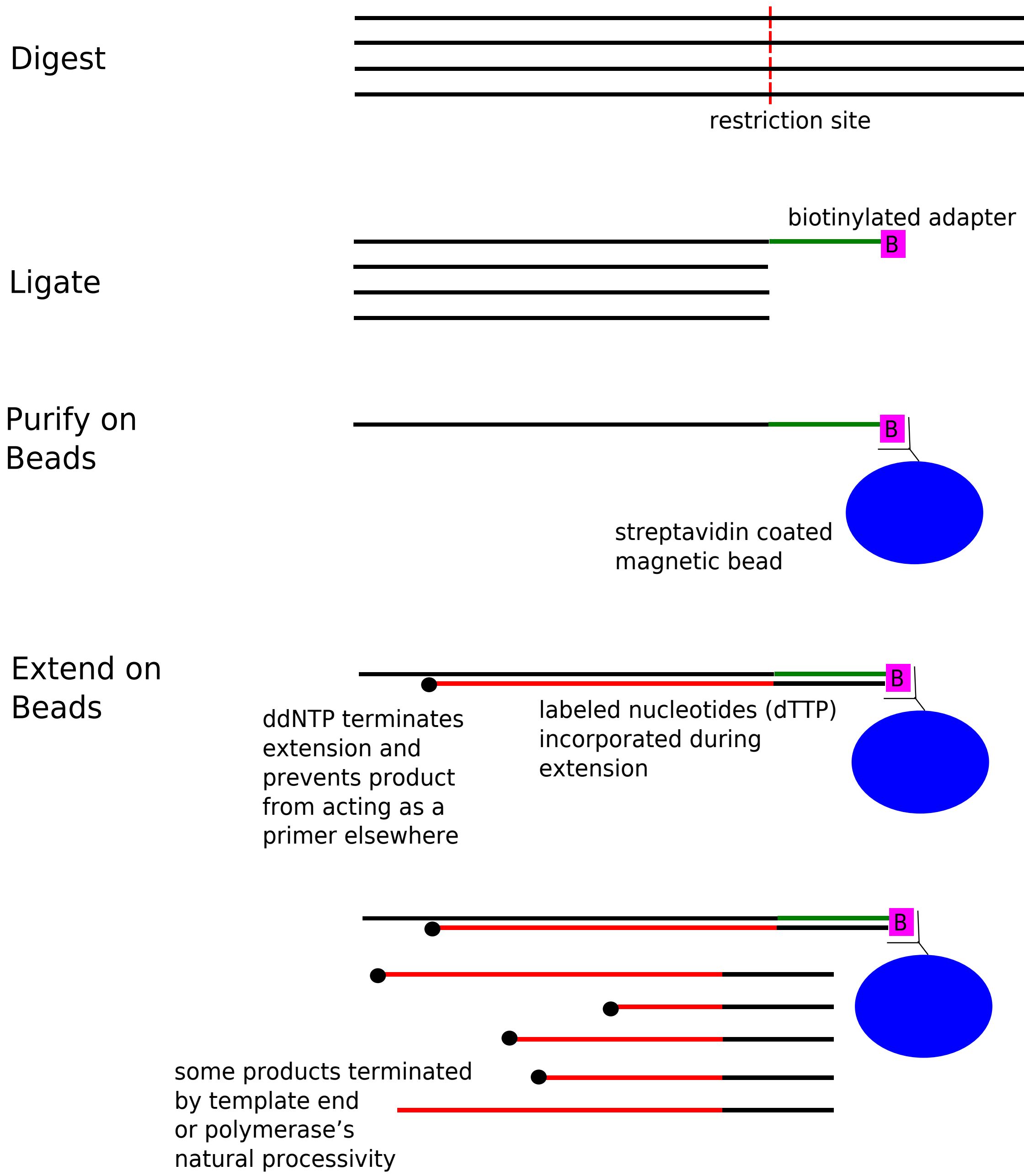

Digest
restriction site
biotinylated adapter
B
Ligate
Purify on Beads
B
streptavidin coated magnetic bead
Extend on Beads
B
ddNTP terminates extension and prevents product from acting as a primer elsewhere
labeled nucleotides (dTTP) incorporated during extension
B
some products terminated by template end or polymerase's natural processivity

Figure 3-1: The Ruler Array laboratory protocol consists of four steps: digestion of the nucleic acid input material by a restriction enzyme, ligation of an adapter molecule onto the resulting sticky ends, generation of labeled material, and hybridization to a microarray

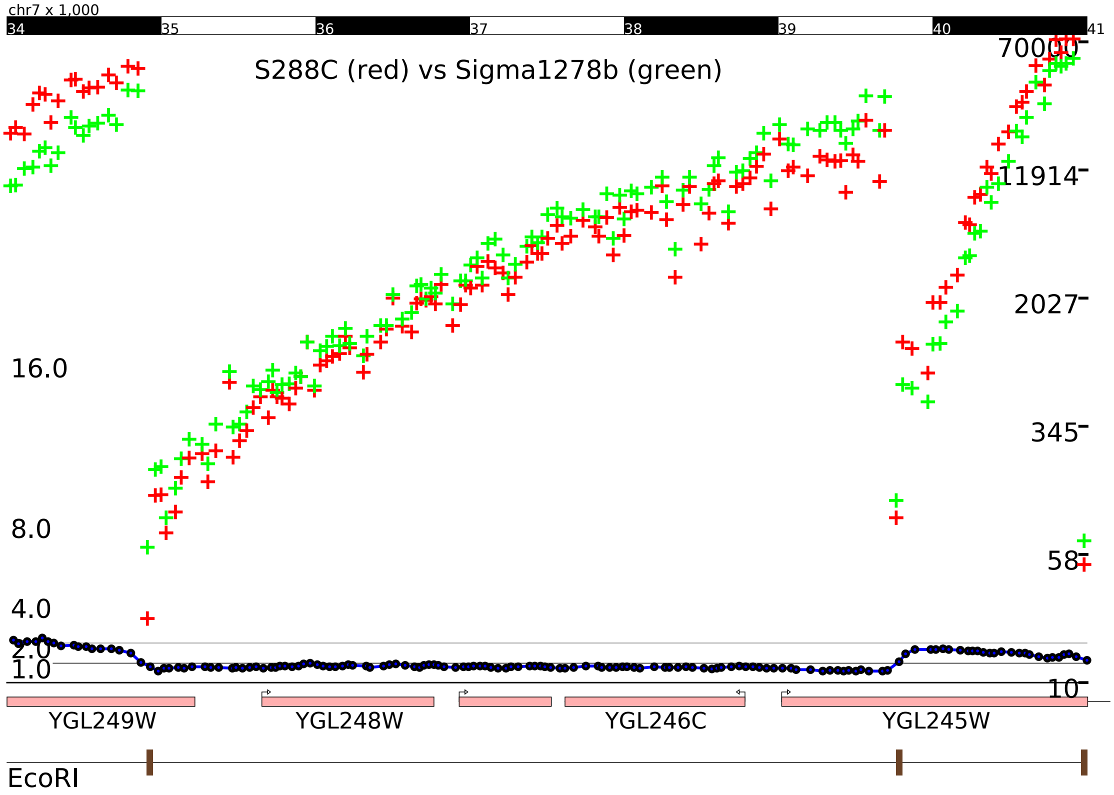


Figure 3-2: Sample Ruler Array data from yeast. The very bottom track shows EcoRI restriction sites as tick marks; these were the labeling sites for this experiment. Red marks in the main track indicate intensities from the S288C sample and green marks indicate intensities from the Σ1278b sample. The intensities are highest at the EcoRI site and fall off gradually over several kilobases (the intensity scale is on the right side). This array tiled only one strand, so the equivalent falloff in material on the opposite strand does not appear. The blue dots under the intensities show the ratio between S288C and Σ1278b (the ratio scale is on the left side).

the sticky ends produced by the restriction enzyme from ligating to each other during the ligation step to create chimeric templates.

3.3 Ligation

The ligation step attaches an adapter molecule to the digested input material (figure 3-3). The adapter carries a 5' phosphate to compensate for the lack of phosphates on the DNA. Since the efficiency of ligation reactions is generally low (perhaps ten percent of the available ends will be attached to an adapter), we want to separate the successfully ligated material from the remaining genomic DNA. A 3' biotinylated adapter molecule allows this separation. After ligation, the sample is hybridized to streptavidin beads and washed. DNA fragments ligated to an adapter remain on the beads because of the interaction between the biotin and the streptavidin. Most of the unligated material washes off.

3.4 Extension and Labeling

A primer complementary to the 3' end of the adapter (that is, complementary to the majority of the adapter that does not correspond to the restriction site) permits initiation of the labeling reaction. The reaction starts at the restriction site and extends outwards, incorporating labeled nucleotides. Since all cutting restriction enzymes recognize palindromes, the extension goes in both directions, though on opposite strands (5'→3' on each strand). The extension reaction typically incorporates Cy-dye labeled nucleotides in addition to plain nucleotides.

With polymerase alone, the length of the labeled product fragments will be determined by the template length and the polymerase's processivity. Careful handling of the input material will produce long template molecules while rougher handling or a treatment such as sonication will produce short templates. Within the bounds set by the template length and extension time, different polymerases will be able to produce longer or shorter fragments.

To control the distribution of extension product lengths more precisely, we use ddNTPs

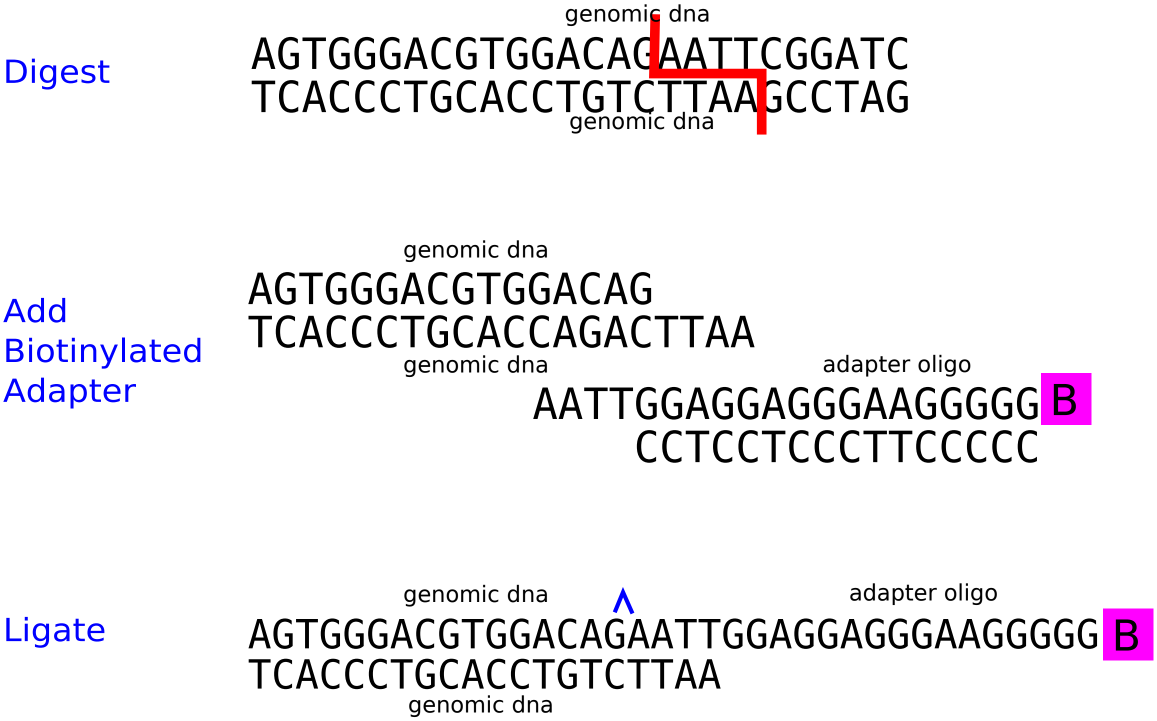

Digest
genomic dna
AGTGGGACGTGGACAGAATTCGGATC
TCACCCTGCACCTGTCTTAAGCCTAG
genomic dna
Add Biotinylated Adapter
genomic dna
AGTGGGACGTGGACAG
TCACCCTGCACCAGACTTAA
genomic dna
adapter oligo
AATTGGAGGAGGGAAGGGGG
B
CCTCCTCCCTTCCCCC
Ligate
genomic dna
adapter oligo
AGTGGGACGTGGACAGAATTGGAGGAGGGAAGGGGG
B
TCACCCTGCACCTGTCTTAA
genomic dna

Figure 3-3: The digestion and ligation steps of the Ruler Array protocol attach an adapter molecule to specific genomic loci. The adapter is biotinylated such that the ligated material can be separated from the unligated material.

in the reaction mix since an extension reaction terminates when a polymerase incorporates a ddNTP instead of a dNTP. A high concentration of ddNTPs produces relatively short fragments while lower concentrations permit longer fragments. This offers a simple and repeatable method to tune the fragment lengths as there is a tradeoff between coverage (long fragments interrogate more of the genome) and sensitivity (shorter fragments may be able detect smaller changes). This tuning allows the researcher to vary the experiment according to what type of change they expect. In practice, the choice of polymerase may limit the utility of ddNTPs as some polymerases ignore these molecules and others perform very poorly in their presence.

ddNTPs offer a second advantage because they prevent a labeled product from acting as a primer later in the labeling reaction. In a complex and repetitive mammalian genome, the 3' end of a product from one genomic locus might match sequence from another genomic locus. If the product hybridizes to the other template and allows an extension reaction to begin at a site other than an expected illumination site, the resulting intensities from the hybridization will not meet our assumptions (e.g., imagine two sequences in a genome: ABC and XBY. If product from AB hybridizes to XBY, it may initiate an extension from B through Y. A product terminated with a ddNTP cannot act as a primer and thus reduces the effects of this type of noise.).

Since incorporating Cy-dye labeled nucleotides may decrease the polymerase's processivity, we have experimented with other labeling techniques. One method uses amine modified dUTP during the extension and then chemically attaches a dye to the modified nucleotides after the polymerase reaction. The ULS labeling system offers another non-enzymatic method to label any nucleic acid without having first incorporated any special nucleotides.

3.5 Hybridization

The labeled material is hybridized to a microarray and scanned to determine the intensity at each probe. The hybridization process is essentially the same way as in expression, ChIP-Chip, or aCGH experiments. Following scanning the probe intensities are mapped to one or more reference genome sequences for display and analysis.

Chapter 4

Predicted Intensity Shape

The labeled DNA fragments produced by the Ruler Array protocol terminate either when the polymerase reaches the end of the template DNA or when the polymerase falls off prematurely. We expect the latter case to end most extensions and assume a uniform probability of the labeling extension terminating at each base and therefore that the product lengths follow an exponential distribution. The Ruler Array data meets this assumption well enough permit accurate analysis.

4.1 Extension Termination

In a perfect experiment, all template DNA fragments would span the entire interval between restriction sites. In practice, the template fragments may break during one of the purification steps. If we assume that the breaks occur uniformly and at random, then the templates contribute some fixed probability to a labeled fragment's termination at each base. However, the template fragments may not break with equal probability at each base; one might imagine that the middle of the template breaks more frequently because it experiences more twisting. Furthermore, long template fragments may break more readily than those spanning short intervals because more twisting or shearing forces may accumulate. For our analysis, we ignore these considerations.

The labeled product extension may terminate before the end of the template either when the polymerase incorporates a terminating nucleotide (e.g. a ddNTP or acyNTP) or when the polymerase falls off for another reason. For example, certain secondary structures in the template DNA (e.g. stemloops) may destabilize the polymerase and cause it to fall off more readily than it might otherwise. Incorporating Cy-dye conjugated dUTP may increase the probability that the polymerase halts due to the bulk of the dye (this would also lead to a sequence-dependent shape in the intensities).

The presence of full length products do not invalidate the resulting analysis method. These products increase the observed intensity uniformly across the entire interval. Since full length products occur more frequently in short intervals, the peak intensity may depend on the interval's length; however, our analysis method accommodates this variation.

4.2 The Shape

Ignoring the various reasons why the probability of termination may not be uniform at each base, we can easily predict the intensity at distance d from a restriction site. Since the polymerase incorporates labeled nucleotides along the fragment, the intensity of a fragment increases linearly with its length. However, the same fragment can hybridize to some number of probes that increases roughly linearly with its length. These two effects cancel, meaning that the intensity observed at any probe due to fragments of some length (remember there will be a population of fragments of any given length, each of which can hybridize independently to any complementary probe) is independent of the fragment's length. The observed intensity at a probe d bases from the labeling site therefore depends only on the number of fragments complementary to the probe, which is the number of fragments longer than d:

$$\text{intensity}(d) = \sum_{i=d}^{D} (1-p)^{i-1} p$$

where D is the length of the interval (i.e. the distance from this restriction site to the next), p represents the probability that the labeling extension terminates at each base, and

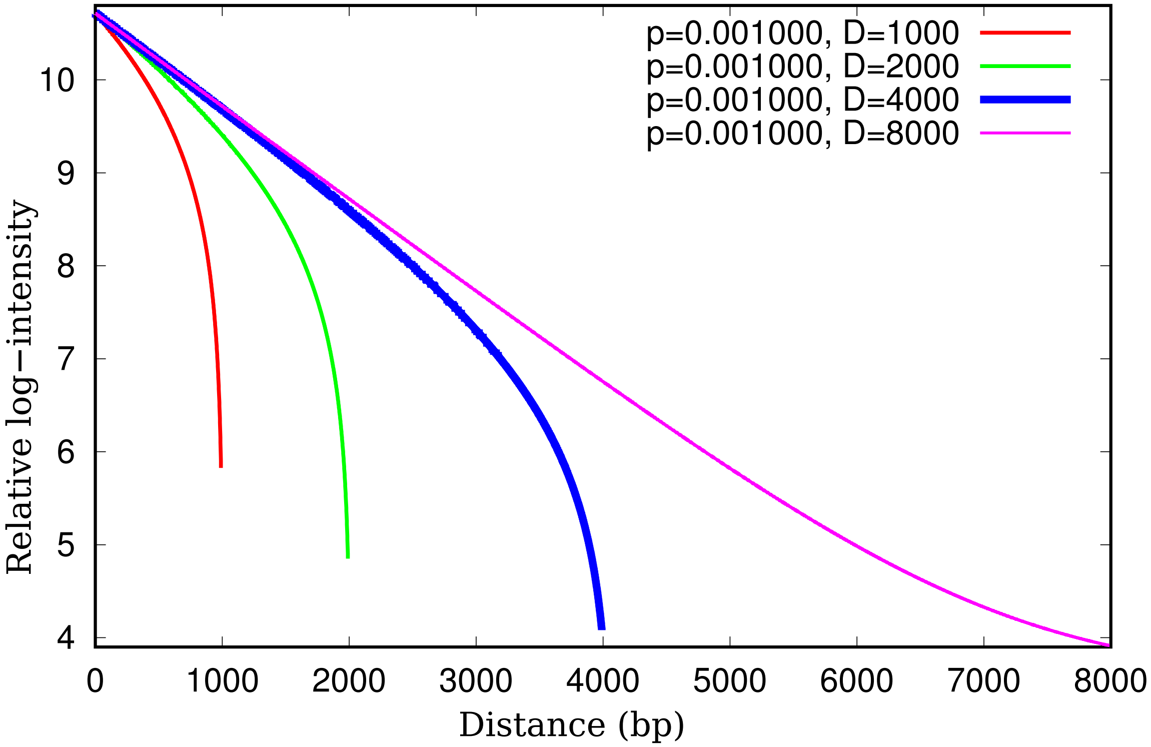


Figure 4-1: Plot of predicted log-intensity vs distance for intervals of size 1000, 2000, 4000, and 8000. The probability of termination at any base is .001 in all four intervals. Note the relatively linear shape over most of the interval followed by a more rapid decrease as the end effects become dominant.

$(1-p)^{i-1}p$ is the probability of a fragment of length exactly i. Figure 4-1 shows the predicted log-intensity over several interval sizes.

Chapter 5

Predicted Ratios from Insertions

We can simulate the intensities and ratios that would result from insertions and deletions using the model described in Chapter 4 for intensity as a function of distance from a restriction site. Examining these synthetic events provides some intuition about which events the Ruler Array should be able to detect and forms the basis for a simplified model to determine whether a given indel will be findable.

Four variables influence our ability to detect insertions and deletions from Ruler Array data:

Indel Size Larger changes are easier to detect because they produce a greater change in intensity and ratio.

Interval Size Longer intervals have shallower slopes than do shorter intervals such that any change in length will yield a smaller relative change in intensity. ddNTPs simulate short intervals by increasing the slope.

Position in Interval The position of the indel relative to the restriction site influences the ability of an analysis method to detect the indel. If the indel is too close to either end there will be too few probes on one to detect the change relative to the probes on the other side of the indel

Microarray Design The microarray tiling density and the amount of experimental noise also influence the effectiveness of an analysis method, but the effects will be specific to the method.

Figures 5-1 and 5-2 show the ratio in 2kb and 4k intervals from simulated insertions of varying sizes. These give some sense of what size indels at what distances from the restriction sites ought to be detectable. Table 5.1 shows the ratio of channel ratios across indels of various sizes (i.e. one is no change while numbers farther from one indicate a greater change in channel ratio across the indel.

In some cases, the expected ratio from the insertion exceeds two, the traditional level of significance for ratio differences in microarray experiments. In other cases, particularly when the change is small or very close to the restriction site, the resulting ratio will be much less than two. As such, our analysis technique will need to aggregate information from all of the probes around the indel to collect a statistically significant signature for the change.

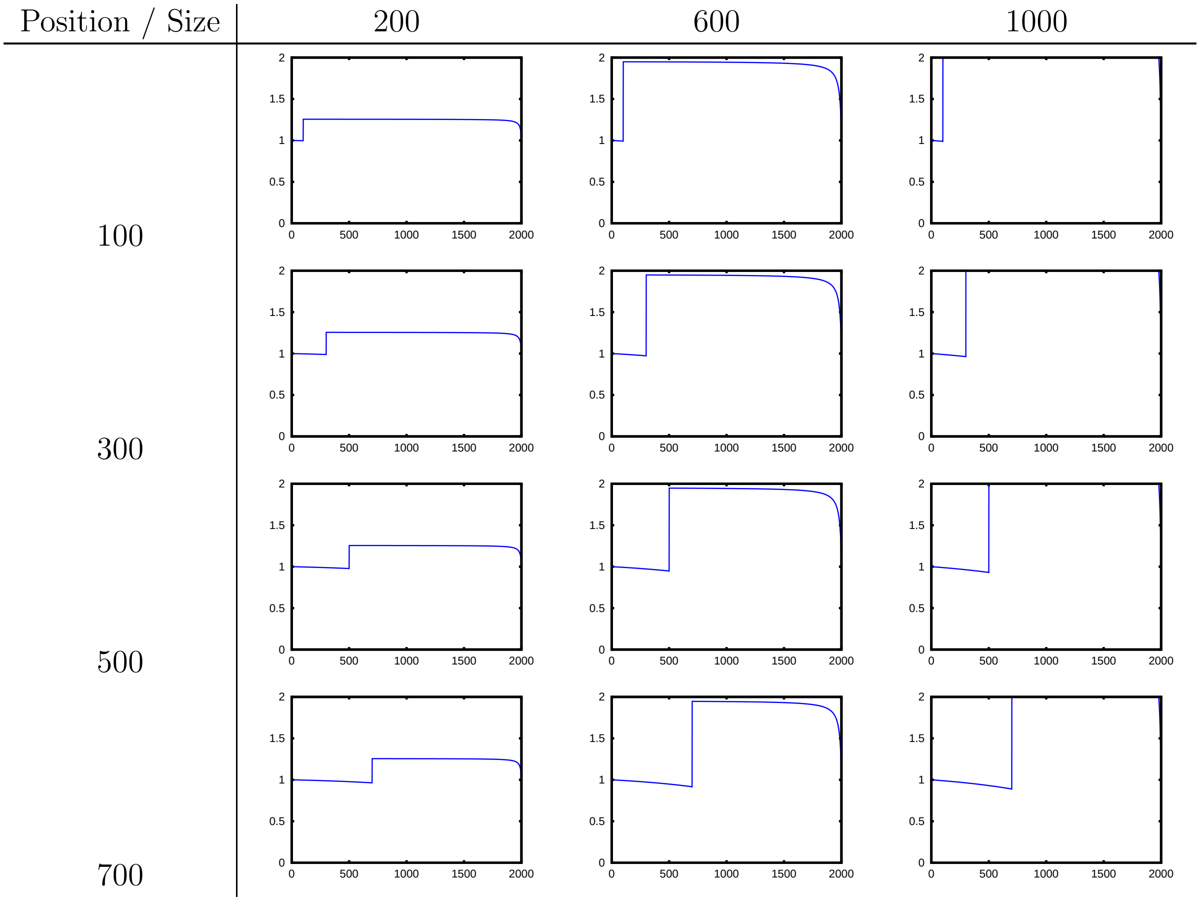


Figure 5-1: Predicted ratios in a 2000bp interval. Each plot shows an insertion of some size (200, 600, or 1000bp) at some distance from the restriction site (100,300,500,or 700bp) when the peak ratio with no insertion was one.

