

Application Note

Software for Amplified Fragment Length Polymorphism (AFLP®*)

Maurisa Riley, ChangSheng Jonathan Liu

SoftGenetics, LLC, 200 Innovation Blvd. Suite 241, State College, PA 16803

Introduction

Amplified Fragment Length Polymorphism (AFLP®) is a polymerase chain reaction (PCR) based genetic fingerprinting technique developed in the early 1990's by Keygene. AFLP uses restriction enzymes to cut genomic DNA, followed by ligation of complimentary double stranded adaptors to the ends of the restriction fragments. A subset of the restriction fragments are then amplified using 2 primers complimentary to the adaptor and restriction site fragments. The fragments are visualized on denaturing polyacrylamide gels through either autoradiographic or fluorescence methodologies¹.

The AFLP technology has the capability to detect various polymorphisms in different genomic regions simultaneously. It is also highly sensitive and reproducible. As a result, AFLP has become widely used for the identification of genetic variation in strains or closely related species of plants, fungi, animals, and bacteria. The AFLP technology has been used in criminal and paternity tests, in population genetics to determine slight differences within populations, and in linkage studies to generate maps for QTL analysis².

GeneMarker™ is an efficient, user-friendly software tool designed for the analysis of data generated by AFLP technology. The software is compatible with electrophoresis systems worldwide, including ABI (Applied BioSystems) files, **MegaBase** files, and **SpectruMedix** files, as well as slab gel output. The software features high efficiency allele calling, adjustable parameters and various reporting options including a trace comparison report.

Procedure

The software settings shown below are recommended to produce the best results for AFLP analysis. The correct settings for AFLP should include a low peak detection threshold and stutter filter, in order to detect small peaks as well as large peaks. The stutter peak filter is designed to remove stutter peaks within 2.5 bp of each detected allele peak. If the user would like to decrease the amount of false positives, then the stutter peak filter percentages can be increased. If the user would like to see all peaks, even those with minimal intensities, then the peak detection thresholds can be set to zero and the stutter peak filter can be turned off.

Suggested Analysis Parameters

1. **Analysis Type:** AFLP
2. **Peak Detection Threshold:** Intensity > 100; Percentage > 1 Max; Local Region % > 1 Local Max.
3. **Stutter Peak Filter (%)**: Left: 5 Right: 5
4. **Allele Evaluation Score:** Reject < 1 Check 7 < Pass; **Unconfidence at Rightside Score** < 30

Trace Comparison

After running the data with a size standard and panel, there are a few different reports and tools available to identify the presence and/or absence of alleles within sample traces. The Trace Comparison tool is designed primarily for AFLP data to identify

SoftGenetics LLC □ 200 Innovation Blvd. □ Suite 241 □ State College, PA 16803

Phone: 814/237/9340 □ Fax 814/237/9343

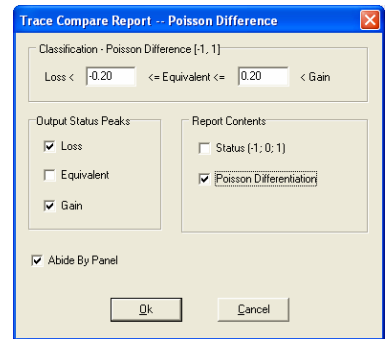
www.softgenetics.com □ email: info@softgenetics.com

length polymorphisms between closely related species.

The Trace Comparison tool is accessible through the Applications menu in the main toolbar. A window will open displaying two quantification methods: Poisson Difference and MLPA (Multiplex Ligation-dependent Probe Amplification) ratio. Please choose Poisson Difference unless working with MLPA data.

The trace comparison function uses Poisson distribution to calculate significant allelic differences between samples or closely related species. The software calculates and displays the allelic differences between a user defined reference and the sample traces in a histogram below each sample electropherogram.

The trace comparison report, which presents the loss, equivalent and gain of allele peaks, can be saved as a text file to be imported into Excel for printing. The Poisson Difference thresholds have default settings of Loss < -.20 < Equivalent < 0.20 < Gain, but may be altered by the user for a narrower or wider range. It is recommended that the user leave the "Abide By Panel" and "Poisson Differentiation" options checked in order to view the allelic probabilities.



Results

The allele report designed for AFLP analysis outputs the numbers 1 and 0 to represent the presence and absence of peaks. The symbol, ?, is used to represent questionable peaks, which the user is advised to check in the electropherogram. The allele report can also output the peak intensity for each detected allele. Both of these reports can be saved as text files to be imported into Excel.

The allele report displaying the presence, absence, and questionable presence of alleles is shown in Figure 1 and the allele report displaying the peak intensities is shown in Figure 2. The peaks with a green symbol are high confidence and those with a yellow symbol are of lower confidence.

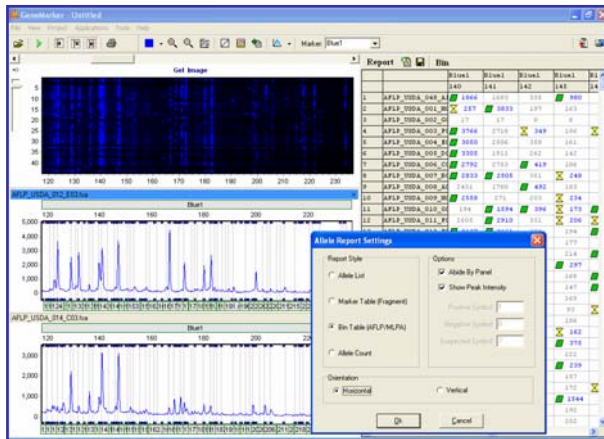


Figure 1. Report displaying presence and absence of alleles

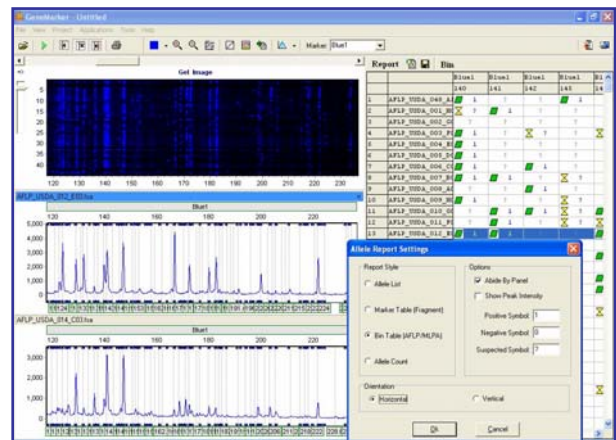


Figure 2. Allele Report displaying peak intensities

The Trace Comparison report shown in **Figure 3** portrays the allelic differences between sample 007_B01 and reference 001_H01. Sample 007 contains alleles at positions 167 and 171 that are not in the reference trace. Accordingly, the Poisson distribution value for both alleles is 1.000 as shown in the allele report, because they exist in the sample but not in the reference trace. When the reference trace contains an allele, which the sample trace does not contain, the Poisson distribution value is -1.

