

Genetic recombinant analysis of TGCE data

by

Philip Michael Maher

A thesis submitted to the graduate faculty
in partial fulfillment of the requirement for the degree of

MASTER OF SCIENCE

Major: Computer Science

Program of Study Committee:
Hui-Hsien Chou (Major Professor)
Xiaoqiu Huang
Patrick Schnable

Iowa State University

Ames, Iowa

2004

Graduate College
Iowa State University

This is to certify the master's thesis of

Philip Michael Maher

has met the thesis requirements of Iowa State University

Signatures have been redacted for privacy

TABLE OF CONTENTS

ABSTRACT.....	vi
CHAPTER 1: INTRODUCTION	1
CHAPTER 2: BIOLOGICAL PROCESSES AND TOOLS BACKGROUND.....	3
CHAPTER 3: TGCE SOFTWARE BACKGROUND	13
CHAPTER 4: METHODS.....	22
CHAPTER 5: PEAK IDENTIFICATION.....	36
CHAPTER 6: CONSENSUS CALLS	42
CHAPTER 7: EXPERIMENTS.....	44
CHAPTER 8: DISCUSSION AND CONCLUSION	51
REFERENCES	53
ACKNOWLEDGEMENTS.....	57

LIST OF FIGURES

Figure 2.1: Development of Recombinant Inbreds.....	5
Figure 2.2: Development of Recombinant Inbreds from Intermated F ₂	6
Figure 2.3: Formation of Homoduplex and Heteroduplex Molecules.....	9
Figure 2.4: Scoring Recombinant Inbreds	12
Figure 3.1: Revelation™.....	13
Figure 3.2: Revelation™ *.dmp View	14
Figure 3.3: Revelation™ *.smr Show All Capillaries	15
Figure 3.4: Revelation™ *.smd Show All Capillaries	17
Figure 3.5: Revelation™ Report View	18
Figure 3.6: Revelation™ Mutation Call Result Window	19
Figure 4.1: GRAMA Initial Window.....	23
Figure 4.2: GRAMA Genetic Recombinant Analysis Window – No Unmixed Plate.....	25
Figure 4.3: GRAMA Genetic Recombinant Analysis Window – With Unmixed Plate.....	26
Figure 4.4: GRAMA Single Well View	29
Figure 4.5: GRAMA Across Mixtures View – No Unmixed Plate	30
Figure 4.6: GRAMA Across Mixtures View – With Unmixed Plate.....	31
Figure 4.7: GRAMA Across Wells View	32
Figure 4.8: GRAMA Genetic Recombinant Analysis Window Displaying Colored Rows..	34
Figure 5.1: Concavity of a Peak.....	37
Figure 5.2: Inflection Points in Electropherogram	38
Figure 5.3: Multiple Peak Markers for Single Peak	39
Figure 5.4: Final Peak Markers Designating Peak.....	40

LIST OF TABLES

Table 6.1: Consensus Call Determination Table - NT Indicates Not Tested.....	43
Table 7.1: Revelation™ and GRAMA Error Rates	46

ABSTRACT

Analysis of recombinant inbred lines is used to generate data needed for the creation of genetic maps of organisms. Using these genetic maps, the functions of genes can be discovered. To be most useful, the genetic map needs to be of high-density. Genetic map density can be increased by using a process called Temperature Gradient Capillary Electrophoresis (TGCE) to detect small polymorphisms between genetic alleles. This thesis introduces a software tool called GRAMA (Genetic Recombinant Analysis and Mapping Assistant) which is able to analyze data from recombinant inbred lines subjected to TGCE and present automated results to the user in an intuitive visual format. Data from multiple TGCE runs are integrated to display all necessary data simultaneously. GRAMA contains its own algorithm to detect peaks from electropherogram data produced by TGCE. Results from GRAMA's algorithm are compared with results from another software package used to evaluate electropherograms and differences are flagged for further analysis. GRAMA produces two sets of consensus scores as output. One set of scores provides the user with very detailed information that encodes all possible experimental results, while the other summarizes these results into mapping scores that can be used as direct input for a genetic mapping program. Experiments reveal that GRAMA generates highly accurate results and boosts user productivity more than two-fold relative to previous methods used to perform recombinant inbred analysis of TGCE data.

CHAPTER 1: INTRODUCTION

GRAMA (Genetic Recombinant Analysis and Mapping Assistant) is a software tool that has been created for the analysis of recombinant inbred (RI) data produced via the Temperature Gradient Capillary Electrophoresis (TGCE) process. By comparing results produced by the TGCE process for differentiable DNA segments, called *genetic markers*, of recombinant inbred DNA, data needed for genetic linkage analysis and genetic mapping are produced. GRAMA facilitates the recombinant inbred analysis by allowing all necessary data to be viewed and edited simultaneously. This simultaneous visualization allows the user to make quick and accurate judgments. In addition, data are presented in a format that makes intuitive sense for this particular problem domain.

This thesis is organized into the following chapters. Chapter 2 provides the background information on recombinant inbreds and TGCE as well as the motivation behind using TGCE to analyze recombinant inbred DNA. An explanation of how this process can be used to produce data useful for genetic linkage analysis and genetic mapping is also included. Chapter 3 introduces the existing commercial software used to analyze TGCE data and its limitations for dealing with recombinant inbred analysis. Chapter 4 describes how to prepare data for input to GRAMA and the features that GRAMA provides for enhanced recombinant inbred analysis. Chapter 5 explains the algorithms used for peak identification and selection in the TGCE electropherogram. Chapter 6 describes how the peak calling consensus scores are produced and what type of biological information each score conveys. Chapter 7 presents several experiments performed to compare the accuracy of GRAMA and the commercial software with regard to their correctness in distinguishing between

homoduplex and heteroduplex DNA. In addition, this chapter describes the benefits of combining both tools to increase the overall accuracy. Finally, Chapter 8 summarizes the contributions of this work and suggests future directions and enhancements.

CHAPTER 2: BIOLOGICAL PROCESSES AND TOOLS BACKGROUND

Genetic mapping is one of the most important means of linking genes to their corresponding functions. Experiments based on a genetic map can identify chromosomal regions associated with mutants or quantitative traits [7]. This is one of the first steps in the pursuit of isolating and cloning a particular gene and identifying the protein it encodes [18]. The process of identifying genes in a chromosomal region is more straightforward when the physical map (nucleotide sequence) is known [25] or if the organism exhibits a high degree of synteny to genomes of other organisms for which the physical map is known [9]. The simple solution would seem to be to sequence the genome of each organism of interest. However, some genomes are too complicated for straight end-to-end sequencing [2]. Thus, the method of aligning the genomes of organisms that have a high degree of conserved gene content is being more closely investigated as an alternative approach for gene discovery.

One such collection of organisms that displays a high degree of synteny among their genomes is the cereals [10, 20]. Nevertheless, there are a number of exceptions to collinearity among the cereals [3, 10, 22]. Thus, a genetic map must be of high density in order to have as many cross-links between the organisms as possible. This allows the genetic map of one organism to be aligned with the physical map of another organism as accurately as possible. The physical map for rice is nearly completed [4], so if high-density genetic maps of the other cereals are generated, many important biological discoveries should result. Similar conclusions can be drawn among any group of organisms that exhibit a high degree of synteny among their genomes.

The natural occurrence of recombination during the meiosis process is used as a tool to gain order and distance information for genetic mapping purposes. Before the first meiotic division, replicated chromosomes of each homologous pair align. At this point, crossover can occur which results in recombination between the chromatids. Crossover points are thought to be fairly evenly distributed along the length of the chromosome, but specific hot and cold spots have been discovered [16]. Because of this, in general “the more frequently recombination occurs between two genes on the same chromosome, the farther apart they are” [18]. Thus, statistical information can be obtained about how often recombination occurs between two specific sites, and the genetic distance between them on the chromosome can be estimated.

Results derived from recombinant inbred lines have been used successfully to gather linkage and recombination statistics for genes [1, 5]. The process for generating a recombinant inbred line is conducted as follows and is illustrated in Figure 2.1. Two unrelated and highly inbred lines of a species are crossed and those offspring are then crossed. The resulting F_2 generation is then selfed or sib-mated until homozygosity is achieved [15]. Homozygosity is eventually achieved when a species is selfed or sib-mated for several generations, excluding the effects of mutations [12]. The resulting population is called a recombinant inbred (RI) population and has two very important properties. The first is that each RI line can be continuously propagated and the progeny’s genetic make-up will be identical to parental DNA, apart from mutations. The second property is that, because multiple rounds of meiosis occur during the inbreeding process, more recombination has taken place. This allows genetic map distance for genes to be determined with higher

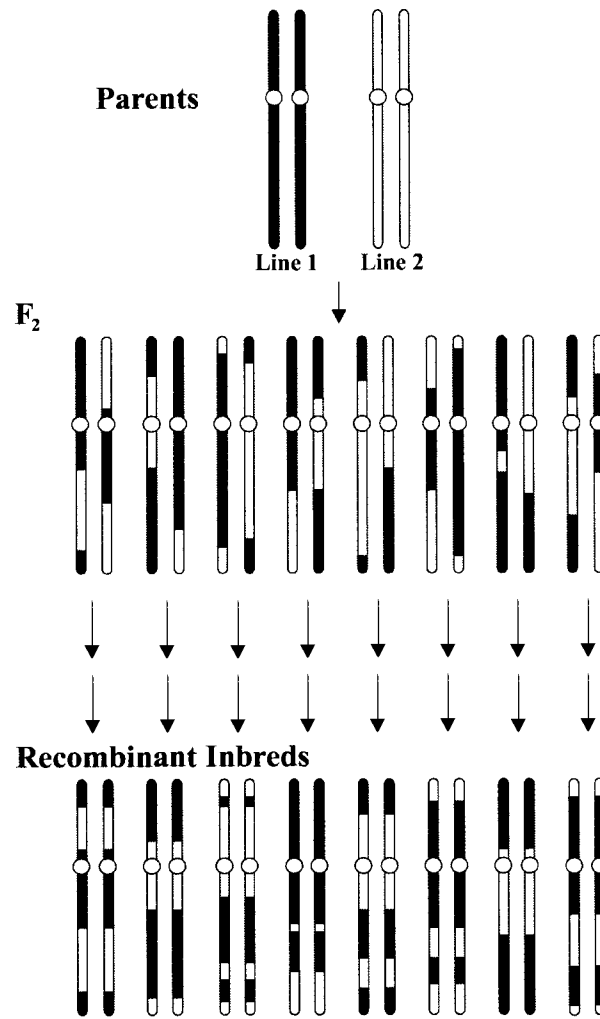


Figure 2.1: Development of Recombinant Inbreds

confidence [5]. The genetic mapping density can be further enhanced by intermating the F_2 population before beginning the development of the RI lines [14]. This process is illustrated in Figure 2.2.

The discovery of genetic markers is an important step in gathering the necessary statistics for estimating genetic map distances. Genetic markers are areas of the genome that contain a difference at the sequence level (polymorphism) between homologous

chromosomes. Many different kinds of polymorphisms can be used as genetic markers. Some examples are restriction fragment length polymorphisms (RFLPs), insertion-deletion polymorphisms (IDPs), simple-sequence repeats (SSRs), and single-nucleotide polymorphisms (SNPs). Sequenced markers that differentiated between alleles from the original parental generation are the most useful for cross-linking purposes. Cross-linking can

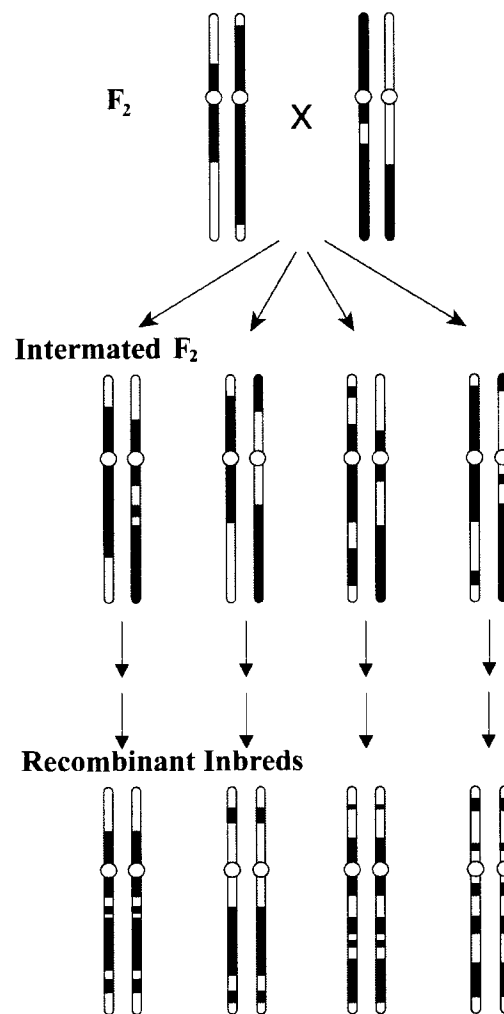


Figure 2.2: Development of Recombinant Inbreds from Intermated F₂

be conducted between the genetic map that is generated and the physical map of another organism whose genome exhibits a high degree of synteny with the target species [9].

Expressed sequence tags (ESTs) are one possible source of such markers. Polymorphisms between homologous chromosomes are most likely to be tolerated in the untranslated regions (UTRs) or the introns because these regions do not contribute directly to the formation of proteins and can therefore more easily tolerate mutations. Since exonic regions are directly related to the formation of proteins, their sequences are typically more highly conserved via natural selection. Primer pairs can be developed that amplify the UTRs and introns via polymerase chain reaction (PCR). Although the actual sequences of the polymorphisms may not be known, agarose gel electrophoresis can detect many IDP polymorphisms with high-throughput. If, after DNA from both parental lines are amplified by the same primer pair and subjected to agarose gel electrophoresis and the PCR products reveal a size polymorphism or only one allele is amplified, then a genetic marker has been discovered that can differentiate between the two different alleles of the original parental lines. This primer pair can then be used to amplify the RI lines. Once the PCR products from the RI lines are subjected to agarose gel electrophoresis, it is possible to determine from which parent the amplified genetic material was inherited.