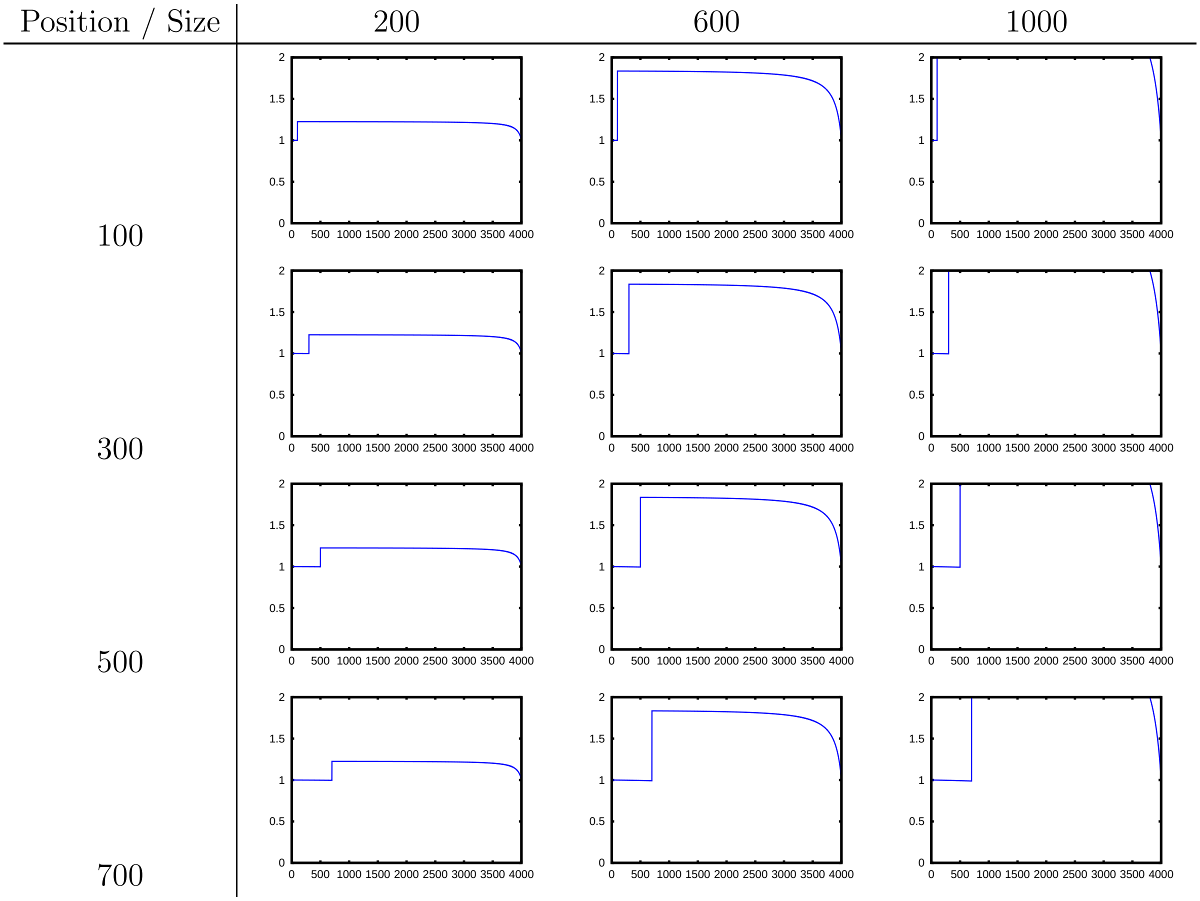


Figure 5-2: Predicted ratios in a 4000bp interval as in figure 5-1.

Interval size 1000

Size / Position	100	200	300	400	500	600
100	0.957856	0.933815	0.910313	0.885451	0.856987	0.821599
200	0.905542	0.866478	0.828633	0.789355	0.745743	0.693897
300	0.849343	0.801187	0.754877	0.707534	0.656192	0.597173
400	0.792729	0.739429	0.688490	0.637048	0.582297	0.520973

Interval size 2000

Size / Position	100	200	300	400	500	600
100	0.979312	0.968207	0.958441	0.949581	0.941301	0.933333
200	0.950148	0.930774	0.913734	0.898294	0.883902	0.870111
300	0.915302	0.889875	0.867511	0.847271	0.828449	0.810479
400	0.876830	0.847086	0.820925	0.797276	0.775332	0.754451

Interval size 4000

Size / Position	100	200	300	400	500	600
100	0.985922	0.978507	0.972198	0.966748	0.961973	0.957739
200	0.964978	0.951648	0.940307	0.930509	0.921928	0.914319
300	0.938741	0.920761	0.905463	0.892247	0.880674	0.870416
400	0.908517	0.886950	0.868598	0.852746	0.838868	0.826571

Interval size 8000

Size / Position	100	200	300	400	500	600
100	0.987066	0.980275	0.974530	0.969610	0.965348	0.961624
200	0.967603	0.955318	0.944927	0.936027	0.928321	0.921586
300	0.942976	0.926308	0.912212	0.900138	0.889686	0.880553
400	0.914349	0.894247	0.877247	0.862689	0.850088	0.839079

Table 5.1: Difference in ratio across an insertion of the specified size at the given position in intervals of 1kb, 2kb, 4kb, and 8kb. The number shown is the channel ratio before the insertion (closer to the restriction site) divided by the ratio beyond the insertion. A value of 1.0 indicates no difference; a smaller value indicates more difference in ratios and therefore a more easily detectable insertion.

Chapter 6

Log Likelihood of the Data

All of the Ruler Array analyses evaluate the fit the observed data to intensities predicted by some model. Determining the best model or comparing two models requires computing the log-likelihood of the data given the model. We develop the basic model of intensity observations normally distributed around their true value to include several other terms to better describe our data and our goals for the model fitting.

6.1 Model for Probe Intensities

A simple way to compute the log-likelihood of the predicted values ($\hat{x}_i$) is to assume a distribution on the probe observations (x_i) and estimate a variance for each probe, σ_i^2. The log likelihood $\mathcal{L}$ will then be

$$\mathcal{L} = \Sigma_i log(p(x_{ij}|\hat{x}_{ij}, \sigma_i))$$

where x_{ij} is the observation of probe i from the jth replicate and $p(x_{ij}|\hat{x}_{ij}, \sigma_i)$ is the probability of the model-predicted value given the observation and the estimate of the standard deviation.

If we assume that the probe observations are normally distributed around their mean, then

$$p(x_{ij}|\hat{x}, \sigma_i) \sim \mathcal{N}(\hat{x}|\frac{1}{J}\Sigma_j x_{ij}, \sigma_i)$$

While we do not have the dozens of replicates of a Ruler Array experiment to justify this assumption, the microarray includes several control probes more than 100 times. In all experiments that we have examined, the intensities of these control probes follow a normal distribution and their means vary by several orders of magnitude from one another.

6.2 Estimating the Variance of Intensity Observations

The best estimate for the variance of a probe's observed values would come from a large set of observations, either by repeating the probe on an array or by repeating the experiment. In practice, however, the number of repeated observations is generally too small to accurately estimate the variance. Instead, we compute the standard deviation (it makes more sense to do a weighted average of standard deviation than of variance) as a combination of the expected standard deviation with the observed standard deviation:

$$\sigma_i = \frac{S \cdot s(\bar{x_i}) + J \cdot \sqrt{\frac{1}{J}\Sigma(x_{ij} - \bar{x_i})^2}}{S + J}$$

This estimate for the standard deviation weights the observed standard deviation based on the number of observations J and weights the prior s according to S, the strength (as a pseudocount) of our prior belief.

In the Ruler Array data, the prior on the standard deviation s should reflect the fact that the standard deviation is generally proportional to the intensity in microarray observations. Figure 6-1 shows the relationship between probe standard deviation and average intensity across many replicates of the control channel of a ChIP-Chip experiment (chosen because we had many replicates of the dataset).

We further limit the effect of apparently noisy probes by adding a term to the standard deviation estimate that accounts for the smoothness of a sequence of observations. More precisely, this term estimates the standard deviation as the difference between an observed value and the value predicted by linear interpolation of the surrounding genomic observations, as shown in figure 6-2. This term makes sense in our data where probes should be part of a

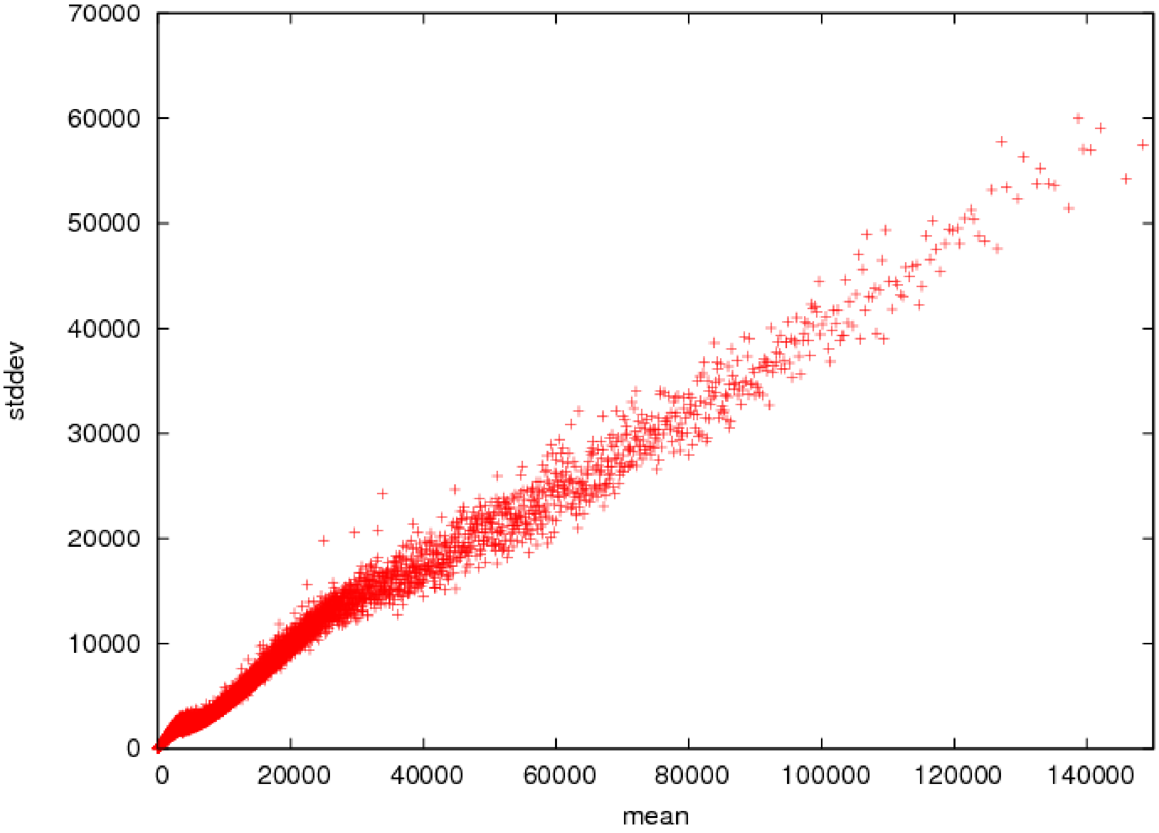


Figure 6-1: Plot of probe standard deviation vs intensity from the whole cell extract channel of ChIP-Chip experiments, measured in fluorescence units as reported by Agilent's scanner and feature extraction software. The experiments were normalized to have the same median intensity. While the noise in a ChIP-Chip experiment may not be identical to that in a Ruler Array experiment, we expect them to be similar and this plot does indicate that the relationship between the standard deviation of a probe's intensity and its intensity is close to linear. We have experimented both with a linear model and a piecewise linear model based on this data but found that the piecewise linear model offers relatively little improvement in the Ruler Array's overall performance.

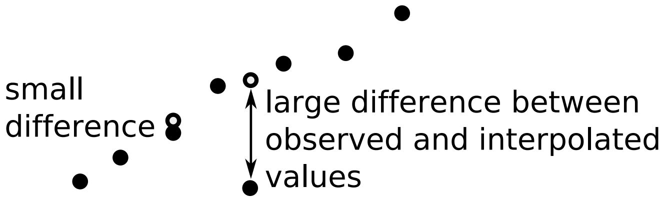


Figure 6-2: The interpolation term in the variance estimate depends on the difference between the observed value for a probe (solid circle) and the value predicted by linear interpolation of the adjacent probes (hollow circle). When a set of probes fall closely along a line, this term of the variance will be small as the probe observations are consistent. When a probe falls far from the predicted value, this term of the variance will be large for that probe to reduce its weight. Computing the interpolated value using only adjacent probes will also downweight the probes on either side of the noisy probe. One might therefore use a larger set of probes near probe i, perform linear regression on those probes without i, and then compute the interpolated value.

smooth falloff from high intensity at a restriction site to low intensity.

The full estimate for the standard deviation is therefore

$$\sigma_i = \frac{L \cdot |x_i - \text{interpolated}(x_i)| + S \cdot s(\bar{x_i}) + J \cdot \sqrt{\frac{1}{J}\Sigma(x_{ij} - \bar{x_i})^2}}{L + S + J}$$

where L is the weight of the standard deviation estimate from the interpolation term.

6.3 Penalizing Systematic Error

The formula for the log likelihood described above treats each probe independently and cannot distinguish, for example, between a model that is a good fit to a noisy dataset and a model of the wrong fundamental shape for a clean dataset, as shown in figure 6-3.

An added term in evaluating the log-likelihood of a model's fit to the data penalizes systematic error based on the residuals of the fit. This term looks at the sign of each residual. Assuming that a positive residual and a negative residual are equally likely in a good fit, each sign should occur with probability one half. If each residual were independent, then all sequences of residuals would be equally likely. However, we assume that the residuals are not independent and that long strings of same-sign residuals are unlikely.

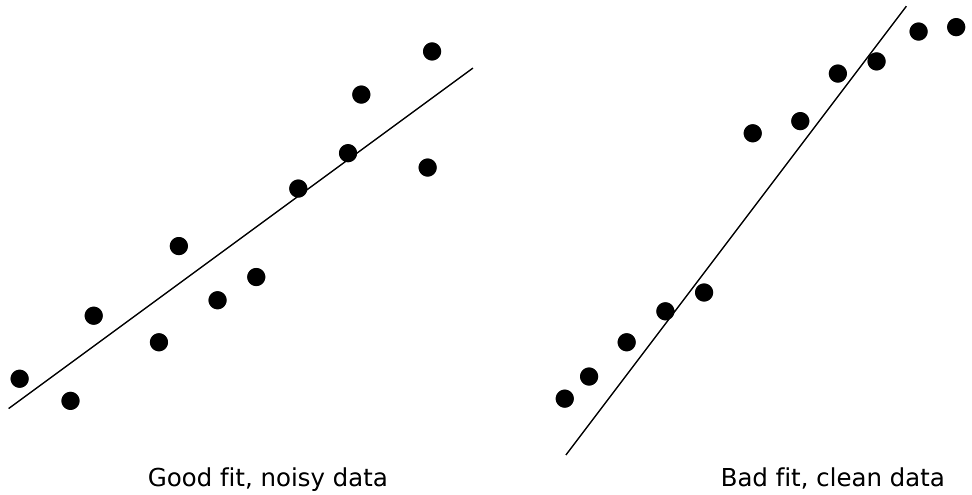


Figure 6-3: Independently scoring probes may fail to distinguish between a good fit to noisy data and systematic error. On the left, the data clearly falls along a single line but seems noisy. On the right, the data is clearly drawn from a two different models; however, when fit with a single line, some probes fit well and others poorly. We use a scoring term that looks at the signs of the residuals to penalize models that seem to consistently over or under predict the true values over large, continuous regions.

x consecutive residuals with the same sign add $log(\frac{1}{2}) \cdot c^x$ to the log likelihood. As x increases, suggesting that the predicted values are too large or too small across a large region, this penalty term increasingly reduces the log likelihood. In practice, the values for c are close to one (e.g. 1.05) to prevent this term from outweighing the first part of the log likelihood formula. Data from a substantially more densely tiled array might benefit from a larger value such that this penalty term increases proportionally to the number of probes per kilobase.

6.4 Limiting Outlier Influence

Microarray datasets sometimes contain extreme outliers, for example from non-specific hybridization or probes that match many genomic locations. We found it important to limit the influence of any single observation on the log-likelihood computation by setting a minimum value for the probability of an observation. This limit expresses the fact that we trust any observation only so much. An empirically determined value between .001 and .0001 seems to work well.

Chapter 7

Theoretical Limits of Detection

Before we explain a functional analysis method for Ruler Array data, we should determine whether the protocol and the analysis will be able to detect changes given our assumptions about the observed intensities and noise in the experiment. This chapter presents two simple models that indicate whether the observed intensities contain enough information to reveal the presence of an indel.

7.1 Highly Simplified Indel Detection

Imagine a highly simplified analysis problem of testing for a proposed indel at a known position in a two channel experiment. This problem can be solved by a simple statistical test for difference of means; we wish to determine whether the mean ratio in one side of the interval is the same as the mean ratio in the other. No difference in mean implies the absence of an indel; a difference of means indicates that a change is present at that location.

While the real analysis method won't know the indel position in advance, this simplified problem should give us some feel for the Ruler Array's ability to detect indels given varying amounts of noise in the data. If this simplified model indicates that the Ruler Array will not be able to detect a change, then we do not expect the actual experiment to detect a change when random noise, systematic bias, and other experimental imperfections cloud the situation.

7.2 Ratio Difference

If the observed ratios are distributed normally about their mean, then we can define a test statistic

$$z = \frac{(\bar{x_l} - \bar{x_r}) - (\mu_l - \mu_r)}{\sqrt{(\frac{\sigma_l^2}{n_l} + \frac{\sigma_r^2}{n_r})}}$$

Since the null hypothesis in this test is no difference in mean between the left and right sides of the interval, $\mu_l = \mu_r$. $\bar{x_l}$ and $\bar{x_r}$ are the mean ratio in the observed data and σ_l and σ_r are the observed standard deviations. n_l and n_r are the number of observations, which will depend on the size of the interval and the probe tiling density.

We can estimate σ under the assumption that the two channel intensities come from normal distributions about their means and are independent (the values are not independent, but the noise may be). Under our standard assumptions about the noise in Ruler Array intensities, the standard deviation of an observation increases linearly with the magnitude of the observation (see section 6.2).

Tables 7.1, 7.2, and 7.3 show the p-values from this test given probe spacings of 55bp, 10bp, and 1bp at a variety of interval sizes, indel positions, and indel sizes. These results indicate that at 55bp tiling density, insertions smaller than 100bp will generally be undetectable while larger changes can be detected in many circumstances.

7.3 Intensity Difference

A different approach to predicting what indels the ruler array may detect compares how well two simple yet ideal models predict the intensities observed in the presence of an insertion. This test assumes that we know which channel contains the indel but that we do not know its location. The channel with the change is the “indel channel” and the other channel is

Interval size 1000

	Distance from Restriction Site			
Size	200	400	600	800
10	0.958224	0.928467	0.881864	0.762793
30	0.874426	0.788061	0.661063	0.393227
70	0.708997	0.531198	0.319916	0.070752
100	0.590398	0.370937	0.162509	0.014993
200	0.265020	0.071149	0.007181	0.000024
300	0.082263	0.005860	0.000073	0.000000
500	0.001597	0.000001	0.000000	0.000000
800	0.000000	0.000000	0.000000	0.000000

Interval size 2000

	Distance from Restriction Site			
Size	400	800	1200	1600
10	0.959109	0.934290	0.898777	0.809744
30	0.876761	0.803731	0.703471	0.481103
70	0.712946	0.558042	0.376197	0.115594
100	0.594537	0.398926	0.206856	0.028863
200	0.266212	0.082276	0.011459	0.000047
300	0.080995	0.007047	0.000128	0.000000
500	0.001379	0.000001	0.000000	0.000000
800	0.000000	0.000000	0.000000	0.000000

Interval size 4000

	Distance from Restriction Site			
Size	800	1600	2400	3200
10	0.951829	0.928419	0.900794	0.837200
30	0.854821	0.785786	0.706905	0.539765
70	0.663392	0.518866	0.374884	0.156197
100	0.528119	0.350506	0.200919	0.043920
200	0.185041	0.050959	0.008288	0.000067
300	0.036573	0.002176	0.000041	0.000000
500	0.000112	0.000000	0.000000	0.000000
800	0.000000	0.000000	0.000000	0.000000

Table 7.1: The p-value for detecting a difference in ratio given a single insertion of the specified size. These values were computing assuming a tiling density of one probe per 55bp and a standard deviation of the ratios of .66.. While the statistical test is performed under the null hypothesis that the mean ratio in each part of the interval is the same, our computational experiment is performed knowing that this is not the case. Hence a small p-value leads us to reject the null hypothesis and correctly detect the indel. In the case of no-indel, the p-value is the probability of a false positive call. These results indicate, for example, that a 300bp indel should be detectable in a 2kb interval but that a 70bp change will probably be missed.

Interval size 1000

	Distance from Restriction Site			
Size	200	400	600	800
10	0.903029	0.835980	0.734042	0.494034
30	0.713188	0.535269	0.316149	0.053033
70	0.385382	0.148708	0.022966	0.000042
100	0.210598	0.039072	0.001407	0.000000
200	0.009530	0.000032	0.000000	0.000000
300	0.000053	0.000000	0.000000	0.000000
500	0.000000	0.000000	0.000000	0.000000
800	0.000000	0.000000	0.000000	0.000000

Interval size 2000

	Distance from Restriction Site			
Size	400	800	1200	1600
10	0.905077	0.849196	0.771183	0.585375
30	0.718332	0.566546	0.384167	0.110387
70	0.392167	0.176735	0.043055	0.000363
100	0.215717	0.051727	0.003903	0.000001
200	0.009711	0.000061	0.000000	0.000000
300	0.000049	0.000000	0.000000	0.000000
500	0.000000	0.000000	0.000000	0.000000
800	0.000000	0.000000	0.000000	0.000000

Interval size 4000

	Distance from Restriction Site			
Size	800	1600	2400	3200
10	0.888261	0.835870	0.775643	0.641497
30	0.670424	0.530788	0.389969	0.164701
70	0.311395	0.136821	0.042484	0.001313
100	0.142282	0.031310	0.003456	0.000005
200	0.002051	0.000007	0.000000	0.000000
300	0.000001	0.000000	0.000000	0.000000
500	0.000000	0.000000	0.000000	0.000000
800	0.000000	0.000000	0.000000	0.000000

Table 7.2: The p-value for detecting a difference in ratio given a single insertion of the specified size with a probe spacing of 10bp.

Interval size 1000

Size	Distance from Restriction Site 200	 400	 600	 800
10	0.727765	0.604740	0.483391	0.313279
30	0.293617	0.121152	0.038659	0.004338
70	0.013136	0.000306	0.000003	0.000000
100	0.000348	0.000000	0.000000	0.000000
200	0.000000	0.000000	0.000000	0.000000
300	0.000000	0.000000	0.000000	0.000000
500	0.000000	0.000000	0.000000	0.000000
800	0.000000	0.000000	0.000000	0.000000

Interval size 2000

Size	Distance from Restriction Site 400	 800	 1200	 1600
10	0.733317	0.634530	0.548551	0.421183
30	0.302757	0.151896	0.072626	0.018580
70	0.014492	0.000732	0.000030	0.000000
100	0.000404	0.000001	0.000000	0.000000
200	0.000000	0.000000	0.000000	0.000000
300	0.000000	0.000000	0.000000	0.000000
500	0.000000	0.000000	0.000000	0.000000
800	0.000000	0.000000	0.000000	0.000000

Interval size 4000

Size	Distance from Restriction Site 800	 1600	 2400	 3200
10	0.688094	0.604496	0.556590	0.492398
30	0.224021	0.117113	0.076167	0.040501
70	0.003825	0.000200	0.000029	0.000002
100	0.000028	0.000000	0.000000	0.000000
200	0.000000	0.000000	0.000000	0.000000
300	0.000000	0.000000	0.000000	0.000000
500	0.000000	0.000000	0.000000	0.000000
800	0.000000	0.000000	0.000000	0.000000

Table 7.3: The p-value for detecting a difference in ratio given a single insertion of the specified size with a probe spacing of 1bp.

the “control channel.”