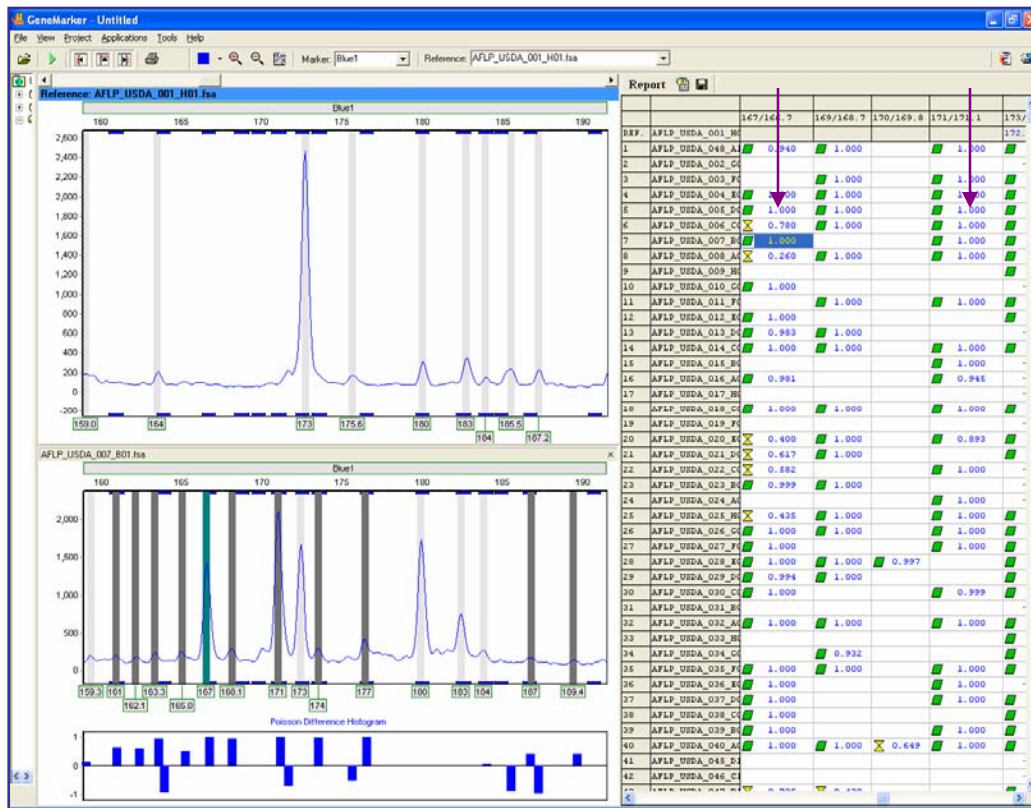


Figure 3. Trace Comparison showing allelic differences between sample and reference at positions 167 and 171.

Discussion

There are many advantages to AFLP when compared to other marker technologies including randomly amplified polymorphic DNA (RAPD), restriction fragment-length polymorphism (RFLP), and micro satellites. AFLP not only has higher reproducibility, resolution, and sensitivity at the whole genome level than other techniques, but it has the capability to amplify between 50 and 100 fragments at one time. In addition, no prior sequence information is needed for amplification. As a result, AFLP has become extremely beneficial in the study of taxa including bacteria, fungi and plants, where much is still unknown about the genomic makeup of various organisms².

AFLP is widely accepted as an effective tool for identifying genomic differences among closely related species, and GeneMarker has the ability to quantify and report these differences. Similar to AFLP, GeneMarker is highly accurate, sensitive and easy to use, making it a successful complement to the genotyping technique.

References

1. AFLP: a new technique for DNA fingerprinting. Vos, P.; Hogers, R.; Bleeker, M.; Reijans, M.; Lee, Th. van der; Hornes, M.; Frijters, A.; Pot, J.; Peleman, J.; Kuiper, M. & Zabeau, M. *Nucleic Acids Research*. 1995. 23(21): 4407-4414
2. AFLP Genotyping and Fingerprinting. Ulrich G. Mueller and L. LaReesa WolfenBarger. *Tree*. October 1999. Vol.14 no.10.

*AFLP is a registered trade mark of KeyGene, N.V.

Application Note

Microsatellite Analysis with Linked Pedigree Tool

Maurisa Riley, ChangSheng Jonathan Liu

SoftGenetics, LLC, 200 Innovation Blvd. Suite 241, State College, PA 16803

Introduction

Microsatellites or variable tandem repeats (VTRs) are successful markers used for various types of genotyping. Microsatellites are short stretches of repeated DNA found in most genomes that are highly polymorphic in humans and most other species. This variability has made microsatellites a popular genetic marker for genotyping applications such as medical genetics, forensics, genetic mapping, and human and plant population studies.

Microsatellites were originally employed for genetic mapping but now are widely used in oncology research. The polymorphic nature of microsatellites makes them useful in linkage studies which attempt to locate genes responsible for various genetic disorders.

Microsatellites also have some advantages, which make them popular markers for research. They are found in large numbers and are relatively evenly spaced throughout the genome. Due to their small size of 2-6 base pairs, microsatellites can be analyzed using polymerase chain reaction (PCR) and accurately sized using electrophoresis systems.

GeneMarker is a unique genotyping tool as it is compatible with files from all major capillary and slab gel electrophoresis systems including ABI files (*.FSA, *.AB1, *.ABI), SCF files, MegaBace files (*.RSD, *.ESD), SpectruMedix files (*.SMD, *.SMR), Beckman files, and Licor files. GeneMarker is a replacement for such software packages as SAGA from LI-COR, TrueAllele from Cybergentics, GeneMapper, Genotyper, and GeneScan from Applied Biosystems. GeneMarker can perform fragment analysis on four or five color data sets from any slab gel or capillary electrophoresis system. Additionally, this software automatically corrects for most instrument and chemistry errors, saving significant analysis time and cost.

Application

Automated Correction

GeneMarker decreases analysis set-up time through automated correction of common genotyping problems including saturated peaks, noisy data, wavelength bleeding, instrument spikes, and stutter peaks. GeneMarker's automated Run Wizard is designed to make analysis quick, easy, and accurate. The Data Analysis window features include:

1. **Saturation Correction:** Analysis of saturated data points by creating a synthetic peak based upon the peak shape before and after saturation.
2. **Baseline Subtraction:** The software removes the baseline so that the Y axis is above the noise level.
3. **Pull-up Correction:** This function removes peaks caused by wavelength bleeding.
4. **Spike Correction:** The software automatically removes peaks from voltage spikes caused by micro-air bubbles or debris in the laser path.
5. **Stutter Peak Correction:** The software automatically filters for stutter peaks caused by PCR slippage.

Analysis Parameters

GeneMarker has flexible settings to allow user manipulation. In order to produce the most accurate results, with low levels of false positives and negatives, the following settings are recommended for fragment analysis.

1. **Analysis Type:** Fragment
2. **Peak Detection Threshold:** Intensity > 100; Percentage > 5 %Max; Local Region > 25 % Local Max.
3. **Stutter Peak Filter[%]:** Left: 90 Right: 30.
4. **Allele Evaluation:** Score: Reject < 1 Check 10 < Pass

Pedigree

The user can either open an existing pedigree file of type ped or pre or create a new pedigree file in GeneMarker's pedigree tool. The pedigree chart is designed to aid identification of inheritance patterns and abnormalities. All individuals in the pedigree with sample files are directly linked to the corresponding electropherograms with a mouse click, and individuals with illogical or abnormal allele calls are highlighted in red. The link between the pedigree and the electropherograms displaying allele calls for each marker make analysis quick and efficient.

Existing Pedigree File: After running sample files in GeneMarker with a size standard and panel, the user must link the sample files to the pedigree file. GeneMarker has incorporated a tool for this function named "Pedigree Filename Match" in the Tools menu. It is necessary to enter the correct character numbers designating family and individual identifiers. In the example, 01_PD_5005_0001_A01_07.fsa, the family identifier 5005 would be from 7 to 10 and the individual identifier 0001 from 12 to 15. After linking the files, open the pedigree file in Applications: Family Tree. Load the original pedigree file in the top panel and the linked .SMP file in the bottom panel.

New Pedigree File: After running samples in GeneMarker go to Applications: Family Tree and click the New Pedigree File icon located in the toolbar. The user may then create a new multigenerational family. After adding the first individual, it is then possible to add mates, siblings, and children by right clicking on the individual node.

Results

In the example shown below for Family 5005, there is an inheritance conflict for individual 0015 in marker Blue 3. Allele 197 is called in individual 0015 but not in either parent (0001, 0002). Based upon the results, there are also problems in markers Green2, 3 and Yellow 5 for this individual as indicated in red.

