ADVANCEMENTS IN FORENSIC DNA-BASED IDENTIFICATION

by

Gina M. Dembinski

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Biological Sciences

Indianapolis, Indiana

August 2017

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Christine Picard, Chair
Department of Biology
Dr. Mark Christie
Department of Biology - West Lafayette
Dr. Susan Walsh
Department of Biology
Dr. Stephen Randall
Department of Biology
Dr. John Goodpaster
Department of Chemistry

Approved by:
Dr. Ted Cummins
Head of the Graduate Program

ACKNOWLEDGMENTS

I would first like to acknowledge all of the people that contributed directly or indirectly to the work presented here. To all the 200 volunteers of the DNA phenotyping sample population and the 5 male volunteers of the age study, especially those willing to be my preliminary test samples from whom I asked for many swabs. To Justina Weiss for all of her work in helping with the digital photo calibrations of all the DNA phenotyping samples, you saved me many valuable hours. To Promega for their donation of the PowerPlex® kits and Rick Frey for his technical assistance in culturing the microbial species for the microbial DNA study. To Carl Sobieralski for his input and support on the microbial DNA study and also for allowing me to collaborate with the Indiana State Police laboratory on the DNA mixture study. To the Walsh lab, especially Krystal Breslin and Charanya Muralidharan, for their assistance and use of reagents in completing the methyl-RADseq sequencing runs. Many thanks also to the Notre Dame Genomics and Bioinformatics Core Facility, especially Brent Harker and Melissa Stephens for their patience and assistance in troubleshooting and assisting with the methyl-RADseq methods. Also for the methyl-RADseq work, the use of the Mason cluster is based upon work supported by the National Science Foundation under Grant Nos. DBI-1458641 and ABI-1062432 to Indiana University. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, the National Center for Genome Analysis Support, or Indiana University.

From a more personal standpoint, I would like to acknowledge my fellow graduate school colleagues, especially Anne Andere and Charity Owings for being awesome lab mates and lending moral support in the struggles of PhD life; it would have been a lot more difficult without your daily encouragements. To Dr. Christine Picard, without whom none of this would have been possible; thank you for your continued encouragement, support, and mentorship for another 4 years. Last but not at all least, I would also like to acknowledge my family for providing their unending support during my journey through graduate school and reassuring me that I can do anything I put my mind to; I would not have made it this far without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Author: Dembinski, Gina, M. Ph.D.
Institution: Purdue University
Degree Received: August 2017
Title: Advancements in Forensic DNA-based Identification
Major Professor: Christine Picard

Modern DNA profiling techniques have increased in sensitivity allowing for higher success in producing a DNA profile from limited evidence sources. However, this can lead to the amplification of more DNA profiles that do not get a hit on a suspect or DNA database and more mixture profiles. The work here aims to address or improve these consequences of current DNA profiling techniques. Based on allele-specific PCR and quantitative color measurements, a 24-SNP forensic phenotypic profile (FPP) assay was designed to simultaneously predict eye color, hair color, skin color, and ancestry, with the potential for age marker incorporation. Bayesian Networks (BNs) were built for model predictions based on a U.S sample population of 200 individuals. For discrete pigmentation traits using an ancestry influenced pigmentation prediction model, AUC values were greater than 0.65 for the eye, hair, and skin color categories considered. For ancestry using an all SNPs prediction model, AUC values were greater than 0.88 for the 5 continental ancestry categories considered. Quantitative pigmentation models were also built with prediction output as RGB values; the average amount of error was approximately 7% for eye color, 12% for hair color, and 8% for skin color. A novel sequencing method, methyl-RADseq, was developed to aid in the discovery of candidate age-informative CpG sites to incorporate into the FPP assay. There were 491 candidate CpG sites found that either increased or decreased with age in three forensically relevant

fluids with greater than 70% correlation: blood, semen, and saliva. The effects of exogenous microbial DNA on human DNA profiles were analyzed by spiking human DNA with differing amounts of microbial DNA using the Promega PowerPlex® 16 HS kit. Although there were no significant effects to human DNA quantitation, two microbial species, *B. subtilis* and *M. smegmatis*, amplified an allelic artifact that mimics a true allele ('5') at the TPOX locus in all samples tested, interfering with the interpretation of the human profile. Lastly, the number of contributors of theoretically generated 2-, 3-, 4-, 5-, and 6-person mixtures were evaluated via allele counting with the Promega PowerPlex® Fusion 6C system, an amplification kit with the newly expanded core STR loci. Maximum allele count in the number of contributors for 2- and 3-person mixtures was correct in 99.99% of mixtures. It was less accurate in the 4-, 5-, and 6-person mixtures at approximately 90%, 57%, and 8%, respectively. This work provides guidance in addressing some of the limitations of current DNA technologies.

CHAPTER 1. INTRODUCTION

Forensic biology encompasses the application of science to the identification and individualization of biological materials found at crime scenes. Although serological techniques are important to be able to identify the types of biological material at hand, most forensic biology analyses focus on DNA. As technology has advanced, the ability to amplify DNA from even a single cell is possible [1]. Increased sensitivity can be an advantage to successfully amplify DNA from even the smallest source, which may be useful in a case that has few items of biological evidence, however, this increased sensitivity has consequences. In some cases, the DNA profile successfully amplified from biological evidence does not get a hit on a suspect or a DNA database and therefore results in a dead end. Also, the co-extraction of other sources of DNA such as bacteria or other individuals that may or may not be associated with the crime can amplify a mixture profile. The work presented here aims to address these consequences by the development of a phenotyping assay that can predict the outward appearance of individual (eye color, skin color, hair color, ancestry, and potentially age) from the DNA as additional intelligence information, an evaluation of the effects of co-extracted microbial DNA on human DNA profiles, and an evaluation on the estimation of the number of contributors from mixture DNA profiles using one of the newly expanded amplification kits.

The use of DNA for individualization has advanced rapidly, especially in the last 30 years. In 1984, Sir Alec Jeffreys developed the DNA fingerprinting technique from which modern tests derive, using restriction fragment length polymorphism (RFLP) on variable number tandem repeat regions (VNTRs), which are unique repeating regions of DNA [2]. DNA fingerprinting was first used in a legal context in an immigration

paternity case in 1985 [3], and first used in a criminal case in 1988, both cases in the UK, leading to wide acceptance of this technique in the forensic community [4]. By late 1986, the DNA fingerprinting was applied to casework in the U.S. [5]. During the 1980s, Kary Mullis developed the polymerase chain reaction (PCR) that could amplify small quantities of DNA [6]. The first PCR-based assay utilized for forensic applications was the HLA-DQA1 locus in 1986 [7, 8]. Short tandem repeats (STRs) were adopted by laboratories in the 1990s [9]. STRs are highly variable markers that are small in length, which increased sensitivity of assays and could successfully be amplified from a range of crime scene conditions. The first national DNA database consisting of STR profiles was established by the U.K. in 1995 [9]. The U.S. closely followed with their own national database in 1998, the Combined DNA Index System (CODIS) and selected 13 STRs that would serve as core loci for DNA profiles within the system [9].

Increased sensitivity leads to a higher probability of amplifying DNA from more than one source, resulting in a mixture. Mixtures can be inherent to the context of a case, such as a sexual assault, however, there are more and more cases with ‘touch’ DNA evidence, meaning DNA collected from a surface that was just touched by an individual [10]. Mixture profiles are a challenge as deducing the number of contributors to a profile, a crucial step for an accurate interpretation, can be difficult. Mixtures between human individuals are one aspect of the challenge; extraneous DNA from the environment, such as bacteria, may also interfere with a DNA profile interpretation. Whenever DNA is extracted from a sample, all sources of DNA are isolated, not just human. The human microbiome has become a target for research [11] spawning studies

into forensic arenas such as using skin bacterial communities for identification [12] and microbial signatures to aid in determining the stages of decomposition [13].