The two models are

1. Null hypothesis: there is no difference between channels (i.e., no indel) and the intensities in both channels have the same mean. In this case, the mean for each probe is the average of the observed intensities in the control and indel channels. We assume that the intensities in the both channels are normally distributed around the mean with a standard deviation that scales linearly with intensity.

2. Alternative hypothesis: there is a difference between the two channels (i.e. presence of an indel). The mean for each probe's observation in each channel is the probe's observed intensity in that channel. As before, we assume that the intensities in the both channels are normally distributed around their mean with a standard deviation that scales linearly with intensity.

Under the null hypothesis, the model will be an imperfect fit to both channels. Under the alternative hypothesis, the model fits both channels perfectly.

Instead of testing for a difference of means, we compute the log-likelihood of the data under both models and compare. A large difference in log-likelihood indicates that the second model is much better. A small difference indicates that the data in the indel channel is similar to the control channel.

To determine a “significant” difference in log-likelihood, we need a prior probability that an interval contains an indel. If we believed, for example, that the probability of an indel at any position is .0001, then the difference in log-likelihood between the two models must exceed $-\log(.0001) = 9.2$ for us to determine from the data than an indel is present. If the difference in log-likelihood is less than this threshold, then the strength of the data does not overcome our prior belief that no indel is present.

Tables 7.4 and 7.5 show the log-likelihood differences for various interval and indel sizes with probe spacings of 55 and 10bp. These simulation results indicate that the Ruler Array will miss some indels, for example a 100bp indel at position 800 of a 1kb interval, and also give some sense of the ease with which various changes might be found. As with the ratio-based approach in the previous section, these results indicate that the Ruler Array will have difficulty finding changes smaller than 100bp with an array tiled at 55bp.

7.4 Conclusion

The models presented here develop the simple ratio changes presented in chapter 5 to account for the information derived from multiple probes. These simulations indicate that the Ruler Array should be able to detect large changes of one hundred basepairs or more but will have difficulty detecting changes of only a few dozen bases.

Interval size 1000

	Distance from Restriction Site			
Size	200	400	600	800
10	0.129368	0.098774	0.078977	0.124359
30	1.148553	0.876061	0.694530	1.040962
70	6.056358	4.611733	3.602645	4.926469
100	12.017085	9.141278	7.075283	9.091845
200	42.588220	32.327887	24.486932	26.637891
300	80.872414	61.376472	46.049552	44.966588
500	134.056126	102.363908	77.881117	71.025276
800	118.213744	92.503262	77.099187	78.735333

Interval size 2000

	Distance from Restriction Site			
Size	400	800	1200	1600
10	0.136845	0.102345	0.072397	0.072502
30	1.226130	0.916722	0.646850	0.633101
70	6.597385	4.929823	3.462952	3.247397
100	13.311676	9.943456	6.964192	6.341594
200	50.418702	37.631006	26.159396	21.951577
300	104.352623	77.874048	53.914983	42.691221
500	219.458475	164.167479	113.986654	85.924688
800	269.857782	204.354090	147.240765	120.141398

Interval size 4000

	Distance from Restriction Site			
Size	800	1600	2400	3200
10	0.200130	0.145474	0.092471	0.051070
30	1.797863	1.306545	0.829940	0.455110
70	9.730225	7.068140	4.484196	2.426934
100	19.729098	14.327521	9.082316	4.872016
200	76.213994	55.316268	34.999334	18.329674
300	161.897514	117.520103	74.330912	38.337600
500	367.812415	267.946166	170.516627	87.810963
800	533.137553	395.251484	260.069783	145.505825

Table 7.4: Maximum difference in log-likelihood for various insertions with a probe spacing of 55bp and $\sigma = .3 \cdot \mu$.

Interval size 1000

	Distance from Restriction Site			
Size	200	400	600	800
10	0.129368	0.098774	0.078977	0.124359
30	1.148553	0.876061	0.694530	1.040962
70	6.056358	4.611733	3.602645	4.926469
100	12.017085	9.141278	7.075283	9.091845
200	42.588220	32.327887	24.486932	26.637891
300	80.872414	61.376472	46.049552	44.966588
500	134.056126	102.363908	77.881117	71.025276
800	118.213744	92.503262	77.099187	78.735333

Interval size 2000

	Distance from Restriction Site			
Size	400	800	1200	1600
10	0.136845	0.102345	0.072397	0.072502
30	1.226130	0.916722	0.646850	0.633101
70	6.597385	4.929823	3.462952	3.247397
100	13.311676	9.943456	6.964192	6.341594
200	50.418702	37.631006	26.159396	21.951577
300	104.352623	77.874048	53.914983	42.691221
500	219.458475	164.167479	113.986654	85.924688
800	269.857782	204.354090	147.240765	120.141398

Interval size 4000

	Distance from Restriction Site			
Size	800	1600	2400	3200
10	0.200130	0.145474	0.092471	0.051070
30	1.797863	1.306545	0.829940	0.455110
70	9.730225	7.068140	4.484196	2.426934
100	19.729098	14.327521	9.082316	4.872016
200	76.213994	55.316268	34.999334	18.329674
300	161.897514	117.520103	74.330912	38.337600
500	367.812415	267.946166	170.516627	87.810963
800	533.137553	395.251484	260.069783	145.505825

Table 7.5: same thing again for probe spacing 10bp

Chapter 8

Analyzing Ruler Array Data Channel Ratios

One technique for analyzing Ruler Array data uses a Hidden Markov Model (HMM) to identify insertions or deletions from regions of high or low ratio in a two channel dataset.

8.1 Hidden Markov Models

A Hidden Markov Model describes data from a sequence of observations. The underlying model specifies how unobserved states produce the observed values and determines the probabilities of transitioning from one state to another at each step. The model is specified by

- a set of states describing the process (e.g. no change between samples, insertion in sample one)
- transition probabilities that specify the probability of changing from state i to state j at each step
- emission probabilities that specify the probability of the visible output tokens from each underlying state

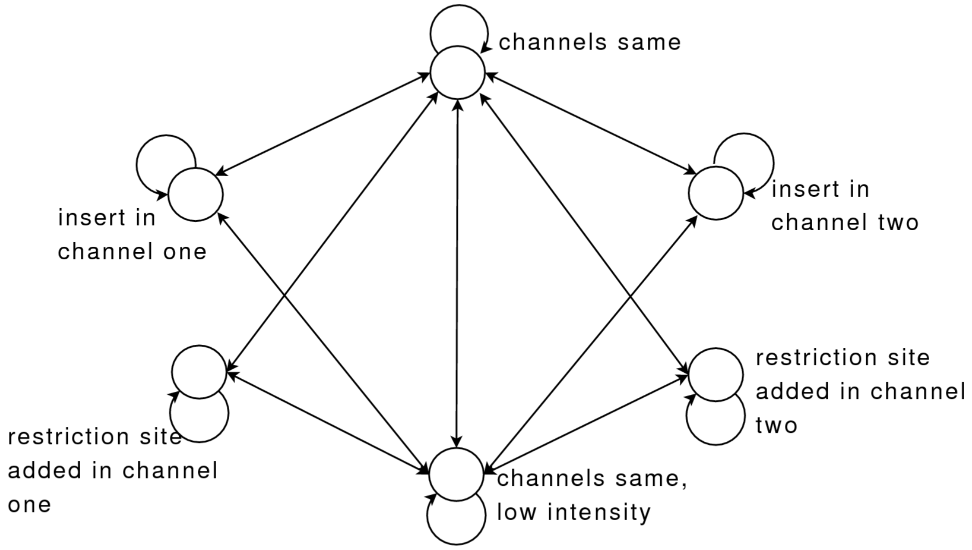

channels same
insert in channel one
insert in channel two
restriction site added in channel one
restriction site added in channel two
channels same, low intensity

Figure 8-1: Six state HMM to model the channel ratio in Ruler Array data. The states represent the underlying sequence around each probe: no change between the samples, an indel, or an added restriction site. The emissions from the HMM are the observed array intensities and the ratio between the two channels. The two "channels same" states represent the most common case in Ruler Array data- no difference between the two samples. The low intensity state permits a different distribution over the channel intensities when no restriction site illuminates a probe. The changed states represent either insertions or single strain restriction sites; both change the ratio between the two channels but will tend to have different intensities- indels can have high intensity in both channels whereas single strain restriction sites yield high intensities only in one channel.

- the probability of the model starting in each state

8.2 Our Model

In our model, the sequence of observations correspond to probes along the genome; the observations include the ratio, individual channel intensities, and slope of the ratios. Regions of high ratios are best explained by the underlying state corresponding to an insertion in sample two whereas regions of ratios near one are best explained by the two samples being the same.

The HMM model uses six underlying states:

1. Channels same
2. Channels same, low intensity
3. Insertion in channel one
4. Insertion in channel two
5. Restriction site added in channel one
6. Restriction site added in channel two

The transition probabilities tune the expected length of the segment with high or low ratios as well as the overall probability of an insertion. In general, however, self transition (from state i to state i) probabilities are large while transitions to other states are less likely.

We modeled the outputs from each state as multivariate Gaussians over the ratio, slope of the ratio, and the two channel intensities. The channel intensities are needed to identify the low intensity state in which even extreme ratios are meaningless.

8.3 HMM Learning

The parameters for the multivariate Gaussians are estimated from a set of training data that has been manually labeled. The **channels same** and **low intensity** states are easiest to learn as they are the most common.

Given a Ruler Array dataset, the output variables (ratio, slope of the ratio, and channel intensities) can be computed quickly and in linear time. The most likely sequence of states can also be found in linear time with the Viterbi algorithm.

8.4 Evaluation

An HMM-based method proved to be a poor technique for analyzing most Ruler Array data. If there are no differences between the two samples in an interval and if the intensities have been properly normalized then we expect a ratio of roughly one across the interval. However, the Ruler Array data contains many examples of intervals in which the observed ratios are not centered at one despite the absence of indels between the two strains. In these cases, the baseline ratio, or “base ratio”, typically peaks at the restriction site and then falls off towards the far end of the interval.

We believe that the non-unity base ratios probably come from exponential amplification of material (rather than the linear amplification expected from our extensions) that occurs with low probability. Since the initiation of PCR is relatively rare but the product has a potentially huge impact on the resulting intensities, one sample’s intensities may be higher than the other’s. Figure 8-2 shows how exponential amplification may begin during the supposedly linear extensions.

While we can predict the ratios in the presence and absence of indels, the HMM cannot easily handle this phenomenon without an explosion of the number of states. We expect that other techniques that also analyze only the channel ratios will suffer from a similar problem. For example, the methods of Erdman and Zhang [14, 57] performed well on CGH and timeseries expression data but seem unfit for the Ruler Array data.

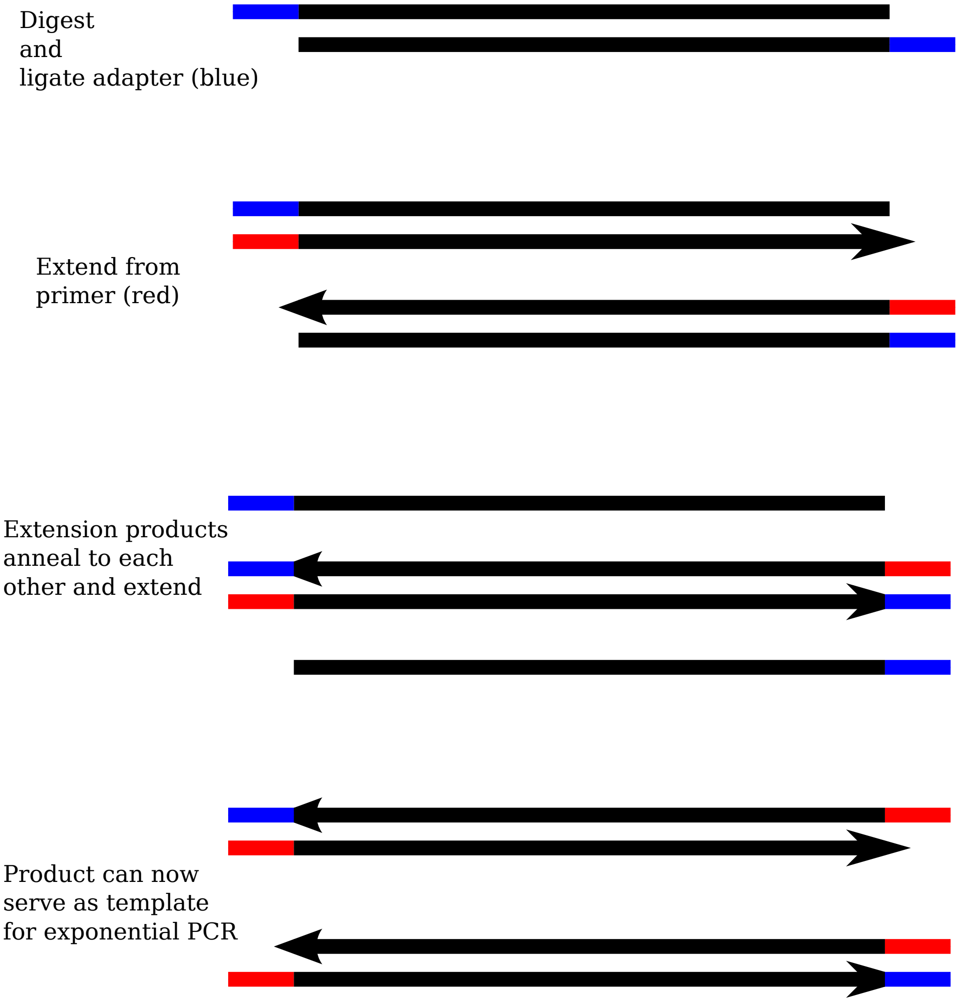

Digest
and
ligate adapter (blue)
Extend from
primer (red)
Extension products
anneal to each
other and extend
Product can now
serve as template
for exponential PCR

Figure 8-2: Ruler Array extensions may result in exponential PCR amplification. In the standard or expected case of Ruler Array extensions, the amount of product increases linearly with the number of cycles- at most one product molecule comes from each genomic template during a cycle. If the products from opposite strands of an interval anneal during a later extension cycle, they may complete the extension to include the complement to the primer (i.e. the adapter) on both ends. In future cycles, this material may begin exponential amplification as in PCR, producing both full length product and partial products. The amount of product in the interval will therefore depend heavily on when the first product-product extension takes place Because the initial annealing of product material is probabilistic, the amount of product is therefore probabilistic and may differ between the two samples in the Ruler Array experiment, leading to an intensity ratio far from one.

Chapter 9

Analyzing Ruler Array Data with Segment Fitting

Both theoretical predictions and empirical observation of numerous ruler array experiment indicate that the log-intensities observed on the microarray should drop linearly with distance from a restriction site (see figures 3-2 and 4-1). An interval between restriction sites that contains no length polymorphisms relative to the reference genome should therefore be modeled as a single line. However, an interval containing an insertion, deletion, or inversion will require two or more lines to appropriately fit the data.

We implemented a segment fitting procedure that finds the optimal fit of the data by lines or some other function. This segmentation defines a set of boundary points between segments; some boundary points will correspond to restriction sites and the remainder are candidate insertions or deletions.

Many previous studies and techniques have addressed similar problems under the labels of change-point analysis or segmented regression. For example, Tishler [48] and Gallant [18] presented methods for finding the split points where the function is continuous. The use of Dynamic Programming to perform the segmentation and fitting goes back to at least Hawkins [23] and has been used in recent microarray analyses such as David [12]. However, we are not aware of previous implementations of the joint segmentation procedure described here

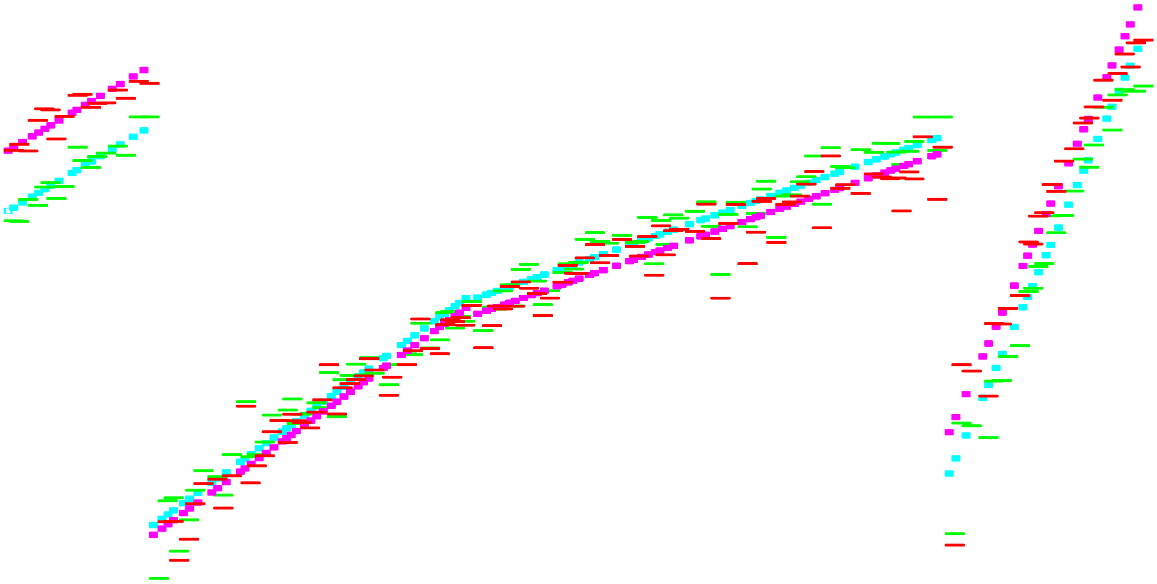


Figure 9-1: Sample segment fitting results from the segment of chromosome seven shown in figure 3-2. The red and green marks indicate the datapoints and the purplish marks show the fitted line segments.

which links the segmentation and splitting of the two channels through the prior probability of using the same split point and prior probability of using the same parameters in both channels.

9.1 Segment Fitting

The goal of segment fitting is to find the optimal fit of a function f to the observations at probes $1..n$ by finding the optimal set of segment boundaries. Each segment is fit by f with a single set of parameters; the parameters may change across segment boundaries. For example, if f is a line, then the parameters are the intercept and slope.

Figure 9-1 shows Ruler Array data and segment fitting results using line segments.

9.1.1 Recursive Solution

Let $f(a, b)$ be the result of fitting the function to the datapoints a through b where $1 \leq a \leq b \leq n$. $\mathcal{L}(f(a, b))$ is the log-likelihood of the observations given the values predicted by the fit. Let $\mathit{opt}(a, b)$ be the optimal fit to the points $a..b$; the optimal fit may be either a single set of parameters for f or it may be a set of split points and the parameters for each interval between split points. The final goal is to find $\mathit{opt}(1, n)$. To find $\mathit{opt}(a, b)$, we use the following recursive procedure:

- First find $f(a, b)$ and its likelihood $\mathcal{L}(f(a, b))$. This is the "fit" outcome.

- For each $k : a \leq k < b$, recursively find the combined likelihood of $\mathit{opt}(a, k)$ and $\mathit{opt}(k + 1, b)$, $\mathcal{L}(\mathit{opt}(a, k)) + \mathcal{L}(\mathit{opt}(k + 1, b))$. Remember the k which gives the highest combined likelihood. This is the "split" outcome.

- Compare the result of fitting a single segment to $a..b$ to the result of fitting two segments with the boundary at the best k and choose whichever has the higher log likelihood. This gives $\mathit{opt}(a, b)$. If fitting gives the best result, then $\mathcal{L}(\mathit{opt}(a, b)) = \mathcal{L}(f(a, b))$. If the interval is split, then $\mathcal{L}(\mathit{opt}(a, b)) = \mathcal{L}(\mathit{opt}(a, k)) + \mathcal{L}(\mathit{opt}(k + 1, b))$.

9.1.2 Correctness

The recursive method for finding $\mathit{opt}(a, b)$ ends up summing $log(p(x_k|\hat{x_k}, \sigma_k))$ over all $a \leq k \leq b$, regardless of the segmentation. The segmentation influences only the parameters of the fitting and therefore the $\hat{x_k}$. Since the $p(x_k|\hat{x_k}, \sigma_k)$ are independent (the observed values are independent given the predicted values), the final probability of the data does not depend on the segment boundaries directly, only on the predicted values.

The choices from which the recursive solution chooses the best log-likelihood are exhaustive: either the interval $a..b$ will be fit with a single segment or it will be split at some point between a and b, inclusive. The segment may be split at more than one point, but that is handled through the recursive use of opt. By choosing the option with the highest log-likelihood from an exhaustive set of options, the algorithm produces the output with the highest log-likelihood for the interval.

9.1.3 Priors on Segment Fitting

As described, the procedure will always choose to split because it achieves a better fit by considering fewer points at a time. The result will be segments of one observation each in which the model, learned from the single datapoint, exactly fits the datapoint. However, we know that the data ought to be fit as intervals with some average length so we include a term that describes that average length as a prior probability on fitting an interval rather than splitting.

Adding a prior on the probability of fitting an interval $a..b$ with a single set of parameters rather than splitting requires only adding another term to the two log likelihoods. Including the priors, the log-likelihood of the "split" and "fit" outcomes are

Outcome	Log-Likelihood
fit	$\mathcal{L}(f(a,b)) + log(p_{\text{fit}})$
split	$\mathcal{L}(f(a,k)) + \mathcal{L}(f(k+1,b)) + log(1-p_{\text{fit}})$

9.1.4 Dynamic Programming Solution

The recursive solution to $opt(1..n)$ is inefficient as it recomputes subproblems many times. For example, $opt(1..n)$ computes $opt(2..n)$, $opt(3..n)$, $opt(4..n)$, etc. $opt(2..n)$ computes $opt(3..n)$, $opt(4..n)$, etc. A more efficient approach computes $opt(i,j)$ only once and stores the result for future use.

The dynamic programming procedure computes $opt(a,b)$ for all a,b starting with the the smallest range of observations $a=b$ and works up to larger intervals. Since the two intervals $a..k$ and $k+1..b$ are smaller than $a..b$ for all k, the solutions and log-likelihoods for those intervals will be ready when the procedure considers $a..b$.

Each solution $opt(a,b)$ is stored in a 2D table as it is computed, taking $O(n^2)$ space. Each element in the table stores either the parameters of the fit or the k at which the interval was split; each element (a, b) also stores $\mathcal{L}(opt(a, b))$. Filling the table will take at least $O(n^3)$ time since each of the $O(n^2)$ entries requires iterating over k; the exact runtime will depend on the time required to compute $f(a, b)$.

9.2 Segment Fitting with Linear Regression

As the log-intensities in a Ruler Array experiment fall roughly linearly as distance from a restriction site increases, we can use a simple linear model as f. For each interval a, b, the fitting procedure performs linear regression on the values from a to b, trying to predict the log intensity observed at position k as

$$\alpha + \beta \cdot \text{distance}(k, b)$$

For probes on the plus strand, the slope of the log intensities is positive and the distances are computed from the genomic position of probe b. For probes on the minus strand, the slope of the log intensities is negative and the distances are computed from a. All other aspects of the computation are the same for two strands.

As mentioned in section 6, we can estimate the standard deviation for an observation based on repeated observations of the same probe, observations of nearby probes, or a prior belief about the reliability of an observation given its intensity. Since our linear model fits the log-intensities, we perform the linear regression on the log-intensities and then exponentiate the predicted values before computing the log-likelihood:

$$\hat{x_i} = e^{\alpha + \beta \cdot \text{distance}(k,b)}$$

Under the assumption that the intensity observations for a probe are normally distributed around their mean, we can compute the probability for the probe intensity that the regression predicts given the observed intensity:

$$\begin{aligned}\mathcal{L}(f(a,b)) &= \Sigma_{k=a..b} log(p(\hat{x}_k|x_k,\sigma_k)) \\ &= \Sigma_{k=a..b} log(\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(\hat{x_k}-x_k)^2}{2\sigma_k^2}})\end{aligned}$$

For the linear regression to maximize the log-likelihood, we use weighted linear regression. To maximize the log-likelihood, we must minimize

$$\Sigma_{k=a..b}\frac{(\hat{x_k}-x_k)^2}{2\sigma_k^2}$$

Linear regression finds the parameters to minimize

$$\Sigma_{k=a..b}(\hat{x_k}-x_k)^2$$

(the sum of the squares of the residuals). Weighted linear regression minimizes

$$\Sigma_{k=a..b}\frac{1}{w_k}(\hat{x_k}-x_k)^2$$

Thus setting the weights $w_k = \frac{1}{\sigma_k^2}$ makes the problems equivalent.