The number of markers on a genetic map can be further increased if smaller IDPs and SNPs can also be distinguished by an efficient high-throughput process. Many methods have been developed to detect SNPs and most of them require specific optimization processes for each particular sequence or prior knowledge of the sequence itself [15, 21]. Temperature Gradient Capillary Electrophoresis (TGCE) is a recent advance in analytical chemistry that is

sensitive enough for efficient detection of small IDPs and SNPs with no prior knowledge about the specifics of the polymorphism [11, 15].

Primer pairs that do not reveal polymorphisms via agarose gel electrophoresis analysis can be used to amplify the same genomic regions to see if there are smaller polymorphisms that may be detectable via TGCE. TGCE works as follows. A mixture of double-stranded DNA (ds-DNA), amplified by the primer pair via PCR, is placed in a well of a microtiter dish (plate). The DNA is heated to a point at which it denatures, leaving a mixture of single-stranded DNA (ss-DNA). As the sample cools, the DNA re-anneals in several combinations. For instance, if the DNA mixture contains two different alleles that contain a SNP, then during the re-annealing process some ss-DNA may re-anneal with ss-DNA of the other allele. In this case the DNA will not re-anneal completely because one base pair will not be complementary. This leaves a small bulge in the new ds-DNA. This is called a heteroduplex molecule. Double-stranded DNA that has re-annealed completely at every base pair is called a homoduplex molecule. The heteroduplex molecules will be formed when ss-DNA from different alleles re-anneals and homoduplex molecules will be formed when ss-DNA from the same allele re-anneals. Because of this, a well containing two different alleles which contain a SNP, will contain both heteroduplex and homoduplex molecules following re-annealing. Figure 2.3 illustrates this process.

When an electrical current is applied, the newly annealed ds-DNA begins to migrate into capillaries which contain gel material. Because of the bulge in the heteroduplex molecules, these molecules move through the gel at a different rate than the homoduplex molecules. Thus, as the DNA migrates through the gel, it separates into different bands.

Conversely, if a well consists entirely of one allele, every molecule will re-anneal completely after the denaturation process and will create a single band of DNA intensity.

Electropherograms are created by graphing the intensity of the DNA as it moves across a fixed point on the capillary. Because a well that contains only homoduplex molecules typically creates a single band in the capillary, the electropherogram will contain a single peak pattern. For a well that contains both homoduplex and heteroduplex molecules, several different bands of DNA are produced in the capillary. Thus, the electropherogram will contain a multiple peak pattern (Figure 2.3).

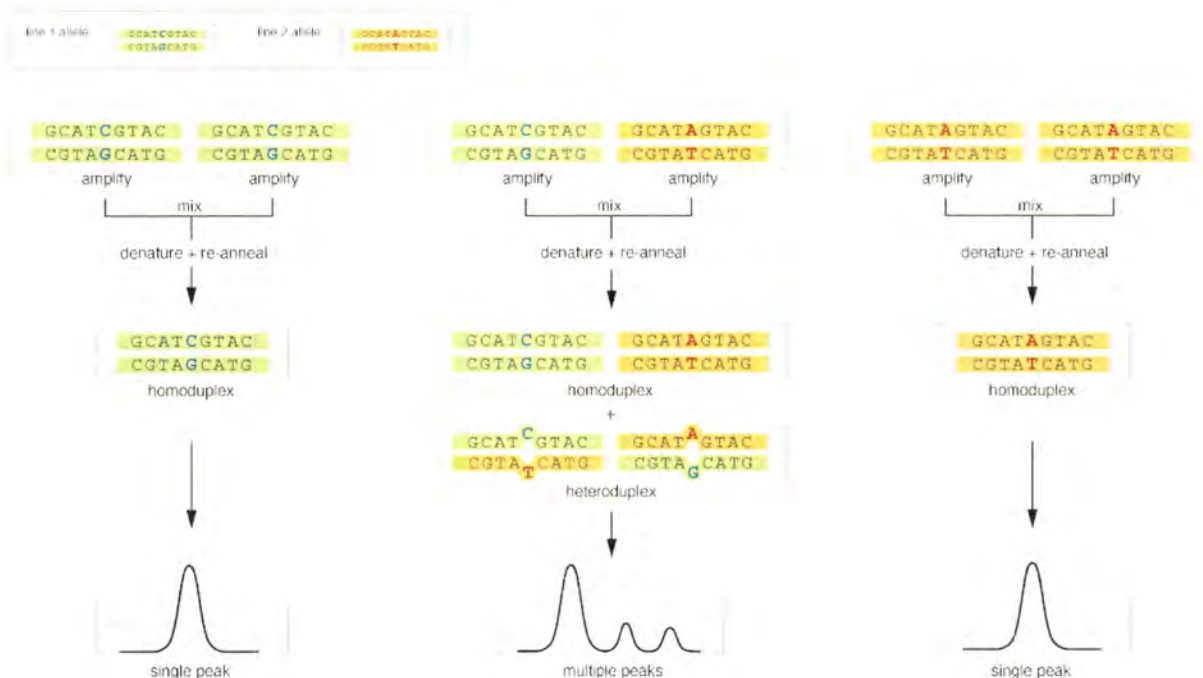


Figure 2.3: Formation of Homoduplex and Heteroduplex Molecules

Again, as in agarose gel electrophoresis, the original lines must be tested first to see if there is a polymorphism present that will be detectable by the TGCE analysis. Let the two original lines from the parental generation be called line 1 and line 2. A sample containing only DNA from line 1, a sample containing only DNA from line 2, and a sample containing a

mixture of DNA from both lines are subjected to the TGCE procedure for each genomic region of interest amplified by a primer pair. In order for this genomic region to be useful in developing mapping information from the RI lines, the sample of line 1 and the sample of line 2 must contain only homoduplex molecules after denaturation and re-annealing have taken place. This is expected because lines 1 and 2 should be homozygous at all sites, but exceptions have been observed. Also, the sample containing DNA from both lines must contain some heteroduplex molecules in addition to the expected homoduplex molecules. If this is not the case, then there is no detectable polymorphism (via TGCE) between lines 1 and 2 in this genomic region.

If results from this analysis indicate that there is a detectable polymorphism between lines 1 and 2 at this site, then the primer pair can be used to amplify the same genomic region for each of the RI lines. Each RI line should contain only genetic material inherited either from line 1 or line 2. The PCR product from each RI line is then mixed separately with the PCR products (generated using the same primer pair) from line 1 and line 2 for analysis. If a particular RI line has genetic content in the region derived from line 1, when its mixture with line 1 is subjected to the TGCE process, only homoduplex molecules should result. In addition, when this same RI line is mixed with line 2 and processed, heteroduplex molecules should form. A reciprocal argument holds if a particular RI line has genetic content that was inherited from line 2. When processed via TGCE, only homoduplex molecules should form in its mixture with line 2 and some heteroduplex molecules should form in its mixture with line 1. Occasionally, a mixture consisting only of the PCR products of a particular RI is subjected to the TGCE analysis to verify that it is homozygous at that particular site. If this

does not hold, no meaningful results can be obtained when mixing the RI with line 1 and line 2.

In the case where the TGCE results of an RI indicate that its genomic region in question is inherited from a particular line, it is possible to assign a mapping “score” to the genetic marker for this RI. The score is either ‘1’ or ‘2’ indicating whether the amplified area is believed to be propagated from line 1 or line 2 respectively. Each of the RI lines for a genetic marker is scored in this manner. The result is a series of ‘1’s and ‘2’s that collectively make up the mapping scores for the genetic marker.

By producing these scores for as many genetic markers as possible, the density of the map will increase and the cross-linking of species will be more accurate. Based on the linear behavior of chromosomes during meiosis, it stands to reason that areas on the chromosome that are closer together will share many of the same mapping scores across all of the RI lines. Some markers will have identical mapping scores and therefore it will not be possible to determine an upper bound on the genetic distance between them. This idea is illustrated in Figure 2.4. Several software packages using a variety of algorithms and heuristics have been written to use mapping score information as input and produce a likely genetic map [6, 13, 17, 19, 26]. Since the TGCE process is capable of detecting polymorphisms that are not detectable using agarose gel electrophoresis, more genetic markers can be scored, which in turn, can produce a higher-density genetic map that is more accurate. Because the genetic map is of higher density, the likelihood of cross-links between this genetic map and a physical map of another similar organism is greatly increased.

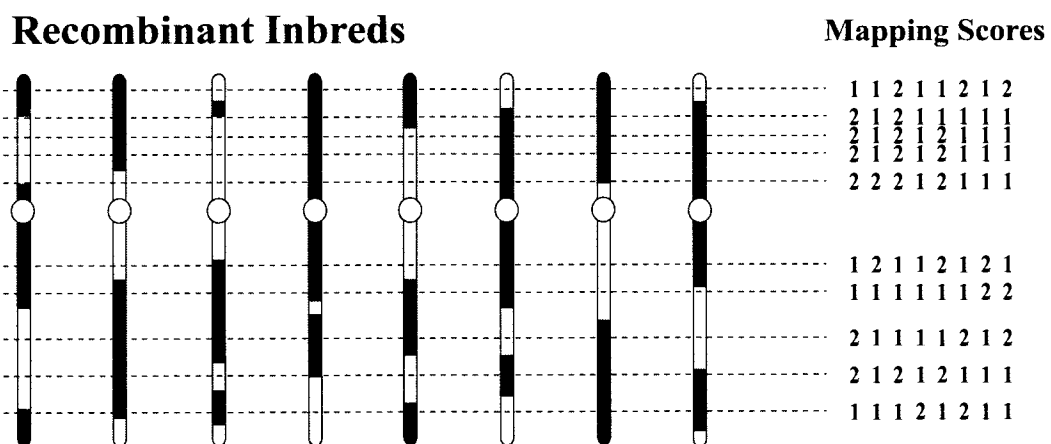


Figure 2.4: Scoring Recombinant Inbreds

CHAPTER 3: TGCE SOFTWARE BACKGROUND

The only existing commercially available TGCE system is produced by the SpectruMedix® Corporation. SpectruMedix® also provides a software solution to analyze the data from the TGCE process. This software package is called Revelation™ (Figure 3.1). In this chapter, the primary functionalities and capabilities of the Revelation™ software package will be discussed. Revelation™ has many features but only those that are primarily used in generating meaningful results for genetic recombinant analysis will be mentioned.



Figure 3.1: Revelation™¹

After the TGCE process is completed, a *.dmp file will be created. The *.dmp file is essentially a digital picture of the capillaries and the DNA band intensities. The first step in

¹ Revelation, SpectruMedix, and the SpectruMedix logo are trademarks of SpectruMedix LLC, State College, PA

the Revelation™ software involves picking the capillaries (Figure 3.2). The program has both automatic and manual capillary picking options. The automatic capillary picking correctly identifies the capillaries in most cases. In some cases, user intervention may be required to make sure the capillary number is aligned correctly with the proper capillary. If this is not done, later analyses will not correspond to the correct capillary.

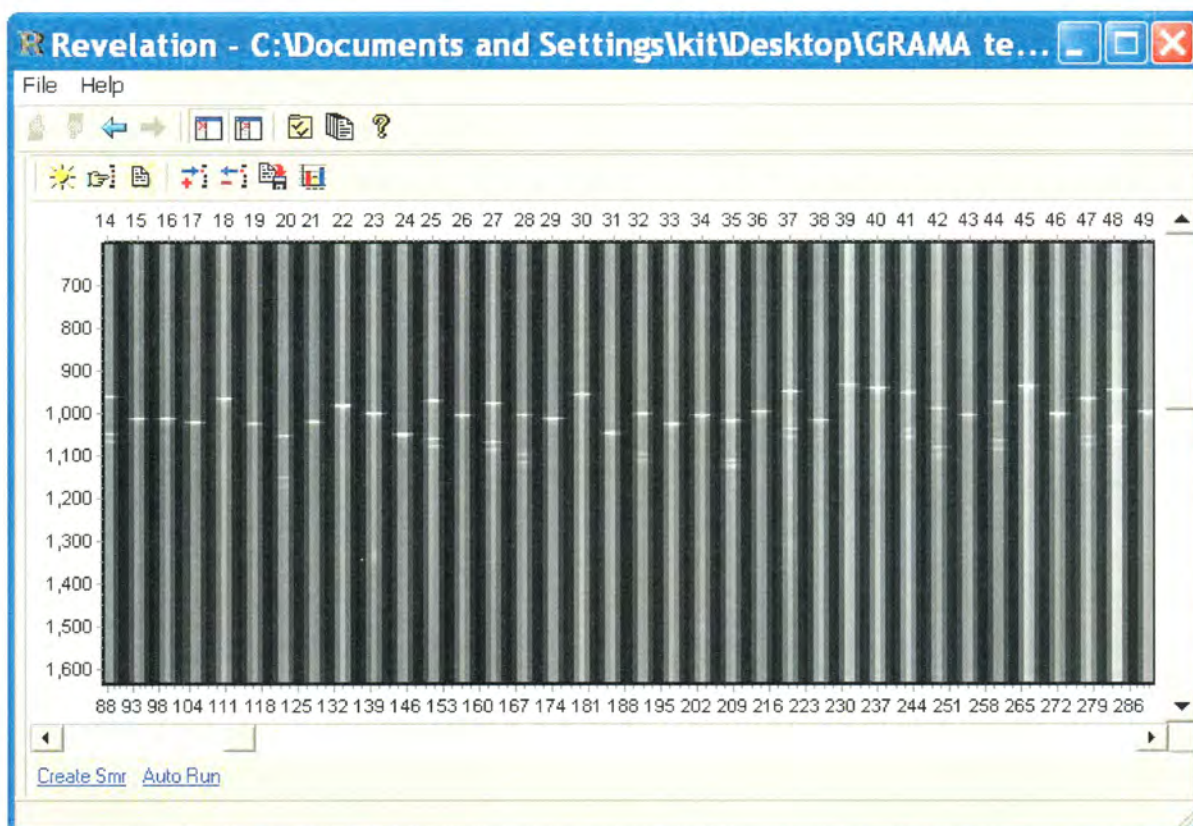


Figure 3.2: Revelation™ *.dmp View

Once the capillary picking process is complete a *.smr file can be created. With the *.smr file electropherograms can be viewed for each capillary (Figure 3.3). The electropherogram for each capillary displays 11 different colors. Ten of which are intensities measured in a particular wavelength, the eleventh (black line) is the average intensity over all measurable wavelengths. The ability to detect DNA intensities at multiple wavelengths is

provided so that multiplexing can be employed by using different fluorescent dye primers [15].