STRs are the current standard for developing DNA profiles in forensic laboratories. As DNA databases are built on these markers, they will never become entirely obsolete; in fact the core STR loci was recently expanded from 13 to 20 to start implementation as of January 2017 [14]. Nevertheless, another type of DNA marker, single nucleotide polymorphisms (SNPs), have become a focus as additional informative markers for identification. More SNPs are required to reach the same level of discrimination as STRs [15], but modern molecular technologies can allow for that level of marker inclusion. Next generation sequencing (NGS), otherwise known as massively parallel sequencing (MPS) technologies, can detect thousands of SNPs simultaneously to allow for a larger portion of the genome to be analyzed in a single run [16, 17]. Although SNPs can be used for individual identification similar to STRs, the biggest advantage of SNPs is that they can convey phenotypic information about an individual. Externally visible traits, such as pigmentation, have been a target area of interest for forensic interests as knowing this type of information can be used to objectively gain intelligence in an investigation. Further advancements are being made to not just limit these studies to the genetic markers of DNA, but are also expanding into epigenetic markers of DNA. Age prediction, for example, has become a recent trending research area and there are already studies showing it is possible to estimate age by analyzing epigenetic markers, specifically methylation, within the human genome [18].

The work described in this dissertation encompasses many of these advanced capabilities of DNA analysis that adds proof to their value for forensic applications:

a) The development of a multiplex forensic phenotype profile (FPP) SNP assay and Bayesian Network (BN) models for prediction evaluation of eye color, hair color, skin color, and ancestry.

b) The development of methyl-RADseq, a novel method for the discovery of candidate CpG sites for age prediction across the 3 main types of forensically relevant biological material: blood, semen, and saliva.

c) An evaluation of the effect of decomposition-related microbial DNA on human DNA profiles.

d) An evaluation on the estimation of the number of contributors from theoretical mixtures of 2-, 3-, 4-, 5-, and 6-person mixtures against the expanded CODIS core STR loci using the PowerPlex® Fusion 6C system (Promega Corp., Madison, WI).

CHAPTER 2. THE DEVELOPMENT OF A FORENSIC PHENOTYPIC PROFILE (FPP) ASSAY FOR PREDICTING PIGMENTATION AND ANCESTRY

2.1 Introduction

The main purpose of forensic DNA analysis is to develop DNA profiles from biological evidence for identification. The ultimate goal is determining the source of that biological material. Conventional DNA profiles are comprised of short tandem repeats (STRs), however, in some cases, an STR profile generated from a crime scene sample does not result in any matches to suspects or to an entry in a DNA database. The crime scene sample donor remains unknown and the evidence cannot be further probative to the case. Forensic DNA phenotyping is the prediction of externally visible characteristics (EVCs) from DNA [19]. DNA phenotyping does not utilize STR markers, but rather, single nucleotide polymorphisms (SNPs), which are typically associated with informative genes of interest for the target phenotypic trait [15]. The inclusion of a DNA-based phenotypic profile would complement STR profiles by providing additional information on the outward appearance of the contributor for investigators to develop a lead. This could also be useful in identifying individuals in cases of missing persons and mass disasters [20]. Essentially, this phenotype profile is an objective, biological eyewitness and could also be used to corroborate or contradict actual eyewitness statements [20]. The EVCs currently being researched and utilized for forensic casework includes traits such as pigmentation (eye, hair, and skin color), ancestry, and more recently, age.

2.1.1 Melanogenesis

The pigmentation of eye, hair, and skin color is the result of melanin production. Melanin is produced by melanocyte cells within specialized vesicles called melanosomes (Figure 2.1) [21]. Melanin is an indole derivative of 3,4 di-hydroxy-phenylalanine (DOPA) formed from a series of enzymatic reactions involving the oxidation of tyrosine [22]. Melanin absorbs and scatters UV radiation (sun), as a mechanism to protect against UV-induced DNA damage [22]. There are two types of melanin, eumelanin (EM) and pheomelanin (PM); the main difference is the dependence of PM on the availability of cysteine or other sulphydryl compounds (Figure 2.1) [23]. EM is a brown/black pigment and PM is a red/yellow pigment. Variations in pigmentation expression can be caused by differing amounts of each type of melanin, differing concentrations, and the shape, size, and transport pathways of melanosomes from melanocytes to the targeted tissues (eye, hair, and skin) [24]. Prota et al.[24] characterized cultured iris melanocytes and found that darker eye colors have greater amounts of EM, intermediate eye colors have greater amounts of PM, and lighter eye colors have little amounts of both pigments [24].

The pathway of the formation of melanin, or melanogenesis, is illustrated in Figure 2.1. The α-melanocyte stimulating hormone (α-MSH) binds to the melanocortin 1 receptor (MC1R), a G-protein coupled receptor with 7 transmembrane domains [22]. Cyclic adenosine monophosphate AMP (cAMP) is activated by the binding of α-MSH to MC1R (G-protein dependent). Increased levels of cAMP activate protein kinase A (PKA), and increased PKA induces the microphthalmia transcription factor (MITF) [22]. MITF upregulates the transcription of tyrosinase by targeting the tyrosinase (*TYR*) gene [21]. Tyrosinase then acts on tyrosine to make dopaquinone, which is the same intermediary in both melanin types [25]. The agouti-signaling protein (ASP), transcribed

by the *ASIP* gene, acts as an agonist to the α-MSH and can bind to MC1R and block α-MSH, which limits cAMP levels and leads to favored production of pheomelanin [25]. *MATP* is involved in tyrosinase trafficking, *OCA2* encodes the P protein, a melanosomal membrane protein, and *SLC24A5* codes for a potassium dependent sodium/calcium exchange also along the melanosomal membrane [25]. It is thought that calcium activates the production of the Pmel17 protein, which is responsible for maturation of eumelanosomes [25].

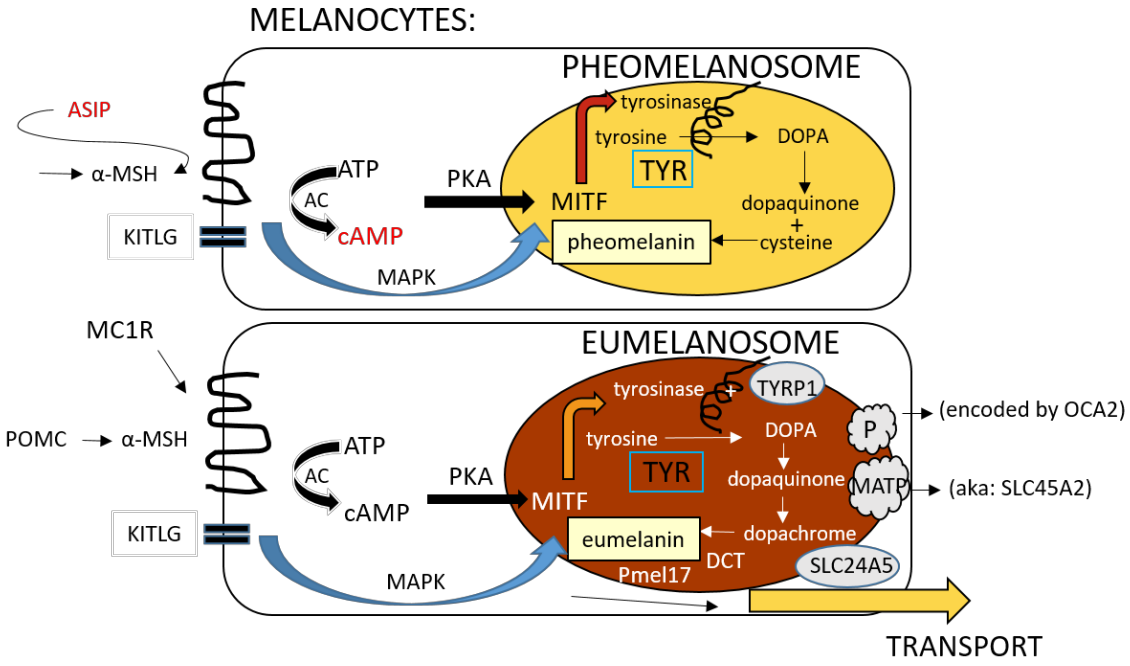


Figure 2.1 Simplified melanogenesis pathway. The production of the two types of melanin, eumelanin and pheomelanin, are illustrated. Adapted from Tully [25].

This pathway of melanin production is the same for eye, hair, and skin pigmentation. Melanocytes are present in the stromal layer of the iris. The melanin gets expressed in melanosomes within the stromal melanocytes. Melanosomes are what are primarily observed when looking at the eye color of an individual [26]. However, melanosomes must be transported from melanocytes for hair and skin cell melanin

expression. Melanocytes are oval dendritic cells that are smaller than keratinocytes, the cells that make up hair and skin [27]. For skin, melanocytes in the basal layer of the epidermis form a melanin unit, where a single unit is comprised of 1 melanocyte associated to 30-40 keratinocytes [27]. Melanosomes are transported to the keratinocytes through the dendritic ends of the melanocyte (Figure 2.2). The exact mechanism of melanosome transfer to keratinocytes is still not fully known [27], however, it is suggested that transport of melanosomes from melanocytes to keratinocytes is very similar in hair follicles. The ratio of melanocyte to keratinocyte is denser in hair follicles (1 melanocyte for 5 keratinocytes), and the dendritic ends extend between the cortical and medullar keratinocytes [27].