If X is the matrix of input variables with one column per variable and one row per datapoint. In the Ruler Array analysis, the first column is the constant one and the second column is the distance from the probe to the end of the interval. W are the weights, an $n \times n$ matrix in which $W_{kk} = w_k$ and all other entries are zero. Y is the vector of output observations. The parameters b are

$$b = (X^\top WX)^{-1}X^\top WY$$

9.3 Runtime of Segment Fitting with Linear Regression

The segment fitting algorithm performs two operations for every interval $a..b$ for $1 \leq a \leq b \leq n$:

1. Weighted linear regression to find the best single model for the entire interval.
2. A search over k to find the best point at which to split the interval.

Performing weighted linear regression on n observations with v variables requires time $O(v^2 n)$. The runtime breaks down as

- WY takes n multiplications
- $X^\top WY$ takes vn multiplications and at most $v(n-1)$ additions
- WX takes vn multiplications
- $X^\top WX$ takes $v^2 n$ multiplications and at most $v^2(n-1)$ additions.
- $(X^\top WX)^{-1}$ takes $O(v^2)$ operations

yielding an overall runtime of $O(v^2 n)$.

Since finding the combined log-likelihood in a split interval is computationally easy (the combined log-likelihood is the sum of the split log-likelihoods plus some prior), the linear regression dominates the work for each interval. The total runtime will thus be

$$\Sigma_{l=1}^{n}(n-l)O(lv^2) = O(v^2 n^3)$$

since there are $n-l$ intervals of length l. In our model, $v = 2$, the constant and the distance from the probe to the end of the interval.

9.4 Handling Experimental Replicates

Replicates of a Ruler Array experiment may present repeated measurements of the same values or they may use a different array design or different restriction enzyme. In either case, the segment fitting procedure requires few modifications.

When presented with replicates that use the same microarray design and the same restriction enzyme, the fitting procedure will run linear regression separately on each replicate, producing one set of parameters for each replicate, but will use the same set of splitting points for all replicates. The overall log-likelihood of an interval is the sum of the log-likelihoods of the replicates in that interval. Replicates on the same array platform increase the runtime linearly since the regression and likelihood computations must be run once per replicate.

Datasets from different array designs will simply increase the number of splitting points that the fitting procedure considers. Observations from different replicates are fit by models with different parameters, but split points between segments apply to all replicates. The increase in the number of split points increases the runtime and memory requirements quadratically (e.g. twice as many probe positions doubles n, quadrupling the runtime).

If the input includes data from different restriction enzymes, then the fitting procedure runs as before. However, the output will now include some split points corresponding to the first restriction enzyme, some split points corresponding to the second restriction enzyme, and the remainder corresponding to indels.

9.5 Independently Handling Two Channel Experiments

A simple extension of the segment fitting procedure to two channel experiments runs the fitting separately on each channel. Segment boundaries identified by fitting one channel but not the other are retained as candidate indels.

While simple to implement, this procedure is extremely sensitive to noise in the data; small changes in probe observations may cause the segment fitting to split at different points. The resulting split points must either be cleaned up with some heuristic method (e.g. split points within d bases are considered to be the same) or the method yields too many false positives to be useful.

9.6 Jointly Handling Two Channel Experiments

We can extend the procedure to simultaneously handle two sets of observations (two channels) by adding more cases from which the choice with the highest log-likelihood is chosen. In particular, we add the option to fit both channels with the same parameters, fit both channels with different parameters, fit one channel but split the other, to split both at the same point, or to split both channels at different points. We use the subscripts 1 and 2 to indicate a fit or solution to data only from that channel.

The new log-likelihood choices are

Outcome	Log-Likelihood of Data	Log-Likelihood of Parameters
fit both same parameters	$\mathcal{L}(f(a,b))$	$2 \cdot log(p_{\text{fit}}) + log(p_{\text{same params}})$
fit both diff. parameters	$\mathcal{L}(f_1(a,b)) + \mathcal{L}(f_2(a,b))$	$2 \cdot log(p_{\text{fit}}) + log(1 - p_{\text{same params}})$
fit one split one	$\mathcal{L}(f_1(a,b)) + \mathcal{L}(opt_2(a,k)) +$ $\mathcal{L}(opt_2(k+1,b))$	$log(P_{\text{fit}}) + log(1 - P_{\text{fit}}) +$ $log(1 - p_{\text{same params}})$
split both	$\mathcal{L}(opt(a,k)) + \mathcal{L}(opt(k+1,b))$	$2 \cdot log(1 - p_{\text{fit}})$ $log(1 - p_{\text{same params}})$

Variants on the fitting procedure might offer more possibilities, such as fitting the intervals with lines of the same slope but different intercepts. In practice, we found this variant important as many ruler experiments include intervals in which the log-intensities in the two channels are the same shape (i.e. the same slope) but have different intercepts.

9.7 A Non-Generative Model for Segment Fitting

Segment fitting with linear regression, or any other function relating intensity to distance, relies on a generative model for the Ruler Array intensities. That model may fail to account for certain features of the data and the experiment, for example by ignoring the effects of certain DNA sequences on the polymerase's processivity. These aberrant locations may result in false positive calls for insertions and deletions.

A different approach to segment fitting makes no assumption about the process that generates the Ruler Array intensities by modeling one channel as a function of the other. When an interval contains no insertions or deletions, the intensities from the first channel can be modeled as some function of the intensities in the second channel, e.g. intensity(Cy5) = $\alpha + \beta \cdot$ (intensity Cy3). However, an indel changes the relation between the channels.

This discriminative segment fitting splits the data into segments for which a single set of parameters holds. The boundaries between segments are the points at which the relationship between the two channels changes. As with the generative segment fitting, most such points will be restriction sites and the remainder represent candidate indels. Aberrant sequence features should not produce false positive indel calls as long as both channels are affected in the same way.

9.8 Segment Fitting Efficiency Hacks

Since the runtime of the segment fitting procedure increases superlinearly with the number of probes n, the overall runtime benefits greatly if we can split a problem into several pieces and run the dynamic programming on each piece individually. Ruler data can be split at

Large gaps between probes If the space between adjacent probes is longer than the maximum illumination distance for a restriction site, the data can be split between those probes. It is possible to split at smaller gaps too at the cost of possibly missing an indel that falls between the probes.

Regions of background intensity The data may be split at a region that is tiled but shows no intensity above background levels in any of the replicates being analyzed. Such regions require knowing what intensity level is background and choosing a threshold for identifying such a region (e.g. 40 consecutive probes of intensity less than 100). These background regions typically occur at large gaps between restriction sites.

The segment fitting procedure also runs faster when fitting is limited to some maximum interval size. For intervals larger than this size, no regression is necessary and the optimization consists solely of searching for the best split point (which takes time proportional to $j - i$ rather than $O((j-i)^2)$.

9.9 Calling Insertions and Deletions from Segment Fitting Output

As mentioned previously, a simple method for detecting insertions and deletions from the segment fitting output looks for the segment boundaries that exist only in one channel. In practice, we have augmented this procedure to recognize other patterns associated with indels.

Our best results have used joint segment fitting on both experimental channels. The "fit both channels with same parameters" outcome actually only fits both channels with the same slope, allowing the intercept to differ. This accounts for the observed phenomenon in which the absolute intensities differ but the shapes are similar; when analyzed in log-space, intensities that differ by some factor will be offset by some constant amount.

In addition to identifying single-channel segment boundaries as indels, the detection procedure identifies transitions from "fit both with same slope" to "fit with different slopes." This identifies regions in which the two channels are no longer the same or similar shapes. Many such points will be restriction sites, as changes in the character of the data often occur on restriction interval boundaries. The remainder ought to be insertions, deletions, or other events that change the data's character.

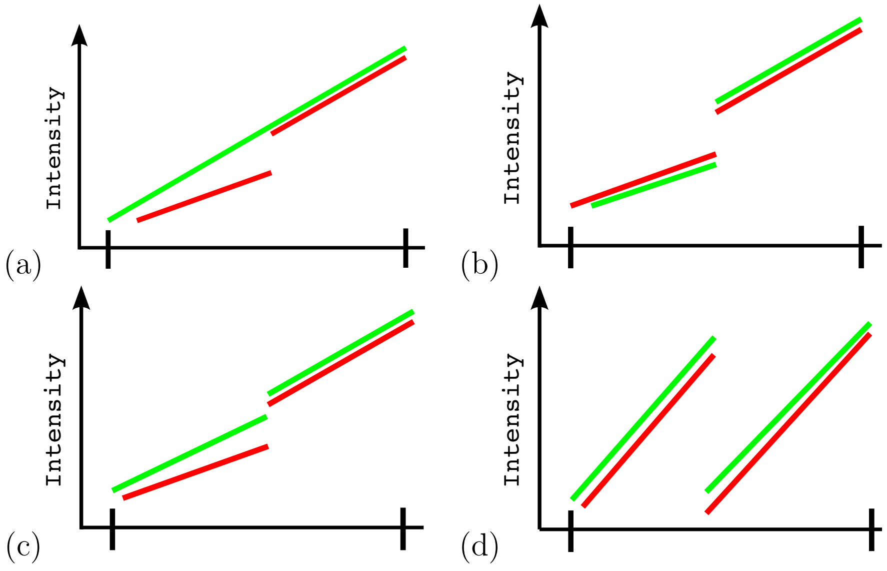


Figure 9-2: The four cases in which the Ruler Array analysis infers the presence of an indel from the segment fitting output. In (a), the segment fitting used one segment to fit the green channel but two segments to fit the red channel; consequently, the analysis makes a call at the split point in the red channel. In (b), the segment fitting used two segments in each channel. The green channel is greater to the right of the break but of lower magnitude to the left. If the change is large enough, the analysis calls this boundary an indel. This change is commonly observed at AT repeat length changes. Example (c) illustrates another change common at repeat length or repetitive element changes. There is a segment boundary in both channels, but the intensities drop much more in one channel than the other. A restriction site, or the insertion of an element that contains a restriction site such as a TY, generates the signature seen in (d).

Finally, the analysis flags boundaries between segments at which the channel ratio changes dramatically or across which the intensity changes as it does at restriction sites.

Figure 9-2 shows the four cases in which the analysis calls an indel from the segment fitting.

9.10 Conclusion

The Segment Fitting analysis combines the simple linear model of intensity vs distance from chapter 4 with the log-likelihood function

Chapter 10

Evaluating the Ruler Array

To evaluate the Ruler Array, we chose as our sample organisms two strains of *S. cerevisiae*, FY4 and Σ1278b. FY4 is very closely related to S288C, the strain sequenced to produce the reference *S. cerevisiae* sequence. Σ1278b was recently sequenced with long-read paired-end sequencing at the Broad Institute and was expected to contain a number of indels with respect to FY4. Since Σ1278b represents a common case of a newly sequenced genome and since we also have short, unpaired reads from a Solexa machine available, we were also able to compare the Ruler Array to several sequencing methods as an indel detection technology.

We have divided the evaluation of the Ruler Array into three parts:

- a technical evaluation in which we test the Ruler Array's ability to detect a set of confirmed indels between two strains of yeast (this chapter)

- a set of biologically-motivated test cases in which we focus on changes in certain classes of changes between the two genomes such as transposable elements and gene duplications (chapter 11)

- a comparison to two sequencing approaches, aCGH, and a hypothetical TIP-Chip experiment (chapter 12)

10.1 Technical Evaluation Method

To evaluate the laboratory protocol and computational analysis, we focused on indels of more than 100bp, a size that would be easy to verify with PCR and that the Ruler Array should be able to find given our expectations for the experimental noise and probe density.

The evaluation begins with four sets of genomic coordinates:

1. The calls made by the segmentation procedure. We used as input a single replicate of the FY4 vs Σ1278b experiment on an Agilent designed microarray with an average probe spacing of roughly 55bp. The array contained roughly 240,000 60mer probes designed against one strand of the S288C genome. Our analysis only used the probes that could be assigned a unique genomic location.

2. The restriction sites. Calls mapped to restriction sites are not counted as indel calls.

3. The "must find" indels larger than 100bp. These must be found or will count as false negatives in the evaluation. The known indels were first identified from the Σ1278b and reference sequences by a Blast-based optimal alignment of S288C chromosomes against Σ1278b and then confirmed either by PCR or CGH.

4. Indels smaller than 100bp and single-strain restriction sites. These may be found by the analysis method and will be counted as true positives but do not count as false negatives if missed by the analysis (the "can find" list). While this definition of true positive seems to skew the results, we use it to avoid penalizing the Ruler Array for detecting small indels that we would like to find but do not necessarily expect to find.

The evaluation first matches the calls to the restriction sites. It then matches the remaining calls to the two lists of indels; each call is matched to at most one indel, selecting one arbitrarily if more than one known indel falls within the distance threshold (multiple Ruler Array calls may map to the same indel; we consider this correct if one imagines that a Ruler Array experiment would be followed by PCR to confirm the calls. Any of the calls identifying the indel would likely lead to confirmation of the change in the PCR step). A call matches the indel if the call falls within 200bp of the indel position. This procedure guarantees that that the Ruler Array will miss indels very near restriction sites (within 100bp) since the call will be matched to the restriction site rather than the indel; this is reasonable given that we cannot easily distinguish between the effects of the indel and the restriction site on the observed probe intensities when they are so close.

Before concluding that a call is a false positive, the analysis aligns the 400bp surrounding the call between the S288C and Σ1278b genomes and looks for SNPs and small indels. If the alignment shows more than four bases inserted or deleted in either direction or more than eight SNPs, the call is considered to have correctly identified a genomic change. Since these calls have not been verified by PCR and rely on a potentially incorrect assembly, some may be wrong (both in classifying an event as a true positive or in classifying it as a false positive); in either case, we expect this method to be correct in most cases and to provide a useful look at the Ruler Array's sensitivity to small changes.

The analysis classifies Ruler Array calls that do not match either list of indels and do not occur over smaller genomic changes as false positives.

10.2 Analysis Parameters

The parameters for the probe intensity variance estimate and the priors for the dynamic programming segment fitting play a key role in determining the accuracy of the Ruler Array method. The complete set of parameters that we used is

variance estimate from mean We used $s(x_i) = .3 \cdot x_i$ as the basic variance estimate. To make the estimate match figure 6-1 more closely and to incorporate our observations about which probes were most informative around errors in the segment fitting, the full form is

- $x_i > 10000 : s(x_i) = \ .3 \cdot 10000$
- $10000 > x_i > \mu_{\text{noise}} : s(x_i) = .3 \cdot x_i$
- $\mu_{\text{noise}} > x_i : s(x_i) = 2 \cdot .3 \cdot x_i$

variance estimate pseudocount We used a pseudo count of 4. That is, $S = 4$ in the equations in section 6.2.

variance weight for linear interpolation We used .1 as the weight for the component of the standard deviation that is the difference between the observed value and the value predicted by linear interpolation of the adjacent points. That is, $L = .1$ in section 6.2.

probability of splitting $log(p_{\text{split}}) = -3$.

probability of splitting at different points $log(p_{\text{different splits}}) = -3$.

probability of fitting with same parameters is .99999999

minimum probability of any observation given the model is .0001

base for systematic error term is 1.02

maximum distance over which to fit is 10kb

10.3 The Test Set

Table 10.1 presents the list of indels that the analysis must find.

The experimental readout used a 244k array from Agilent Technologies that tiled the Watson strand of the S288C genome with an average probe spacing of roughly 55bp. This array design does not tile many of the repetitive regions of the genome or the telomeres.

A single replicate of the Ruler Array experiment using EcoRI as the restriction enzyme allowed us to find 29 of the 35 events. In addition, the algorithm found 211 more events classified as true positives (single-strain restriction sites, smaller indels, etc) for a total of 240 true positives. Two-hundred and four events were classified as false positives for a true positive to false positive ratio of 1.18:1.

Position	Confirmed By	Found?	Description
1:180839-180880	PCR	Y	400bp insertion in sigma
1:182580-182925	PCR		300bp deletion in sigma
1:198200-203000	CGH	Y	4kb deletion in sigma
2:29655-35551	PCR	Y	2kb insertion in sigma
2:428094-429944	PCR	Y	1kb deletion in sigma
2:643485-643861	PCR		500bp deletion in sigma
2:644926-644926	PCR		TY insertion in sigma
2:801000-805000	CGH	Y	MAL32 duplication. 3kb or so added in sigma
4:434441-435118	PCR		600bp deletion in sigma
4:437140-438315	PCR	Y	400bp deletion in sigma
4:462154-462154	PCR	Y	1kb insertion in sigma
4:523000-527000	CGH	Y	(several small indels) 500bp gone in sigma
4:871500-885000	PCR	Y	300bp insertion in sigma
4:957500-958000	PCR	Y	100bp deletion in sigma
4:1023149-1023496	PCR		300bp deletion in sigma
5:207100-207400	PCR	Y	100bp insertion in sigma
6:30048-30048	assembly/blots	Y	telomere moved to another chromosome
8:175482-175482	PCR	Y	100bp insertion in sigma
8:86065-91139	CGH	Y	5kb deletion in sigma
8:93500-95000	CGH	Y	800bp of deletion in 2 pieces in sigma
9:349999-349999	PCR	Y	100bp deletion in sigma
9:434645-436741	CGH	Y	2.5kb deletion in sigma
10:21000-24500	CGH	Y	3kb deletion in sigma
11:310683-310883	CGH	Y	100bp gone in 288c
11:388778-388778	PCR		200bp insertion in sigma
11:513003-513603	PCR	Y	TY insertion in sigma?
14:34470-34470	PCR	Y	200bp deletion in sigma
14:429700-430000	PCR	Y	100bp insertion in sigma
14:546700-547100	PCR	Y	100bp deletion in sigma
14:765200-772500	CGH	Y	7kb deletion in sigma
14:777000-779000	CGH	Y	2kb deletion in sigma
15:30388-30388	PCR	Y	400bp insertion in sigma
16:926900-927900	CGH	Y	1kb deletion in sigma
16:928300-931300	CGH	Y	3kb deletion in sigma
16:932800-941700	CGH	Y	9kb deletion in sigma

Table 10.1: The 35 indels that must be found by the Ruler Array analysis. Alignments of the curated Σ1278b assembly to the S288C reference sequence predicted each indel, which we then confirmed with PCR, CGH, or chromoblot.

10.4 Evaluating the False Negatives

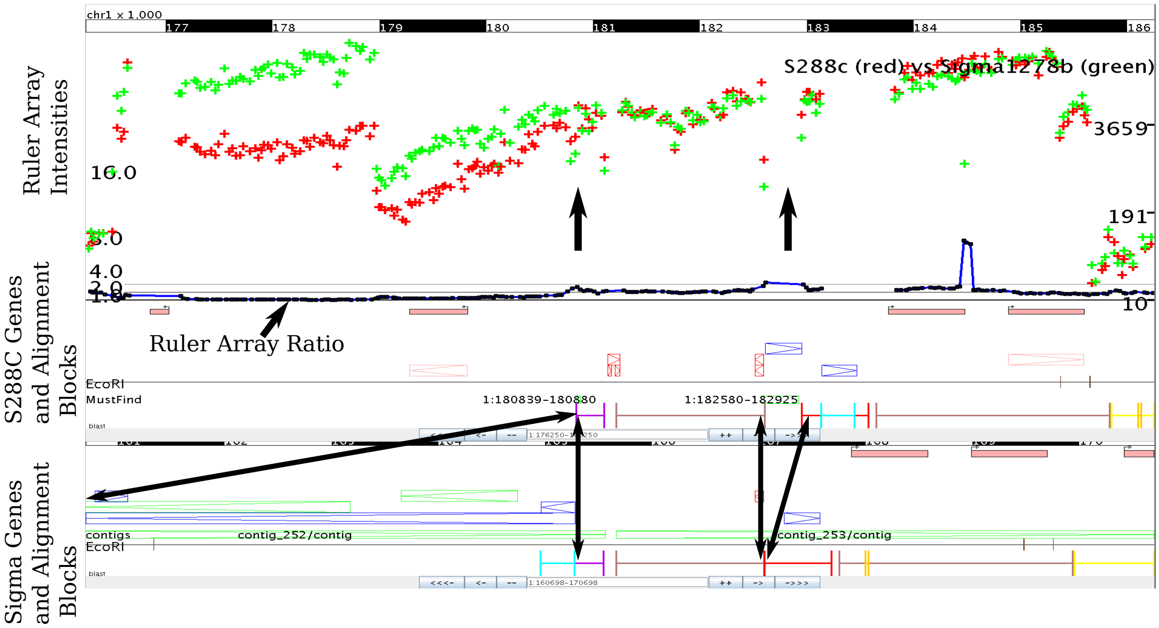
To evaluate why the Ruler Array analysis failed to detect an indel, we look at the ruler probe intensities, channel ratio, and the linefitting output. For each event, we show two plots. The first shows, from top to bottom

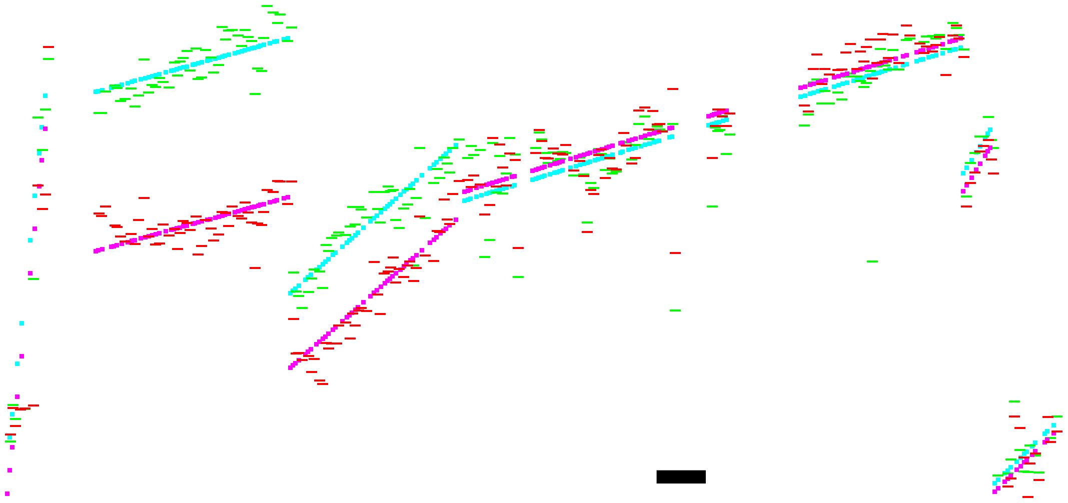
- the chromosomal coordinates for S288C
- the log-intensities from S288C (red) and Σ1278b (green) mapped to the S288C genome
- the channel ratio (blue) mapped to the S288C genome
- genes and other annotations for S288C
- the chromosomal coordinates and genes for Σ1278b
- the alignment between the strains (according to Blast[1]) as colored bars at the bottom of the plots. A blue bar, for example, shows a particular region that aligns between the two strains.

We also include a second plot that shows the log-intensities again as well as the line fitting output. The red and green marks are the same as in the first plot and a large black rectangle marks the missed indel.