Several tools are provided in the *.smr view to prepare the data for later mutation (heteroduplex) detection. The data edit options define specific regions to be considered for further analyses. There are three different data edit options. The first is Set Interest Region. This limits the analysis area to the user-selected region, and data below the user-selected intensity value will be ignored. The Erase Noise option attempts to remove all noise in the user-defined region. The Data Trim option is similar to the Set Interest Region option except that intensity values are not considered.

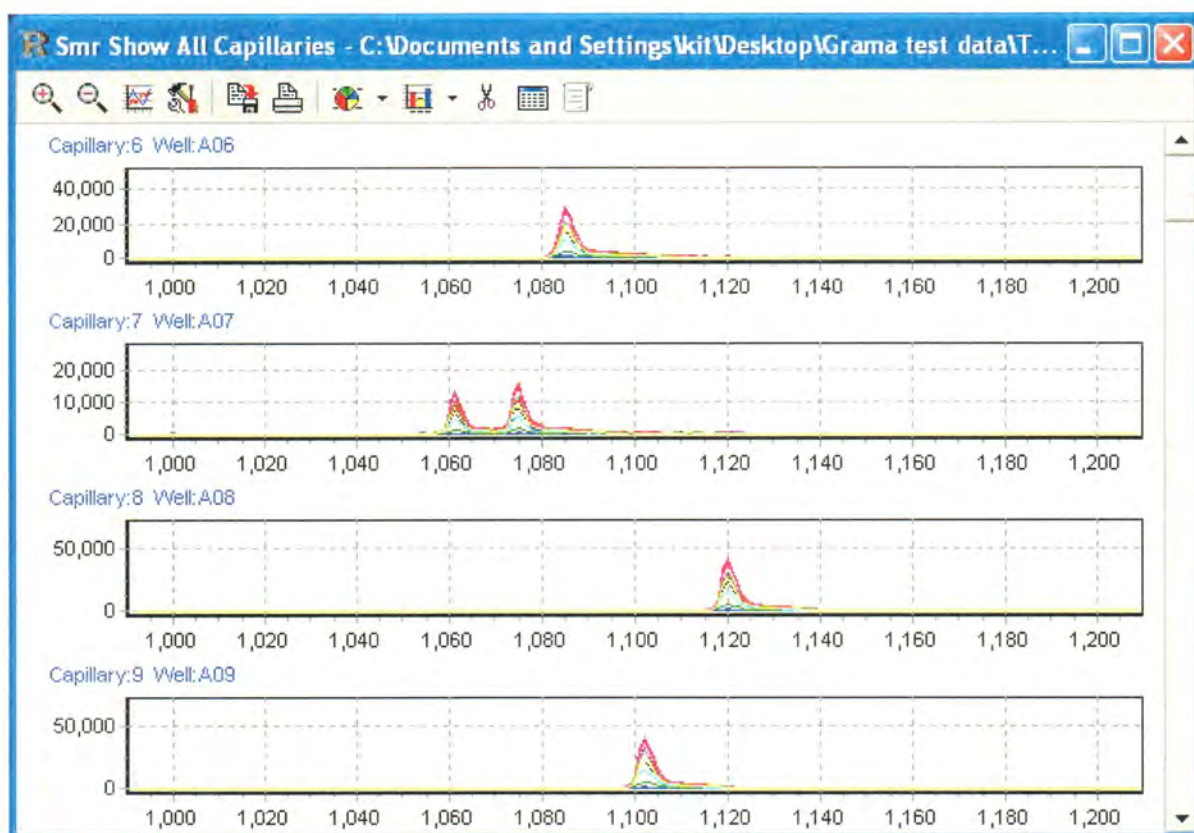


Figure 3.3: Revelation™ *.smr Show All Capillaries

Another option provided is the smoothing option. This attempts to remove all jagged edges from the peaks. The software allows the user to choose one of three smoothing algorithms (Stavitzky-Golay, Stavitzky-Golay Adjust Window, or Fast Fourier Transformation).

The third and most important data editing option is Baseline Subtract. Four options are available for subtracting the baseline: Minimal Intensity, Minimal Deviation, Average of Low Point, and Medium of Low Point. It is suggested that baseline subtraction always be performed. Minimal intensity baseline subtraction is appropriate in most cases [23].

A specific dye color must be selected by the user for further analyses. In cases where no fluorescent dyes have been used, the dye color that yields the highest intensity peak should be selected. Once all data editing options are completed a *.smd file will be created.

It is from the *.smd file where final settings are made in preparation for mutation detection. The electropherograms for the particular wavelength selected for analysis can be viewed for all of the capillaries (Figure 3.4). These electropherograms will only contain data from the regions selected by the user during the previous step. At this point, the user is expected to select which well is to be used as the control. The control well should be a homogeneous DNA solution and not a mixture, and contain only homoduplex molecules.

The presence of a mutation (indicated by the presence of heteroduplex molecules) is detected by calculating a match factor. The match factor is based on how similar the electropherogram for each capillary is to the electropherogram for the capillary selected as the control. Several parameters can be tweaked using the Mutation Call Threshold Setting option. These include the Normalization Window Size, Comparison Window Size, and thresholds to be placed on the match factor values that will be used in deciding whether the

well contains a mutation, does not contain a mutation, or if its status can not be determined. There are also tools available to assist with size multiplexing as long as the PCR products are separated by at least 50 base pairs. After the control is selected and the parameters are set as desired, the electropherograms will be analyzed for mutations and a report will be generated.

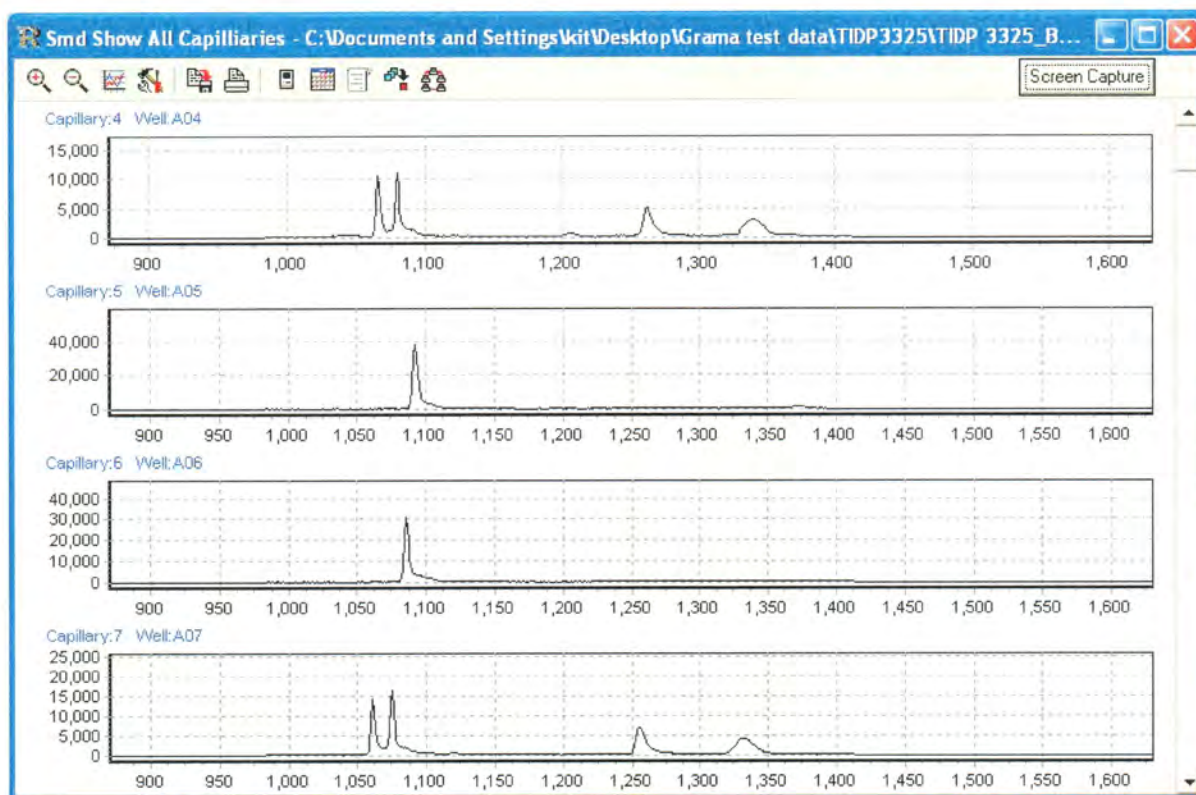
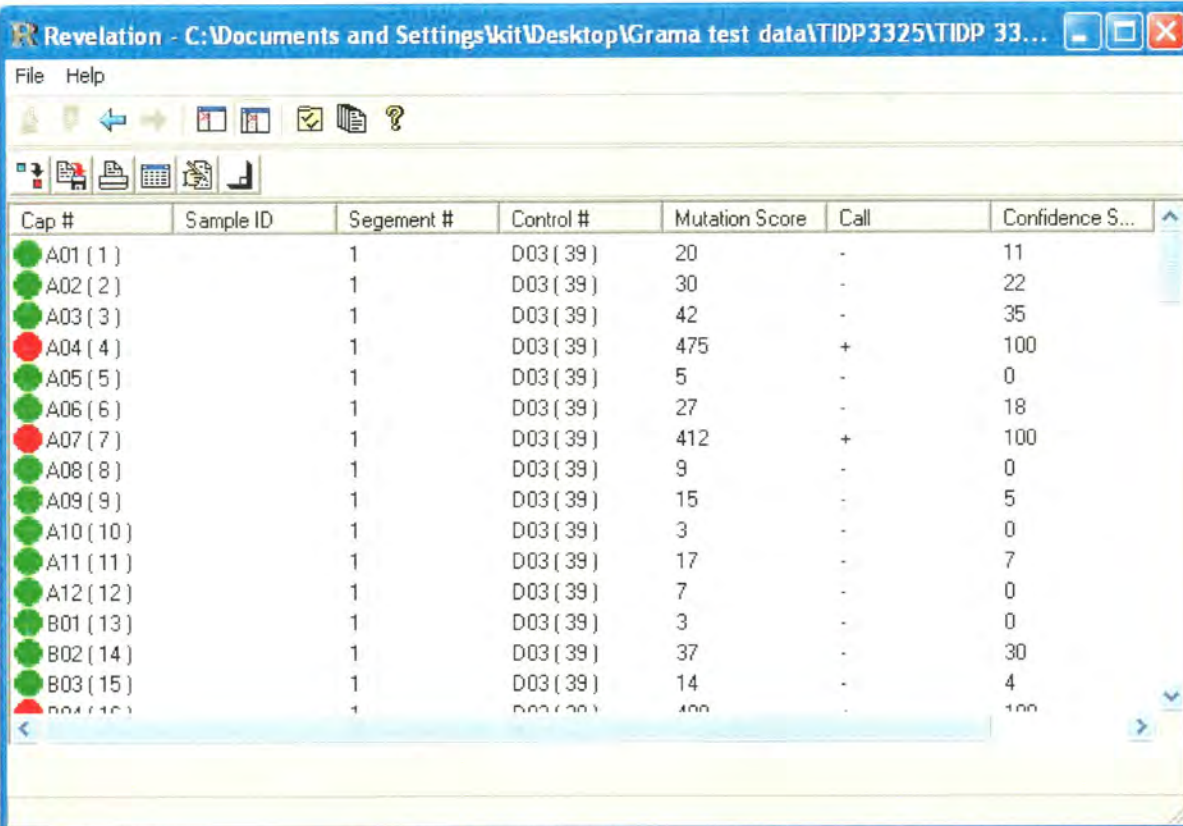


Figure 3.4: Revelation™ *.smd Show All Capillaries

The report (Figure 3.5) shows information about each capillary. The well that is used as the control when comparing each of the other wells is listed. A mutation score is assigned to each capillary: the higher the value, the more likely it is a mutation. The final call is reported in the Call column and indicated by a circle of a corresponding color in the Cap # column. In the Call column, mutations are indicated by '+', non-mutations are indicated by



The screenshot shows the 'Revelation' software window with a menu bar (File, Help) and a toolbar. Below the toolbar is a table with the following columns: Cap #, Sample ID, Segment #, Control #, Mutation Score, Call, and Confidence S... The table contains 15 rows of data, each representing a capillary. The 'Cap #' column lists capillaries A01 through B04. The 'Sample ID' column lists sample IDs from 1 to 15. The 'Segment #' column lists segment numbers from 1 to 15. The 'Control #' column lists control IDs from D03 to D04. The 'Mutation Score' column lists scores from 20 to 475. The 'Call' column lists calls from '-' to '+'. The 'Confidence S...' column lists confidence scores from 11 to 100. The table is scrollable, and the bottom of the table is partially obscured by a blue bar.

Cap #	Sample ID	Segment #	Control #	Mutation Score	Call	Confidence S...
A01 (1)	1	1	D03 (39)	20	-	11
A02 (2)	1	1	D03 (39)	30	-	22
A03 (3)	1	1	D03 (39)	42	-	35
A04 (4)	1	1	D03 (39)	475	+	100
A05 (5)	1	1	D03 (39)	5	-	0
A06 (6)	1	1	D03 (39)	27	-	18
A07 (7)	1	1	D03 (39)	412	+	100
A08 (8)	1	1	D03 (39)	9	-	0
A09 (9)	1	1	D03 (39)	15	-	5
A10 (10)	1	1	D03 (39)	3	-	0
A11 (11)	1	1	D03 (39)	17	-	7
A12 (12)	1	1	D03 (39)	7	-	0
B01 (13)	1	1	D03 (39)	3	-	0
B02 (14)	1	1	D03 (39)	37	-	30
B03 (15)	1	1	D03 (39)	14	-	4
B04 (16)	1	1	D03 (39)	400	-	100

Figure 3.5: Revelation™ Report View

‘-’, undecided is indicated by ‘N’, oversaturation is indicated by ‘?’, and no data or low intensity data is indicated by a blank. The confidence score indicates how confident the software is that heteroduplex molecules were present in a particular well.