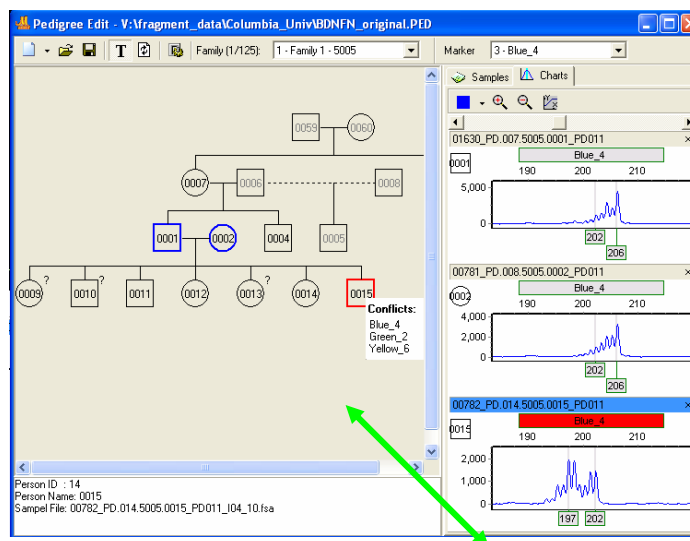


Figure 1. Pedigree chart showing inheritance conflicts in Family 5005.

Discussion

The reasoning behind these conflicts shown in the results may be due to instrument or paternity identification errors. The pedigree chart makes analysis quick by identifying problematic marker regions in red. The user may add, delete, or change alleles in the electropherogram by right clicking at the allele location.

The integration of the pedigree tool into genotyping software is very beneficial to track the inheritance of alleles over generations, especially in known genetic disease regions. The identification, presence and/or absence, of alleles is a crucial initial step in the study of disease inheritance patterns within families. GeneMarker's pedigree is a useful tool for the identification and genetic tracking of alleles spanning human chromosomes.

SoftGenetics LLC □ 200 Innovation Blvd □ Suite 241 □ State College, PA 16803

Phone: 814/237-9340

Fax: 814/237-9343

www.softgenetics.com email: info@softgenetics.com

GeneMarker[®] Software for Terminal-Restriction Fragment Length Polymorphism (T-RFLP) Data Analysis

David Hulce, PhD, ChangSheng (Jonathan) Liu, PhD, SoftGenetics LLC

Introduction

One culture-independent molecular method for fingerprinting microbial populations is terminal-restriction fragment length polymorphism (T-RFLP) (1). T-RFLP has been used to demonstrate "species diversity in activated sludge, bioreactor sludge, aquifer sand and termite guts" (2); to distinguish populations of denitrifiers, *Bacteria* and *Archaea* in marine sediments (3); to show "temporal changes in the diversity of the bacterial communities" in polluted marine environments (4); and to study fungal populations in soil (5). This sensitive high-throughput method has the potential to generate large sets of data. Improving speed and efficiency of terminal restriction fragment (T-RF) analysis will further understanding the diversity, richness and dynamics of microbial populations and its relationship to ecological processes.

T-RFLP is a polymerase chain reaction (PCR) based genetic fingerprinting technique. DNA is extracted from a microbial community. One of the primers of a primer pair is labeled with a fluorescent dye and used to amplify a selected region of a gene of interest by PCR. The resulting PCR fragment is digested with one (sometimes two or more) restriction endonuclease and the T-RFs are separated with an automated DNA analyzer. Microbial diversity in a community can be estimated by analyzing the number and peak heights of T-RF patterns (2).

GeneMarker can perform fragment analysis and genotyping on four or five color data sets from any slab gel or capillary electrophoresis system. This software automatically corrects for many common problems—instrument spike, color pull-up, peak pull-up, noisy data, saturated peaks and stutter peaks—saving significant analysis time and cost, efficiently analyzing raw fragment data within seconds. GeneMarker is robust software to analyze DNA fragment data labeled with MegaBACE™ dyes (Amersham), Big Dye® (Applied Biosystems Inc.) or Beckman dyes from a variety of platforms: ABI DNA Analyzer or Genetic Analyzer, Amersham instruments or Beckman instruments. GeneMarker is compatible with files from all major capillary and slab gel electrophoresis systems including ABI files (*.FSA, *.AB1, *.ABI), SCF files, MegaBace® files (*.RSD, *.ESD), SpectruMedix files (*.SMD, *.SMR), Beckman files and Li-Cor files. The software features high efficiency allele calling, adjustable parameters and a variety of reporting options.

T_RFLP Fragment Data Analysis

The software settings shown below are suggested to produce highly sensitive and reproducible results for T-RFLP analysis. To detect low intensity peaks, T-RFLP analyses should include a low peak detection threshold with stutter filter turned off or set to zero. The peak detection threshold intensity value may be estimated based on the peak with the maximum intensity from the negative control in the sample set. All parameters can easily be adjusted to optimize the analysis for your specific requirements.

The following analysis parameters are adjustments made to the default settings for AFLP analysis using GeneMarker® software.

Suggested Analysis Parameters

I. Template Selection

1. Panel: None
2. Analysis Type: AFLP

II. Data Process

1. Raw Data Analysis: select Smooth
2. Allele Call: deselect Auto Range and set start to 60, end to 500
3. Peak Detection Threshold: Intensity > 40
4. Stutter Peak Filter: deselect or set left and right values to 0

III. Additional settings

1. Peak Score: Reject < 0 check 1 < Pass
2. AFLP-Unconfidence at Rightside Score: 1

Following size calling the Analysis Display window will activate (figure 1). The samples listed on the far left are marked with green symbols, indicating that size calling was successful. Sample electropherograms are displayed in the center of the window. An allele/fragment report records the presence (1) or absence (0) of a fragment for the samples (upper right of figure 1). The Allele Report Settings dialog box presents the option to display peak areas and peak intensities in the Report. The report can be saved as a *.txt file.

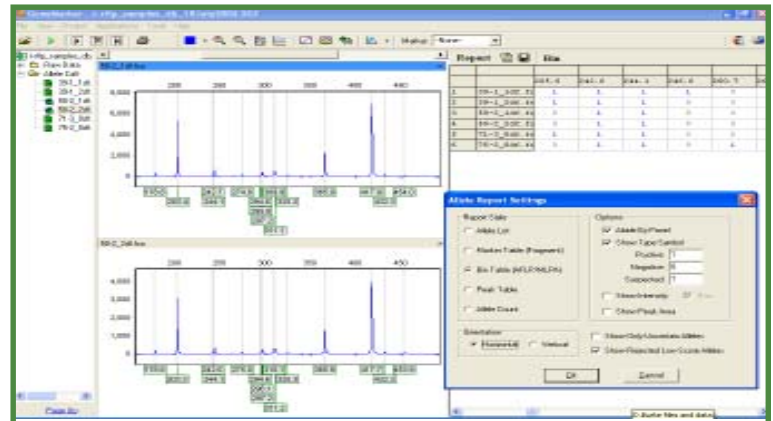


Figure 1. GeneMarker analysis window and Allele Report Setting dialog box: The samples are listed on the far left. Two sample file electropherograms are displayed in the center. The allele report table is displayed on the right. The Allele Report Settings dialog box is shown below the Report

A peak table summarizing fragment size, peak height and peak area, along with a comments field, can be activated (clicking the third icon from right in the toolbar) and used to record comments, edit alleles and confirm fragments. The peak table can be saved as a *.txt file (clicking second icon from right in the toolbar). Text files can be imported into other applications for further analysis.

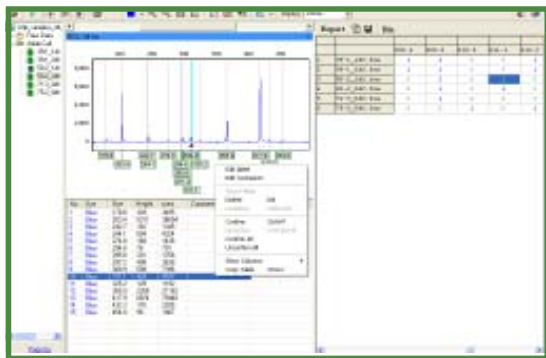


Figure 2. Peak table displayed below a sample electropherogram showing the option menu for editing items in the peak table (clicking the right mouse button with the cursor on a highlighted cell will activate the editing menu).