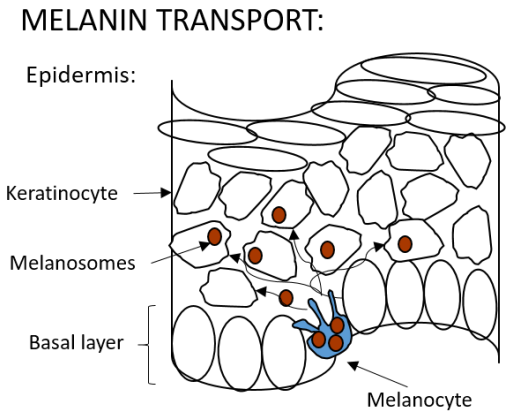


Figure 2.2 Example of Melanosome Transport. This illustrates melanosomes being transported via the dendritic ends of the melanocyte into skin keratinocytes. Adapted from Cichorek et al. [27].

2.1.2 Predicting Pigmentation: Hair, Eye and Skin Color

There have been numerous studies focused on discovering biological markers that explain normal variation in human phenotypes, specifically, DNA markers associated with pigmentation. Most DNA markers (typically associated with disease) have been

discovered through genome-wide association studies (GWAS) looking for significantly correlated SNPs [28-34]. Grimes et al. [35] published the first example of a phenotype prediction test in 2001 in which variants in the *MC1R* gene were found to be informative for the red hair phenotype [35]. *MC1R* encodes for a transmembrane G-coupled protein receptor that activates the melanin synthesis in melanocytes (see 2.1.1). Polymorphisms in *MC1R* predominantly result in the phenotype consisting of red hair, blue eyes, and fair skin [36]. Most of the pigmentation informative SNPs are associated with genes whose functions are known to be involved in the synthesis of melanin and its transport. One SNP in particular, rs12913832, is located in a conserved intronic region of the *HERC2* gene upstream from the *OCA2* gene promoter, and was found to have the highest association to predicting pigmentation traits despite the lack of the functional understanding of the *HERC2* gene [37]. This SNP has a regulatory effect on the expression of *OCA2*. *OCA2* was thought to be the highest associated gene related to pigmentation up until the discovery of the *HERC2* silencing interaction. Specifically, rs12913832 has been found to be the main causative SNP for the lack of brown eye color expression [38].

Initial successful work in predicting eye color was based on SNPs analyzed using a multinomial logistic regression model to predict eye color [39]. A multiplex assay was developed using statistically correlated SNPs isolated from a GWAS for eye color, termed IrisPlex [20], which accurately predicted blue and brown eye colors in a small European population (N = 40, 91.6% and 87.5%, respectively). The predictions were based on six SNPs: rs12913832 (gene: *HERC2*), rs1800407 (*OCA2*), rs12896399 (*SLC24A4*), rs16891982 (*SLC45A2*), rs1393350 (*TYR*), and rs12203592 (*IRF4*). We

evaluated the IrisPlex assay in a Midwest United States population, a population consisting of a greater proportion of ancestral descent variation than the original study (N=200), and genotypes using the same model suffered in accuracy with respect to brown eye color prediction, with 91% and 79% for blue and brown eye color, respectively [40, 41]. The reduction in accuracy of brown eye color predictions is likely due to an increase in the heterozygosity of the *HERC2* SNP (rs12913832) in a more admixed population. Furthermore, a weakness is inherent in the prediction of the intermediate eye color (not brown or blue), thus additional SNPs are necessary to improve prediction in those eye color categories [42]. Towards that goal of improving intermediate eye color predictions, a study by Pośpiech et al. [43] focused on the interactions between genes that could contribute to eye color expression and found an additional SNP, rs1408799 (*TYRP1*), that was associated with green eye color when considering its interaction with *HERC2*. Following the discovery of these eye color SNPs, Branicki et al. [44] developed a DNA-based model for genotypes of a set of 13 SNPs to predict hair color in 4 categories: black, brown, red, or blonde, with prediction accuracies of 87.5%, 78.5%, 80%, and 69.5%, respectively. This study led to the inclusion of these significantly associated hair color SNPs in a multiplex assay combining the eye and hair color SNPs, termed HIrisPlex [45]. HIrisplex includes the original 6 IrisPlex SNPs along with 18 additional SNPs that are correlated to hair color variation. Among these 24 SNPs, 11 are associated with the *MC1R* gene, polymorphisms that interact with each other through what is called compound heterozygosity; and some SNPs have a stronger effect on the phenotype (penetrance) than others [44]. The initial results of the HIrisPlex genotyping assays revealed accuracy for hair color prediction in European populations.

Because phenotypes resulting from pigmentation genes are related, many of the above-mentioned SNPs can be used in tandem to predict skin pigmentation. Han et al. [31] found that rs12203592 (*IRF4*) and rs12896399 (*SLC24A4*), both included in the IrisPlex panel for eye color, are associated with the tanning response of skin. Another study by Lamason et al. found rs1426654 (*SLC24A5*) to be correlated to light skin pigmentation for Europeans [46]. Interestingly, there is evidence that light skin pigmentation derived separately in European and East Asian populations, and a SNP associated in the *OCA2* region, rs1800414, was found to be only informative of light skin in East Asian population [47]. A study by Maroñas et al.[48] screened African and European populations from the 1000Genomes database and found 10 significant SNPs for skin color differences based on allele frequencies to test in their own dataset of individuals: rs1426654 (*SLC24A5*), rs6119471 (*ASIP*), rs6058017 (*ASIP*), rs1408799 (*TYRP1*), rs1448484 (*OCA2*), rs16891982 (*SLC45A2*), rs10777129 (*KITLG*), rs3829241 (*TPCN2*), and rs13289 (*SLC45A2*) [48]. A SNP located upstream of the *KITLG* gene, rs642742, is found to be in a conserved region and shows skin pigmentation differentiation between Europeans and West Africans [49]. Another study had already found the same *ASIP* SNP, rs6119471, to be informative for the expression of darker skin (greater melanin) [50]. The *ASIP* SNP along with 6 additional SNPs were used to predict both eye and skin color in diverse populations, although accuracy was observed in the European samples and only when homozygous genotypes were used for skin color predictions (without evaluating heterozygote genotypes) [50]. Hart et al. [51] developed an assay using 8 SNPs for eye and skin color prediction which included 4 of the IrisPlex SNPs: rs12913832 (*HERC2*), rs12203592 (*IRF4*), rs12896399 (*SLC24A4*), rs1426654

(*SLC24A5*), rs16891982 (*SLC45A2*), rs885479 (*MC1R*), rs6119471 (*ASIP*), and rs1545397 (*OCA2*). These results were impressive with 1% error in their test set for skin color prediction (light vs dark), however, this was not an accurate measure of error, similarly, as in the study by Spichenok et al. [50], it is based on homozygous genotypes only; they reported inconclusive results for heterozygous outcomes.

2.1.3 Predicting Ancestry

There has been a wealth of research in finding ancestry informative SNPs, also known as ancestry informative markers (AIMs), for determining the genetic ancestry of an individual [52-59]. One of the earlier ancestry assays to be considered for forensic use was the SNP*for*ID assay that incorporated 34 AIMs for the distinction of three major population groups: Europeans, Africans, and East Asians [56]. Phillips et al. [58] developed a 23-SNP assay, termed Eurasiaplex, which distinguishes European and South Asian ancestries. For a more comprehensive analysis of genetic ancestry, Kidd et al. [57] found a set of 128 AIMs to be useful in discriminating between samples from 119 different populations. The Kidd lab [60] created and maintains an online frequency database, the ALlele FRequency Database (ALFRED) that contains SNP frequency data for many populations, including AIMs and other phenotypic informative SNPs [61].