Also available from the report is the Mutation Call Result window (Figure 3.6). This window quickly summarizes the calls for each well via a colorized grid. In addition, by clicking on a particular capillary, its electropherogram is displayed next to the control capillary’s electropherogram. These are the electropherograms used to determine match values and mutation scores. The electropherograms are normalized, and all data outside of the normalization region is ignored. Viewing these two electropherograms allows the user to quickly determine if the call is correct.

The user can edit the automatic calls produced by the program. After the user is satisfied with the calls, the report can be output to text. Particular columns can be selected for output or rearranged to allow flexibility in the reports produced.

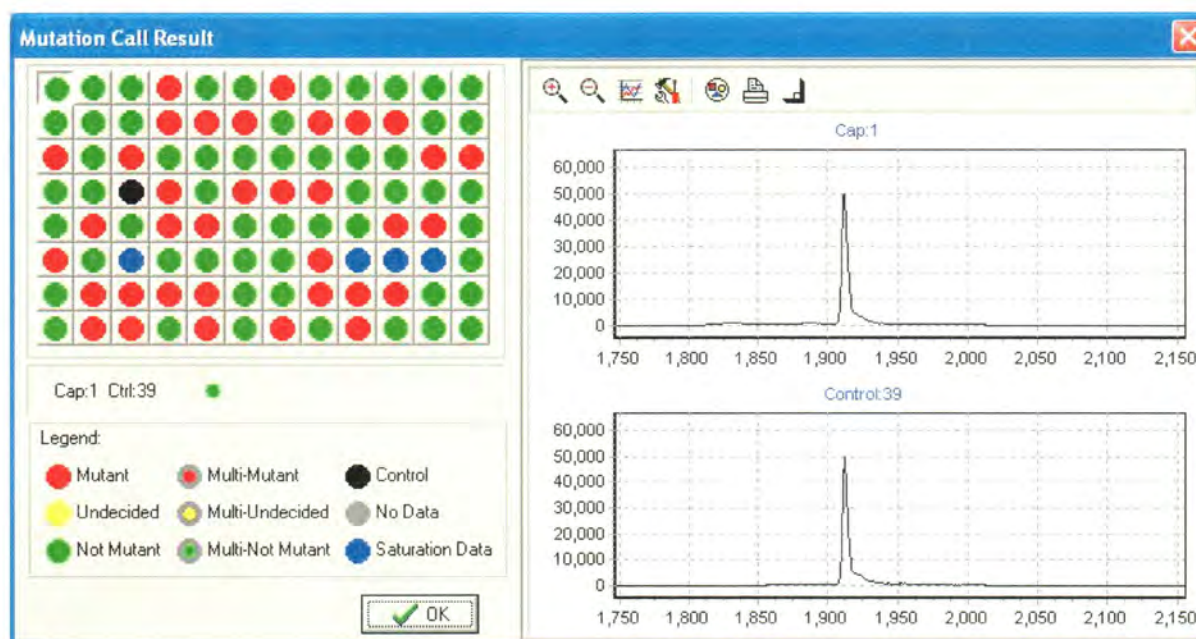


Figure 3.6: Revelation™ Mutation Call Result Window

The Revelation™ software was developed specifically to detect mutations between two DNA variants. The ability for the TGCE process to detect SNPs and small IDPs makes it equally useful for detecting genetic variation between two different alleles. However, recombinant inbred (RI) analysis is more complex than simply identifying which wells contain heteroduplex molecules.

Each RI line must be mixed with each of its progenitor lines for each genetic marker. Let these two original parent lines be line 1 and line 2. If heteroduplex molecules are discovered in the RI and line 1 mixture, then this indicates that the RI likely has genetic content inherited from line 2 at this genetic marker. However, if heteroduplex molecules are

discovered in the RI and line 2 mixture, then this indicates that the RI likely has genetic content inherited from line 1. Thus, simply marking heteroduplex molecules as a '+' has no straightforward meaning when working in this particular problem domain. The same problems arise for wells in which homoduplex molecules are discovered. The conclusions drawn from detecting homoduplex molecules in a well are different depending on with which of the original progenitor lines the RI line is mixed.

The ability to easily compare the results from both mixtures for a particular RI line would be helpful. The expectation is that no heteroduplex molecules will be present in one mixture when present in the other. If mixtures with line 1 and line 2 are both placed in wells on the same plate and run at the same time, lots of documentation would be required to make sure that interpretation of the results is done correctly. In addition, extreme care would have to be taken when choosing the controls and selecting the wells to compare to the controls. This added complexity increases the chance of user errors.

An alternative solution is to run one plate with RI lines mixed with line 1 and another plate with RI lines mixed with line 2. In this situation, only one control needs to be selected and the results can be interpreted the same way for every well. The only way to view results simultaneously in this scenario is to open two instances of the Revelation™ software so that both plates' data are observable. This is an inelegant and inefficient solution. It is possible to view calls simultaneously by outputting a report for each run and importing these results into a spreadsheet or a database. However, if the calls do not indicate that one mixture produced strictly homoduplex molecules and the other mixture produced some heteroduplex molecules, then the Revelation™ software must be re-opened for each mixture so the electropherograms for that well can be reviewed by the user in an attempt to correct the

discrepancy. In many cases, it is easier for the user to identify wells where the calls do not agree by changing the '+'s and '-'s to '1's and '2's based on the mixture for which the call was made. This means the user must carefully interpret the data.

Regardless of the method used to perform recombinant inbred analysis with the existing software, the probability of user error is increased due to the complex result interpretation, difficult data management, and extensive data manipulation issues. Elimination of these issues will not only reduce the probability of user error, but also increase the rate at which the analysis can be completed. Since the ultimate goal of detecting SNPs in RI lines is to produce a high-density genetic map, there will be many genetic markers to score. Thus, the time it takes to complete the analysis accurately is critical to reaching the goal.

As discussed, there are many different parameters that can be adjusted and data editing options that can be performed in the Revelation™ software. This flexibility allows the software to be effective in detecting the differences between homoduplex and heteroduplex electropherograms under many different circumstances. Because different homoduplex and heteroduplex molecules move through the gel at different rates, the electropherograms vary. Another factor that may affect the appearance of the electropherograms is the intensity of the DNA bands. Random "noise" may also be present in the electropherograms due to a variety of factors. Due to these variations, the default parameters in Revelation™ are more accurate in detecting mutations in some data sets than others. Adjusting the parameters or changing the data editing options that are used can increase the accuracy of the mutation calls. The challenge is to find the best data editing options and parameters for a particular set of data in an expeditious manner.

CHAPTER 4: METHODS

GRAMA (Genetic Recombinant Analysis and Mapping Assistant) is a software tool that has been developed specifically for recombinant inbred analysis. It allows the user to view all data simultaneously. This enables the user to decide quickly and accurately from which of the progenitor lines the genomic content of an RI line for a particular genetic marker is inherited (or if it is undeterminable).

Revelation™ is used to generate the raw data that the GRAMA software uses as input. The default parameters are used and the data edit options recommended by the Revelation™ software manual are performed. Through testing, this has been found to generate fairly accurate calls. This is not to say that the accuracy could not be improved by modifying parameters or using additional data editing options, but because of how recombinant inbred lines are generated, many errors that are made in the call results can be detected later. If the genetic material from an RI line forms only homoduplex molecules when mixed with one of the parent lines, it is expected to form heteroduplex molecules when mixed with the other parent line. If this is not the case with the automatic calling processes then the electropherograms of the RI can be reviewed later to determine if one of the calls was incorrect.

To prepare the data for input to GRAMA, the standard process of operating the Revelation™ software is followed. The data editing options performed include subtracting the baseline of minimal intensity and setting the interest region to the area where peaks occur in the electropherograms. In addition to generating a report, the data for the electropherograms from all of the capillaries is output. Using this information, GRAMA can

accurately reproduce the electropherograms for each capillary. There is no longer a need to edit any of the calls at this time because more convenient editing functionality is provided in GRAMA where more data is at the user's disposal to determine if and how the calls should be edited. This process is followed for each run of the TGCE system that contains DNA being analyzed for a particular genetic marker. There will typically be two or three runs in order to capture all necessary data: one run in which the RI DNA is mixed with parent 1 DNA, one run in which the RI DNA is mixed with parent 2 DNA, and another optional run containing only unmixed RI DNA.

GRAMA's opening screen (Figure 4.1) prompts the user to open the report files, also called score files, produced by Revelation™. Score files must be provided for both of the DNA mixtures. If an unmixed RI plate was run it can also be included.

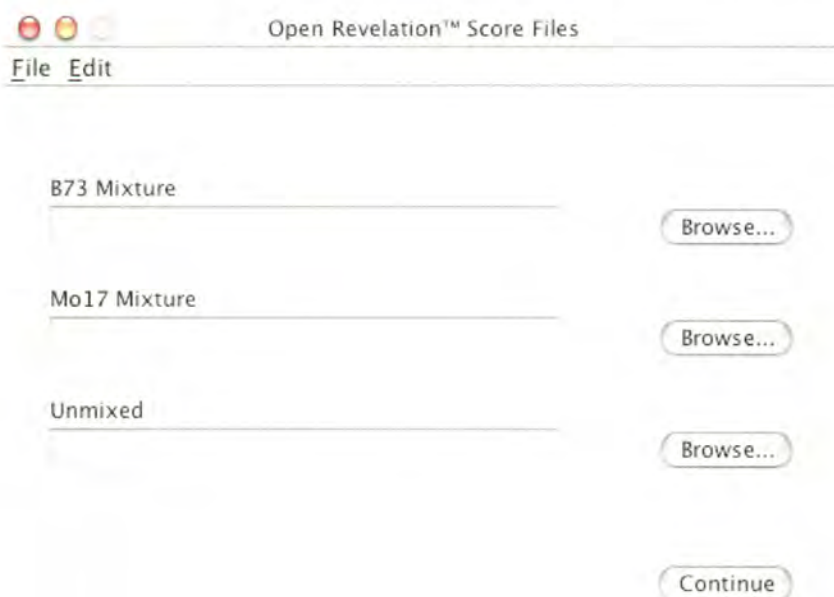


Figure 4.1: GRAMA Initial Window

When the Continue button is clicked, massive amounts of data processing takes place. Each of the score files is read, and the '+' and '-' scores are changed to '1' and '2' to represent which of the original progenitors each of the RIs resemble at a particular genetic marker. When reading the score file for the line 1 mixture, a '-' indicates that only homoduplex molecules (typically a single peak) were detected. Thus, the RI DNA was likely inherited from line 1 at the position of the genetic marker. On the other hand, a '+' indicates that heteroduplex molecules (multiple peaks) were detected. The RI DNA was therefore likely inherited from line 2. Similarly, when processing the score file for the mixture with line 2, a '-' indicates that the RI genetic material was inherited from line 2, and a '+' indicates that the RI DNA was inherited from line 1. The '+'s and '-'s are transformed to '1's and '2's according to these criterion. If a score file from an unmixed plate is included, the '+'s and '-'s for this score file are also transformed. In this case, no information is gained about from which progenitor line the RI DNA originated, so it is simply noted whether any heteroduplex molecules are formed in each well. If no heteroduplex molecules were present then a single peak appears on the electropherogram. However, if there were heteroduplex molecules, then multiple peaks appear in the electropherogram. Because of this, '-' calls are changed to 'S' for single and '+' calls are changed to 'M' for multiple. In the unmixed sample, if multiple peaks are detected, it indicates that the RI line DNA may not be homogeneous and may cause the results of the mixture plates to be inaccurate. All other calls from Revelation™, i.e., 'N' for undetermined, '?' for over saturated, and no call for no or low intensity data, are converted to a score of '0' inside GRAMA to indicate that Revelation™ was unable to determine the correct call.

In addition to the reading and translating of the score files, the electropherograms for each capillary and plate combination are analyzed at the same time. GRAMA includes its own peak detection algorithm, which will be discussed in detail in the next chapter. GRAMA uses this algorithm to analyze each electropherogram and count the number of peaks. For the unmixed plate, the well is scored as an 'S', 'M', or 'O'. For the mixed plates, the well is scored as a '1', '2', or '0'. The criterion for the scoring assignments is the same as transforming the Revelation™ calls.