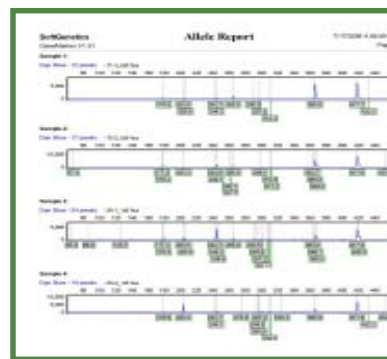


Figure 3. Print Report: Comparison of electropherograms from four samples (only one sample from a duplicate pair is shown).

Results

Samples 50-2_2df and 50-1_1df are replicates (Figure 1). The electropherograms for these duplicate samples display a nearly identical pattern of fragment sizes, all within 0.4 bp. The peak heights for both samples display comparable intensities.

The pattern of fragment sizes and peak heights displayed in the electropherograms presents a genetic fingerprint for each of the sample populations (figure 3). Sample 39-1_df contains two unique fragments, 86.8 and 125.7, with fragment 244.2 displaying greatest intensity among these samples. Fragment 203.4 +/- 0.1 is common among the four samples, but shows greatest intensity in sample 50-2_1. The differences among these genetic fingerprints may reflect the relative abundance and diversity of organisms that are contained in these populations.

Discussion

There are many molecular methods for investigating microbial ecology: denaturing gradient gel electrophoresis (DGGE), microarrays, quantitative polymerase chain reaction (Q-PCR), restriction fragment length polymorphism (RFLP) and T-RFLP, to name a few. Changing a few parameters for AFLP analysis quickly and efficiently adjusts fragment detection for T-RFLP analyses. T-RF fragment size and abundance is easily visualized on the electropherograms, quickly showing differences and similarities among the samples. Saving the fragment size and peak information as a *.txt file provides the opportunity to export the analytical results to a data base.

T-RFLP analyses provide genetic fingerprints of microbial communities. GeneMarker® is an accurate, sensitive and easy to use software tool which can simplify DNA fragment analysis.

References:

1. Christopher B. Blackwood, Terry Marsh, Sang-Hoon Kim and Eldor A. Paul. 2003. Terminal restriction fragment length polymorphism data analysis for quantitative comparison of microbial communities. *Appl. Environ. Microbiol.* 69:926-932.
2. Wen-Tso Liu, Terence L. Marsh, Hans Cheng and Larry J. Forney. 1997. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Appl. Environ. Microbiol.* 63:4516-4522.
3. Gesche Braker, Héctor L. Ayala-del-Rio, Allan H. Devol, Andreas Fesefeldt and James M. Tiedje. 2001. Community structure of denitrifiers, Bacteria and Archaea along redox gradients in Pacific Northwest marine sediments by terminal restriction fragment length polymorphism analysis of amplified nitrite reductase (nirS) and 16S rRNA genes. *Appl. Environ. Microbiol.* 67:1893-1901.
4. R. Denaro, G. D'Auria, G. Di Marco, M. Genovese, M. Troussellier, M. M. Yakimov and L. Giuliano. 2005. Assessing terminal restriction fragment length polymorphism suitability for the description of bacterial community structure and dynamics in hydrocarbon-polluted marine environments. *Environ. Microbiol.* 7:78-87.
5. Ian C. Anderson and John W. G. Cairney. 2004. Diversity and ecology of soil fungal communities: increased understanding through the application of molecular techniques. *Environ. Microbiol.* 6:769-779.

Acknowledgement:

We would like to thank David Burke, Ph.D., The Holden Arboretum, for collaborating with developing the analysis parameters for this application.

SoftGenetics LLC 200 Innovation Blvd. Suite 235 State College, PA 16803 USA

Phone: 814/237/9340 Fax 814/237/9343

www.softgenetics.com email: info@softgenetics.com

GeneMarker[®] Software for SNPWave[™] Analysis

David Hulce, Wan Ning, ChangSheng Jonathan Liu

Introduction

Single nucleotide polymorphisms (SNPs) occur every 100 to 300 bases along the human genome and make up to 90% of human genetic variation **(1)**. Functional SNPs—classified as non-synonymous SNPs (nsSNPs) that occur in the coding region of a gene or as regulatory SNPs (rSNPs) that occur in the promoter region of a gene—are often associated with altered protein function or gene expression **(2, 3)**. Intronic or intergenic SNPs may not alter gene or protein function, but can be used to address questions in evolutionary biology **(4)** or in association studies with complex diseases, drug response, environmental insults **(1)** quantitative trait loci (QTL) **(5)** or genotyping plants and animals **(6)**.

One high-throughput method to determine SNP genotypes is SNPWave (Keygene N.V.). SNPWave uses multiplex oligonucleotide ligation amplification of allele-specific probes coupled with AFLP[®]-primer selective amplification. SNPWave can detect up to 100 SNPs **(6)**.

Circularizing padlock ligation probes are constructed that are specific to the SNP and flanking sequences. Locus-specific probes will hybridize to complementary denatured genomic DNA. Allele-specificity is determined by the SNP at the 5' end of the padlock probe. Probes that contain a 5' nucleotide complementary to the SNP will be ligated and amplified by PCR in subsequent reactions. Probes that do not contain a 5' nucleotide complementary to the SNP will not be ligated and will not be amplified **(6)**.

The padlock probes contain stuffer regions and primer binding sites for AFLP-selective amplification. The ligated padlock probe is amplified using fluorescently-labeled +2-selective and unlabeled non-selective AFLP primers. The stuffer region provides length discrimination between alleles and among loci. SNPs are separated by 2 base pairs and loci are separated by 3 bp. The fragments are separated by size using capillary electrophoresis. Fragment dye color and length indicate SNP locus and allele.

To fully utilize the investigative potential of SNPWave, a robust genotyping and data analysis system should be employed. GeneMarker genotyping software is designed for fast, accurate and efficient analysis and reporting of SNP data generated by capillary electrophoresis.

GeneMarker is user-friendly software containing a robust size calling algorithm that resolves fragment lengths to less than one base pair with high efficiency allele calling, essential for data generated by SNPWave. The software is capable of analyzing data files from major capillary electrophoresis systems including ABI, MegaBACE[™] and Beckman. GeneMarker, in combination with our JelMarker[™], can also be used for analysis of slab gel outputs from Li-Cor and FMBIO.

Procedure

The software settings shown below are recommended to produce the best results for SNPWave analysis. The process from raw data to review of the analysis report occurs in three steps: size calling of the data, creating the SNPWave panel, running the sized data against the panel.

Import the data files (up to 1000 lanes) into GeneMarker using the Open Files icon.

Step 1: Suggested Analysis Parameters for Size Calling

Click the **Run Project** icon to activate the *Run Wizard* for size calling of the data.

Template Selection

1. Panel: none
2. Size Standard: ET400-R
3. Standard Color: red
4. Analysis Type: *SNPlex*

Data Process-*SNPlex* Analysis

1. Accept defaults

Step 2: Create Panel

After the data is processed, create a panel: **Tools** '!' **Panel Editor** '!

Create New Panel.

1. The **Panel Editor** will create a panel based on the called peaks for each dye color of the selected samples.
2. Rename any of the SNPs as necessary to improve readability of final SNP report. The details for editing the panel can be located in the manual.
3. Click OK and the panel is automatically uploaded to the **Panel Editor**.
4. Reprocess the data with the new panel (see Step 3).

Step 3: Compare Sized Fragments to Panel

1. Under *Panel* in the *Template Selection* window of the *Run Wizard* use the drop-down menu to select the appropriate panel for the analysis (use all previously selected options).
2. To activate the SNP analysis report, click **Applications** in the menu bar and select **SNPlex/SNaPshot**.