Previous ancestry studies have either focused on specific population discrimination (e.g., Nigerian vs. Kenyan) or more broad continental distinctions (e.g., European vs. African). Halder et al. [62] conferred a panel of 176 AIMs for admixture discrimination at the major four continental origins: European, West African, Indigenous American, and East Asian, and found it to be most informative for discriminating African vs non-African and least informative for Indigenous American and East Asian. Nassir et

al. [63] evaluated 93 AIMs for determining continental origin, and found this panel was particularly useful in providing admixture proportions in even the largest admixed population groups – for example, African-American and Mexican-American in the United States. Nievergelt et al. [55] developed a discriminative AIM panel for geographical origin using 41 SNPs with distinction between seven continental populations: Africa, Middle East, Europe, Central/South Asia, East Asia, the Americas, and Oceania. Kosoy et al. [64] found that using a subset of 24 SNPs, originally isolated from a panel of 128 SNPs, had the same power of discrimination for continental ancestry groups, highlighting that more SNPs are not necessarily more powerful for such a broad level of ancestral discrimination. In general, there are few SNPs that overlap between published panels, mostly because of the methodology used in selecting SNPs, and because the purpose of panels may differ (regional vs. more specific) [57]. Kidd et al. recently published a panel of 55 SNPs combining the data from previous published panels, efficient for distinction of eight biogeographic regions [65]. This is further evidence that panels with a smaller number of SNPs continue to be developed with similar powers of discrimination.

As previously mentioned, pigmentation traits are sometimes associated with ancestral origins, geographically speaking, and some of the same genes from pigmentation studies may also be useful for ancestry studies, therefore reducing the number of overall SNPs that would need to be included into a multiplex assay. This relationship exists because the differences in melanosome qualities can be distinctive for different ethnic groups (Figure 2.3) [66]. Generally, the number of melanocytes are the

same, but the manner in which melanosomes are transported and distributed to the keratinocytes, as well as the quantity and shape, differs [66].

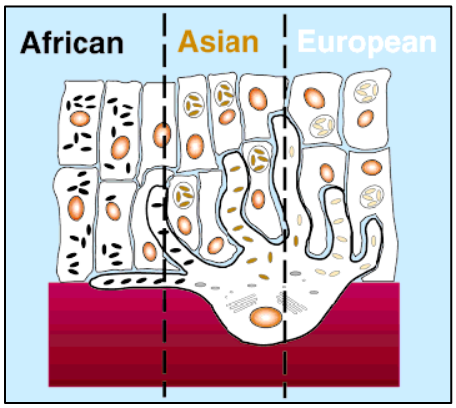


Figure 2.3 Differences in melanosome distribution attributed to ethnic groups. Reprinted with permission from Sturm et al. [66].

For example, rs1426654 (*SLC24A5*) is informative for European ancestry as it is an informative marker for light skin derived in Europeans [56]. The SNP rs16891982 (*SLC45A2*) is used in IrisPlex for eye color and is also useful for European ancestry [58], and rs6119471 (*ASIP*) is informative for African ancestry and relates to dark skin expression [50]. Because of the inter-relatedness of each of these traits to genes specifically related to pigmentation, it may be possible to develop a multiplex SNP assay that is capable of predicting each of these traits with some degree of accuracy without the need for hundreds of SNPs.

2.1.4 Quantitative Color Classification

Most of the studies discussed in the previous sections had based their color phenotype predictions on qualitative determinations – where human observers visually evaluated the samples to determine the color. A more objective way to evaluate each phenotype color is to measure it quantitatively; this will reduce the visual perception

differences of human observers. Color is a function of 3 variables, known as matching stimuli or tristimulus values, and can be designated as: Red (R), Green (G), and Blue (B). This is partly based on the fact that there are 3 types of spectrally different cones in the retina of the eye for perceiving color [67]. Brightness is the attribute of visual perception in which something appears to exhibit more or less light, and hue is the attribute to which something appears to be similar to red, yellow, green, and blue – the basic chromatic colors [67]. The relative colorfulness of an object in proportion to the brightness is called chroma [67]. The Commision Internationale de l'Éclairage (CIE) is an international commission responsible for standardization of lighting, color, imaging, and vision [68]. This commission creates different standardized color models that can describe color quantitatively. Three CIE color models were considered to objectively measure the color in this work: RGB, XYZ, and LUV. The RGB space is device dependent and therefore its range can vary depending on which type of display is transmitting the color [69]; however, it can be advantageous to be able to represent colors on technological displays such as computer or television screens. RGB color is similar to the XYZ color space, although XYZ is device independent. These color spaces can be normalized to reduce three variables to two, for example, XYZ can be normalized to create the xyY space which corresponds to relative tristimulus values, or chromaticity coordinates. Y remains the same as it still correlates with the brightness or luminance [67]. A chromaticity diagram shows the proportion of the tristimulus values in two dimensions and therefore a plot of all possible color chroma (Figure 2.4). Light and dark colors plot at the same point in a chromaticity diagram if they have the same ratio [67].

Also for this reason, white, black, and all shades of gray are represented by the same point in the center (S_E in Figure 2.4).

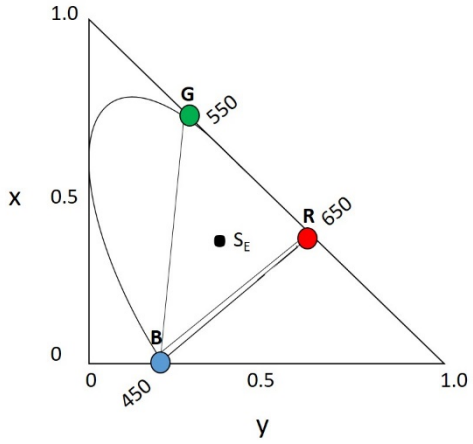


Figure 2.4 A Simple Chromaticity Diagram. The xy and RGB spaces are shown in relation to each other. Adapted from Hunt [67].

All color coordinates can be converted between color spaces, so why consider different color models? Some color spaces have certain advantages or disadvantages depending on the proposed use. For example, the xy chromaticity diagram does not have a uniform distribution of colors (the green gamut is larger than purple), and therefore the normalized LUV (Lu'v'), diagram was developed as it reduces this distortion [67]. The u'v' diagram is useful for showing the relationships between colors and their discrimination [67]. And as mentioned, the RGB space represents color on digital displays, which are an important tool for viewing digital photographs and images.

In this work, the RGB, XYZ, and LUV color models were evaluated as possible models to quantitatively measure color from digital eye, hair, and skin photos of the sample population. Consensus qualitative (visual) determinations were also collected and used to gauge the accuracy of the quantitative measurements to human perception.

If a quantitative color model can accurately reflect predicted color, it eliminates human perception bias and can then reliably corroborate or contradict actual eye witness accounts.

2.2 Methods

2.2.1 Sample Collection

Buccal swabs, hair, digital photos, and a self-reporting survey were collected from 200 volunteers who were randomly assigned 4 digit numerical identifiers (Indiana University IRB Approved Protocol #1407693464). Digital photos included a profile picture and an upper inner arm picture taken with a SpyderCHECKR 24 color calibration card (Datacolor, Lawrenceville, NJ) from a fixed distance of 130 cm using a Canon EOS Rebel T5 digital camera using aperture setting of F5.6, 1/25 shutter speed, ISO 800, and enabling flash. Hair samples included plucking or cutting of 2-3 hair strands near the crown of the head and stored in individual plain white envelopes labeled with the sample number. Buccal swabs were stored at -20°C until DNA extraction. A copy of the self-reporting survey can be found in Appendix A.

2.2.2 DNA Extraction and Quantitation

An organic DNA extraction protocol was used to isolate DNA from the collected buccal swabs. Each buccal swab tip was incubated in a 1.5mL tube at 56°C in 500 µL of ChargeSwitch lysis buffer (Thermo Fisher Scientific, Waltham, MA) with 25 µL of proteinase K (Thermo Fisher Scientific) for 8-12 hours. After incubation, the swab was placed into a spin basket in the tube and centrifuged to remove all liquid from the swab; the swab was subsequently discarded. To the tube, 500 µL of phenol: chloroform:

isoamyl alcohol (25:24:1) (Thermo Fisher Scientific) was added and mixed by inversion. The tube was centrifuged for 1 minute at 13,000rpm. The aqueous layer was removed into a new 1.5 mL tube and the remaining layer was discarded. To the new tube, 500 µL of chloroform: isoamyl alcohol (24:1) (Thermo Fisher Scientific) was added and inverted to mix. The tube was centrifuged for 1 minute at 13,000 rpm. The aqueous layer was removed into a new 1.5mL tube and 25 µL of 0.2M NaCl and 500 µL of cold 95% ethanol was added. The tube was centrifuged at 4°C for 15 minutes at 15,000 rpm. The liquid was removed, being careful not to disturb the pellet on the bottom of the tube. The pellet was washed with 500 µL of 70% ethanol and centrifuged at 4°C for 5 minutes at 15,000rpm. The liquid was removed and the pellet was allowed to air dry for 20-30 minutes. The DNA was re-suspended in 50 µL of TE buffer (Thermo Fisher Scientific) and stored at -20°C. Quantitation of the extracted DNA was performed with the Quantifiler® Human DNA quantification kit (Thermo Fisher Scientific) following the manufacturer's protocol. The DNA samples were diluted to working solutions of 1 ng/µL.