10.4.1 1:182380-183125

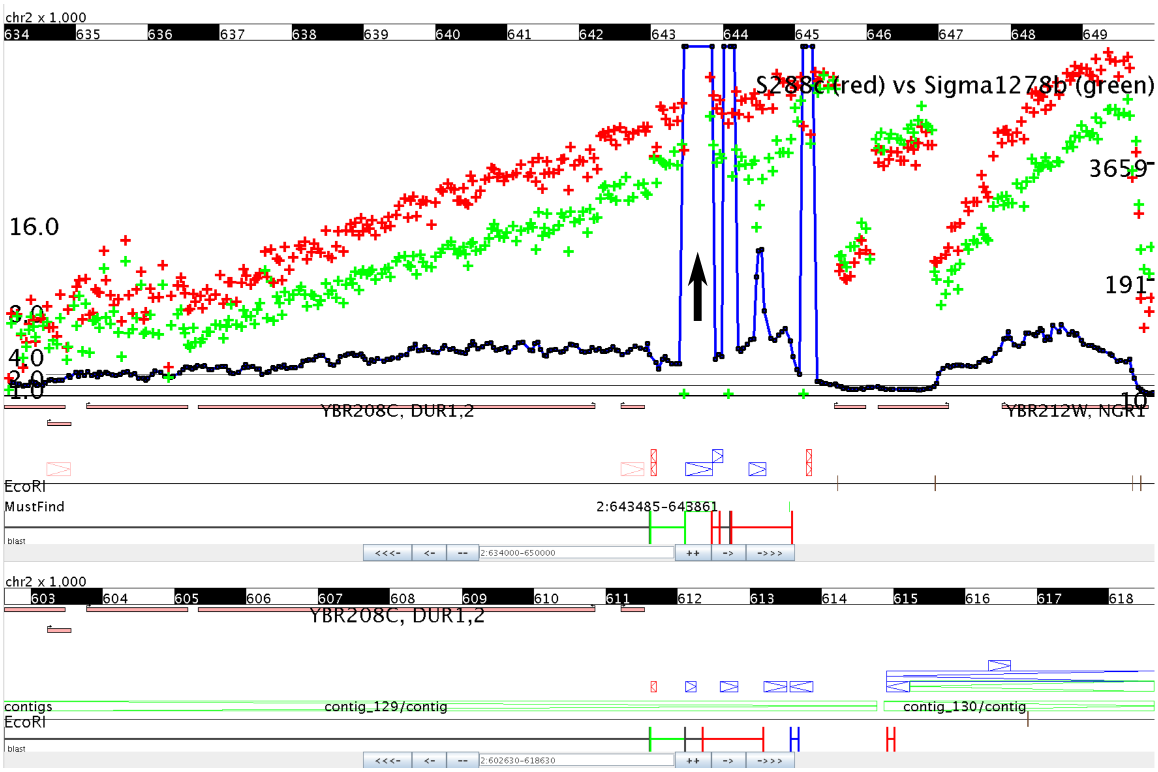
1:182380-183125 is a Σ element deletion in Σ1278b. The analysis misses 1:182380-183125 because of low probe coverage (a second, nearby Σ element is not tiled).

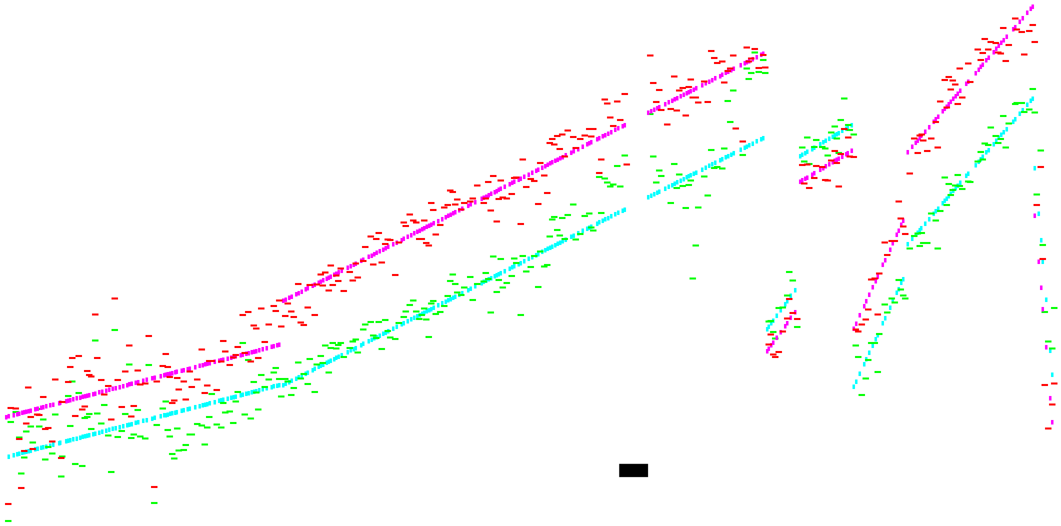




10.4.2 2:643285-644061 and 2:644726-645126

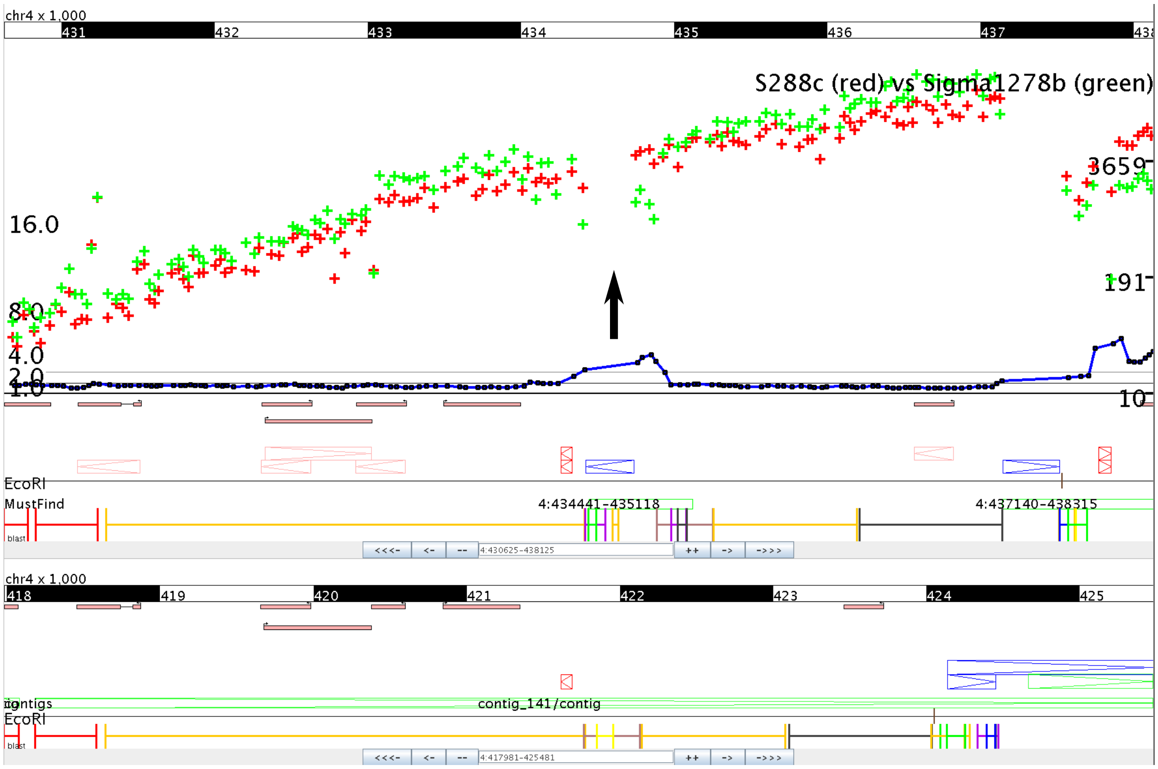
The two small deletions, 500bp and 100bp in sigma, 2:643285-644061 and 2:644726-645126 are near a much larger insertion of a TY1 element in sigma. The channel ratio changes substantially at 2:644726-645126, though it is too close to the EcoRI site to force the linefitting to use two segments. The intensity changes from the TY insertion also obscure the effects of the smaller changes.

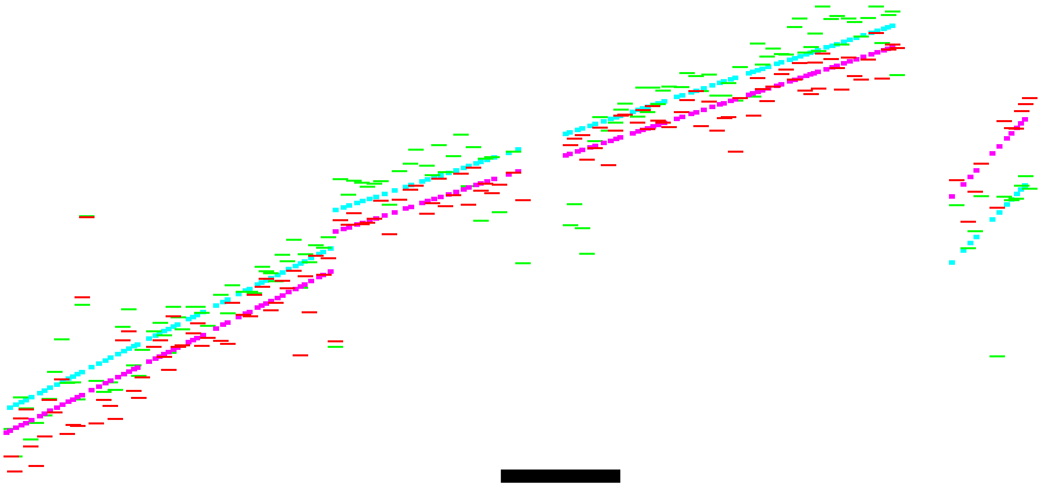




10.4.3 4:434241-435318

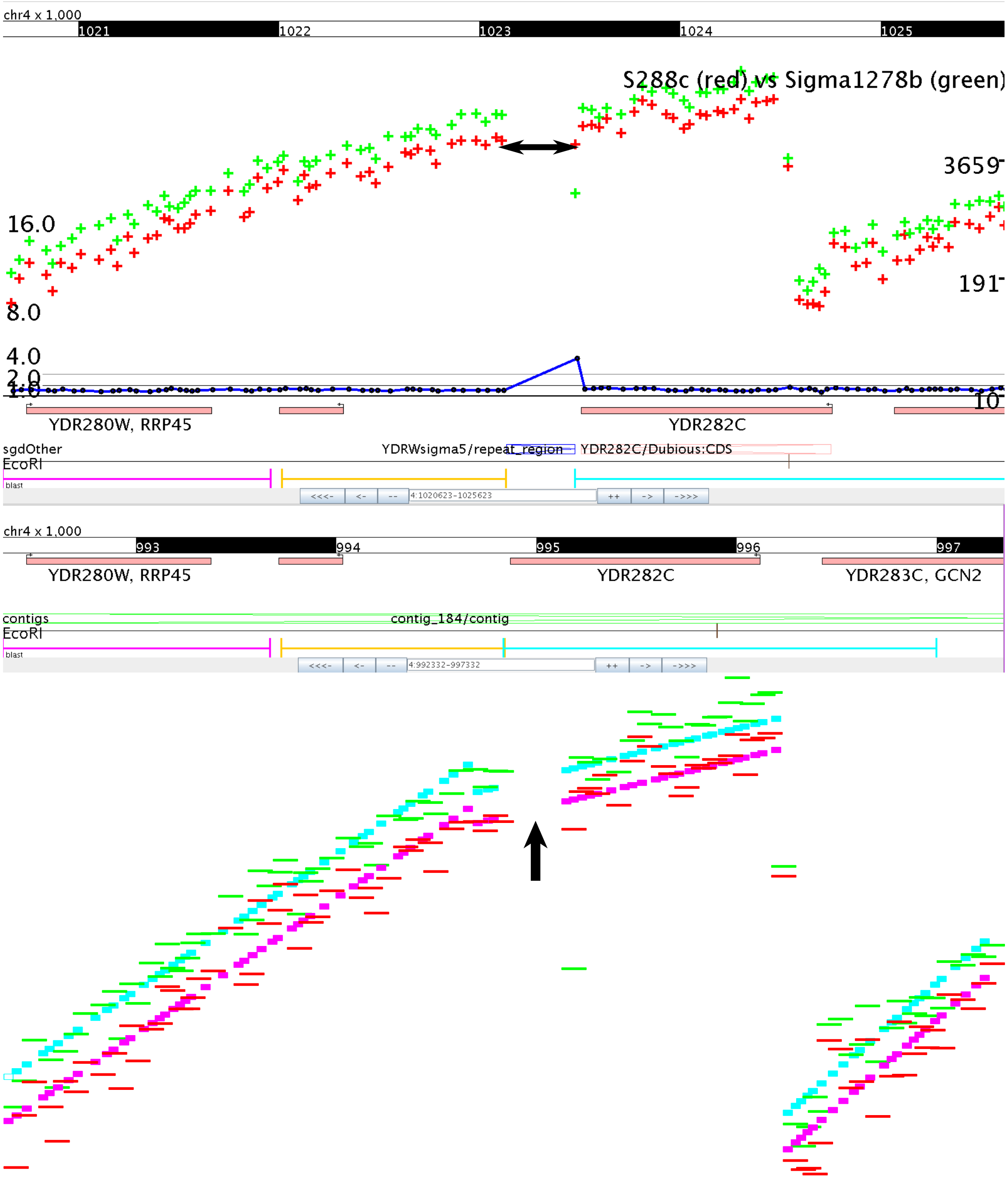
This Σ element is absent from Σ1278b. The intensity drop at 433k causes the model to split the line segments and puts this indel into a shorter segment; the short segment allows a looser fit (fewer observations to justify generating an indel call).


chr4 x 1,000
431
432
433
434
435
436
437
S288c (red) vs Sigma1278b (green)
16.0
8.0
4.0
2.0
1.0
3659
191
10
EcoRI
MustFind
4:434441-435118
4:437140-438315
blast
<<<-
<-
--
4:430625-438125
++
->
->>>
chr4 x 1,000
418
419
420
421
422
423
424
425
contigs
contig_141/contig
EcoRI
blast
<<<-
<-
--
4:417981-425481
++
->
->>>



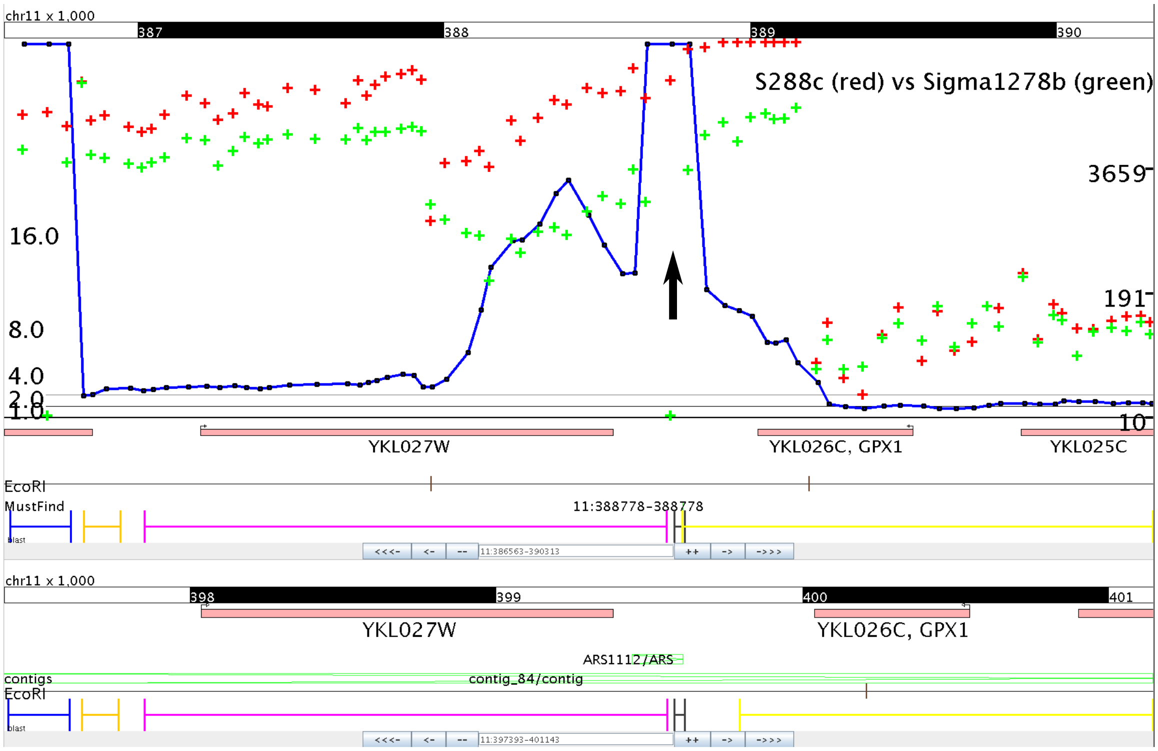
10.4.4 4:1023149-1023496

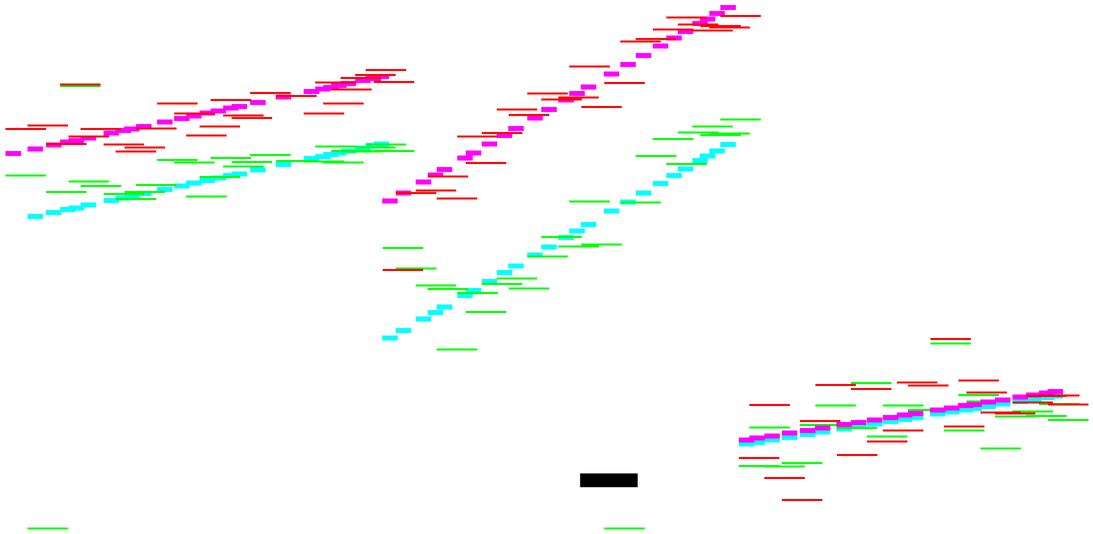
This Σ element is absent from Σ1278b. This Ruler Array misses this call since both channels exhibit similar intensity drops and the ratio changes very little.



10.4.5 11:388578-388978

The Ruler Array misses this 200bp insertion in Σ1278b because there are too few probes in this 1.2kb interval to force the linefitting to fit separate segments to the two sides around the indel.


chr11 x 1,000
387
388
389
390
S288c (red) vs Sigma1278b (green)
3659
16.0
8.0
4.0
2.0
1.0
191
10
YKL027W
YKL026C, GPX1
YKL025C
EcoRI
MustFind
11:388778-388778
blast
<<<- <- -- 11:386563-390313 ++ -> ->>>
chr11 x 1,000
398
399
400
401
YKL027W
YKL026C, GPX1
ARS1112/ARS
contigs
contig_84/contig
EcoRI
blast
<<<- <- -- 11:397393-401143 ++ -> ->>>



10.4.6 Lessons from the False Negatives

The confirmed indels that the Ruler Array misses demonstrate the importance of probe coverage and restriction site spacing. The Ruler Array misses some changes because there aren't enough probes nearby to observe the effects of the change; in the array design we used, this is typically the case when the change occurs near Σ, Δ, or τ elements or TY elements. Future array designs might tile repetitive areas more aggressively as observations from not-perfectly-unique probes may still be useful in the Ruler Array analysis. An array design for Ruler experiments might also tile the borders of repetitive regions more heavily that the rest of the genome such that the high density of observations at the edge of the repetitive sequence counteracts the nearby lack of probes.

Performing multiple replicates of the Ruler Array experiment using different restriction enzymes should also decrease the false negative rate. We expect such replicates to help in two ways. First, the second experiment will help distinguish between experimental noise and underlying biological effects (either indels or other sequence features that influence polymerase processivity). Second, the experiment with a different, carefully chosen restriction enzyme will provide good coverage in many of the places that are too close to or too far from the first enzyme's sites.

10.5 Synthetic Diploid Experiment

While our experiments compared two strains of haploid yeast, we can use the same data to simulate the performance of the Ruler Array on a hypothetical diploid organism. The first channel in the synthetic diploid experiment is the S288C channel from an existing Ruler Array experiment and the second channel is the average of the S288C and Σ1278b channels.

This experiment simulates a completely heterozygous diploid, the hardest case for the Ruler Array. A heterozygous change should show only half the intensity difference between channels of a homozygous change and a ratio of x would become $\frac{x+1}{2}$ (or the average of x and 1).

We used the same analysis parameters for this experiment as for the previously described results rather than trying to optimize the parameters to increase the analysis's sensitivity. The resulting set of calls was roughly 30% smaller and contains 220 correct calls, 18 false negatives (compared to six in the haploid experiment), and 138 false positives. These results indicate that the Ruler Array technique is applicable to diploids but, as expected, less effective than in haploids. We expect that, as with the other cases in which the Ruler Array fails to detect a change, more replicates with different enzymes should improve the detect rate.

10.6 Technical Evaluation of the Ruler Array

Using a PCR-confirmed set of insertions and deletions of varying size and position, we have confirmed that the Ruler Array can detect over 80% of the changes in our set of confirmed indels while producing fewer than 50% false positives. We expect that many if not all of the changes missed by the Ruler Array in this evaluation would be found in a second experiment using a different restriction enzyme. Furthermore, the total number of calls made is small enough compared to the number of genomic changes that PCR confirmation of each or a large subset of the calls can be done easily. Finally, our simulated experiment shows that the Ruler Array can work even in diploid genomes to find heterozygous changes.

Chapter 11

Biological Test Cases

Our biological test cases each evaluate the Ruler Array in the light of a particular type of genomic change between S288C and Σ1278b. Transposable element movement, changes in di- and tri- nucleotide repeat length, gene family copy number changes, and gross rearrangements represent major classes of insertions and deletions with known biological significance. For each type of change, we evaluate the Ruler Array results using the same data as in the previous chapter as if this type of change were the only change of interest. Since most of the differences between strains presented here have not been confirmed with PCR, we have used the genome assemblies, the CGH data, and whatever low-throughput data is available to determine the ground truth against which we evaluate the Ruler Array.

11.1 TY Elements

The TY elements in yeast include several families of long (roughly six kilobases) transposable elements found throughout the yeast genome[7, 32]. The S288C genome contains 50 TY elements; given their potential importance to gene regulation and their utility as genetic signposts, the presence of TY elements in another strain such as Σ1278b is of great interest.

We compiled the list of the 25 elements that appear to be present in Σ1278b but absent in S288C. TY elements present in S288C but absent from Σ1278b are not an interesting test case; the presence of a TY element at a known location can be easily and inexpensively tested with PCR. Detecting the locations of novel elements in Σ1278b is the interesting case for the Ruler Array.

Testing Ruler Array results to look for TY changes is fairly straightforward using PCR. To confirm TY elements present in the reference strain (the strain against which the microarray was designed; S288C in this case) and missing from the experimental strain (Σ1278b in this case) according to the Ruler Array, one can design primers around the TY element in the reference strain and compare the product size. The number of primer pairs is the number of Ruler Array calls at TY elements, perhaps one or two dozen. Testing for TY insertions in the experimental strain may require many more primer pairs; unless the Ruler Array calls are narrowed (e.g. by knowing candidate insertion sites for the transposable element), one might need to design one primer pair for every Ruler Array call. Our experience shows this to be perhaps two hundred primer pairs, still a manageable quantity.

Of the 25 elements present in Σ1278b but not in S288C, the single replicate of the Ruler Array detects twelve of the changes using the same parameters as used for the analysis in the previous chapter. Of the remaining thirteen changes, the Ruler Array missed eight because they were too close to the EcoRI site (less than $\sim$ 300bp or fewer than five probes). In several other cases, the Ruler Array failed to detect the TY elements presence because the TY contains EcoRI sites such that the distance between adjacent probes and the nearest EcoRI changes only slightly between strains.

As with the false negatives in chapter 10, we expect that performing Ruler Array experiments with several restriction enzymes and then combining the results should substantially decrease the false negative rate.

11.2 Repeat Length Changes

While examining an early set of apparent false positive calls from the Ruler Array, we picked a small set of calls over AT repeats or other repetitive sequences for sequencing. We hypothesized that small changes in the repeat length might lead to the large differences in intensity change and hence to the Ruler Array calls.

Figures 11-1 and 11-2 show two examples in which relatively small indels (14 and 2 bp, respectively) caused large intensity differences. Of the ten ruler array calls that we sequenced to determine whether repeats changed length, seven occurred over a repeat length change. Those seven included one single nucleotide change in a poly-A repeat, two two-nucleotide changes, and four changes of eight or more nucleotides. These repeat length changes show that the Ruler Array can detect indels far smaller than 100bp under the right circumstances.