Results

The SNP analysis report window displays a synthetic gel image, list of samples, cluster plot to analyze peak information and assign SNP genotypes, and sample electropherogram (**Figure 1**). A SNP genotyping report is available (on right, **Figure 1**) and can be saved as Excel (*.xls) or Text (*.txt) files. The information contained in the report is interlinked. Double-clicking on the cell in the report containing SNP 107/109 for locus SNP_06 of sample E01 highlights the cell, corresponding sample ID in the sample list, data point in the cluster plot and locus in the electropherogram.

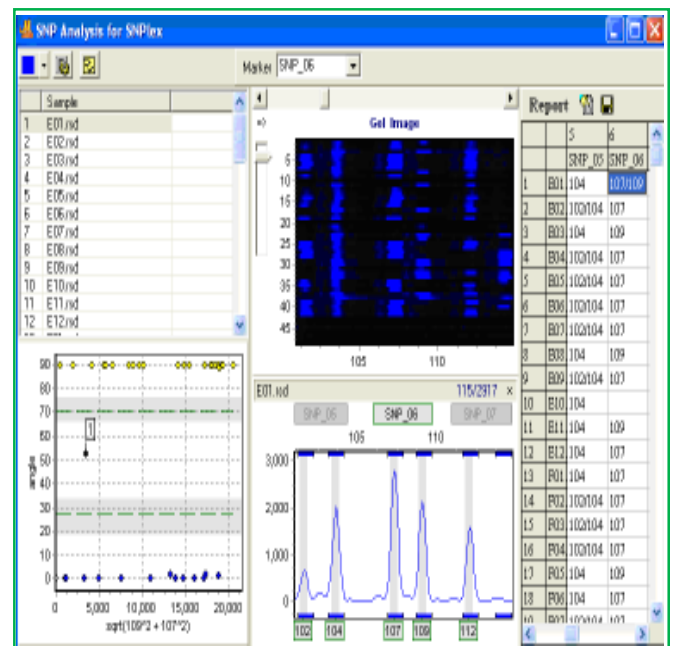


Figure 1: SNP Analysis for SNPWave report

The electropherogram displays peak intensity, fragment size (in box along bottom of trace) and SNP name (along top of trace). Low intensity peaks (<200 rfu) are detected with high sensitivity using derivative to remove influence of the baseline. The intensity threshold may be set to filter out low intensity peaks that contribute to incorrect SNP calling. A SNP is represented by the fragment size, height and color of the peaks. SNP_06 for sample E02 is heterozygous for alleles 107 and 109 (**Figure 1**).

The default cluster plot layout is polar view. Polar view plots the angle of the vector connecting the peak intensities of the SNP and the origin. This vector is plotted against the square root of the sum of the square of the longer fragment and the square of the shorter fragment. A Cartesian view (**Figure 2**) may be selected by clicking the Layout Settings icon on the toolbar. The number of samples, the mean and the standard deviation for each group can be displayed under the cluster plot.

Data points within the gray region of the cluster plot indicate SNP calls of low confidence. The clustering algorithm calculates a t-value for each data point in each cluster. Samples that do not meet the 95% confidence level for the cluster will appear in the gray region. Editing SNPs can be done by right-clicking on a data point in the cluster plot or on the cell of the report table.

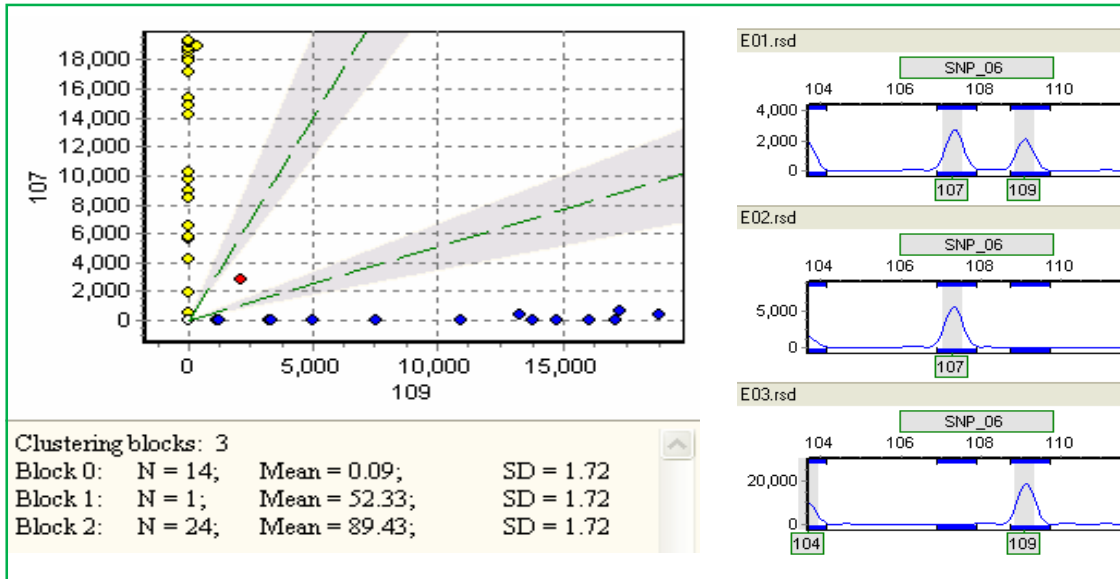


Figure 2: Cartesian view of cluster plot displaying statistics (left) and examples of heterozygous 107/109, homozygous 107 and homozygous 109 tomato lines (right).

Discussion

The data presented in Figure 1 was obtained from a collection of tomato lines. This data set shows efficient allele calls for *.rsd files from MegaBACE DNA Analysis System. One can quickly confirm allele calls using the cluster plot and SNP report. For SNP_06 the majority of the samples are homozygous for either allele 107 or allele 109. One sample is heterozygous for both alleles (**Figure 2**, right).

Various techniques have been developed to interrogate SNPs, including single nucleotide primer extension (SNUPE), SNPlex™, SNaPshot™, SNUpe™, SNPWave®, SNP chips and DNA sequence analysis. SNPWave has the advantage of accurate high-throughput genotyping of up to 100 SNPs.

GeneMarker is highly accurate and easy to use. The robust sizing and efficient allele calling along with advanced clustering algorithms can quickly and efficiently determine SNP genotypes while maintaining the advantages of SNPWave technology.

References

1. SNP Fact Sheet. Human Genome Project Information. U.S. Department of Energy-Office of Science, Office of Biological and Environmental Research. http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml.
2. Salim Mottagui-Tabar, MA Faghihi, Y Mizuno, PG Engström, B Lenhard, WW Wasserman and C Wahlestedt. 2005. Identification of functional SNPs in the 5-prime flanking sequences of human genes. *BMC Genomics* 6:18. <http://www.biomedcentral.com/1471-2164/6/18>.
3. Yan H, W Yuan, VE Velculescu, B Vogelstein, KW Kinzler. 2002. Allelic variation in human gene expression. *Science* 297(5584):1143.
4. Tiedemann R, K Moll, KB Paulus, M Scheer, P Williot, R Bartel, J Gessner and F Kirschbaum. 2007. Atlantic sturgeons (*Acipenser sturio*, *Acipenser oxyrinchus*): American females successful in Europe. *Naturwissenschaften* 94(3): 213–217.
5. Pot D, J-C Rodrigues, P Rozenberg, G Chantre, J Tibbits, C Cahalan, F Pichavant, C Plomion. 2006. QTLs and candidate genes for wood properties in maritime pine (*Pinus pinaster* Ait.). *Tree Genetics & Genomes* 2: 10–24.
6. van Eijk, MJT, JLN Broekhof, HJA van der Poel, RCJ Hogers, H Schneiders, J Kamerbeek, E Verstege, JW van Aart, H Geerlings, JB Buntjer, AJ van Oeveren and P Vos. 2004. SNPWave™: a flexible multiplexed SNP genotyping technology. *Nucleic Acids Research* 32(4): e47.