2.2.3 Digital Color Measurement

All collected digital photos were color calibrated using SpyderCHECKR 1.2.1 software (Datacolor) in Adobe Lightroom 5 (Adobe Systems, San Jose, CA). The collected hair strands were cut to mount 1-2 hairs onto microscope slides using Permount® fixative (Thermo Fisher Scientific). Microscope photos were taken with Leica DFC290 HD digital camera (Leica Microsystems Inc, Buffalo Grove, IL) on an Olympus BX51 microscope (Olympus, Waltham, MA) at 100X magnification. A ChromaCal™ microscope calibration slide (Datacolor) was also used to apply color

correction to the photos, as well as ChromaCal™ monitor calibration (Datacolor). This monitor-calibrated computer was further used by 6 anonymous photo raters, 3 male and 3 female, to rate the visual color of all digital photos for hair, eye, and skin color for each of the 200 samples to determine a visual consensus rating. Where there was a split rating (3 votes for one color, 3 for another), the self-reported color was used as a 7th rating. RGB values were measured from the extracted digital photos using Adobe Photoshop Elements (Adobe Systems). For eye color, only the iris portion was extracted and the average color of the entire extracted iris was digitally measured. For skin color, an approximate 1x1 inch square was extracted from the inner arm area and the average color was measured. For hair color, an approximate 1 inch section was cut from 2 of the hair strands near the root from each individual (3 of the samples had 1 strand collected) and the average color was measured. The RGB values were averaged between the two strands (if two were able to be collected). Color model transformations were performed in R v3.1.3 [70] using the *grDevices* package. The following equations were used to transform the xy to u'v' coordinates:

$$u' = \frac{4X}{X+15Y+3Z} \qquad \text{Equation 2.1}$$

$$v' = \frac{9Y}{X+15Y+3Z} \qquad \text{Equation 2.2}$$

2.2.4 Multiplex Phenotypic Assay

The 24 pigmentation and ancestry informative SNPs chosen from literature for inclusion in this study are shown in Table 2.1 and Table 2.2, respectively. The eye and

hair SNPs were primarily chosen from the HIrisPlex assay [45], but as a reduced set where some of the less influential SNPs were not included (i.e., rs1393350 for eye color was excluded). And to conserve panel space for other trait SNPs, only one of the higher penetrating *MC1R* SNPs was chosen from the 11 *MC1R* SNPs included in HIrisPlex. As previously discussed, some SNPs are informative for more than one trait, especially the skin and ancestry SNPs. For the ancestry SNPs, at least 2 SNPs were chosen for the main continental groups being evaluated: African, European, Asian, and Amerindian (Hispanic). The additional SNPs used for ancestry prediction were in common with the skin informative SNPs. The chosen SNPs were also based on successful primer design and allele discrimination. The first step in building the assay was to get successful single SNP discrimination with designed allele-specific primers. Allele-specific PCR (AS-PCR) involves the design and use of 3 primers and utilizes size discrimination for genotyping (Figure 2.5). Two of the primers are allele-specific (AS) and therefore mutually exclusive for amplification of the two possible bases at the SNP site. The allele specific primers are complementary to the DNA directly adjacent to the SNP site, with the 3' base of the corresponding to the target SNP. A poly T-tail of 4 or 10 base pairs (bp) was added to one of the AS primers for size discrimination between the two alleles. The third primer is located further upstream or downstream of the SNP site as a common primer that pairs with both allele-specific primers. Therefore, for each forensic phenotypic profile (FPP), each homozygous genotype is represented by a single peak discriminated by different fragment sizes, differing by either 4 or 10 bp, and a heterozygous genotype is represented as two peaks separated by 4 or 10 bp (Figure 2.5). PCR products in the assay range from 120-365 bp.

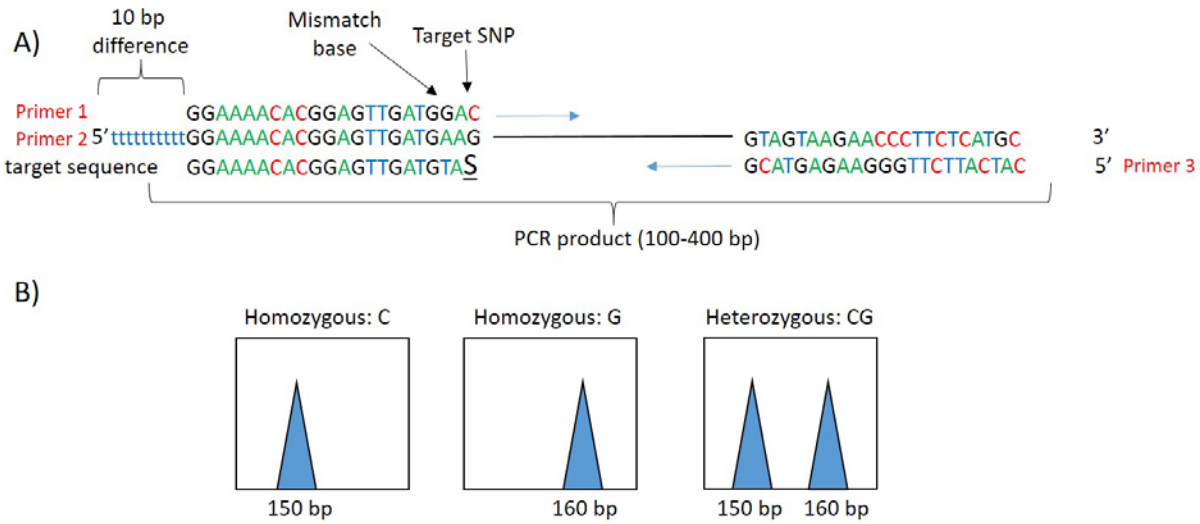


Figure 2.5 Allele-specific PCR design. A) This illustrates the design and location of the allele-specific primers (Primer 1, Primer 2) in relation to the common primer (Primer 3) along the target SNP sequence. B) The electropherogram peak output interpretation from this primer design.

Table 2.1 Pigmentation informative SNPs used in this study.

SNP	Gene	Pigmentation Trait	Previous Studies
rs12913832	*HERC2*	Eye, Hair, Skin	[20, 37-39, 44, 45, 71-73]
rs1800407	*OCA2*	Eye, Hair	[20, 39, 43-45, 71, 72, 74]
rs12896399	*SLC24A4*	Eye, Hair	[20, 39, 44, 45]
rs16891982	*SLC45A2*	Eye, Hair, Skin	[20, 39, 44, 45, 48, 71-73]
rs12203592	*IRF4*	Eye, Hair	[20, 39, 44, 45]
rs3829241	*TPCN2*	Eye, Hair, Skin	[43, 44, 48]
rs1408799	*TYRP1*	Eye, Hair, Skin	[43, 44, 48, 75]
rs1805007	*MC1R*	Hair	[44, 45, 74, 76]
rs28777	*SLC45A2*	Hair	[45]
rs12821256	*KITLG*	Hair	[44, 45, 77]

Table 2.1 continued

rs4959270	*EXOC2*	Hair	[45]
rs2378249	*ASIP*	Hair	[44, 45]
rs683	*TYRP1*	Hair	[44, 45, 76]
rs6119471	*ASIP*	Skin	[48, 50, 51]
rs1800414	*OCA2*	Skin	[47, 78]
rs10777129	*KITLG*	Skin	[48]
rs1426654	*SLC24A5*	Skin	[48, 73, 79, 80]

Table 2.2 Ancestry informative SNPs used in this study.