Comparing the Σ1278b assembly to the S288C reference sequence indicates that 51 ruler array calls occur over a change in repeat length (generally AT or poly-A). Of these, 35 occur within 100bp of a stop site ($p < .00001$) as indicated by tiling microarray expression data for these two strains and 15 occur within 100bp of a convergent stop site ($p < .0005$).

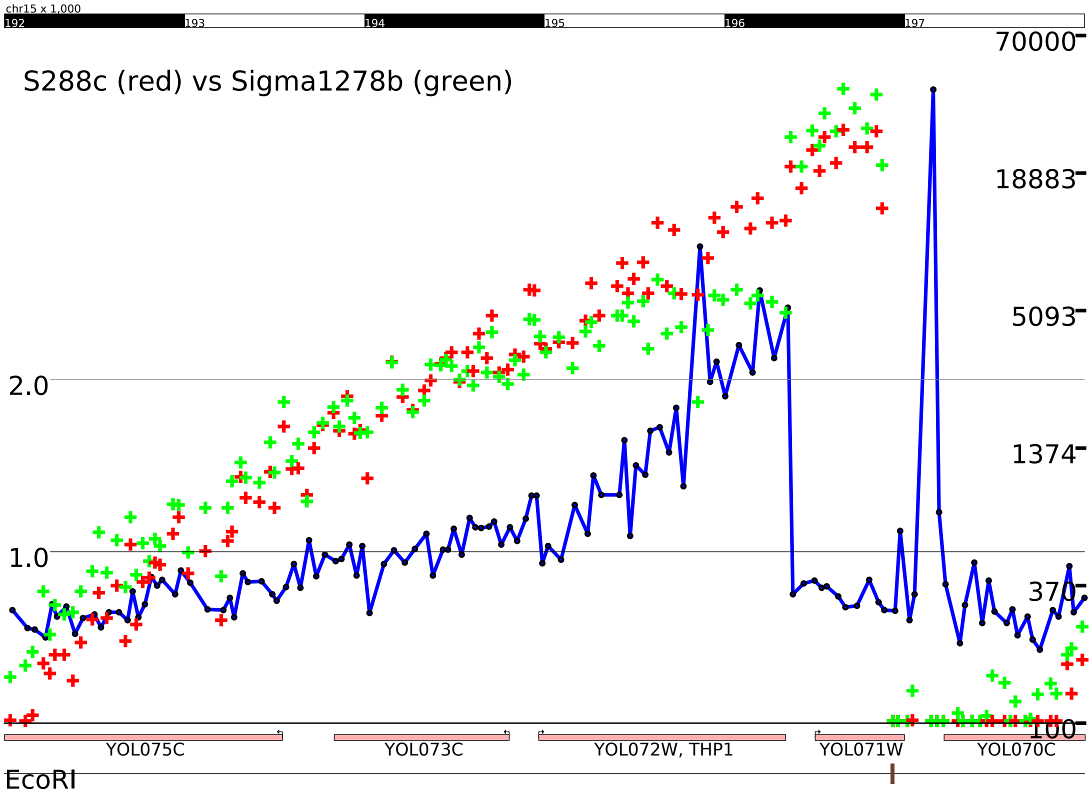
While we have not fully evaluated biological significance of these repeat length changes, their correlation with transcript stop sites is certainly intriguing. Furthermore, it seems that the Ruler Array can effectively assay small repeat changes that play biologically important roles in other settings.

11.3 Gene Family Expansions and Contractions

Changes in the size of gene families or in the copy number of extremely similar genes represents an important type of change in evolution and between strains. The sequencing and assembly of Σ1278b detected several putative changes in gene families that present an excellent test case for the Ruler Array. In particular, the *Cup*, *Pho*, and *Mal* families contain genes related to copper resistance, phosphorus metabolism, and maltose fermentation.

11.3.1 Cup

The *Cup1-1* and *Cup1-2* genes are separated by an uncharacterized ORF in S288C. In Σ1278b, this locus has expanded such that there are five copies of *Cup1* and four copies of the intervening ORF. A four kilobase region around these genes in S288C is untiled because



```
Sigma  TTTGGTGATATGTAGATATATATATATATATATATATATAGGAAATAGAAGAGAAGGAGCGA
S288C  TTTGGTGATATGTAGATATATATATA--------------GGAAATAGAAGAGAAGGAGCGA
```

Figure 11-1: The top panel shows the Ruler Array data (S288C in red, Σ1278b in green) over part of chromosome 15. The Σ1278b intensities fall suddenly over an AT repeat at the transcription stop of the THP1 gene whereas the S288C intensities continue a linear decline. The bottom panel shows the sequencing results for this locus; each strain was sequenced in both directions. The 14bp expansion of the AT repeat in Σ1278b seems the likely cause of the sudden intensity drop.

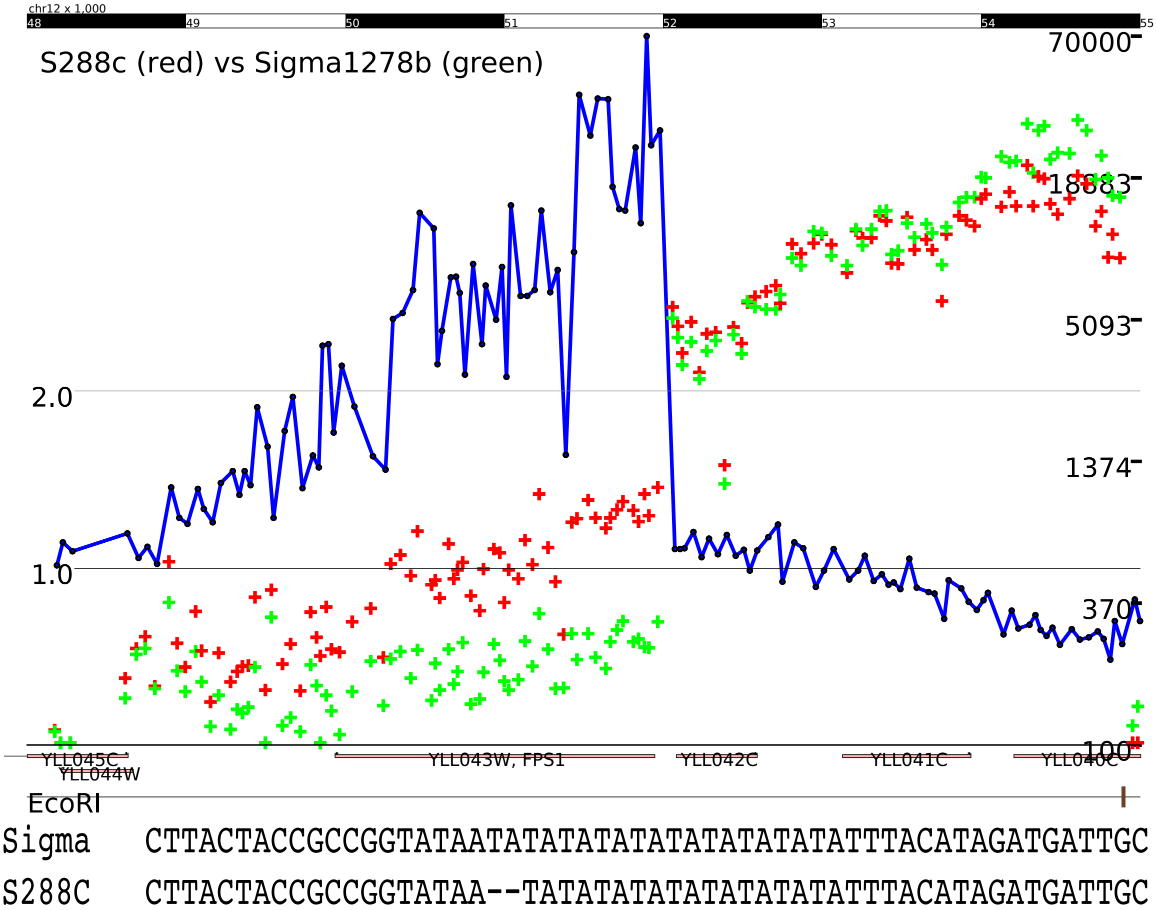


Figure 11-2: This example is similar to the previous AT repeat length change, though in this case the repeat expands by only 2 base pairs (one AT unit). Interestingly, the magnitude of the difference between the log-intensity drops across this repeat is greater than in the previous example.

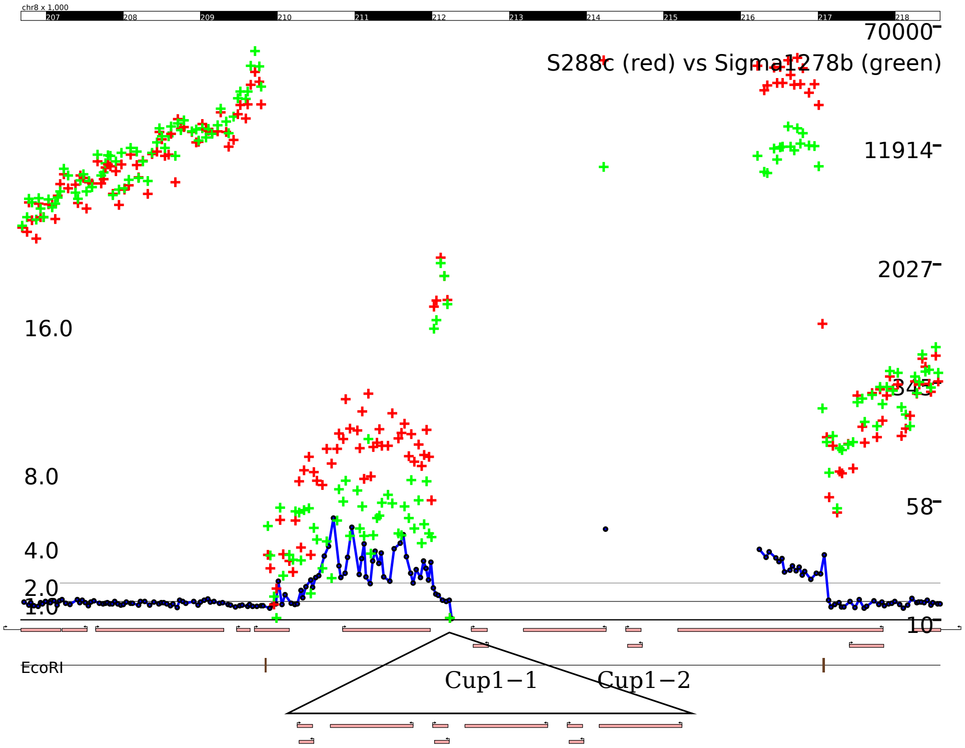


Figure 11-3: The *Cup1-1*, *Cup1-2* locus. While these genes and the region around them are not tiled, the Ruler Array shows evidence of a change by the high ratio observed to the left of *Cup1-1*.

of its repetitive nature.

While the Ruler Array fails to produce a call at this locus, the data does reflect some of the underlying change by a relatively high ratio downstream of the duplication as shown in figure 11-3. We believe that a more aggressively tiled microarray design and an analysis that incorporates non-unique probes would detect this insertion.

11.3.2 Pho

In S288C, *Pho3* and *Pho5* are separated by roughly 500bp on chromosome two. The *Pho* family is repetitive in general, causing misleading aCGH results. While the aCGH data indicate a duplication of both *Pho3* and *Pho5* in Σ1278b, *Pho3* is actually missing from this locus and *Pho5* is unchanged. The Ruler Array correctly detects the location of the *Pho3*

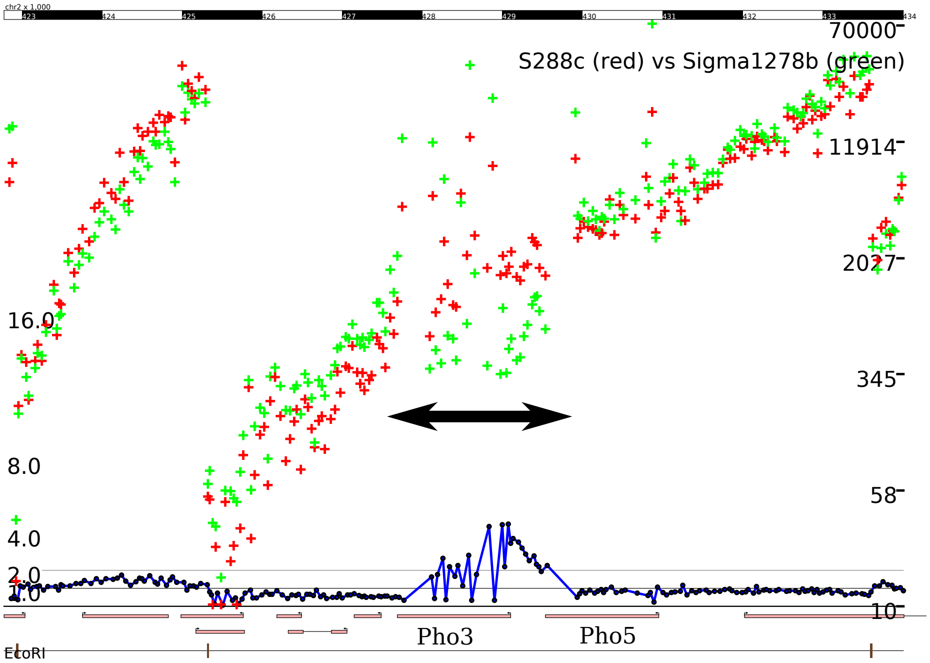


Figure 11-4: The *Pho3*, *Pho5* locus. The Ruler Array intensities over *Pho3* seem noisy and don't follow the expected falloff pattern. Since many genes in the *Pho* family exhibit high similarity, the intensities are not uniformly low in Σ1278b; however, enough probes detect the deletion to allow the analysis to identify the deletion of *Pho3* in Σ1278b.

deletion as shown in figure 11-4.

11.3.3 Mal

The *Mal* family of genes form a maltose metabolism pathway. In S288C, *Mal33*, *Mal31*, and *Mal32* are adjacent on chromosome two. In Σ1278b, there are three extra copies of *Mal* genes at the same locus between *Mal33* and *Mal31*. The Ruler Array correctly detects a change at *Mal33*, roughly the beginning of the inserted sequence as shown in figure 11-5.

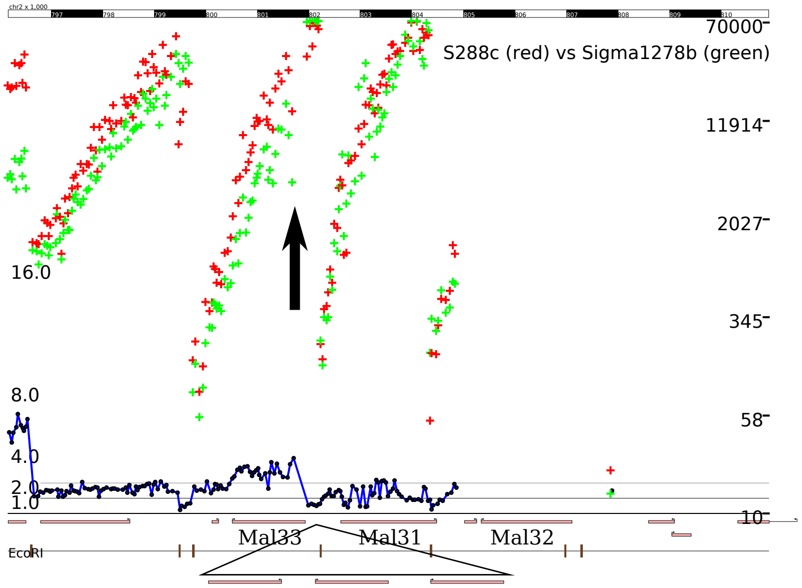


Figure 11-5: The *Mal33*, *Mal31*, *Mal33* locus. Three copies of *Mal* genes have been inserted in Σ1278b between *Mal33* and *Mal31*. The sudden change in ratio over *Mal31* reveals the change.

11.4 Gross Rearrangements

The Ruler Array detected one of the few gross sub-telomeric rearrangements that the array covered. Figure 11-6 shows Ruler Array data near the left arm of chromosome 6 (top) and the left arm of chromosome 10 (bottom) in S288C. The peak at 30kb is not a Σ1278b specific restriction site but rather the result of the rearrangement; in the Σ1278b genome, the sequence to the left of the peak is adjacent to a restriction site.

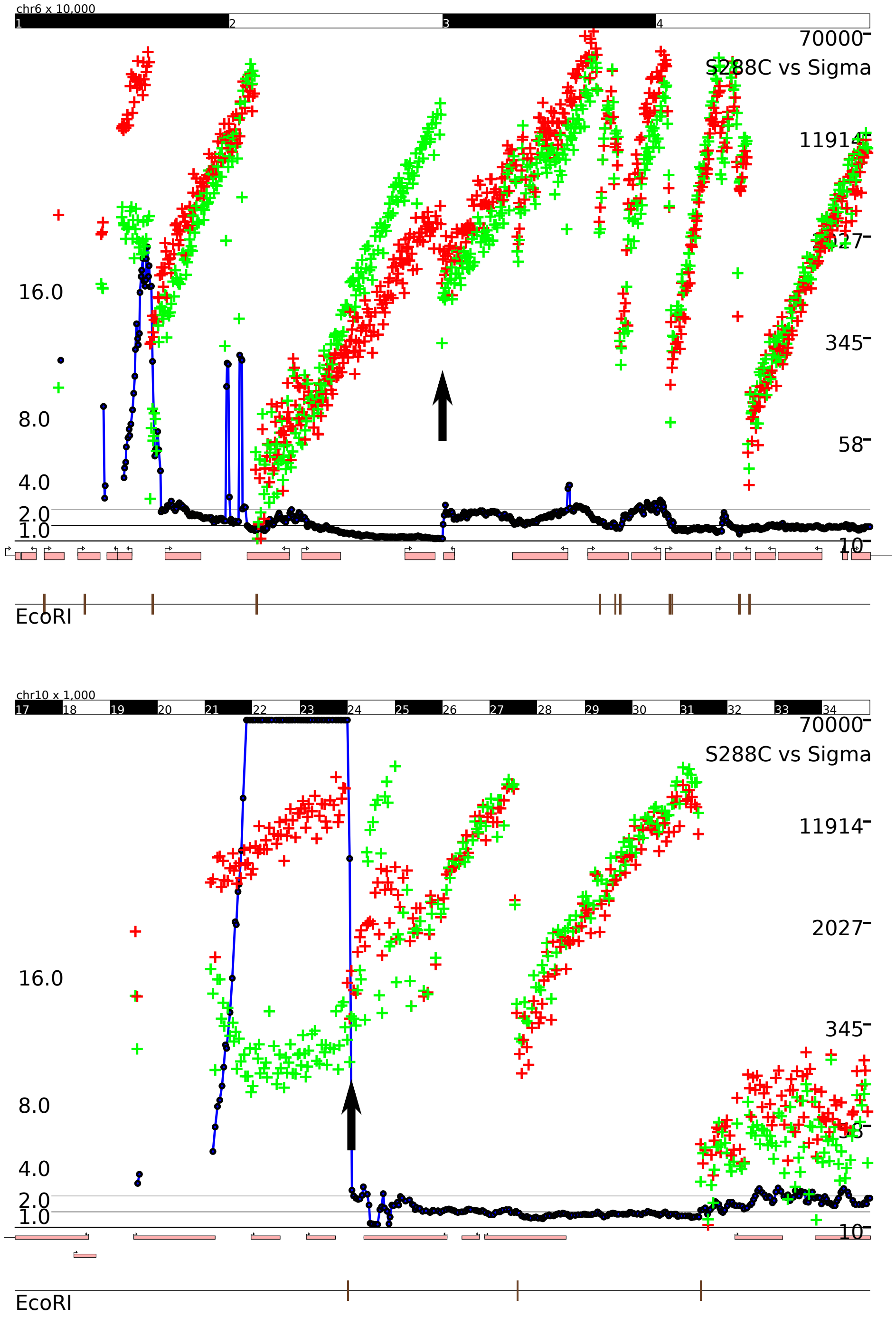


Figure 11-6: The left arm of chromosome six has moved to the left arm of chromosome ten between S288C and Σ1278b. While the Ruler Array can't determine what moved where, it does make evident the sites at which some change occurred. The upper panel shows chromosome six; the break point is at 30kb. The lower panel shows the break around 24kb on chromosome ten.

Chapter 12

Comparison to Other Indel Detection Techniques

We compared the Ruler Array to four other indel detection techniques on the S288C vs Σ1278b test case. Our TIP-Chip results are theoretical; the aCGH, short read assembly, and long read assembly results represent experiments that we performed on these strains. For each technique, we present the results of the technical evaluation (as done for the Ruler Array in chapter 10) and a brief look at the biological test cases in chapter 11.

12.1 Comparison to TIP-Chip

We have not performed a TIP-Chip experiment so instead assume that, as its inventors claim, it can detect essentially all TY1 elements with few false positives[53]. As such, TIP-Chip offers a simpler technique to discover TY insertions but offers no information about other changes (it would have missed all but two of the 35 "must find" examples). On the other hand, the Ruler Array may require several experiments to detect as many TY insertions as TIP-Chip but in the process it produces information about a wide variety of other genomic changes.

12.2 Comparison to aCGH

We used a single replicate of an aCGH experiment between FY4 and Σ1278b to compare aCGH's performance against that of the Ruler Array. The experimental protocol used the non-enzymatic ULS labeling system to avoid amplification or dye incorporation biases.

Our HMM analysis of the aCGH experiment produced 183 calls. Twelve appear incorrect given the two genome assemblies and 33 are confirmed by the assemblies. The remainder occur in repetitive regions (e.g. TY, sigma, tau, and delta elements) such that both the CGH data and the assembly are likely to be incorrect.

The aCGH experiment found 21 of the 35 "must-find" indels and missed the remaining 14. Thirteen of the 35 were originally added to our list of known indels because of the aCGH experiment, so their detection is not surprising. Figures 12-2 and 12-3 shows an examples of insertions that the aCGH experiment misses because there is no change in the unique probes surrounding the changes. Figure 12-4 shows the large region on chromosome 16 where both the Ruler Array and the aCGH experiment detect several large deletions. Finally, figure 12-1 shows the left arm of chromosome 6 where the aCGH experiment fails to detect a translocation because there is no change in copy number.

To more accurately compare the aCGH experiment to the Ruler Array experiment, we re-ran the analysis using only array probes with a unique genomic location; this excludes probes that map to TY or other repetitive elements. By only including unique probes, we now know the location of any change that the aCGH experiment detects. On this input, the same HMM analysis produced only 18 calls and found 6 of the 35 "must find" events.

12.3 Comparison to Sequencing Methods

Given the history of paired-end, long read sequencing as the gold standard for detecting changes between nucleic acid sequences, we wanted to compare the Ruler Array to both traditional long read sequencing and to a newer short read technique.

A key parameter of this comparison is the amount of data available for each technique-

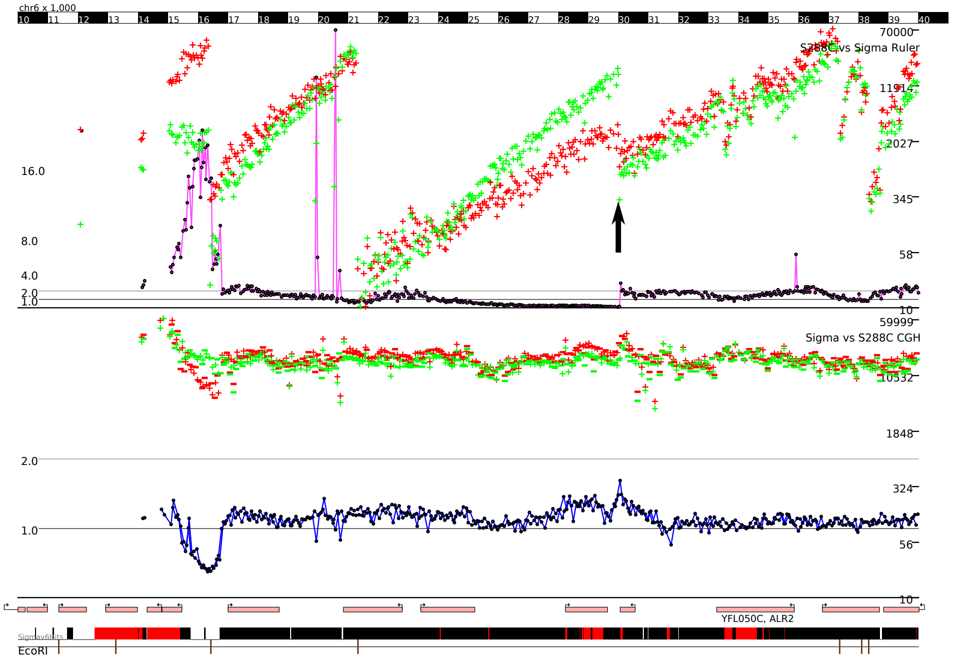


Figure 12-1: While the Ruler Array data (top track) over the left arm of chromosome 6 clearly shows the location of the translocation between chromosomes 6 and ten in Σ1278b (at 30kb, marked with a black arrow), the aCGH data in the bottom track shows no difference. In the aCGH plot, the FY4 intensities are green and the Σ1278b intensities are red; the ratio is shown in blue. Both methods clearly show a deletion in Σ1278b at the left edge of the plot.