GeneMarker[®] software for SNPlex[™] analysis

David Hulce, Jonathan Liu

Introduction

Single nucleotide polymorphisms (SNPs) occur frequently and with a relatively even distribution in the human genome (1, 2). Functional SNPs (non-synonymous or nsSNPs occur in the coding region of a gene, regulatory or rSNPs occur in the promoter of a gene) are associated with altered protein function or gene expression (3). Intronic or intergenic SNPs may not alter gene or protein function, but may be associated with complex diseases, drug response or environmental insults (4).

To fully utilize the diagnostic potential of SNPs, a robust high-throughput genotyping and data analysis system should be employed. SNPlex[™] Genotyping System (Applied Biosystems) can interrogate 48 SNPs simultaneously and has been used to investigate SNPs in 92 cancer-related genes in breast cancer (5) and to genotype plants (6). GeneMarker genotyping software is designed for fast, accurate and efficient analysis of SNPlex data.

GeneMarker is user-friendly software containing a robust size calling algorithm that detects fragment lengths to less than one base pair and high efficiency allele calling, essential for data generated by SNPlex technology. The software is compatible with electrophoresis systems worldwide, including ABI files, MegaBACE[™] and Beckman files, as well as slab gel outputs.

Procedure

The software settings shown below are recommended to produce the best results for SNPlex analysis. The process from raw data to review of the analysis report occurs in three steps: size calling of the data, importing and adjusting the SNPlex panel, running the sized data against the panel.

Import the data files into GeneMarker using the Open Files icon.

Step 1: Suggested Analysis Parameters for Size Calling

Click the Run Project icon to activate the Run Wizard for size calling of the data.

Template Selection

1. Panel: none
 2. Size Standard: SNPlex_48plex_v1
 3. Standard Color: orange
 4. Analysis Type: SNPlex
- Data Process-SNPlex Analysis
1. Accept defaults

Step 2: Importing SNPlex Panel Information

The panel information can be imported using the Panel Editor (Tools→Panel Editor).

1. Under Files in the menu bar select Import ABI Panels.
2. In the Panel File field navigate to the directory that contains the panel information) file format *Panels.txt).
3. In the Bins field navigate to the directory that contains the bin information (file format *Bins.txt).

4. Click OK and the Panel is automatically uploaded to the Panel Editor
5. Rename any of the SNPs, if necessary, to improve readability of final SNP report.
6. To adjust the panel, hold down Shift key and left mouse button, and drag the gray bars over blue or green peaks.

Step 3: Compare Sized Fragments to SNPlex Panel

1. Under Panel in the Template Selection window of the Run Wizard use the drop-down menu to select the appropriate SNPlex panel for the analysis (use all previously selected options).
2. To activate the SNPlex analysis report, click Applications in the menu bar and select SNPlex/SNaPshot.

Results

The SNPlex report window displays a synthetic gel image, list of samples, cluster plot to analyze peak information and assign SNP genotypes, sample electropherogram and SNP genotyping report (Figure 1). The information contained in the report is interlinked. Double-clicking on the cell containing alleles 58/60 for SNP locus P37 of sample A14 highlights the cell, corresponding sample ID in the sample list, data point in the cluster plot and locus in the electropherogram.

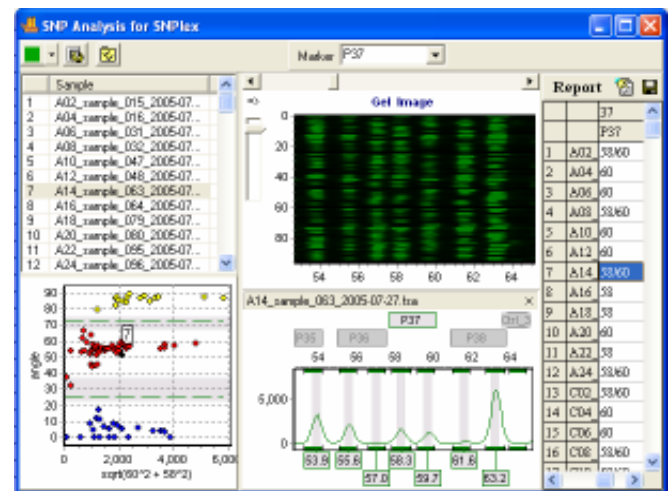


Figure 1: SNPlex report from GeneMarker

The electropherogram displays peak intensity, fragment size (in box along bottom of trace), allele name (along top of trace) and locus identification. Low intensity peaks (<200 rfu) are detected with high sensitivity using derivative to remove influence of baseline. The intensity threshold may be set to filter out low intensity peaks that contribute to incorrect allele calling. A specific SNP is represented by the size and color of the fragments contained in the locus.

The default cluster plot layout is polar view. A Cartesian view (Figure 2) may be selected by clicking the Layout Settings icon on the toolbar. Polar view plots the angle of the vector connecting the peak intensities of the SNP—shorter fragment on y-axis and longer fragment on x-axis—and the origin against the square root of the sum of the square of the longer fragment and the square of the shorter fragment. The number of samples, mean and standard deviation for each group can be displayed under the clustering plot.

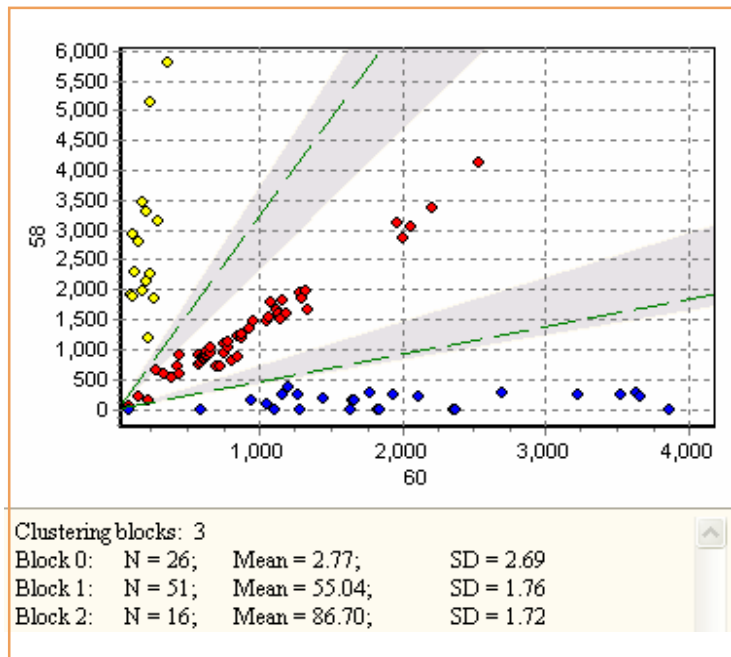


Figure 2: Cartesian view of cluster plot from GeneMarker

Data points within the gray region of the cluster plot indicate SNP calls of low confidence. The clustering algorithm calculates a t-value for each data point in each cluster. Samples that do not meet the 95% confidence level for the cluster will appear in the gray region. Editing SNPs can be done by double right-clicking on a data point in the cluster plot or on the cell of the report table.

The SNP analysis report can be saved as Excel (*.xls) or Text (*.txt) files.

Discussion

Various techniques have been developed to interrogate SNPs, including single nucleotide primer extension (SNuPe), SNPlex™, SNPWave™ (KeyGene), SNP chips and DNA sequence analysis. SNPlex has the capacity of high-throughput genotyping.

GeneMarker is highly accurate and easy to use. High efficiency allele calling and low intensity peak detection along with robust size calling and clustering algorithms can quickly and efficiently determine SNP genotypes, maintaining the high-throughput advantages of SNPlex genotyping.

Acknowledgments

We would like to thank James Knowles and Oleg V. Evgrafov of Columbia University for providing the SNPlex data.