SNP	**Gene**	**F_{ST} Value[65]**	**Ancestry Trait**	**Previous Studies**
rs1426654	*SLC24A5*	0.73	European	[56, 65, 79-82]
rs3827760	*EDAR*	0.71	E. Asian	[59, 65, 80, 82, 83]
rs16891982	*SLC45A2*	0.69	European	[56, 65, 79, 80, 82]
rs3916235	*CD226*	0.63	African	[65, 82]
rs4918664	*intergenic*	0.53	Amerindian/Asian	[55, 65, 82]
rs12498138	*GOLGB1*	0.48	Amerindian	[55, 65]
rs3737576	*S1PR1*	0.44	Amerindian	[64, 65, 80]
rs1229984	*ADH1*	0.43	E. Asian	[64, 65, 82]
rs6119471	*ASIP*	--*	African	[50, 51]
rs1800414	*OCA2*	0.57	E. Asian	[65, 78, 80]
rs12913832	*HERC2*	0.52	European	[65, 80, 82]
rs7657799	*TACR3*	0.44	African	[64, 65]

* = not found

Primer concentrations were optimized for each SNP set within each dye channel. There were 3 PCR multiplex reactions per sample, one for each dye color set. Qiagen Multiplex PCR Master Mix (Qiagen, Hilden, Germany) was used in the PCR setup for the AS-PCR reactions. Touchdown PCR was used to perform PCR, which can help increase specificity of primer annealing, where the annealing temperature in the first 6 cycles differs and decreases by 1°C. The optimized PCR conditions are listed in Table 2.3. All PCR was performed on an Eppendorf Mastercycler Pro thermal cycler (Eppendorf, Hamburg, Germany). PCR products were separated on an ABI 3500 Genetic Analyzer (Thermo Fisher Scientific) following the same injection parameters as listed by the manufacturer for the Identifiler Plus kit (Thermo Fisher Scientific). The FPP profile electropherograms generated for all 200 samples are in Appendix B.

To ensure accuracy of the FPP assay genotypes, a subset of 20 samples were sequenced at all loci. Initial PCR for the sequencing was setup according the PCR cycling conditions listed in Table 2.3 and the primer sequences used are listed in Appendix C. Samples were then treated with ExoSAP-IT (Thermo Fisher Scientific) before cycle sequencing was performed. Cycle sequencing reactions were set up using the BigDye Terminator v3.1 Cycle Sequencing kit (Thermo Fisher Scientific) following the manufacturer's protocol. The sequencing products were cleaned with the BigDye XTerminator purification kit (Thermo Fisher Scientific) following manufacturer's protocol. Cycle sequencing was detected on an ABI 3500 Genetic Analyzer (Thermo Fisher Scientific).

Table 2.3 PCR thermal cycling conditions.

PCR programs						
AS-PCR:	Temperature, Time Cycles (#):	95°C, 15 min	94°C, 30s 61-56°C, 40s 72°C, 1:35 min 1 cycle each	94°C, 30s 55°C, 40s 72°C, 1:35 min 25 cycles	72°C, 15 min	4°C, hold
Sequencing PCR:	Temperature, Time Cycles (#):	95°C, 10 min	95°C, 30s 59°C, 30s 33 cycles		59°C, 5 min	4°C, hold
Sequencing PCR (MC1R):	Temperature, Time Cycles (#):	95°C, 10 min	95°C, 30s 59°C, 35s 72°C, 40s 32 cycles			4°C, hold

2.2.5 Statistical Models

The goal of DNA phenotyping is to determine a genotype that can accurately infer the phenotype of the individual. These inferences are determined from statistical models. Model building methods can be used to find the best fitting model to describe the relationship between the outcome and predictor variables [84]. Three types of statistical analyses were performed: discriminant analysis; Bayesian networking; and multinomial logistic regression.

2.2.5.1 Discriminant Analysis

Linear discriminant analysis (LDA) is a supervised technique that was first described by Fisher in 1936 [85]. It is a method that describes, classifies, and predicts multivariate data into separated qualitative groups. A new set of axes that best separates the data into groups is created, which are also called canonical variates, and are just linear combinations of the original values [85]. LDA places members of the same group as close as possible together while at the same time moving all other groups as far as possible by maximizing the variance between groups by the variance within groups [85].

LDA was performed on the RGB, xyY, and Lu'v' color coordinates using the consensus color ratings as the classifications for the samples. Discriminant analysis was performed using JMP Pro 12 (SAS Institute Inc., Cary, NC).

2.2.5.2 Bayesian Networks

A Bayesian Network (BN) is a specific type of directed acyclic graph (DAG). It is a pictorial representation of the probabilistic relationship between variables based on a set of conditional probability distributions [86]. Nodes represent the variables and the directed arcs represent the dependencies between them. There is a conditional probability distribution associated with each node. Simple Bayesian classifiers (naïve Bayes) assign a class label based on the set of attributes, in this case, the SNP genotypes. The probability theory used to compute the posterior probabilities in these models is based on Bayes' Theorem [87, 88]. BNs have been applied to clinical studies to analyze variables in disease developments [89] and have been applied in forensic phenotype predictions, for example, eye color [40, 90]. Figure 2.6 shows the basic equation of Bayes' Theorem as it applies to phenotype predictions. Discrete BNs were designed and built within R v3.1.3 [70] using the *bnlearn*, and *gRain* packages. Each of the 24 SNPs was a child node of the parent trait (eye color, hair color, skin color, ancestry). Additionally, alternative models where ancestry was considered as an additional child to each of the pigmentation traits were also tested.

For the continuous RGB coordinate predictions, hybrid BN models and functions were designed and performed by using the *HydeNet* package within R v3.1.3[70]. The SNP inputs were binary coded as 0, 1, and 2 corresponding to each possible genotype (0 and 2 as the homozygotes, and 1 as the heterozygote). The functions generated were

used in Microsoft Excel to calculate the predicted R, G, and B values. Absolute error was calculated between the predicted and actual digitally measured RGB values using the distance formula as follows:

$$\sqrt{\{(predR-R)^2 + (predG- G)^2 + (predB- B)^2 \}} \qquad \text{Equation 2.3}$$

Percentages were calculated out of the total possible error. RGB values range from 0-255, and therefore the maximum error is 441.673.

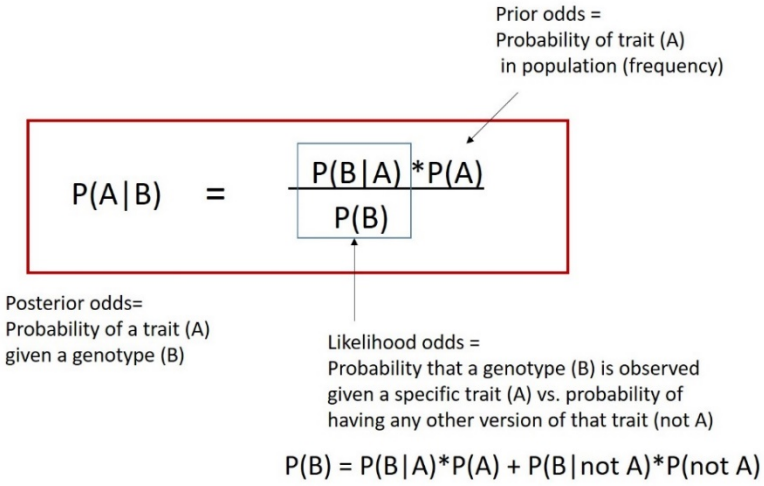


Figure 2.6 Simplified Bayes' Theorem applied to phenotype predictions. The posterior odds are the frequency of the trait given the observed genotype based on the prior odds of the trait within the population and the likelihood of all the other possible versions of that trait.

2.2.5.3 Multinomial Logistic Regression

Logistic regression involves relating a categorical response (e.g., eye color) to a set of explanatory variables (e.g., SNPs). Multinomial logistic regression (MLR) involves more than 2 categories as possible outcomes. A set of parameters are estimated that fit the maximum likelihood of the observed data [84]. These parameters can then be used to estimate the probability of classification of new observations into each of the

possible categories. The probability outcomes may be evaluated using a certain threshold to estimate the goodness of fit of the model to the data. MLR has been previously used in pigmentation prediction models, especially in the IrisPlex and HIrisPlex assays [20, 45] and it does not need prior known data, e.g., ancestry, to generate model parameters. MLR was performed using MATLAB R2016a (The MathWorks Inc., Natick, MA) and macros were modified as designed by Walsh et al. [20, 45] to adjust for the sample population allele frequencies using Microsoft Excel to determine the probabilities of the validation set for each pigmentation category using similar models as with the BN analysis for comparison.