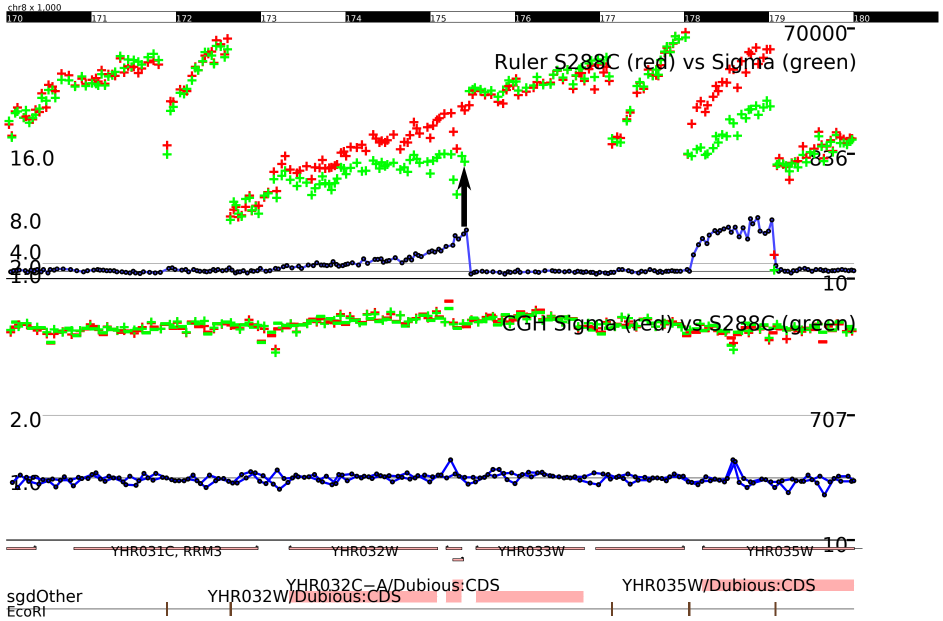


Figure 12-2: The Ruler Array (data in top track) successfully detects the insertion of roughly 100bp on chromosome eight while the unique probes in the aCGH data show no difference.

the number of replicates of the Ruler Array, the number of paired-end long reads, and the number of lanes of and number of short reads. We performed our comparison using

- the same single replicate of the Ruler Array as was used previously
- the S288C reference assembly from the Stanford Genome Database, retrieved in October 2006.
- 114,528 long reads (average length 909bp) from Σ1278b assembled with the Arachne assembler and manually curated to include information from chromoblots and other experiments. The reads assembled into 358 contigs, 49 scaffolds, and eventually 16 chromosomes.
- three lanes of Solexa 25bp reads from Σ1278b for a total of 20.4 million 36bp reads

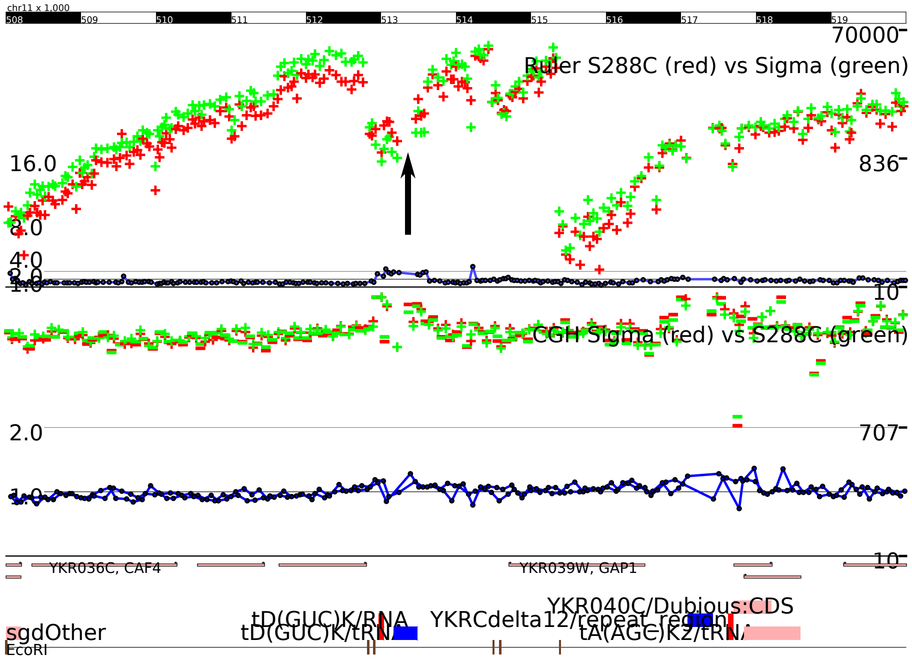


Figure 12-3: The Ruler Array (data in top track) successfully detects the insertion of a TY element on chromosome eleven while the unique probes in the aCGH data show no difference. While the CGH data does show a difference in ratio over repetitive elements such as the TY family, it cannot localize the changes to particular insertion sites such as this one.

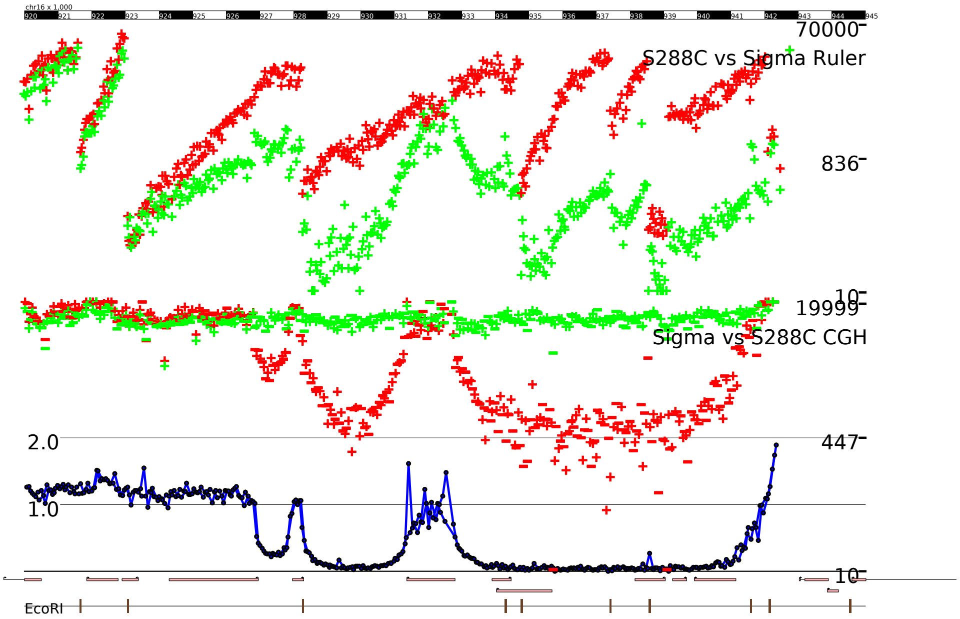


Figure 12-4: Both the Ruler Array and aCGH correctly detect the deletion of parts of the right arm of chromosome sixteen in Σ1278b (note that the channels are reversed between the two experiments). The low intensities and low ratio make the deleted regions obvious in both experiments.

12.3.1 Long Read Assembly

The paired read assembly started with 114,528 reads providing 7.3X coverage. The paired reads allowed the 358 contigs to be grouped into 49 scaffolds by the Arachne assembler[3]. Each inter-contig gap represents a region of unknown length and unknown sequence. The scaffolds were manually curated using chromoblot information, synteny with S288C, and other techniques to produce 16 chromosomes; in some cases, scaffolds were broken and rejoined.

An alignment of the original 49 scaffolds to S288C using the Fast Statistical Alignment program (FSA) produced a list of 2132 indels of more than 50bp and 1685 of more than 100bp[5]. The 50bp list would have missed between six of the 35 "must find" indels and the 100bp list would have missed ten. These false negatives typically occur because of assembly errors (three Arachne mis-assembled three scaffolds that needed to be broken by hand) or alignment difficulties around repetitive sequences.

Aligning the curated assembly tends to predict indels at all contig boundaries because the assembly uses a default size of 100bp of N's to fill gaps between contigs. In general, the actual amount of missing sequence will not be 100bp, resulting in an indel when the Σ1278b sequence is aligned to S288C. We selected 106 indels predicted by alignment of the curated Σ1278b genome against S288C to be more than 50bp for PCR validation. Of the 106 indels tested, only 35 (33%) produced a change in PCR product size that was visible on an agarose gel (i.e. a change of roughly 10bp or more). The subset of these 35 changes that were more than 100bp form the core of the "must find" set of indels used in chapter 10 (those determined to be less than 100bp were included in the "can find" set rather than the "must find" set).

The results from examining both the initial assembly and the curated assembly demonstrate that indel detection from sequencing is not a push-button operation. Assemblers make mistakes, even with large amounts of high quality input. Furthermore, alignment results can be confusing or wrong around repetitive elements, especially when the divergence between elements within a genome is similar to the divergence between genomes or even to the error rate of the sequencing process. Given these difficulties, the Ruler Array offers an independent method to asses the physical distances between genomic locations and can be used effectively in concert with a draft genome assembly to produce a finished, curated assembly.

12.3.2 Short Read Assembly

The short read assembly combined 20.4 million 36bp reads from a Solexa machine (265 million base pairs of sequence for an expected 22X coverage) into an 11.3Mb assembly. We used Velvet[56] to assemble the reads and filtered the output to include only contigs with 5X coverage and a minimum length of 100bp, leaving 5419 contigs.

We mapped the contigs to the S288C reference genome with Blat[25] and then produced detailed alignments with FSA[5] to produce a list of 24680 indels of more than 50bp and 21424 indels of more than 100bp. Either indel list represents an unreasonable number of candidate changes to confirm with PCR and a false-positive rate of roughly 90% if the set of changes predicted by the long read assembly is complete.

Since FSA produces an unreasonably large number of indel calls, we used the indels as predicted by BLAT for the remainder of this comparison. We included indels predicted by the best alignment of the short read contig to the S288C reference sequence. Fifty-six of the 75 indels called by the short read assembly agree with the long read assembly for a true positive rate of 75%. Fourteen calls were clearly wrong and five could not be evaluated because of problems with the long read assembly or because highly repetitive sequence made it too hard to determine whether the short read assembly was accurate.

To determine whether the short read assembly can detect changes in TY presence or location, we determined whether any contigs spanned the location of a TY insertion (a TY present in S288C but not in Σ1278b) or included both unique sequence and a TY that was present in Σ1278b but not in S288C (the TY must include the TY itself and not just the surrounding LTRs). Using this method, the short read assembly detected three of the 30 TYs present in S288C but not in Σ1278b and one of the 25 present in Σ1278b but not in S288C.

Coordinates	Description
1:198200-203000	4kb deletion in sigma
2:644926-644926	TY insertion in sigma
2:428094-429944	1kb deletion in sigma
2:801000-805000	MAL32 duplication. 3kb or so added in sigma
4:523000-527000	(several small indels) 500bp gone in sigma
4:957500-958000	100bp deletion in sigma
5:207100-207400	100bp insertion in sigma
11:513003-513603	TY insertion in sigma
14:429700-430000	100bp insertion in sigma
16:928300-931300	3kb deletion in sigma
16:932800-941700	9kb deletion in sigma
14:777000-779000	2kb deletion in sigma
10:21000-24500	3kb deletion in sigma
9:434645-436741	2.5kb deletion in sigma
4:462154-462154	1kb insertion in sigma
8:175482-175482	100bp insertion in sigma
9:349999-349999	100bp deletion in sigma
14:34470-34470	200bp deletion in sigma
15:30388-30388	400bp insertion in sigma

Table 12.1: Indels not found by comparing the short read assembly to the S288C reference sequence.

Only 16 of the 35 indels used to evaluate the Ruler Array were found when we used the same assembly and method to determine a list of indels of more than 50bp. Table 12.1 lists the confirmed indels not found by the short read assembly.

The short read assembly suffers from two key weaknesses compared to the long read assembly. First, the short reads cannot span most repetitive elements (e.g. TY, Σ, or Δ elements) so this assembly relies on SNPs to differentiate instances of these elements from each other. Second, the lack of paired reads makes assembly across larger distances extremely difficult since any repetitive sequence may prevent the assembler from joining two small contigs. While continued technology improvement will increase read length, read count, and offer paired reads, assemblies and alignments will continue to suffer from inherent difficulties at repetitive elements and other non-unique sequences. The Ruler Array will continue to offer a second opinion at these loci.

12.4 Evaluation Summary

The technical evaluation, biological use cases, and comparisons to other indel detection technologies show that the Ruler Array offers and effective new technique to screen for genomic changes. While aCGH excels at detecting copy number variation in some circumstances, it can easily miss changes involving novel sequence or changes in repetitive sequences. TIP-Chip accurately assays for transposon positions but is blind to all other types of changes. Finally, sequencing techniques promise to detect all changes but encounter difficulties at repetitive sequences; furthermore, producing a high quality genome assembly that accurately predicts indels requires substantial manual effort, even for the relatively small yeast genome.

The Ruler Array expands the toolkit for screening for insertions and deletions. We have shown that a single experiment finds over 80% of the changes in our test set as well as nearly 200 other changes of varying size. The Ruler Array works on all types of changes, unlike aCGH and TIP-Chip, and does not depend on the sequence being measured, allowing it to work across the repetitive sequences that can trip up assemblies.

Chapter 13

Extensions of the Ruler Array Technique

We have developed several extensions to the Ruler Array technique to take advantage of high-throughput short-read sequencing machines.

13.1 Ruler Seq

Ruler Seq generates material in the same way as does the Ruler Array but end-sequences the fragments rather than hybridizing them to an array. Figure 13-1 shows the steps in the protocol.

The Ruler Seq experimental results can be analyzed by generating synthetic array intensities or through assembly. In the synthetic intensities method, each read is extended back to the restriction site; probe intensities count the number reads extended underneath some point, as shown in figure 13-2.

We have run the linefitting analysis on the virtual array intensities produced by two lanes of Solexa short read sequencing (read length 36bp, one lane per strain). This method produced 23 false negatives, over twice as many as the array method but similar to the synthetic diploid. Unlike the synthetic diploid, however, the Ruler Seq results produced

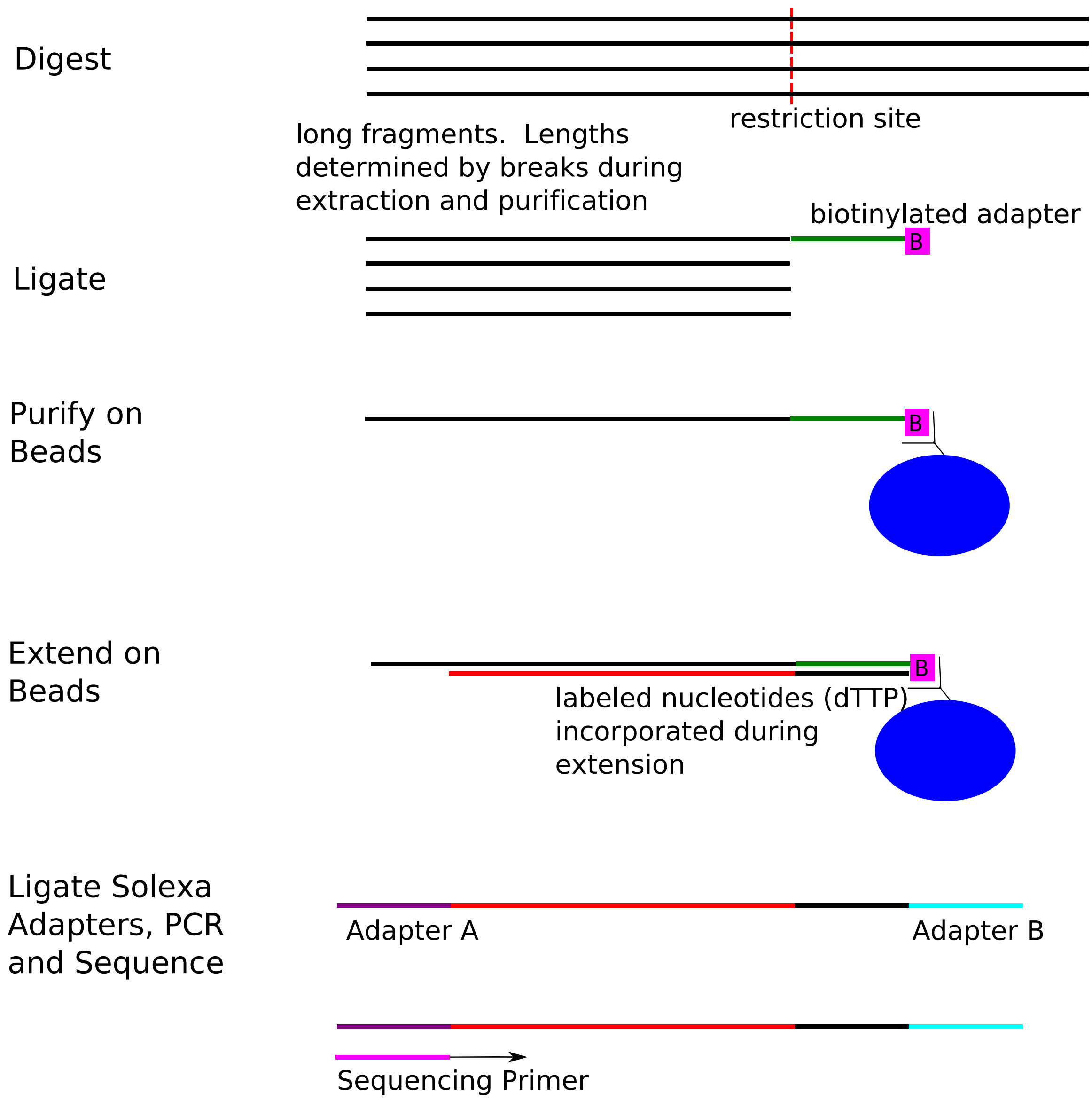

Digest
restriction site
long fragments. Lengths determined by breaks during extraction and purification
biotinylated adapter
B
Ligate
Purify on Beads
B
Extend on Beads
B
labeled nucleotides (dTTP) incorporated during extension
Ligate Solexa Adapters, PCR and Sequence
Adapter A
Adapter B
Sequencing Primer

Figure 13-1: The Ruler Seq protocol generates fragments in the same way as the Ruler Array Protocol. However, instead of labeling the fragments with fluorescent dyes, the Ruler Seq protocol ligates adapters to both ends of the fragment and then sequences from the 3' end. Mapping the read sequences to the genome produces the location and strand of the read.

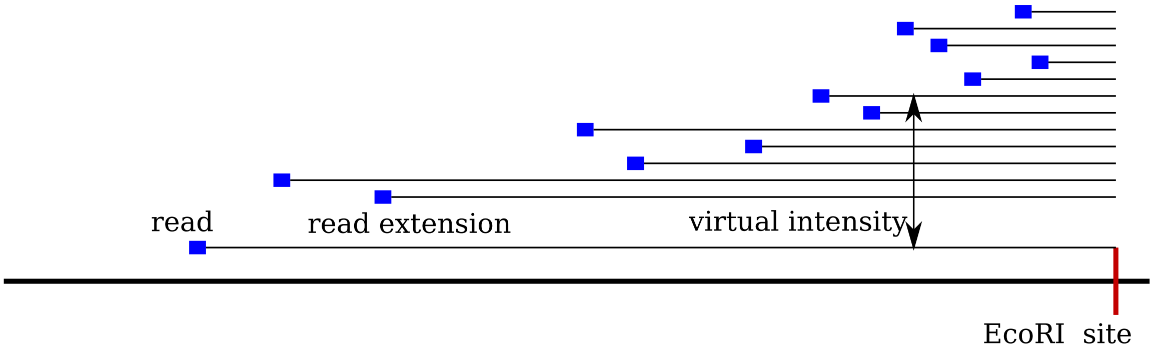


Figure 13-2: Mapping the Ruler Seq reads to the genome produces stranded locations. After each read is extended back to the restriction site from which its fragment came, the number of reads crossing any point is the virtual intensity at that point. While one could generate an intensity at every base pair, those intensities would be repetitive. Instead, we generate a virtual intensity measurement at every position to which one or more reads align.

nearly 900 false positives.

A more sophisticated analysis of Ruler Seq data combines the virtual array intensities with an assembly of the read sequences. Short read sequences provide excellent coverage of small indels and SNPs that the Ruler Array might miss. The read assembly should also confirm many larger indels either by providing the full sequence of the change (if the change is near a restriction site) or hopefully including one or more reads on the edge of the change (if the read is farther from the restriction site) that would confirm the change's presence.

While the Ruler Seq technique offers several advantages over the Ruler Array protocol because of the sequence information returned, we do not expect it to perform well in large genomes. To produce meaningful virtual intensities or to enable a useful assembly, the sequencing runs should produce thousands of reads per interval. In a yeast-sized genome with about 8000 intervals, a single Solexa lane can satisfy this requirement. However, dozens of lanes would be necessary to adequately cover a mammalian genome.

13.2 Targeted Assembly

A potential application of the Ruler Seq protocol that we have not explored allows genomic sequencing of the subset of a genome near the restriction sites. Instead of Whole Genome Shotgun sequencing, one might run one or more lanes of Ruler Seq with one restriction enzyme and assemble the resulting reads. The expected distribution of reads from the Ruler protocol ensures excellent coverage around the restriction sites. One would then use a second, complementary enzyme to generate a second set of contigs. By using two (or more enzymes), one can target the coverage of each sequencing run to the parts of the genome that need it most.

13.3 TIP-Seq

Previous work by Wheelan[53] and Gabriel[17] developed techniques to profile the insertion sites of transposable elements using a microarray readout. These techniques, known as TIP-Chip, employ a primer designed against the transposable element of interest and some biochemical trickery to PCR amplify and label the region around the primer's annealing site (and therefore around the transposable element). This illuminates the corresponding probes on the microarray to make the locations of the transposable element clear.

As is the case with other microarray techniques, TIP-Chip may not scale well to larger genomes due to the large number of arrays required to tile a full genome. One might address this limitation by using short-read sequencing (e.g. Solexa) for the readout. The TIP-Seq protocol uses a biotinylated primer designed against the element of interest. After performing linear extensions against genomic DNA, the template is purified away by extracting the product with streptavidin beads. The material on the beads can then be amplified and sequenced. The analysis would detect a transposon insertion by the presence of several reads mapping to nearby genomic locations.

Chapter 14

Future Work

Our work so far on the Ruler Array demonstrates its ability to detect insertions and deletions between two strains of yeast and suggests several areas for future technical development and uses for the technique.

14.1 Polymerase Characterization

Our Ruler Array experiments have determined that two polymerases, ExTaq and Vent Exo- produce high quality data. However, we suspect that other enzymes will work well and that some might work better in some circumstances by, for example, producing longer fragments in cases where one wants to use a smaller set of restriction sites. One might also choose a polymerase that exhibits less sensitivity to sequence features such as AT repeats when one is only interested in larger indels.

Further work might characterize a polymerase's probability of terminating as a function of total bases incorporated as well as the current nucleotide, dinucleotide, or other sequence feature. Such a model would permit intensity normalization based on the reference sequence and lead to more accurate line fitting.

14.2 Labeling Sites

Instead of relying on restriction sites to define the initiation points for the labeling reactions, one might want to choose arbitrary sequences to achieve better coverage. One approach would use short sequences (e.g. a fixed hexamer or octamer) to produce a reasonable number of sites; another approach would use a large number of long primers, perhaps oligos sheared off a microarray, to achieve near-optimal coverage. Either case requires re-optimizing the conditions for the extension reaction. Short primers and a low temperature extension might require a different polymerase and a different noise model to account for the higher probability of spurious initiation.