References

1. Kruglyak L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* 22:139-144.
2. Venter JC, MD Adams, EW Myers, PW Li, RJ Mural, GG Sutton, et al. 2001. The sequence of the human genome. *Science* 291:1304-1351.
3. Salim Mottagui-Tabar, MA Faghihi, Y Mizuno, PG Engström, B Lenhard, WW Wasserman and C Wahlestedt. 2005. Identification of functional SNPs in the 5-prime flanking sequences of human genes. *BMC Genomics* 6:18. <http://www.biomedcentral.com/1471-2164/6/18>.
4. SNP Fact Sheet. Human Genome Project Information. U.S. Department of Energy-Office of Science, Office of Biological and Environmental Research. http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml.
5. A Vega, A Salas, C Phillips, B Sobrino, B Carracedo, C Ruiz-Ponte, R Rodríguez-López, G Rivas, J Benítez, A Carracedo. 2005. Large-scale single nucleotide polymorphism analysis of candidates for low-penetrance breast cancer genes. *Breast Cancer Research* 7(Suppl 2):P1.14.
6. Massimo Pindo, M Troglio, D Cartwright, A Gutin and R Velasco. 2007. SNPlex™ Genotyping System to Develop A Dense SNP-Based Genetic Linkage Map of Grapevine (*Vitis vinifera* L.). Plant & Animal Genomes XV Conference; San Diego, CA. January 13-17:P74.

Trademarks are property of their respective owners.

Clustering Algorithms for Genetic Analysis with GeneMarker®

Tamela Serensits, Wan Ning, Haiguo He, Jonathan Liu, Ph.D.

Introduction

Biological applications of data clustering include phylogeny analysis and community comparisons in ecology, gene expression pattern, enzymatic pathway mapping, and functional gene family classification in the bioinformatics field. (1) It has also been successfully paired with the AFLP analysis technique for a variety of applications. (2)

Amplified Fragment Length Polymorphisms (AFLP®) is a polymerase chain reaction (PCR)-based genetic fingerprinting technique developed in the early 1990's by Keygene*. AFLP technology has the capability to detect polymorphisms in different genomic regions simultaneously. It is also highly sensitive and reproducible. As a result, AFLP has become widely used for the identification of genetic variation in strains or closely related species of plants, fungi, animals, and bacteria. AFLP technology has also been used in criminal and paternity tests, population genetics and linkage studies. (3)

As the results of AFLP are obtained, some researchers turn to novel statistical tools for analyzing the data. An example of this was in 2005 when Fearnley et al. applied a clustering algorithm to AFLP data. (4) The results helped differentiate the relationship between closely related strains of *Yersinia enterocolitica*, a bacterium that infects several species including humans, pigs, sheep, and cattle. The study found clustering analysis of AFLP data to be highly discriminatory.

GeneMarker is an easy-to-use, accurate fragment analysis tool and can perform analysis on up to 1,000 lanes of four or five color data sets generated by either slab gel or capillary electrophoresis. It is a unique genotyping tool as it is compatible with files from all major capillary and slab gel electrophoresis systems including ABI files (*.FSA, *.AB1, *.ABI), SCF files, MegaBace files (*.RSD, *.ESD), SpectruMedix files (*.SMD, *.SMR), and Beckman files. *AFLP is a registered trademark of KeyGene, N.V.

Procedure

There are two types of data clustering: hierarchical and partitional. Partitional clustering includes the K-means and Self-Organizing Map methods. Hierarchical clustering is the second method of clustering and is the method that is implemented in GeneMarker. Hierarchical Clustering treats each data point as a single cluster and successively merges clusters until all points have been merged into a single remaining cluster. Hierarchical clustering is often represented as a dendrogram. In GeneMarker, the hierarchical algorithm is agglomerative and establishes clusters from the bottom up.

Distance Measure

The first step in hierarchical clustering is to select a distance measure. GeneMarker distance options include Euclidean Distance, Correlation Coefficient, and Percentage of Same Genotypes. Euclidean Distance is the straight line distance between two points in two or three dimensional space. The equation is essentially the same as that for determining the length of the hypotenuse of a triangle – computed by finding the square of the distance between each variable, summing the squares, and finding the square root of that sum. We have simplified this equation (below) in GeneMarker. The Correlation Coefficient is based on the Pearson Correlation equation and is a statistical concept that quantifies the level of relationship between two sets of measurements. It is a measure of similarity where two values that are perfectly correlated have a distance of 1.00. Percentage of Same Genotypes is simply the number of similar genotypes divided by the total number of genotypes.

The following are GeneMarker's clustering algorithms:

$$\text{Euclidean Distance: } \sum_{i=1}^n |(x_i - y_i)| \quad \text{Correlation Coefficient: } r \equiv \sqrt{bb'} = \frac{n \sum xy - \sum x \sum y}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

Linkage

In addition to a distance measure, the type of linkage needs to be applied. GeneMarker has three options: Single, Complete, and Average linkage. Single linkage measures the minimum distance between two clusters. Clustering using single linkage tends to produce an effect called chaining where single genes are added to clusters one at a time. Complete linkage is the opposite of single linkage. It measures the distance between the farthest two points in the clusters. Complete linkage performs well when the clusters are well defined with minimal noise. Average linkage defines the distance between two clusters as the mean distance between all points in the clusters. It is important to note that choosing different linkage measures results in different cluster diagrams. We demonstrate in the Results section how different distance measure and linkage analysis settings have an effect on how the data are analyzed.

Results

Notice how when just the distance measure is changed (Fig 1 & 2), the basic overall structure is similar, however; on closer examination the fine structure of ordering within the main clusters differs. The samples with "3" as the first character in the file name are grouped, as are the samples with the number "4". The sole "7" sample is grouped in its own cluster in both examples. These results are as expected.

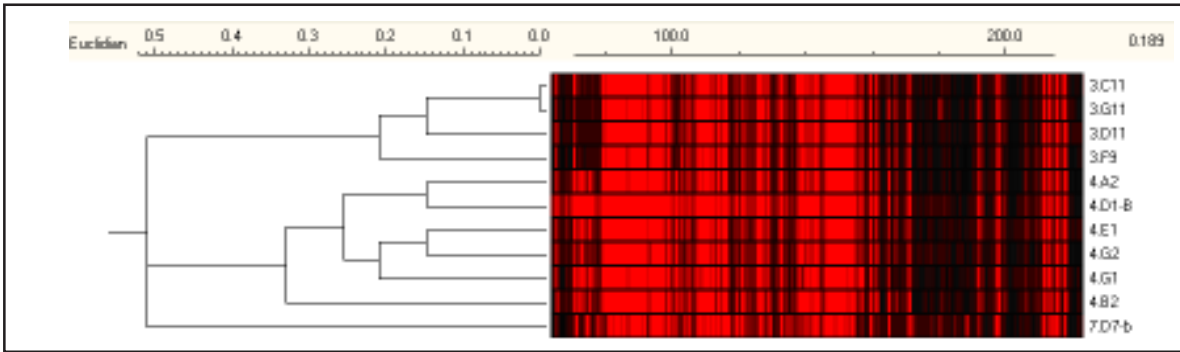


Fig. 1 Euclidean Distance Single Linkage

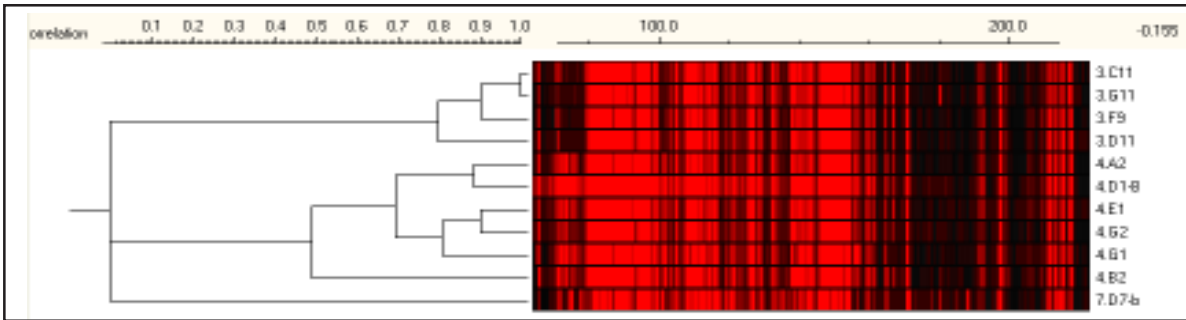


Fig. 2 Correlation Coefficient Single Linkage

When altering the analysis based just on linkage type and holding the distance measure constant (Figs 2-4), we see that the overall structure remains the same, however; the finer structure is greatly affected. Notice how in single linkage (Fig 2) the three main groups are independent of one another; where in complete linkage (Fig 3), the "4.B2" sample is separated from the main group. This is representative of complete linkage's tendency to form smaller, compact clusters. It can also be seen from this example how average linkage (Fig 4) is an amalgam of single and complete linkage.