2.2.5.4 Model Evaluation

Evaluation of the each prediction model performance was done by generating confusion matrices of the validation set to measure the area under the receiver operating characteristic (ROC) curve (AUC), which utilizes the sensitivity and specificity of the model performance. Sensitivity is the true positive rate and specificity is the true negative rate [91]. A ROC curve shows the true positive rate at varying false positive thresholds (which is 1- specificity) [91]. The AUC of this curve is a value usually between 0.5 and 1; AUC values near 1 are accurate predictors whereas AUC values near 0.5 are characteristic of a model with no predictive value and therefore not any more informative for prediction than random guessing [91].. Model evaluations were measured using the *caret* and *ROCR* packages in Rv3.1.3 [70]. A model prediction threshold of 70% was selected for two reasons: 1) the highest optimal average threshold across all pigmentation models (optimal being the point balanced between a high true positive rate and low false positive rate) was determined to be approximately 65%, and 2)

a 70% threshold was found optimal in previously designed and tested pigmentation models (e.g., IrisPlex [20, 40] and HIrisPlex [45]) and using the same threshold here would allow for direct comparison.

2.3 Result and Discussion

2.3.1 Sample Collection

An analysis of the 200 individuals based on the self-reported categories was performed to examine the overall composition of each phenotype for each trait being evaluated for predictions. Table 2.4 shows the distribution of the sample population. There was a high proportion (68%) of individuals who identified as European. Additionally, when looking at the distributions of the possible phenotypes, Europeans have the most eye and hair color diversity; they are the only individuals who reported having blue eye color and red or blonde hair (Table 2.4). This is not surprising as blue eye color is proposed to have originated from Europe [37]. FPP profiles of all 200 samples can be seen in Appendix B.

Table 2.4 Sample population trait frequencies based on self-reported data for pigmentation and ancestry.

Ancestry (N)	Eye			Hair				Skin		
	BLU	BRN	INT	BRN	RED	BLD	BLK	LIT	INT	DRK
European (135)	35%	27%	38%	70%	7%	21%	2%	90%	10%	--
African (16)	--	91%	9%	55%	--	--	45%	9%	36%	55%
Other Asian (12)	--	100%	--	8%	--	--	92%	--	75%	25%
Hispanic (15)	--	87%	13%	27%	--	--	73%	26%	67%	7%
East Asian (7)	--	100%	--	29%	--	--	71%	43%	57%	--

Legend: BLU= blue, BRN = brown, INT= intermediate, BLD= blonde, BLK= black, RED= red

2.3.2 Discriminant Analysis Color Measurement

Three color models were considered: RGB, xyY, and Lu'v'. LDA was performed on the RGB coordinates extracted from the digital photos of the eye, hair, and skin color. Three categories were evaluated for eye color – blue, intermediate, or brown. Intermediate colors are those not classified as blue or brown and would include colors such as green or hazel. Skin color also had three categories – light, intermediate, and dark, where intermediate color was similar to eye color in that it represents those not classified as light or dark. Hair color was evaluated in four categories – black, blonde, brown, or red. For eye and skin color, the training set was composed of 150 of the 200 samples, and the remaining 50 samples were used as the validation set, no overlap of samples occurred between the two sets. Because there were some samples without hair samples (bald individuals) or where the hair collected was gray or white, those samples were not considered and therefore for hair color, there were 148 samples in the training set, and 46 in the validation set. The RGB coordinates were classified with 2 different color classification groups: self-reported (SR) color and consensus color. The consensus color had fewer misclassifications than SR color: 12% vs. 22% for eye color, 8% vs. 24% for skin color, and 30% vs. 38% for hair color (Consensus-RGB vs SR-RGB in Tables 2.5-2.7). Evaluating concordance of the consensus ratings from the 6 individuals, there was 82% concordance for hair color, 80% for eye color, and 89% for skin color; this generally equated to 4 or 5 out of 6 individuals rating the color in the same category. Therefore the consensus classifications were used to evaluate the RGB color model against the xyY and Lu'v' color models. The results are summarized in Tables 2.5-2.7 for eye color, skin color, and hair color, respectively. Hair color was the most difficult category to classify, especially between brown and red. This may be due to the method

of hair and digital photo collection as only 1 or 2 strands of hair were used for the color ratings. Comparing all 3 color models between all the traits, the Lu'v' model was chosen as the optimal quantitative classification scheme as it had the similar percentage of overall misclassifications across all traits as the RGB model, but is device independent and a more uniform color scale than RGB (see section 2.1.4) (Tables 2.5-2.7). The LDA Lu'v' classifications were henceforth used as the known color categories for the samples for evaluating the prediction models.

2.3.3 Pigmentation Prediction Model Evaluation

Bayesian networks were built for each pigmentation trait and ancestry using the discrete Lu'v' categories as determined by LDA. There were 4 different BN models developed and tested. The first BN model consists of a naive Bayes classifier using all 24 SNPs (all SNPs) as child nodes with arrows directed from each trait parent node, an example of which can be seen in Figure 2.7. The second and third BN models have a reduced number of SNPs: the second model (pigment SNPs only) uses the 17 SNPs as found informative for any of the pigmentation traits (Figure 2.8) and the third model was further reduced in the number of SNPs to those only found informative for each pigmentation trait in the literature (trait SNPs only) (Figure 2.9). There were 10, 7, and 6 SNPs for hair, eye, and skin color, respectively.

As previously discussed, pigmentation can be influenced by ancestry. Therefore the fourth model considered for the pigmentation predictions included ancestry as an additional factor (pigment + ancestry) (Figure 2.10). The directionality of arrows differs slightly in this classifier to allow the ancestry SNP information to influence the pigmentation trait SNPs through the inclusion of an ancestry node, the arrows are

reversed as according to the Bayes-Ball algorithm [92]. In addition to the BN models, multinomial logistic regression (MLR) was performed using the all SNPs model. The MLR parameters calculated for the all SNPs prediction model are listed in Table 2.8. All prediction models were evaluated and compared based on the AUC and classification performance in terms of correct, incorrect, and inconclusive predictions in each category using the same 70% threshold (Table 2.9).

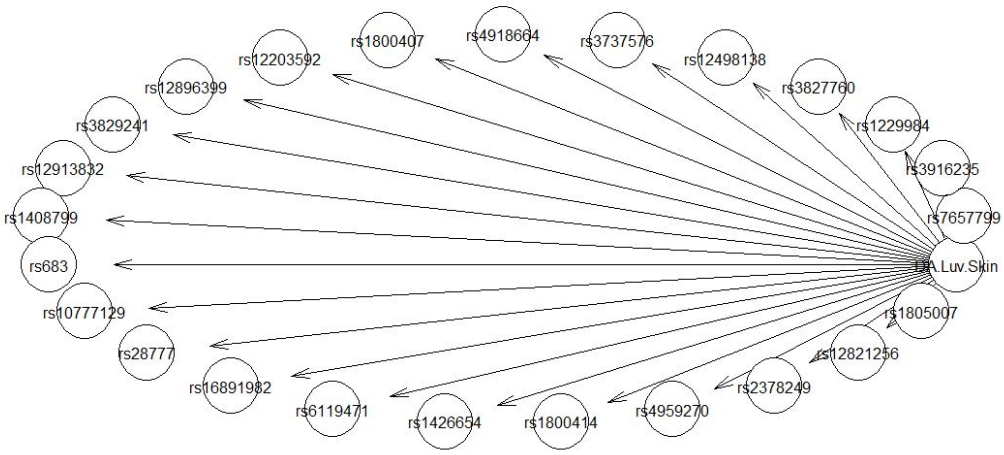


Figure 2.7 Example BN with all 24 SNPs for one trait (skin color).

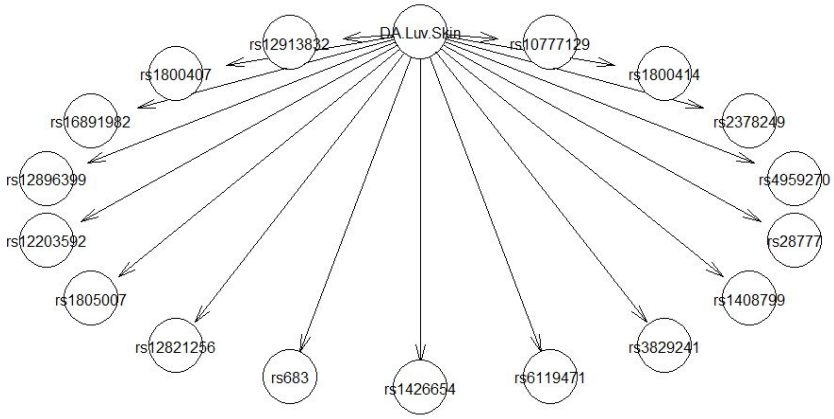


Figure 2.8 Example BN with the 17 pigmentation informative traits (skin color).