Genetic Recombinant Analysis - TIDP2752 - Mon Sep 22 12:52:33 CDT 2003

Capillary	Revelation™	B73 Mixture			Well	Revelation™	Mo17 Mixture			Well	Consensus 1	Consensus 2
		GRAMA	Final				GRAMA	Final				
1	1	1	1	A01	1	1	1	1	A01	1	1	1
2	1	1	1	A02	1	1	1	1	A02	1	1	1
3	2	2	2	A03	2	2	2	2	A03	2	2	2
4	2	2	2	A04	2	2	2	2	A04	2	2	2
5	2	2	2	A05	2	2	2	2	A05	2	2	2
6	2	2	2	A06	2	2	2	2	A06	2	2	2
7	2	2	2	A07	2	2	2	2	A07	2	2	2
8	2	2	2	A08	2	2	2	2	A08	2	2	2
9	2	2	2	A09	2	2	2	2	A09	2	2	2
10	2	2	2	A10	2	2	2	2	A10	2	2	2
11	1	1	1	A11	1	1	1	1	A11	1	1	1
12	1	2	1	A12	2	2	2	2	A12	0	0	0
13	2	2	2	B01	2	2	2	2	B01	2	2	2
14	1	1	1	B02	1	1	1	1	B02	1	1	1
15	1	1	1	B03	1	1	1	1	B03	1	1	1
16	2	2	2	B04	2	2	2	2	B04	2	2	2
17	1	1	1	B05	1	1	1	1	B05	1	1	1
18	2	2	2	B06	2	2	2	2	B06	2	2	2
19	2	2	2	B07	2	2	2	2	B07	2	2	2
20	2	2	2	B08	2	2	2	2	B08	2	2	2
21	1	1	1	B09	1	1	1	1	B09	1	1	1
22	1	1	1	B10	1	1	1	1	B10	1	1	1
23	1	1	1	B11	1	1	1	1	B11	1	1	1
24	2	2	2	B12	2	2	2	2	B12	2	2	2
25	1	1	1	C01	1	1	1	1	C01	1	1	1
26	2	2	2	C02	2	2	2	2	C02	2	2	2
27	1	1	1	C03	1	1	1	1	C03	1	1	1
28	2	2	2	C04	2	2	2	2	C04	2	2	2
29	1	1	1	C05	1	1	1	1	C05	1	1	1
30	2	2	2	C06	2	2	2	2	C06	2	2	2

B73 Peak Range: 681 - 988 Notes:	Mo17 Peak Range: 678 - 948 Notes:	Overall Peak Range: 678 - 988 Notes: Missing data for capillary 61 and 62 for both mixtures
---	--	---

Finalize

Figure 4.2: GRAMA Genetic Recombinant Analysis Window – No Unmixed Plate

The Genetic Recombinant Analysis Window contains two sections (Figure 4.2 and Figure 4.3). The topmost part of the window consists of a tabular display of all of the data collected and calculated about each of the RIs for a particular genetic marker. The bottom

portion of the window contains summary information for this genetic marker and each of the plates. The title bar displays the name of the genetic marker being analyzed and the date in which the mixture with line 1 was run. It is assumed that the other plates were run around the same time. This information is obtained from the Revelation™ score files.

The contents of the Genetic Recombinant Analysis Window vary slightly depending on if the unmixed data is included. This difference can be seen by comparing Figure 4.2 and Figure 4.3. Figure 4.3 contains data from an unmixed plate and has four additional columns in the table and an additional summary section in the bottom portion of the window.

Genetic Recombinant Analysis – TIDP2752 – Mon Sep 22 12:52:33 CDT 2003														
Capillary	Unmixed				B73 Mixture				Mo17 Mixture				Consensus 1	Consensus 2
	Revelation™	GRAMA	Final	Well	Revelation™	GRAMA	Final	Well	Revelation™	GRAMA	Final	Well		
1	S	S	S	A01	1	1	1	A01	1	1	1	A01	1	1
2	S	S	S	A02	1	1	1	A02	1	1	1	A02	1	1
3	S	S	S	A03	2	2	2	A03	2	2	2	A03	2	2
4	S	S	S	A04	2	2	2	A04	2	2	2	A04	2	2
5	S	S	S	A05	2	2	2	A05	2	2	2	A05	2	2
6	S	S	S	A06	2	2	2	A06	2	2	2	A06	2	2
7	S	S	S	A07	2	2	2	A07	2	2	2	A07	2	2
8	S	S	S	A08	2	2	2	A08	2	2	2	A08	2	2
9	S	S	S	A09	2	2	2	A09	2	2	2	A09	2	2
10	S	S	S	A10	2	2	2	A10	2	2	2	A10	2	2
11	S	S	S	A11	1	1	1	A11	1	1	1	A11	1	1
12	S	S	S	A12	1	2	1	A12	2	2	2	A12	0	0
13	S	S	S	B01	2	2	2	B01	2	2	2	B01	2	2
14	S	S	S	B02	1	1	1	B02	1	1	1	B02	1	1
15	S	S	S	B03	1	1	1	B03	1	1	1	B03	1	1
16	S	S	S	B04	2	2	2	B04	2	2	2	B04	2	2
17	S	S	S	B05	1	1	1	B05	1	1	1	B05	1	1
18	S	S	S	B06	2	2	2	B06	2	2	2	B06	2	2
19	S	S	S	B07	2	2	2	B07	2	2	2	B07	2	2
20	S	S	S	B08	2	2	2	B08	2	2	2	B08	2	2
21	S	S	S	B09	1	1	1	B09	1	1	1	B09	1	1
22	S	M	S	B10	1	1	1	B10	1	1	1	B10	1	1
23	S	S	S	B11	1	1	1	B11	1	1	1	B11	1	1
24	S	S	S	B12	2	2	2	B12	2	2	2	B12	2	2
25	S	S	S	C01	1	1	1	C01	1	1	1	C01	1	1
26	S	S	S	C02	2	2	2	C02	2	2	2	C02	2	2
27	S	S	S	C03	1	1	1	C03	1	1	1	C03	1	1
28	S	S	S	C04	2	2	2	C04	2	2	2	C04	2	2
29	S	S	S	C05	1	1	1	C05	1	1	1	C05	1	1
30	S	S	S	C06	2	2	2	C06	2	2	2	C06	2	2
Unmixed				B73				Mo17				Overall		
Peak Range: 673 – 801				Peak Range: 681 – 988				Peak Range: 678 – 948				Peak Range: 673 – 988		
Notes:				Notes:				Notes:				Notes:		
				Strange peak pattern for capillary 43								Missing data for several capillaries		
Finalize														

columns are grouped. There is one column group for each of the plates containing DNA mixtures. An additional column group is present if unmixed data is also included. Each of these column groups contains four subcolumns. The first column is the Revelation™ score column. This reports the Revelation™ score as translated from the score file provided. The second column in a group is the GRAMA score column. The GRAMA score column reports the score that GRAMA calculated by using its own internal peak determination algorithm. A group's third column reports the final score determined for the RI on a particular plate. The final score can initially be set to default to the Revelation™ score or the GRAMA score, but the user can edit this column to correct a score if necessary. The fourth and final column in each group is the Well column. Clicking on any cell in this column yields a graph where the user can review the electropherogram for the RI on the corresponding plate. Immediately following all column groups are two more columns that contain consensus calls. Consensus calls are calls automatically generated from the final scores of each plate. More about how they are calculated and their interpretation will be given in Chapter 6. In general, the Consensus 1 column contains more descriptive consensus scores that provide more insight into the final scores that produced them. The Consensus 2 column contains more general scores that are intended to be used as the input to a genetic mapping program. The scores that the Consensus 2 column reports indicate from which inbred line the genetic content for the RI was most likely originated.

The bottom portion of the Genetic Recombinant Analysis window contains summary information. Summary information is provided for each of the plates. In addition, an overall summary section for the genetic marker is included. The peak range that is reported is calculated by finding the beginning of the leftmost peak and ending of the rightmost peak

across all of the capillaries of a particular plate. The overall peak range is simply the beginning of the leftmost and ending of the rightmost peak over all of the capillaries and all of the plates. A small margin is also added to each end of the range. This will provide the user important information regarding how DNA moves through the gel when size multiplexing is being considered. The summary section also contains a textfield for the user to enter notes about each of the plates as well as an overall note about all of the plates run for the genetic marker.

More information can be gained from the table provided in GRAMA than is initially displayed. Its cursor changes when above a portion of the table in which further information can be gained. As previously mentioned, when a user clicks on any cell in the Well columns, an electropherogram will appear. An example of this window appears in Figure 4.4. Its title bar reports the genetic marker, mixture plate, well, and corresponding capillary of the electropherogram being viewed. Red and blue lines appear on the electropherogram with the red lines indicating the beginning of a GRAMA detected peak and the blue lines indicating the end of a peak. These indicators allow the user to see clearly how GRAMA arrived at its scoring decision. The vertical scale of the graph is automatically adjusted based on the size of the tallest peak. This gives the user the ability to focus on the overall shape of the graph and not on the local intensity. The graph is also automatically adjusted horizontally to zoom into the region where peaks were identified. This presents a clear view of the area that is of most importance for decision-making. The coordinate labels on both axes are automatically adjusted to represent the area being viewed.

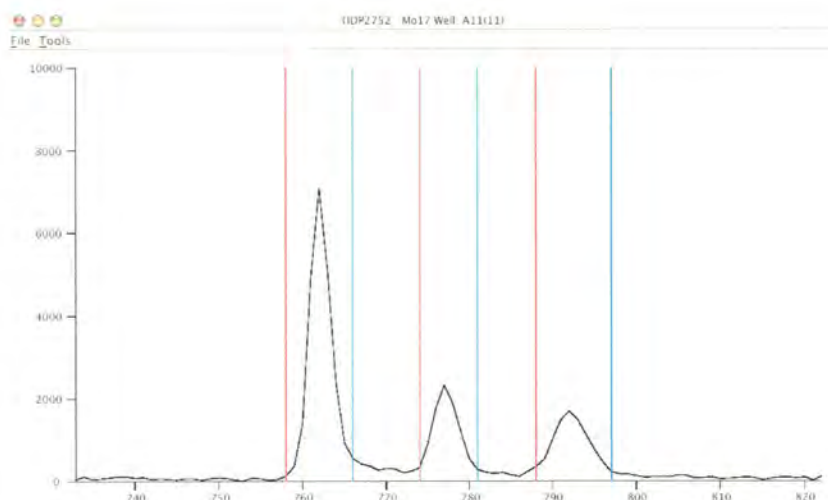


Figure 4.4: GRAMA Single Well View

The horizontal viewing area can be adjusted by using the mouse to draw a box in the graphing area. If the mouse is released to the right of where the initial mouse click is made the graph will zoom into the specified area. The initial mouse click will be the new leftmost point on the graph and the mouse release will be the new rightmost point on the graph. If, however, the mouse release point is to the left of the initial mouse click point the graph will zoom out an amount proportional to the length of the box. This zooming style emulates the behavior of Revelation™'s software and allows the user to move seamlessly between the two applications without having to consider which application they are using.

The Tools Menu in this window contains two options. The first is “Zoom to Interest Region.” This focuses the graph to the region where all of the detected peaks are located. This is the region that the graph initially displays when the window is opened. The second option is “Hide/Show Peak Indicators.” This toggles the presence of the red and blue indicator lines on and off as desired.

By clicking on any cell in the capillary column, the user can obtain another view (Figure 4.5 and Figure 4.6). This view allows the user to view the electropherograms of a particular capillary for all of the plates simultaneously. The absence of this feature is a major shortcoming of the Revelation™ software for use in genetic mapping experiments. In Figure 4.6, the unmixed plate is also included when the information for this run is available. The title bar of this window includes the genetic marker name, the well identifier, and the capillary number. Each electropherogram is labeled to indicate from which plate it is originated. Each of the graphs can be adjusted as in the single electropherogram view, but the graphs initially display the region of interest selected in Revelation™ as opposed to the region where GRAMA detects peaks. Therefore, the region is much larger because it typically has to be wide enough to include the peaks from each capillary. Zooming for each graph is handled separately so that each graph can be adjusted to focus on the desired region. The Tools Menu again contains the “Zoom to Interest Region” option and the “Hide/Show Peak Indicators” option, but when one of these options is selected, the action is applied to all of the electropherograms. Here, “Zoom to Interest Region” focuses each graph to the area where GRAMA detected peaks. By viewing all of these electropherograms at the same time, the user can quickly determine the best score for the RI at a particular genetic marker.

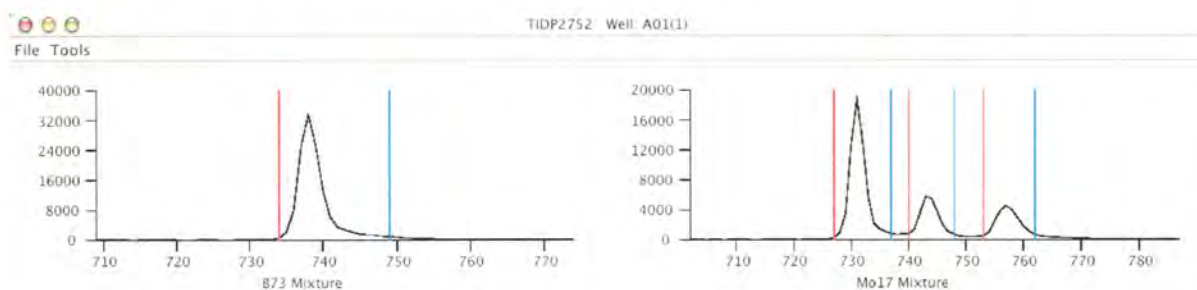


Figure 4.5: GRAMA Across Mixtures View – No Unmixed Plate

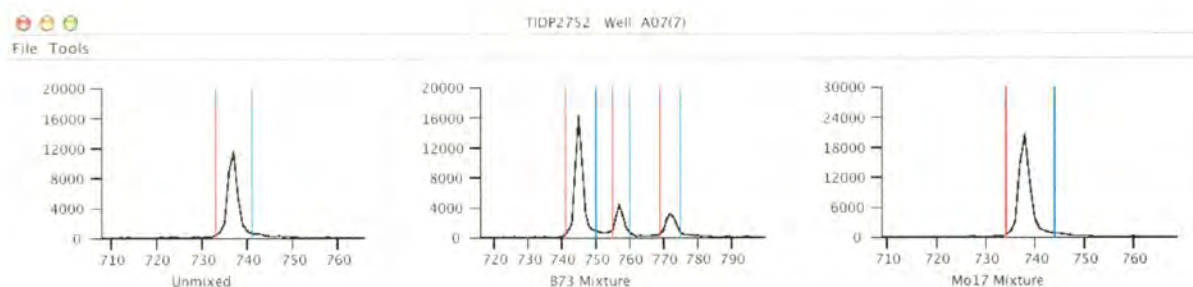


Figure 4.6: GRAMA Across Mixtures View – With Unmixed Plate

By clicking on the Well column heading for any of the plates, a new window (Figure 4.7) appears displaying each capillary's electropherogram for that plate. Its title bar informs the user of the genetic marker's name as well as from which plate the currently viewable electropherograms originated. Each electropherogram is labeled with its well identifier and capillary number. A scrollbar is provided so that all electropherograms for the plate can be quickly viewed and accessed. For the most part, these graphs function as those previously described, but in this window, all graphs display the same region. The region displayed is the area in which GRAMA detected peaks across all of the capillaries. Therefore, this may be a wider region than would be expected by opening each electropherogram individually. The zooming functionality works slightly differently for this view as well. While zooming was independent of the electropherograms when viewed separately, in this window, zooming for all electropherograms is linked.