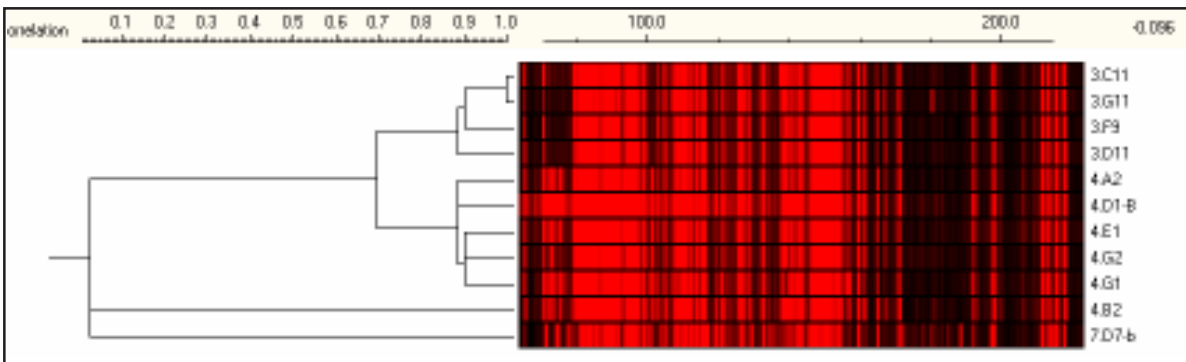


Fig. 3 Correlation Coefficient Complete Linkage

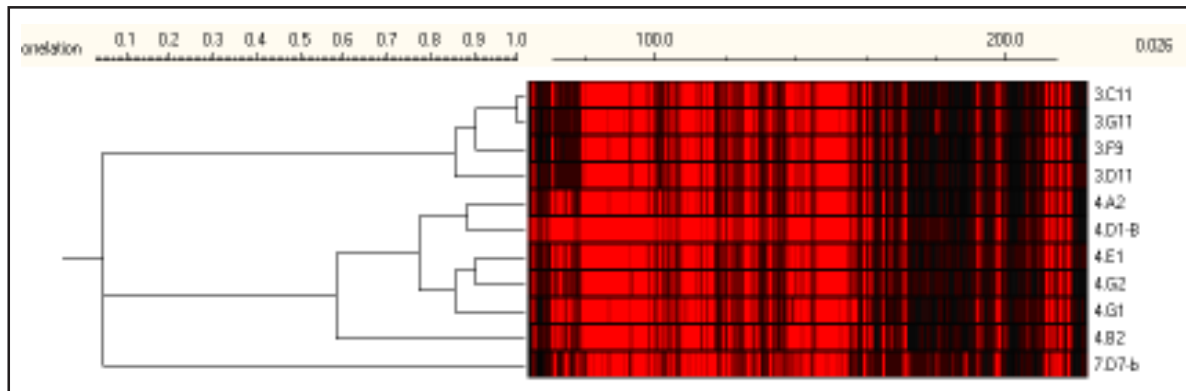


Fig. 4 Correlation Coefficient Average Linkage

In addition to dendrograms, GeneMarker outputs a Matrix Report to save as a Text (.txt) file. As mentioned in the Introduction, when the Correlation Coefficient distance measure is applied to the data, a value of 1.00 indicates a perfectly correlated pair. This can be observed in the Matrix Report where the row and column of the same sample meet (Fig 5).

	1	2	3	4	5	6	7	8	9	10	11
1	1.000	1.000	0.902	0.884	0.636	0.551	0.487	0.551	0.487	0.327	0.393
2	1.000	1.000	0.902	0.884	0.636	0.551	0.487	0.551	0.487	0.327	0.393
3	0.902	0.902	1.000	0.797	0.568	0.487	0.617	0.694	0.617	0.280	0.333
4	0.884	0.884	0.797	1.000	0.452	0.388	0.339	0.388	0.339	0.388	0.265
5	0.636	0.636	0.568	0.452	1.000	0.884	0.797	0.884	0.797	0.636	0.062
6	0.551	0.551	0.487	0.388	0.884	1.000	0.694	0.776	0.694	0.551	0.024
7	0.487	0.487	0.617	0.339	0.797	0.694	1.000	0.902	0.808	0.694	-0.007
8	0.551	0.551	0.694	0.388	0.884	0.776	0.902	1.000	0.902	0.551	0.024
9	0.487	0.487	0.617	0.339	0.797	0.694	0.808	0.902	1.000	0.487	0.163
10	0.327	0.327	0.280	0.388	0.636	0.551	0.694	0.551	0.487	1.000	0.024
11	0.393	0.393	0.333	0.265	0.062	0.024	-0.007	0.024	0.163	0.024	1.000

Figure 5 Clustering Report

Discussion

As we have seen, the linkage method and distance metric chosen produce different clustering results. The following tables demonstrate the strengths and weaknesses of each parameter. (5)

	<u>Euclidean Distance</u>	<u>Correlation Coefficient</u>
<u>Strengths</u>	Geometric interpretation	Powerful
	Retains up/down regulation scaling	Detects positive and negative correlations
	Detects magnitude of changes without scaling	Scale invariant on centered data
<u>Weaknesses</u>	Results depend on scaling used	Assumes linearity
	Cannot detect negative correlations	Susceptible to outliers

	<u>Single Linkage</u>	<u>Complete Linkage</u>	<u>Average Linkage</u>
<u>Strengths</u>	Simple analysis	Highly informative	Most commonly used
	Useful when data clusters well defined but have an irregular shape	Produces small, compact, well-defined clusters	Middle-road between the extremes of single and complete linkage
<u>Weaknesses</u>	Chaining effect - single clusters added one at a time	Does not perform well on noisy data	Measure is an average, not an actual distance, making analysis more difficult
	Sensitive to outliers	Forms many clusters	

Which parameters you choose are up to you – there is no right answer. We recommend that you apply all distance measures and linkage algorithms to your cluster analysis and look at the results to determine which method is right for your data.

Acknowledgement

We would like to thank Heidi Meudt PhD at the Museum of New Zealand, Te Papa Tongarewa, New Zealand, and Andrew Clarke at Massey University, New Zealand for their collaboration. We would also like to acknowledge Haiou Hu for her contribution in developing GeneMarker's clustering algorithms.

References

1. **Data clustering in life sciences.** Y Zhao, G Karypis. *Molecular biotechnology*. 2005. **31** (55-80).
2. **Almost Forgotten or Latest Practice? AFLP applications, analyses, and advances.** HM Meudt, AC Clarke. *Trends in Plant Science*. (in press).
3. **AFLP: a new technique for DNA fingerprinting.** P Vos, R Hogers, M Bleeker, M Reijans, T Lee, M Hornes, A Frijters, J Pot, J Peleman, M Kuiper, M Zabeau. *Nucleic Acids Research*. 1995. **23** (4407-4414).
4. **Application of Fluorescent Amplified Fragment Length Polymorphism for Comparison of Human and Animal Isolates of *Yersinia enterocolitica*.** C Fearnley, SLW On, B Kokotovic, G Manning, T Cheasty, DG Newell. *Applied and Environmental Microbiology*. 2005. **71** (4960-4965).
5. **Microarray Bioinformatics.** D Stekel. *Cambridge University Press*. 2003. (139-182).