14.3 Screening Closely Related Strains for Indels

Several recent studies have grown yeast under a particular stress condition for hundreds of generations, tested the resulting strains for enhanced growth under that condition, and then screened for genotypic changes to explain the fitness phenotype[19, 51, 41, 47]. As the Ruler Array can detect transposable element changes, gene family copy number changes, certain repeat length changes, and other length polymorphisms, one could easily imagine screening the evolved strains from these experiments with the Ruler Array, perhaps in addition to the aCGH, SNP arrays, and low coverage sequencing that previous studies employed.

14.4 Checking Genome Assemblies

While we have offered a comparison of Ruler Arrays to two sequencing techniques, we see the technologies as complementary. We expect that a key use for Ruler Arrays will be to support and proofread assemblies of novel genomes. We expect that the sequencing and assembly of the Σ1278b genome will be a common case; sequencing closely related species allows studies to link phenotype and genotype since the number of genomic changes is relatively small.

14.5 Biological Significance of Repeat Length Changes

Previous studies have looked at repeat length changes as drivers of pathogen evolution to avoid immune response and other evolutionary functions by altering the coding sequence of cell surface genes[33]. The A, AT, and ATT repeat length changes discussed in chapter 11 tend to occur in intergenic regions and generally at transcription stop sites. Given that the different repeat lengths caused different effects on the polymerase in our Ruler Array experiments, we wonder whether the repeat length changes also indicate some difference in the transcriptional boundaries, levels, or regulation between S288C and Σ1278b.

14.6 Ruler Arrays Expand Toolset for Discovering Genomic Differences

Ruler Arrays expand the researcher's ability to detect insertions and deletions at their genomic locus and will lead to a greater understanding of the relationship between genotype and phenotype. By itself, the Ruler Array certainly does not solve all problems in the search for genomic changes. Rather, the Ruler Array expands the arsenal of high throughput techniques to detect changes. Since past studies have been limited to evaluating the changes they could discover- aCGH (detection of copy number variation but not the locus of insertions), SNPs technologies (SNPs can be discovered by low coverage sequencing and assayed by microarray, and TIP-Chip (transposon changes)- most understanding of genomic changes centers on SNPs and copy number variation.

Appendix A

Ruler Array Laboratory Protocol for *S. cerevisiae*

A.1 Growing Cells

Grow one liter of yeast to an OD600 of roughly 1.0. This will give roughly 3×10^{10} yeast cells. If you are worried that OD600 is not an accurate measurement of cell density you can count colonies to determine actual number of yeast cells. Pellet yeast cells and store at -80 until you are ready to isolate the DNA.

A.2 DNA Extraction

Once you are ready to isolate DNA thaw cell pellets at RT. Resuspend 3×10^{10} cells in 12ml TE. Pellet for 5 minutes at 3000RPM 4 degrees. Remove supernatant. The rest of the DNA prep is a variation of Qiagen 250 DNA prep kit. I have found that the amount of time necessary to fracture the cell walls of *S. cerevisiae* can vary quite a bit. I have done everything from just letting incubate with lyticase for 30 minutes at 30 degrees, For other strains I have let sit for 2hours and include some time shaking to achieve more lysis. The next step after this in the Qiagen protocol can also be modified by letting it go for longer than the minimum 30 minutes that they recommend. These are both protocol changes that are discussed in the Qiagen book. After these steps proceed with the Qiagen instructions for column purification, although the columns can run very slow this is fine. I have left the columns to run over a number of days. I have found that when I try to force out the DNA using air pressure I tend to get less DNA recovered.

A.3 Digestion

Once DNA has been purified from the Qiagen column I digest the DNA with an endonuclease. For purposes of trouble shooting I have been using EcoRI. Protocol for this is as follows:

1. Bring each 20ug sample of DNA to a volume of 230uL, add 23ul Buffer 3, and 10 ul EcoRI.
2. Put at 37-36 C for 2-3 hours.
3. Add 5ul Calf Intestinal Phosphatase leave for 1 hour, CIP works well in buffer 3.
4. At this point the DNA can be frozen over night or you can begin the next steps.

Phenol Extraction and Ethanol Precipitation of DNA:

1. Add 1 vol Phenol, recover aqueous top layer
2. Add 1 vol Phenol Chloroform, recover aqueous top layer
3. Ethanol Precipitate by adding 1/10 volume NaoAC (roughly 26ul) and 2.5 volumes Ethanol (roughly 700ul)
4. Let sit for 30 minutes at -80 or ON at -20
5. Centrifuge max speed for 30 min
6. Pipette off supernatant

7. Wash with 800uL -20C 70% ethanol
8. Let pellet dry
9. Resuspend in 135ul H2O (should be enough DNA that Glycogen is not needed as a carrier)

A.4 Ligation of Biotin Linker to Digested DNA

A.4.1 Making Biotin Linker

Tris pH 7.9 (1M) 250uL

Oligo1 (40 uM stock) 375uL

Oligo2 (40 uM stock) 375uL

Heat at 95C for 5 min then at 70C and let the heat block cool to room temp gradually, I just turn off the heat block and let it sit 4 C overnight.

Aliquot into 40uL aliquots so I do not need to thaw and freeze the stock repeatedly.

A.4.2 Ligation

- 135uL digested DNA
- 20uL T4 DNA Ligase Buffer (make sure the ATP is dissolved)
- 40uL Biotin Linker from above
- 5uL T4 DNA ligase

Let sit overnight at 14 C

A.4.3 Cleanup and Binding to Beads

- Aliquot 30uL Magnetic bead slurry to eppendorf tube and wash 2x w/ binding buffer
- Resuspend the beads in 200uL binding solution
- Add this 200uL to the ligation reaction
- Shake at RT for 3.5 hours
- Wash beads 2x w/ wash buffer and 2x w/ H20
- Resuspend Beads in 75 uL H2O

A.5 Polymerase Extensions

There are four variations: the basic reaction uses Cy3 and Cy5 with ExTaq. You can also do aminoallyl-dUTP or ULS labeling or use Vent Exo- as the polymerase.

A.5.1 Cy3/Cy5 and ExTaq

- 10uL ExTaq 10x Buffer
- 6uL dNTPs 2.5 mM each
- 2uL Ex Taq
- 4uL Primer (40uM)
- 75uL H20 and beads
- 3uL Cy 3 or Cy 5

A.5.2 ULS

Don't use the labeled nucleotides. Instead, use 10uL (2.5mM) of each dNTP.

A.5.3 Aminoallyl-dUTP

- 10uL ExTaq 10x Buffer
- 10uL dNTPs 2.5mM each
- 2uL ExTaq
- 4uL Primer (40uM)
- 15uL Amine modified dUTP
- 60uL H2O and Beads

A.5.4 Vent Exo-

Swap ExTaq buffer w/ Thermo Pol buffer and Vent exo- for ExTaq.

A.5.5 Extensions

Polymerase Program:

1. 1: 94 C for 2 min
2. 2: 94 C for 1 min
3. 3: 62 C for 30 sec (this will change depending upon the linker you are using)
4. 4: 72 C for 2 min
5. 5: Go to 2 39 times
6. 6: 74 C for 5:00
7. 7: 4 C overnight

A.5.6 Isolating DNA

From any of the polymerase reactions:

- Place eppendorf tube on magnetic rack
- Pipette off supernatant leaving beads
- Purify supernatant using Qiagen spin column
- If the length of the fragments is fine then you can elute in 40ul component C

A.5.7 ULS Labeling

Purify DNA with qiagen column eluting in component C. Heat DNA to 95 C for 5 minutes then put on ice for 5 minutes to ensure ssDNA. Incubate DNA with Dye for 20 minutes at 80 degrees C. Purify excess dye and labeling reagents away using qiagen column purification eluting in H20.

A.5.8 Aminoallyl Labeling

Ethanol precipitate DNA and resuspend in 5ul H20. Add 2uL labeling buffer (25 mg/ml Na Bicarbonate) Add amine modified DNA to the reactive dye and leave at room temperature 1 hour. Purify excess dye and labeling reagents away using qiagen column purification eluting in H20.

A.6 Hybridization

Use equamolar amounts of dye in each channel. For Agilent 244k arrays, we typically use 80-100pmol per channel. Incubate array spinning at 65C for 40 hours and wash.

Bibliography

[1] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *J Mol Biol*, 215:403–10, December 1990.

[2] Swaroop Aradhya, Melanie A Manning, Alessandra Splendore, and Athena M Cherry. Whole-genome array-cgh identifies novel contiguous gene deletions and duplications associated with developmental delay, mental retardation, and dysmorphic features. *Am J Med Genet A*, 143:1431–41, June 2007.

[3] Serafim Batzoglou, David B Jaffe, Ken Stanley, Jonathan Butler, Sante Gnerre, Evan Mauceli, Bonnie Berger, Jill P Mesirov, and Eric S Lander. Arachne: a whole-genome shotgun assembler. *Genome Res*, 12:177–89, January 2002.

[4] David R Bentley. Whole-genome re-sequencing. *Curr Opin Genet Dev*, 16:545–52, November 2006.

[5] Robert K. Bradley, Adam Roberts, Michael Smoot, Sudeep Juvekar, Jaeyoung Do, Colin Dewey, Ian Holmes, and Lior Pachter. Fast statistical alignment. 2008.

[6] Mikael Brandstrom and Hans Ellegren. The genomic landscape of short insertion and deletion polymorphisms in the chicken (gallus gallus) genome: a high frequency of deletions in tandem duplicates. *Genetics JT - Genetics*, 176:1691–701, July 2007.

[7] J R Cameron, E Y Loh, and R W Davis. Evidence for transposition of dispersed repetitive dna families in yeast. *Cell*, 16:739–51, September 1979.

[8] J H Cha and L S 4th Dure. Trinucleotide repeats in neurologic diseases: an hypothesis concerning the pathogenesis of huntington's disease, kennedy's disease, and spinocerebellar ataxia type i. *Life Sci JT - Life sciences*, 54:1459–64, June 1994.

[9] X Chen, J A Knauf, R Gonsky, M Wang, E H Lai, S Chissoe, J A Fagin, and J R Korenberg. From amplification to gene in thyroid cancer: a high-resolution mapped bacterial-artificial-chromosome resource for cancer chromosome aberrations guides gene discovery after comparative genome hybridization. *Am J Hum Genet JT - American journal of human genetics*, 63:625–37, December 1998.

[10] S W Cheung, P V Tishler, L Atkins, S K Sengupta, E J Modest, and B G Forget. Gene mapping by fluorescent in situ hybridization. *Cell Biol Int Rep*, 1:255–62, May 1978.

[11] Haiping Dai, Yongquan Xue, Jinlan Pan, Yafang Wu, Yong Wang, Juan Shen, and Jun Zhang. Two novel translocations disrupt the runx1 gene in acute myeloid leukemia. *Cancer Genet Cytogenet*, 177:120–4, September 2007.

[12] Lior David, Wolfgang Huber, Marina Granovskaia, Joern Toedling, Curtis J Palm, Lee Bofkin, Ted Jones, Ronald W Davis, and Lars M Steinmetz. A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci U S A*, 103:5320–5, April 2006.

[13] J. C. Engert, M. Lemire, J. Faith, D. Brisson, T. M. Fujiwara, N. M. Roslin, C. G. Brewer, A. Montpetit, C. Zwaig, Y. Renaud, C. Dore, S. D. Bailey, A. Verner, G. Tremblay, J. Pierre, C. Betard, J. Platko, J. D. Rioux, K. Morgan, T. J. Hudson, and D. Gaudet. Identification of a chromosome 8p locus for early-onset coronary heart disease in a french canadian population. *Eur J Hum Genet*, September 2007.

[14] Chandra Erdman and John W Emerson. A fast bayesian change point analysis for the segmentation of microarray data. *Bioinformatics*, 24:2143–8, September 2008.

[15] Levy et al. The diploid genome sequence of an individual human. *PLoS Biol*, 5:e254, September 2007.

[16] C A Feener, F M Boyce, and L M Kunkel. Rapid detection of ca polymorphisms in cloned dna: application to the 5' region of the dystrophin gene. *Am J Hum Genet JT - American journal of human genetics*, 48:621–7, April 1991.

[17] Abram Gabriel, Johannes Dapprich, Mark Kunkel, David Gresham, Stephen C Pratt, and Maitreya J Dunham. Global mapping of transposon location. *PLoS Genet*, 2:e212, August 2007.

[18] A R Gallant and Wayne A Fuller. Fitting segmented polynomial regression models whose join points have to be estimated. *Journal of the American Statistical Association*, 68:144–147, March 1973.

[19] David Gresham, Michael M Desai, Cheryl M Tucker, Harry T Jenq, Dave A Pai, Alexandra Ward, Christopher G DeSevo, and David Botstein a nd Maitreya J Dunham. The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS Genet*, 4:e1000303, December 2008.

[20] J G Hacia, J B Fan, O Ryder, L Jin, K Edgemon, G Ghandour, R A Mayer, B Sun, L Hsie, C M Robbins, L C Brody, D Wang, E S Lander, R Lipshutz, S P Fodor, and F S Collins. Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat Genet JT - Nature genetics*, 22:164–7, June 1999.

[21] Kevin C Halling and Benjamin R Kipp. Fluorescence in situ hybridization in diagnostic cytology. *Hum Pathol*, 38:1137–44, July 2007.

[22] J Hasler, T Samuelsson, and K Strub. Useful 'junk': Alu rnas in the human transcriptome. *Cell Mol Life Sci*, 64:1793–800, July 2007.

[23] Douglas M Hawkins. Point estimation of the parameters of piecewise regression models. *Applied Statistics*, 25:51–57, 1976.

[24] A Kallioniemi, O P Kallioniemi, D Sudar, D Rutovitz, J W Gray, F Waldman, and D Pinkel. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, 258:818–21, December 1992.

[25] W James Kent. Blat–the blast-like alignment tool. *Genome Res*, 12:656–64, April 2002.

[26] Niels G F Klito, Qihua Tan, Mette Nyegaard, Klaus Brusgaard, Mads Thomassen, Charlotte Skouboe, Jesper Dahlgaard, and Torben A Kruse. Arrayed primer extension in the "array of arrays" format: a rational approach for microarray-based snp genotyping. *Genet Test JT - Genetic testing*, 11:160–6, July 2007.

[27] Jan O Korbel, Alexander Eckehart Urban, Jason P Affourtit, Brian Godwin, Fabian Grubert, Jan Fredrik Simons, Philip M Kim, Dean Palejev, Nicholas J Carriero, Lei Du, Bruce E Taillon, Zhoutao Chen, Andrea Tanzer, A C Eugenia Saunders, Jianxiang Chi, Fengtang Yang, Nigel P Carter, Matthew E Hurles, Sherman M Weissman, Timothy T Harkins, Mark B Gerstein, Michael Egholm, and Michael Snyder. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318:420–6, October 2007.

[28] J Kraus, R G Weber, M Cremer, T Seebacher, C Fischer, C Schurra, A Jauch, P Lichter, A Bensimon, and T Cremer. High-resolution comparative hybridization to combed dna fibers. *Hum Genet JT - Human genetics*, 99:374–80, March 1997.

[29] A Kuklin, K Munson, D Gjerde, R Haefele, and P Taylor. Detection of single-nucleotide polymorphisms with the wave dna fragment analysis system. *Genet Test JT - Genetic testing*, 1:201–6, September 1999.

[30] E Leich, E Haralambieva, A Zettl, A Chott, T Rudiger, S Holler, H-K Muller-Hermelink, G Ott, and A Rosenwald. Tissue microarray-based screening for chromosomal breakpoints affecting the t-cell receptor gene loci in mature t-cell lymphomas. *J Pathol*, 213:99–105, August 2007.

[31] Emmanuelle Lerat and Marie Semon. Influence of the transposable element neighborhood on human gene expression in normal and tumor tissues. *Gene*, 396:303–11, June 2007.

[32] P Lesage and A L Todeschini. Happy together: the life and times of ty retrotransposons and their hosts. *Cytogenet Genome Res*, 110:70–90, August 2005.

[33] Emma Levdansky, Jacob Romano, Yona Shadkchan, Haim Sharon, Kevin J Verstrepen, Gerald R Fink, and Nir Osherov. Coding tandem repeats generate diversity in aspergillus fumigatus genes. *Eukaryot Cell*, 6:1380–91, August 2007.

[34] Victoria V Lunyak, Gratien G Prefontaine, Esperanza Nunez, Thorsten Cramer, Bong-Gun Ju, Kenneth A Ohgi, Kasey Hutt, Rosa Roy, Angel Garcia-Diaz, Xiaoyan Zhu, Yun Yung, Lluis Montoliu, Christopher K Glass, and Michael G Rosenfeld. Developmentally regulated activation of a sine b2 repeat as a domain boundary in organogenesis. *Science*, 317:248–51, July 2007.

[35] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B Dewell, Lei Du, Joseph M Fierro, Xavier V Gomes, Brian C Godwin, Wen He, Scott Helgesen, Chun Heen Ho, Gerard P Irzyk, Szilveszter C Jando, Maria L I Alenquer, Thomas P Jarvie, Kshama B Jirage, Jong-Bum Kim, James R Knight, Janna R Lanza, John H Leamon, Steven M Lefkowitz, Ming Lei, Jing Li, Kenton L Lohman, Hong Lu, Vinod B Makhijani, Keith E McDade, Michael P McKenna, Eugene W Myers, Elizabeth Nickerson, John R Nobile, Ramona Plant, Bernard P Puc, Michael T Ronan, George T Roth, Gary J Sarkis, Jan Fredrik Simons, John W Simpson, Maithreyan Srinivasan, Karrie R Tartaro, Alexander Tomasz, Kari A Vogt, Greg A Volkmer, Shally H Wang, Yong Wang, Michael P Weiner, Pengguang Yu, Richard F Begley, and Jonathan M Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–80, September 2005.

[36] Yael P Mosse, Sharon J Diskin, Nora Wasserman, Katherine Rinaldi, Edward F Attiyeh, Kristina Cole, Jayanti Jagannathan, Karishma Bhambhani, Cynthia Winter, and John M Maris. Neuroblastomas have distinct genomic dna profiles that predict clinical phenotype and regional gene expression. *Genes Chromosomes Cancer*, 46:936–49, August 2007.

[37] Jan Mrazek, Xiangxue Guo, and Apurva Shah. Simple sequence repeats in prokaryotic genomes. *Proc Natl Acad Sci U S A JT - Proceedings of the National Academy of Sciences of the United States of America*, 104:8472–7, May 2007.

[38] Alysson R. Muotri, Vi T. Chu, Maria C.N. Marchetto, Wi Deng, John V. Moran, and Fred H. Gage. Somatic mosaicism in neural precursor cells mediated by l1 retrotransposition. *Nature*, 435:903–912, June 2005.

[39] G Ng, D Winder, B Muralidhar, E Gooding, I Roberts, M Pett, G Mukherjee, J Huang, and N Coleman. Gain and overexpression of the oncostatin m receptor occur frequently in cervical squamous cell carcinoma and are associated with adverse clinical outcome. *J Pathol*, 212:325–34, June 2007.

[40] Lilia Perfeito, Lisete Fernandes, Catarina Mota, and Isabel Gordo. Adaptive mutations in bacteria: high rate and small effects. *Science JT - Science (New York, N.Y.)*, 317:813–5, August 2007.

[41] Nadege Philippe, Estelle Crozat, Richard E Lenski, and Dominique Schneider. Evolution of global regulatory networks during a long-term experiment with escherichia coli. *Bioessays*, 29:846–60, August 2007.

[42] D Pinkel, R Segraves, D Sudar, S Clark, I Poole, D Kowbel, C Collins, W L Kuo, C Chen, Y Zhai, S H Dairkee, B M Ljung, J W Gray, and D G Albertson. High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nat Genet JT - Nature genetics*, 20:207–11, October 1998.

[43] E Rouchleau, C Lefol, S Tozlu, C Andrieu, C Guy, F Copigny, C Nogues, I Bieche, and R Lidereau. High-resolution oligonucleotide array-cgh applied to the detection and characterization of large rearrangements in the hereditary breast cancer gene brca1. *Clin Genet*, 72:199–207, August 2007.

[44] Neeraj Salathia, Hana N Lee, Todd A Sangster, Keith Morneau, Christian R Landry, Kurt Schellenberg, Aditi S Behere, Kevin L Gunderson, Duccio Cavalieri, Georg Jander, and Christine Queitsch. Indel arrays: an affordable alternative for genotyping. *Plant J JT - The Plant journal : for cell and molecular biology*, 51:727–37, August 2007.

[45] Mh Shen, K Mantripragada, Jp Dumanski, I Frayling, and M Upadhyaya. Detection of copy number changes at the nf1 locus with improved high-resolution array cgh. *Clin Genet JT - Clinical genetics*, 72:238–44, August 2007.

[46] M M Shi. Enabling large-scale pharmacogenetic studies by high-throughput mutation detection and genotyping technologies. *Clin Chem JT - Clinical chemistry*, 47:164–72, February 2001.

[47] Douglas R Smith, Aaron R Quinlan, Heather E Peckham, Kathryn Makowsky, Wei Tao, Betty Woolf, Lei Shen, William F Donahue, Nadeem Tusneem, Michael P Stromberg, Donald A Stewart, Lu Zhang, Swati S Ranade, Jason B Warner, Clarence C Lee, Brittney E Coleman, Zheng Zhang, Stephen F McLaughlin, Joel A Malek, Jon M Sorenson, Alan P Blanchard, Jarrod Chapman, David Hillman, Feng Chen, Daniel S Rokhsar, Kevin J McKernan, Thomas W Jeffries, Gabor T Marth, and Paul M Richardson. Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res*, 18:1638–42, October 2008.

[48] Asher Tishler and Israel Zang. A new maximum likelihood algorithm for piecewise regression. *Journal of the American Statistical Association*, 76:980–987, December 1981.

[49] Eray Tuzun, Andrew J Sharp, Jeffrey A Bailey, Rajinder Kaul, V Anne Morrison, Lisa M Pertz, Eric Haugen, Hillary Hayden, Donna Albertson, Daniel Pinkel, Maynard V Olson, and Evan E Eichler. Fine-scale structural variation of the human genome. *Nat Genet*, 37:727–32, July 2005.

[50] Jose A J M van den Hurk, Iwan C Meij, Maria del Carmen Seleme, Hiroki Kano, Konstantinos Nikopoulos, Lies H Hoefsloot, Erik A Sistermans, Ilse J de Wijs, Arijit

Mukhopadhyay, Astrid S Plomp, Paulus T V M de Jong, Haig H Kazazian, and Frans P M Cremers. L1 retrotransposition can occur early in human embryonic development. *Hum Mol Genet*, 16:1587–92, June 2007.

[51] Gregory J Velicer, Gunter Raddatz, Heike Keller, Silvia Deiss, Christa Lanz, Iris Dinkelacker, and Stephan C Schuster. Comprehensive mutation identification in an evolved bacterial cooperator and its cheating ancestor. *Proc Natl Acad Sci U S A*, 103:8107–12, May 2006.

[52] Amy J Vogler, Christine Keys, Yoshimi Nemoto, Rebecca E Colman, Zack Jay, and Paul Keim. Effect of repeat copy number on variable-number tandem repeat mutations in escherichia coli o157:h7. *J Bacteriol JT - Journal of bacteriology*, 188:4253–63, June 2006.

[53] Sarah J Wheelan, Lisa Z Scheifele, Francisco Martinez-Murillo, Rafael A Irizarry, and Jef D Boeke. Transposon insertion site profiling chip (tip-chip). *Proc Natl Acad Sci U S A JT - Proceedings of the National Academy of Sciences of the United States of America*, 103:17632–7, November 2006.

[54] R M Williams. The yeast lifecycle and dna array technology. *J Ind Microbiol Biotechnol JT - Journal of industrial microbiology & biotechnology*, 28:186–91, June 2002.

[55] Kelly M Winterberg and William S Reznikoff. Screening transposon mutant libraries using full-genome oligonucleotide microarrays. *Methods Enzymol JT - Methods in enzymology*, 421:110–25, March 2007.

[56] Daniel R Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Res*, 18:821–9, May 2008.

[57] Yi Zhang, Kim A Hatch, Lorenz Wernisch, and Joanna Bacon. A bayesian change point model for differential gene expression patterns of the dosr regulon of mycobacterium tuberculosis. *BMC Genomics*, 9:87, March 2008.