The "Zoom to Interest Region" and "Hide/Show Peak Indicators" options also operate on all electropherograms simultaneously. "Zoom to Interest Region" resets all electropherograms to the region where GRAMA detected peaks across all capillaries. This electropherogram view allows the user to identify the peak patterns produced by a particular

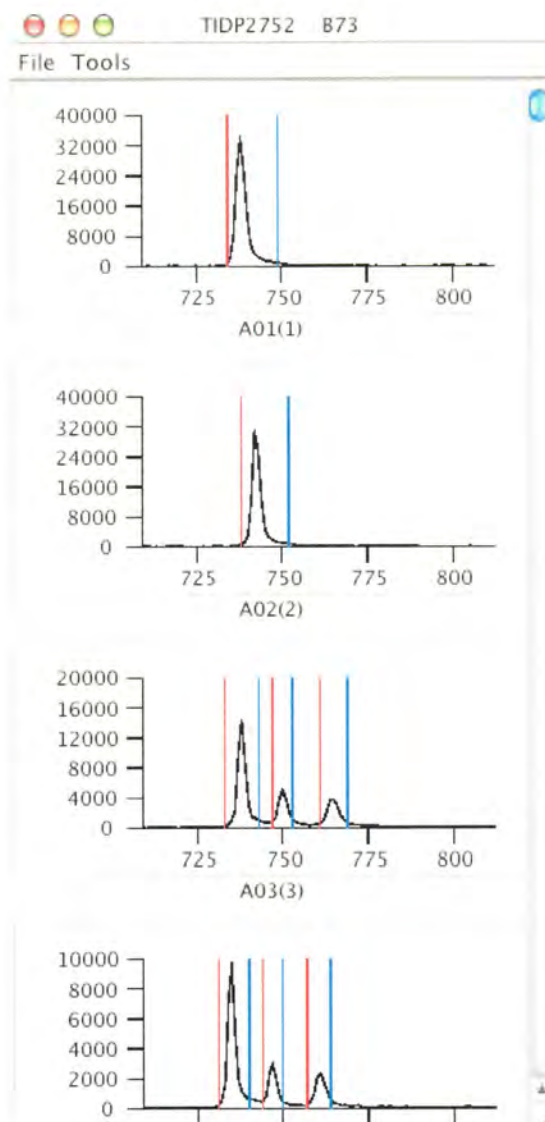


Figure 4.7: GRAMA Across Wells View

genetic marker. If one electropherogram seems to have multiple peaks but does not match the multiple peak pattern found in the other electropherograms, the user may not wish to score the RI the same way. This window gives the user the ability to easily observe such occurrences.

The remaining user action that can be performed on the table is the ability to change the values in the final call column for each plate. By double clicking on the cell which

contains the value to be edited, the cell changes to the edit mode. At this point, the user is restricted to entering only legal values. For example, only 'S', 'M', or '0' can be entered for the unmixed plate and only '1', '2', '0' can be entered for the other plates. If the final call cell is edited, the consensus calls are automatically updated to reflect the change.

Another important aspect of the table is its color-coding of the rows (Figure 4.8). Each row can be in one of four colors: green, yellow, red, or blue. Green indicates that GRAMA and RevelationTM agree on the score for each plate for the RI, that the final scores for both mixture plates match, and that, if the unmixed plate data is included, its final score is an 'S'. In short, this means that it is highly probable that the RI genetic material inherited at the genetic marker is from the original line indicated by the score of both mixture plates. A row is colored yellow if the initial final scores for both mixture plates agree, but GRAMA disagrees with the RevelationTM score for any of the plates analyzed. This tells the user that the electropherograms should be reviewed to ensure that the correct final call is made. A row that is colored red indicates that either the final score from the line 1 mixture plate does not match the final score from the line 2 mixture plate, that the unmixed plate's final score is an 'M', or that the final score for one of the plates is 0. This indicates to the user that their intervention is required in order for the RI to be scored correctly at this genetic marker. A row is color-coded blue once the user edits any final score for the row. By so doing, the user can easily identify which rows have already been analyzed and corrected.

Once the user is satisfied that all of the correct scores have been entered in all of the final score columns, the Finalize button in the lower right corner of the Recombinant Inbred Analysis window should be clicked. This outputs all data for this particular genetic marker

Genetic Recombinant Analysis - TIDP2752 - Mon Sep 22 12:52:33 CDT 2003

Capillary	Revelation™	Unmixed	Final	Well	Revelation™	B73 Mixture	Final	Well	Revelation™	Mo17 Mixture	Final	Well	Consensus 1	Consensus 2
56	S	S	S	E08	2	2	2	E08	2	2	2	E08	2	2
57	S	S	S	E09	2	2	2	E09	2	2	2	E09	2	2
58	S	S	S	E10	1	1	1	E10	1	1	1	E10	1	1
59	S	S	S	E11	2	2	2	E11	2	2	2	E11	2	2
60	S	S	S	E12	2	2	2	E12	2	2	2	E12	2	2
61	0	0	0	F01	0	0	0	F01	0	0	0	F01	30	0
62	S	0	S	F02	0	0	0	F02	0	0	0	F02	28	0
63	S	S	S	F03	2	2	2	F03	2	2	2	F03	2	2
64	M	S	M	F04	1	1	1	F04	1	1	1	F04	12	1
65	S	M	S	F05	2	2	2	F05	2	1	2	F05	2	2
66	S	S	S	F06	1	1	1	F06	1	1	1	F06	1	1
67	S	S	S	F07	1	1	1	F07	1	1	1	F07	1	1
68	S	S	S	F08	2	2	2	F08	2	2	2	F08	2	2
69	S	S	S	F09	2	2	2	F09	2	2	2	F09	2	2
70	S	S	S	F10	1	1	1	F10	1	1	1	F10	1	1
71	S	S	S	F11	1	1	1	F11	1	1	1	F11	1	1
72	S	S	S	F12	2	2	2	F12	2	2	2	F12	2	2
73	S	S	S	G01	1	1	1	G01	1	1	1	G01	1	1
74	S	S	S	G02	1	1	1	G02	2	1	1	G02	1	1
75	0	S	S	G03	1	1	1	G03	1	1	1	G03	1	1
76	S	S	S	G04	2	2	2	G04	2	2	2	G04	2	2
77	0	0	0	G05	1	1	1	G05	2	2	2	G05	21	0
78	S	S	S	G06	2	2	2	G06	2	2	2	G06	2	2
79	S	S	S	G07	2	2	2	G07	2	2	2	G07	2	2
80	S	S	S	G08	1	1	1	G08	1	1	1	G08	1	1
81	S	S	S	G09	1	1	1	G09	1	1	1	G09	1	1
82	S	S	S	G10	1	2	1	G10	2	2	2	G10	0	0
83	S	S	S	G11	2	2	2	G11	2	2	2	G11	2	2
84	S	S	S	G12	2	2	2	G12	2	2	2	G12	2	2
85	S	S	S	H01	1	1	1	H01	1	1	1	H01	1	1

Unmixed
Peak Range: 673 - 801
Notes:

B73
Peak Range: 681 - 988
Notes:

Mo17
Peak Range: 678 - 948
Notes:

Overall
Peak Range: 673 - 988
Notes:

Finalize

Figure 4.8: GRAMA Genetic Recombinant Analysis Window Displaying Colored Rows

to a tab-delimited text file. This allows easy importation of the data into a spreadsheet or a database. From there, results from multiple genetic markers can be collected and formatted for input to a genetic mapping program.

The use of two algorithms to score each electropherogram so that their results can be compared is very important for improved correctness. The correctness is further enhanced by the fact that the final scores from both mixture plates agree because of the design of the experiment. If a row is colored green, the user can safely assume with a high probability that this consensus score is correct and expeditiously move on to investigate other rows, because it means that two different algorithms have agreed on two different calls, or three if an unmixed plate is also included. It also means that both of the algorithms, via analysis of the

electropherograms, detected only homoduplex molecules when the RI was mixed with one of the lines and the presence of heteroduplex molecules when mixed with the other line. Thus, the scores for both mixture plates are identical and are scored either as a '1' or a '2'. The consensus 2 score will therefore also be scored as a '1' or a '2' reflecting the final score agreed upon for both mixture plates. The only way the row can be colored green yet the consensus score is incorrect would be when both final scores are '1's when they should have been '2's or vice versa. This indicates that both algorithms have to score both electropherograms incorrectly at the same time, which is an unlikely occurrence for the following reasons.

Both algorithms are very unlikely to incorrectly identify multiple peaks as a single peak. SpectruMedix®'s website claims that this type of error rate is less than 5% with its algorithm [24]. GRAMA is also very unlikely to make this type of error, which will be explained in Chapter 7. Thus, if a single peak is detected by either of the algorithms, it is highly probable that this is correct. If a row is colored green, both algorithms must have detected a single peak for one of the mixture plates. The user can therefore assume with a high confidence that the consensus code generated for green-colored rows is correct. This significantly reduces the amount of manual inspection the user must perform. In addition, combining both algorithms allows data whose scoring is questionable to be easily identified (yellow or red color-coding). Subsequently, the user can use this information to check the relevant electropherograms and make sure the correct score has been determined. To summarize, combining both algorithms not only helps identify scoring errors, but it also increases the overall user efficiency by limiting attention to the yellow and red rows.

CHAPTER 5: PEAK IDENTIFICATION

Humans can easily identify peaks when observing a graph. However, for a computer to automatically detect peaks without user intervention, generic rules must be established that work for the majority of situations. The rules that have been adopted for the peak identification algorithm in GRAMA are based on changes in the slope of a graph from point to point.

Although the electropherograms in GRAMA appear to be continuous on screen, the data used to create the electropherogram actually consists of a set of discrete points. The electropherogram graph is created by connecting these discrete points. Each electropherogram has a baseline from which the peaks arise. This baseline is not completely flat, in fact, it can be quite uneven because of noise and other factors. Yet, the fact that in general a baseline exists is sufficient for the GRAMA algorithm to operate properly. Moving from left to right along the electropherogram, if the beginning of a standard peak is encountered, the slope of the graph between consecutive pairs of discrete points will begin to increase. The region of the graph in which the slope between consecutive points is increasing is said to be concave upward. At some point while continuing from left to right, the slope between consecutive points will cease to increase and begin to decrease. The region of the graph where this occurs is said to be concave downward. The maximum point of the peak occurs in this area barring imperfections in the shape of the peak. Continuing from left to right, the slope between consecutive points will begin to increase again resulting in another concave upward region. These regions can be seen in Figure 5.1. By this point the maximum value of the peak has already been encountered, so this occurs on the

downward slope of the peak. This increase in slope must occur so the curvature of the peak can eventually reconverge with the baseline. The points where the change in slope switches from increasing to decreasing or from decreasing to increasing are called inflection points. These are indicated by the red and blue lines on the electropherogram example in Figure 5.2.

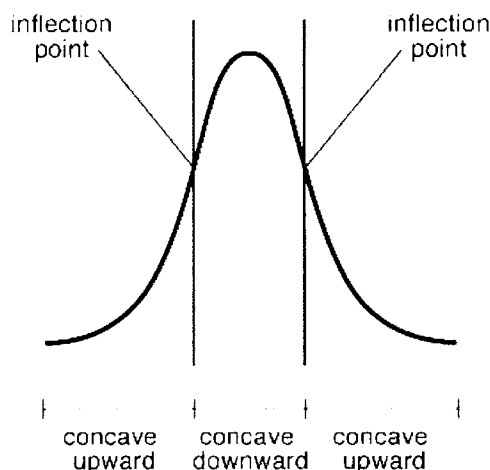


Figure 5.1: Concavity of a Peak

The initial step of the GRAMA algorithm is to locate all of these inflection points. The algorithm searches the graph from left to right checking the slopes between consecutive pairs of points. It first looks for the point where the graph switches from being concave upward to being concave downward. GRAMA marks this as the beginning of a peak. Then it searches for the point where the graph switches from being concave downward to being concave upward and marks this as the end of a peak.

The idea is that between every point marked as the beginning of a peak and the following point marked as the end of a peak the maximum point for a peak should occur. Because of this, GRAMA stores the maximum point in each of these intervals. The

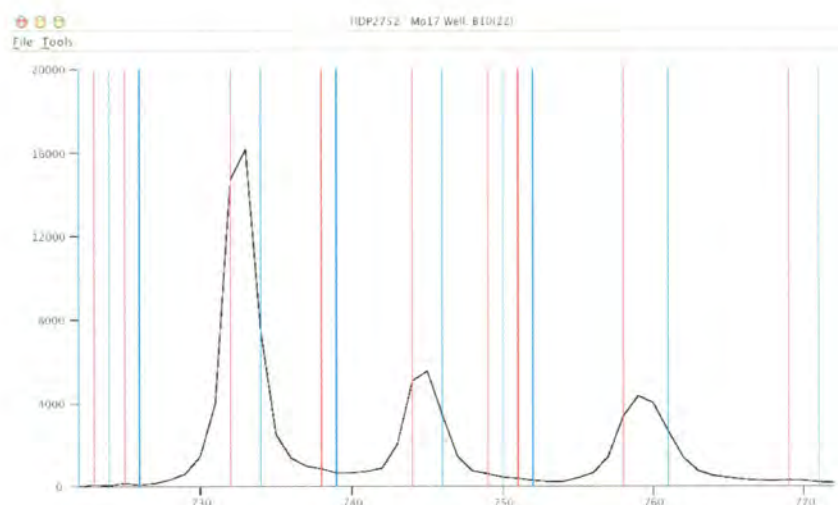


Figure 5.2: Inflection Points in Electropherogram

maximum point for a peak does not always occur in these intervals, however. Bumps or ridges on the surface of the peak will cause changes in concavity, yielding points that are incorrectly marked as beginnings and endings of peaks. An example of this is shown in Figure 5.3. To combat this problem GRAMA “slides” the beginning and ending markers along the graph. The beginning markers are slid to the left and the ending markers are slid to the right. The slope on the left side of the peak will be positive so the left marker is allowed to slide left until it encounters a sufficiently large negative slope as long as the value of the graph at that point is above a certain threshold. By allowing the marker to continue sliding left even after encountering a small negative slope, minor bumps on the peak are ignored. After the marker reaches a point on the side of the peak where the graph value is less than the threshold value it continues to move left until the slope is near 0. After this process, the beginning peak marker should be in a location that is the very beginning of the peak. The ending peak marker is adjusted via a similar process except that, because the slope on the

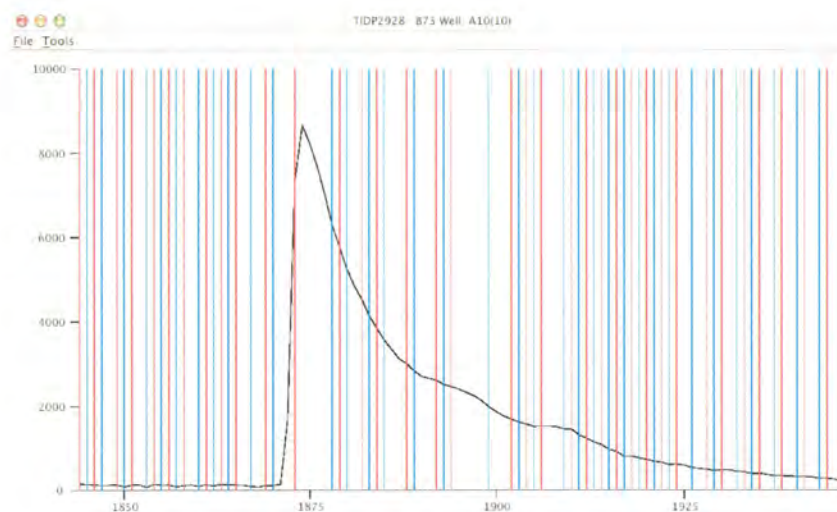


Figure 5.3: Multiple Peak Markers for Single Peak

right side of the peak is negative, it is allowed to continue sliding after encountering a small positive slope so that minor bumps on the right side of the peak are ignored. When the inflection points are flanking a bump or ridge on the peak, either the beginning or ending marker will not be able to slide far as it will be sliding toward the maximum point of the peak.

Once the sliding has completed, the GRAMA peak detection algorithm checks whether the peak flanked by each pair of markers is tall enough to be considered relevant. The difference in height is calculated between the maximum point of the peak in the interval between the beginning and ending peak markers and the height of the graph at the beginning peak marker. The difference in height is also measured between the maximum point of the peak and the height of the graph at the ending peak marker. If both of these distances are greater than a specified percentage of the tallest peak in the graph, then the peak is considered relevant.

As a final step, GRAMA analyzes the interval between each beginning peak marker and ending peak marker to see if another beginning and ending peak marker are contained within. As aforementioned, the markers flanking bumps and ridges are not allowed to slide along both sides of the peak, but the peak markers that actually flank the maximum point of the peak will slide along both sides. The end result is that the markers flanking the bump or ridge are contained within the interval defined by the beginning and ending peak marker that flanked the maximum of the peak. When the algorithm detects peak markers contained within an interval defined by another set of peak markers it simply removes the internal set of markers from the overall set of relevant peaks. The electropherogram from Figure 5.3 is again shown in Figure 5.4 after the sliding has taken place and irrelevant and interior peak markers have been removed. The effectiveness of these three steps is quite obvious from this example. In practice the GRAMA peak detection algorithm proves to be quite successful. The experimental results will be reported in Chapter 7.

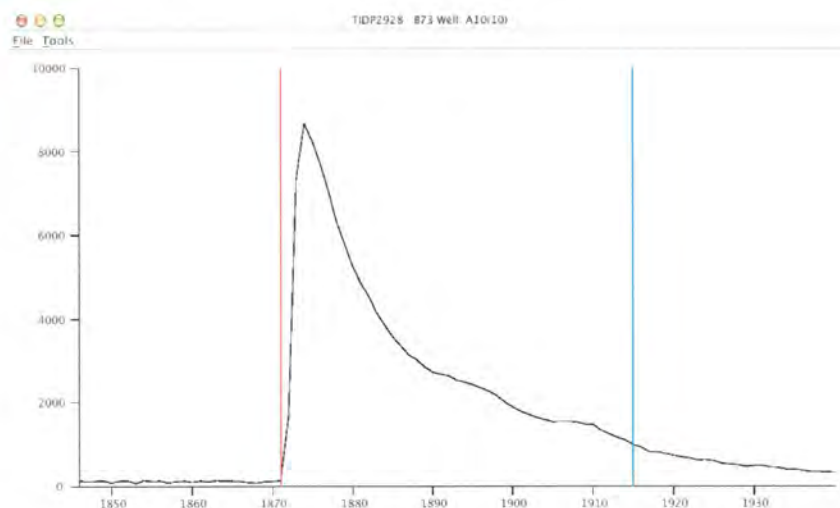


Figure 5.4: Final Peak Markers Designating Peak

The primary weakness of the GRAMA peak detection algorithm is in detecting homoduplex molecules under low data intensity. When this occurs, many peaks that are simply noise may be classified as relevant and counted toward the final total. Noise might also affect high intensity data in some cases. If a peak generated from “noise” is considered relevant for a well actually containing only homoduplex molecules, then the well could be scored incorrectly. However, the reason for the “noise” should actually be investigated by the user. As a result, rows containing these false peaks will most likely not be color-coded green, and the user will have a chance to take a closer look at the electropherograms and identify any serious problems that may potentially exist.

CHAPTER 6: CONSENSUS CALLS

The consensus calls (scores) are based on the final scores of each plate for a particular RI. There are two consensus calls for each RI. Consensus 1 calls are unique for nearly every possible combination of final scores. Thus, when armed with a consensus 1 call and a lookup table, the final score combination that produced a particular consensus 1 call can be determined. Consensus 2 calls are more generic as the same score can be produced from several different combinations of final scores. Consensus 2 calls represent the decision on which of the parent lines the genetic content of a RI at the genetic marker most likely originated. In short, Consensus 1 calls encode all obtainable experimental results, while Consensus 2 calls summarize the results for direct input to a genetic mapping program. Table 6.1 details how the two consensus scores are generated.

As can be seen from the table, consensus 2 calls are '1' or '2' only if the final scores from both of the mixture plates agree. This indicates that there is sufficient evidence that the RIs DNA in the marker region did indeed originate from a particular line. A consensus 2 call of '0' means that the progenitor line of the DNA in this region of the RI cannot be determined given the current available data.

Consensus 1 calls provide the user with more information about the exact circumstances that resulted in the corresponding consensus 2 call. There may be occasions when this extra information is useful. Sometimes one of the final scores cannot be determined because of no data or low intensity data, but it may be acceptable to use the result obtained from just one mixture as a mapping score. In another scenario, the consensus 2 call for a particular capillary may always be 0. There are a number of different situations in which this can happen, but, if the consensus 1 call is also always 30 for this capillary, a possible inference

may be that the capillary is no longer working correctly. In both examples, as well as in many others situations not mentioned here, consensus 1 calls provide a level of detailed information that cannot be gathered otherwise.

Table 6.1: Consensus Call Determination Table - NT Indicates Not Tested

Unmixed	Line 1 Mix	Line 2 Mix	Consensus 1	Consensus 2
S	1	1	1	1
S	1	2	0	0
S	1	0	8	0
S	2	2	2	2
S	2	1	0	0
S	2	0	9	0
S	0	1	10	0
S	0	2	11	0
S	0	0	28	0
M	1	1	12	1
M	1	2	13	0
M	1	0	14	0
M	2	2	15	2
M	2	1	16	0
M	2	0	17	0
M	0	1	18	0
M	0	2	19	0
M	0	0	29	0
0	1	1	20	1
0	1	2	21	0
0	1	0	22	0
0	2	2	23	2
0	2	1	24	0
0	2	0	25	0
0	0	1	26	0
0	0	2	27	0
0	0	0	30	0
NT	1	1	1	1
NT	1	2	0	0
NT	1	0	1b	0
NT	2	2	2	2
NT	2	1	0	0
NT	2	0	2b	0
NT	0	1	1a	0
NT	0	2	2a	0
NT	0	0	ab	0

CHAPTER 7: EXPERIMENTS

In this chapter, the results of several experiments performed to investigate the accuracy of the GRAMA and Revelation™ peak determination algorithms are discussed. The accuracy based on the results of using both algorithms at the same time is also reviewed. Throughout this chapter, the implications of these findings also will be highlighted.

Hundreds of insertion-deletion polymorphisms (IDPs) that distinguish two inbred lines of maize (B73 and Mo17) and can be detected via the Temperature Gradient Capillary Electrophoresis (TGCE) process have been discovered by the Schnable Plant Genomics Laboratory at Iowa State University. These genetic markers were used to amplify recombinant inbred (RI) lines from an intermated B73xMo17 (IBM) population developed by Mike Lee and his co-workers [14]. Using B73 as line 1 and Mo17 as line 2 the genetic recombinant inbred analysis process previously described has been applied by the Schnable Plant Genomics Team funded by National Science Foundation award DBI 0321711. This includes running the resulting data from the TGCE process through both the Revelation™ and GRAMA software packages. All of the data output from GRAMA has been placed in a database from which statistics can be gathered.

For every plate run through the TGCE process, the following operations are performed on the resulting data in Revelation™. Once the capillaries have been aligned by the user and a *.smr file created, the baseline is subtracted via the minimal intensity option. The interest region is then selected. After this, the *.smd file is created and the electropherogram data is output to text for use by GRAMA. The control well is then selected and a report is generated. The report is immediately output to text without any user

modification so that Revelation™ scoring information can be used by GRAMA. All other parameters are left at their default values. Using GRAMA, members of the Schnable Plant Genomics Lab conduct recombinant inbred analysis of the data. The results are being stored in a database.

So far, 529 different genetic markers have been analyzed via this process. From these 529 markers, the statistical calculations will be derived. Since each plate contains 96 wells, 529 markers provide 50,784 mapping score results. For 37 of the 529 genetic markers, an unmixed plate was also run and its results included. So for 37 genetic markers, 3 plates were run and for the remaining 492 markers only 2 plates were run. Because each plate contains 96 wells, a total of 105,120 electropherograms were evaluated by both programs.

The first property analyzed is each program's ability to distinguish between wells containing homoduplex and heteroduplex molecules. It is assumed for these statistical calculations that the final score determined by the user is correct. In this analysis, if a '0' is entered as a final score, it is of no interest because it means that the user was unable to determine if the well contained only homoduplex molecules or a combination of homoduplex and heteroduplex molecules. Therefore, a computer algorithm is not expected to score the well correctly either.

Revelation™ incorrectly scored a well that contained only homoduplex molecules as a well containing heteroduplex molecules 2,162 times. A total of 51,739 wells were scored as containing only homoduplex molecules by the user for those wells that Revelation™ scored. Thus, Revelation™ had a 4.18% false positive error rate on this sample. GRAMA incorrectly scored a well containing only homoduplex molecules as containing heteroduplex molecules 3,180 times. For those wells where GRAMA attempted to make a

call, the user scored the well as containing only homoduplex molecules 53,989 times. Thus, GRAMA has a 5.89% false positive error rate. The difference in the total number of wells containing homoduplex molecules as scored by the two different programs is due to the fact that Revelation™ has the option to categorize a well as undeterminable. Thus, there were 2,250 times where GRAMA attempted to make a call but Revelation™ did not. The accuracy of GRAMA's peak determination algorithm on wells uncalled by Revelation™ will be analyzed later in this chapter.

Revelation™ has a 1.28% false negative rate on the sample set. False negatives occur when the peak determination algorithm fails to identify heteroduplex molecules and scores the well as containing only homoduplex molecules. There were 621 wells out of 48,698 that were scored as containing heteroduplex molecules by the users that Revelation™ scored incorrectly. GRAMA's false negative rate was very similar at 1.29%. A total of 48,877 wells which GRAMA attempted to score were classified as containing homoduplex molecules by the user. GRAMA incorrectly classified 630 wells as containing only homoduplex molecules which the user had classified these wells as containing heteroduplex molecules. Again, the difference in the total number of wells that were scored between the programs results from the fact that Revelation™ scores some wells as undeterminable while GRAMA attempts to score all wells.

Table 7.1: Revelation™ and GRAMA Error Rates

Algorithm	Heteroduplex Calls		Homoduplex Calls	
	Revelation™	GRAMA	Revelation™	GRAMA
Total Calls	51,739	53,989	48,698	48,877
Incorrect Calls	2,162 (4.18%)	3,180 (5.89%)	621 (1.28%)	630 (1.29%)

As aforementioned, Revelation™ at times does not make a definitive call because of a detected saturation or because the calculated match factor falls in the default undeterminable match factor range. On the other hand, when provided with data that is not all at low intensity, GRAMA still tries to make a correct call for the well. Out of the 2,458 instances where Revelation™ did not make a call and GRAMA did, GRAMA made the correct call 2,376 times. Therefore, GRAMA has a 96.7% accuracy rate on wells that Revelation™ opted not to score.

Since GRAMA reports results from both its own algorithm and the Revelation™ algorithm, the user can be alerted to a possible mistake in the call made by either algorithm if the other algorithm is able to make the correct determination. In other words, if either algorithm makes a mistake, their results will disagree, and the user is then alerted so that they can take a closer look. Next we discuss how accurate the algorithms are when their results are combined together to detect mistakes.

Again, wells that have a score of '0' will not be considered since this indicates that the user was unable to determine the correct score from the electropherogram. Cases where either Revelation™ or GRAMA had a score of '0' will not be considered either, since these cases will be flagged for a closer look by the user anyhow. There are a total of 100,408 wells in the sample set that fit these criteria. Out of these 100,408 wells GRAMA and Revelation™ both scored wells correctly on 94,561 occasions or about 94.2% of the time. For 3,077 of these wells Revelation™ was correct and GRAMA was incorrect. For 2,119 wells, GRAMA was correct and Revelation™ was incorrect. Thus, the two algorithms disagreed on about 5.17% of the wells. On the wells where they disagreed, Revelation™ made the correct call 59.2% of the time, while GRAMA was correct for the other 40.8% of

the wells. The number of wells where both algorithms made the same mistake is 651. This is only 0.65% of the total number of wells for this study. Therefore, by using both algorithms, GRAMA is able to alert the user to 76.5% of the mistakes that would have been made had only Revelation™ been used, and 82.5% of the mistakes that would have been made had GRAMA used only its own algorithm.

From the above results, it can be seen that both algorithms are highly accurate in distinguishing between wells containing only homoduplex molecules and wells containing both homoduplex and heteroduplex molecules and, in turn, alerting the user of possible mistakes made by one of the algorithms. Because of the very nature of recombinant inbred analysis, by comparing the results from multiple plates together, the ability to detect mistakes and flag them for the user are further improved. For a pair of mixture plates, the expectation is to only find homoduplex molecules in one plate and both heteroduplex and homoduplex molecules in the other for a particular RI. As discussed earlier, the scores are translated so that they reflect from which of the progenitor lines the genetic material of the RI in the region of the genetic marker most likely originated. Thus, the scores for both of the mixtures should either match or the user should be alerted. Since one of the mixture plates should always contain strictly homoduplex molecules and both algorithms have very low false positive error rates on this, it is highly likely that any mistake made will be discovered and the well will be flagged for further observation by the user.

This process of increasing accuracy by comparing results from multiple plates could be done by using just one of the algorithms, but if the algorithm were to make mistakes scoring both mixtures, the mistakes would not be detected. By harnessing the power of the two algorithms to catch each other's mistakes, coupled with comparing results from both

mixtures, nearly all mistakes in scoring can be detected and referred to the user for further evaluation. In the following the number of mistakes that would have gone undetected by using each algorithm, even after comparing results from both mixtures, as well as the number of mistakes that were undetected by using both algorithms together are discussed.

The sample contains 50,784 mapping scores. Of these Revelation™ made a mistake for both mixtures 23 times. GRAMA made a mistake scoring both mixtures a total of 28 times. What is interesting, however, is that these mistakes did not occur for the same RI and genetic marker combination. For each of the 23 cases where Revelation™ made a mistake in scoring both mixtures, GRAMA was able to correctly score at least one of the mixtures, so this inconsistency was brought to the user's attention. Revelation™ in the same way was able to bring GRAMA's 28 mistakes to the user's attention by correctly scoring at least one of the mixtures. There were no cases found in which both Revelation™ and GRAMA scored both mixtures incorrectly for the same RI and genetic marker. Since the user does not typically look at the electropherograms in the cases where the row is colored green, it is still possible that the final score was incorrect, though from the statistics reported above the possibility of this is extremely rare.

By flagging all of the potential mistakes, it may seem as if this approach would create more work for the user. Nevertheless, it was found that for 76.1% of the mapping scores both algorithms agreed for all mixtures (and the unmixed plate if included), and the final scores agreed for both plates. Thus, 76.1% of the time the user can simply accept the results of GRAMA's combined analysis and quickly move on to mapping scores that have been flagged. It may also seem that the small number of mistakes caught by using both algorithms is not worth the added number of mapping scores that the user must evaluate; however, it is

extremely important that the mapping scores are as accurate as possible in order for the genetic mapping programs to produce the most accurate results. In addition to improving accuracy, GRAMA saves end users a large amount of time. Users from the Schnable Plant Genomics Team report greater than a two-fold increase in productivity by using a combination of GRAMA and Revelation™ to conduct genetic recombinant analysis as opposed to only using Revelation™ [Elizabeth Hahn, per. comm.].

CHAPTER 8: DISCUSSION AND CONCLUSION

In this thesis, the need for high-throughput recombinant inbred analysis has been established. The existing software tool that can be used to do this type of analysis as well as its limitations when dealing with this specific type of data has been discussed. The GRAMA software package developed under this thesis work for high-throughput recombinant inbred analysis has been introduced and its peak determination algorithm has been examined at length. The development and use of combined consensus scores to provide more accurate mapping scores and the detailed information about how the consensus mapping scores are generated have been discussed. Finally, an experimental section was included to statistically evaluate GRAMA's ability to assist users in efficiently producing highly accurate mapping scores. The results of the experiments reveal that by using both the Revelation™ and GRAMA peak identification algorithms, user efficiency and final accuracy are both greatly enhanced.

The main contribution of this thesis work is the development of the GRAMA software tool for high-throughput recombinant inbred analysis. While this type of analysis can be done using existing software tools not specifically designed for this purpose, it is both tedious and unintuitive for the end user. GRAMA solves this problem by allowing the user to view all relevant information needed to make decisions at the same time and in a visual format that most accurately represents the problem to be solved. By combining the results of the Revelation™ and GRAMA peak determination algorithms, both accuracy and user efficiency are significantly increased. Because of this, GRAMA is an excellent new software package for dealing with high-throughput genetic recombinant inbred analysis.

There are a few ways that the GRAMA software tool can be enhanced in the future. Currently, in order for GRAMA to function correctly, each RI's genetic content must be contained within the same well on each plate. In addition, every well on a specific plate must contain the same kind of mixture. It would be nice if some flexibility were available for a user to run half of the wells on a plate with one mixture or track the RIs in different wells among the plates.

The ability to save and reopen all of the information stored in GRAMA at once is another useful addition. This can give the user the ability to quit the program while in the process of analyzing a specific genetic marker and come back to it directly to resume the work, instead of having to finish the whole analysis all at once. This also allows the user to easily review and/or change a final score at a later time using GRAMA.

GRAMA's peak determination algorithm can be enhanced by giving the user the ability to choose a control or representative electropherogram that clearly displays the homoduplex molecule electrophoretic pattern and the combination of heteroduplex and homoduplex molecules electrophoretic pattern. Many "noise" peaks can be eliminated from the set of relevant peaks if they occur in areas where no peaks were located in the representative electrophoretic patterns. It will also be convenient if GRAMA has the same curve smoothing and baseline subtraction abilities as Revelation™. These options can then be applied directly in GRAMA as needed instead of in Revelation™. This will also give GRAMA the ability to work with *.smr files directly.

REFERENCES

1. D.W. Bailey, Recombinant-inbred lines: An aid for finding identity, linkage, and function of histocompatibility and other genes, *Transplantation*, vol. 11, no. 3, pp. 325-327, 1971.
2. J.L. Bennetzen, V.L. Chandler, and P. Schnable, National Science Foundation-sponsored workshop report: Maize genome sequencing project, *Plant Physiology*, vol. 127, no. 4, pp. 1572-1578, 2001.
3. J.L. Bennetzen and W. Ramakrishna, Numerous small rearrangements of gene content, order and orientation differentiate grass genomes, *Plant Molecular Biology*, vol. 48, no. 5-6, pp. 821-827, 2002.
4. C.R. Buell, Current status of the sequence of the rice genome and prospects for finishing the first monocot genome, *Plant Physiology*, vol. 130, no. 4, pp. 1585-1586, 2002.
5. B. Burr and F.A. Burr, Recombinant inbreds for molecular mapping in maize: theoretical and practical considerations, *Trends in Genetics*, vol. 7, no. 2, pp. 55-60, 1991.
6. D. Curtis and H. Gurling, A procedure for combining two-point lod scores into a summary multipoint map, *Human Heredity*, vol. 43, no. 2, pp. 173-185, 1993.
7. R.W. Doerge, Mapping and analysis of quantitative trait loci in experimental populations, *Nature Reviews Genetics*, vol. 3, no. 1, pp. 43-52, 2002.
8. B. Ewing, L. Hillier, M.C. Wendl, and P. Green, Base-calling of automated sequencer traces using Phred, *Genome Research*, vol. 8, no. 3, pp. 175-194, 1998.

9. M.D. Gale and K.M. Devos, Comparative genetics in the grasses, *Proceedings of the National Academy of Sciences USA*, vol. 95, no. 5, pp. 1971-1974, 1998.
10. M.D. Gale and K.M. Devos, Plant comparative genetics after 10 years, *Science*, vol. 282, no. 5389, pp. 656-659, 1998.
11. Q. Gao and E.S. Yeung, High-throughput detection of unknown mutations by using multiplexed capillary electrophoresis with poly(vinylpyrrolidone) solution, *Analytical Chemistry*, vol. 72, no. 11, pp. 2499-2506, 2000.
12. J.B.S. Haldane and C.H. Waddington, Imbreeding and linkage, *Genetics*, vol. 16, no. 4, pp. 357-374, 1931.
13. G.M. Lathrop and J.M. Lalouel, Easy calculation of lod scores and genetic risks on small computers, *American Journal of Human Genetics*, vol. 36, no. 2, pp. 460-465, 1984.
14. M. Lee, N. Sharopova, W.D. Beavis, D. Grant, M. Katt, D. Blair, and A. Hallauer, Expanding the genetic map of maize with the intermated B73 x Mo17 (IBM) population, *Plant Molecular Biology*, vol. 48, no. 5-6 pp. 453-461, 2002.
15. Q. Li, Z. Liu, H. Monroe, and C.T. Culiat, Integrated platform for detection of DNA sequence variants using capillary array electrophoresis, *Electrophoresis*, vol. 23, no. 10, pp. 1499-1511, 2002.
16. M. Lichten and A.S.H. Goldman, Meiotic recombination hotspots, *Annual Review of Genetics*, vol. 29, pp. 423-444, 1995.

17. S. Lincoln, M.J. Daly, and E.S. Lander, Constructing genetic linkage maps with MAPMAKER/EXP version 3.0: A tutorial and reference manual, *A Whitehead Institute for Biomedical Research Technical Report*, 3rd ed.,
http://www.broad.mit.edu/genome_software/other/mapmaker.html, 1993.
18. H. Lodish, A. Berk, S.L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, *Molecular Cell Biology*, 4th ed., New York: W.H. Freeman, 2001.
19. D.I. Mester, Y.I. Ronin, Y. Hu, J. Peng, E. Nevo, and A.B. Korol, Efficient multipoint mapping: Making use of dominant repulsion phase markers, *Theoretical and Applied Genetics*, vol. 107, no. 6, pp. 1102-1112, 2003.
20. G. Moore, K.M. Devos, Z. Wang, and M.D. Gale, Grasses, line up and form a circle, *Current Biology*, vol. 5, no. 7, pp. 737-739, 1995.
21. M.L. Nickerson, M.B. Warren, B. Zbar, and L.S. Schmidt, Random mutagenesis-PCR to introduce alterations into defined sequences for validation of SNP and mutation detection methods, *Human Mutation*, vol. 17, no. 3, pp. 210-219, 2001.
22. R. Song, V. Llaca, and J. Messing, Mosaic organization of orthologous sequences in grass genomes, *Genome Research*, vol. 12, no. 10, pp. 1549-1555, 2002.
23. Spectrumedix® Corporation, Revelation™ 2.4 Manual v. 2.0, 2001.
24. Spectrumedix® Corporation, The Reveal System™,
<http://www.spectrumedix.com/Reveal.htm>, 2002.
25. J.I. Spiegelman, M.N. Mindrinos, C. Fankhauser, D. Richards, J. Lutes, J. Chory, and P.J. Oefner, Cloning of the Arabidopsis RSF1 gene by using a mapping strategy based on high-density DNA arrays and denaturing high-performance liquid chromatography, *The Plant Cell*, vol. 12, no. 12, pp. 2485-2498, 2000.

26. P. Stam, Construction of integrated genetic linkage maps by means of a new computer package: JoinMap, *The Plant Journal*, vol. 3, no. 5, pp. 739-744, 1993.
27. R.T. Swank and D.W. Bailey, Recombinant inbred lines: Value in the genetic analysis of biochemical variants, *Science*, vol. 181, no. 4106, pp. 1249-1252, 1973.

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks and gratitude to those who have helped with my work during the course of my Masters program. First and foremost, Dr. Hui-Hsien Chou for the guidance, support, understanding, and encouragement he has provided me towards the completion of this thesis. It would be impossible for me to describe the many ways that his assistance has helped me to achieve my goals. I would additionally like to thank Dr. Patrick Schnable and Dr. Xiaoqi Huang for serving on my committee.

I would like to thank Dr. Schnable and members of the Iowa State University Center for Plant Genomics whose research drove the need for the functionalities provided by GRAMA. I would like to thank Drs. Schnable, Tsui-Jung Wen, and An-Ping Hsia for their guidance in helping me to understand the biological processes that form the foundations of this thesis and for their assistance in gathering materials and resources. Thanks to Hsin “Debbie” Chen for her initial assistance in describing the functionalities required by the software program that was to become GRAMA. Thanks to Yi-Yin “Rita” Chen for her willingness to always offer assistance and her perpetually cheerful attitude. Thanks to Karthik Viswanathan for his technical assistance in many manners and for helping to coordinate the database connectivity. Most of all I would like to thank Elizabeth Hahn who has helped me in nearly every way imaginable. Her primary assistance was provided by performing the role of “customer” and relaying what functionalities GRAMA needed to provide and what problems needed to be addressed. However, she has given me an abundance of valuable input and advice beyond the “customer” role. Largely in part to Elizabeth’s involvement, the development of GRAMA has remained enjoyable throughout.

I also want to thank members of the Complex Computation Laboratory; Song Li, Ye Lin, Denise Mooney, Sunyoung Park, Kent Weber, Yue "Annie" Zhao, Yukari Ikuno, and Jia-Zhen Lee; for their support and assistance. I would especially like to thank Sunyoung Park for her assistance in the creation of some of the figures used in this thesis. Lastly, I would like to thank my family and my friends for their unwavering support throughout the years.

This research was funded by National Science Foundation award DBI 0321711.