Table 2.5 Eye color LDA confusion matrices for the self-reported (SR) classifications with the RGB model, the consensus rating classifications by the 6 independent individuals with the RGB model, the consensus rating classifications with the LUV model, and the consensus rating classifications with the XYZ model. Correct classifications are shown in bold.

EYE	SR-RGB			Consensus-RGB			Consensus-Lu'v'			Consensus-xyY		
	Count	Number misclassified N (%)		Count	Number misclassified N (%)		Count	Number misclassified N (%)		Count	Number misclassified N (%)	
Validation Set (N=50)	50	11 (22)		50	6 (12)		50	7 (14)		50	11 (22)	
	Predicted			Predicted			Predicted			Predicted		
Actual	BLU	BRN	INT	BLU	BRN	INT	BLU	BRN	INT	BLU	BRN	INT
BLU	**11**	0	1	**14**	0	0	**14**	0	0	**14**	0	0
BRN	0	**23**	2	0	**22**	0	0	**21**	1	2	**19**	1
INT	4	4	**5**	2	4	**8**	3	3	**8**	4	4	**6**

Legend: SR= self-reported, consensus= consensus rating, BLU= blue, BRN= brown, INT= intermediate

Table 2.6 Skin color LDA confusion matrices for the self-reported (SR) classifications with the RGB model, the consensus rating classifications by the 6 independent individuals with the RGB model, the consensus rating classifications with the LUV model, and the consensus rating classifications with the XYZ model. Correct classifications are shown in bold.

SKIN	SR-RGB			Consensus-RGB			Consensus-Lu'v'			Consensus-xyY		
	Count	Number misclassified N (%)		Count	Number misclassified N (%)		Count	Number misclassified N (%)		Count	Number misclassified N (%)	
Validation Set (N=50)	50	12 (24)		50	4 (8)		50	3 (6)		50	5 (10)	
	Predicted			Predicted			Predicted			Predicted		
Actual	DRK	INT	LIT	DRK	INT	LIT	DRK	INT	LIT	DRK	INT	LIT
DRK	**3**	2	0	**3**	1	0	**3**	1	0	**2**	2	0
INT	0	**1**	10	0	**1**	3	0	**2**	2	0	**1**	3
LIT	0	0	**34**	0	0	**42**	0	0	**42**	0	0	**42**

Legend: SR= self-reported, consensus= consensus rating, LIT= light, DRK= dark, INT= intermediate

Table 2.7 Hair color LDA confusion matrices for the self-reported (SR) classifications with the RGB model, the consensus rating classifications by the 6 independent individuals with the RGB model, the consensus rating classifications with the LUV model, and the consensus rating classifications with the XYZ model. Correct classifications are shown in bold.

HAIR	SR-RGB				Consensus-RGB				Consensus-LUV				Consensus-XYZ			
	Count	Number misclassified N (%)			Count	Number misclassified N (%)			Count	Number misclassified N (%)			Count	Number misclassified N (%)		
Validation Set (N=50)	46	18 (38)			46	14 (30)			46	15 (33)			46	16 (35)		
	Predicted				Predicted				Predicted				Predicted			
Actual	BLD	BLK	BRN	RED	BLD	BLK	BRN	RED	BLD	BLK	BRN	RED	BLD	BLK	BRN	RED
BLD	**5**	0	3	0	**10**	0	1	0	**8**	0	3	0	**7**	0	4	0
BLK	0	**5**	4	0	0	**9**	4	1	0	**8**	5	1	0	**8**	5	1
BRN	4	3	**20**	0	1	0	**12**	0	1	0	**12**	0	1	0	**12**	0
RED	0	0	4	**0**	0	1	6	**1**	0	1	4	**3**	0	1	4	**3**

Legend: SR= self-reported, consensus= consensus rating, BLD= blonde, BLK=black, BRN= brown, RED= red

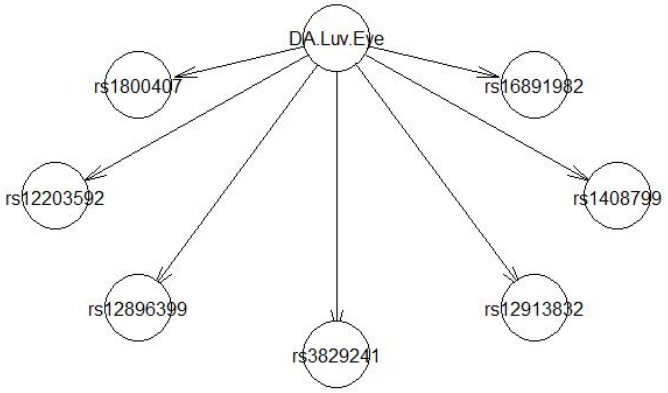


Figure 2.9 Example BN with selective trait SNPs only (eye color).

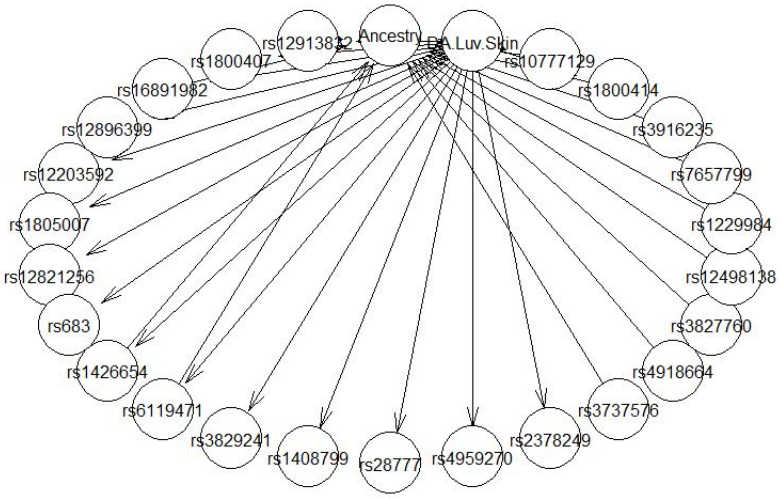


Figure 2.10 Example of pigment + ancestry BN model (skin color).

Table 2.8 Multinomial logistic regression parameters. a) The alpha intercept values (α), and b) the beta coefficients (β) for the all SNPs model.

Intercept		**EYE COLOR**		**HAIR COLOR**			**SKIN COLOR**	
α1		-2.23		17.01			-7.03	
α2		-1.05		15.31			-12.9	
α3		--		19.61			--	
SNP	**Minor Allele**	**β1**	**β2**	**β1**	**β2**	**β3**	**β1**	**β2**
rs12913832	C	4.52	2.94	-2.25	-3.76	-3.19	0.14	2.82
rs16891982	C	-3.57	-9.80	-13.54	-7.23	-10.76	1.70	8.44
rs12203592	A	1.87	2.55	-4.30	-4.54	-4.41	-1.64	-0.43
rs1800407	T	2.11	2.03	-0.53	-4.24	-1.55	-0.74	-2.13
rs3829241	A	-0.15	-0.74	-1.91	-3.39	-0.93	-0.20	-11.63
rs1805007	T	-1.43	-0.70	-2.48	-2.92	-2.84	0.18	-6.02
rs1408799	A	-1.12	-1.14	-3.03	-3.96	-3.51	-1.64	-4.38
rs683	T	-0.47	-0.27	-2.02	-2.89	-2.39	-1.53	4.37
rs3737576	T	0.82	-0.19	-3.48	-3.79	-4.10	-0.54	1.80
rs1229984	A	0.02	1.89	3.80	5.91	4.02	0.35	1.80
rs12498138	A	-2.16	-2.51	-10.20	-8.69	-9.29	0.37	-4.66
rs7657799	C	-4.60	-4.19	16.63	19.15	18.93	2.65	0.53
rs3916235	T	-3.67	-3.71	-1.91	-3.00	-1.66	2.57	7.28
rs4918664	G	1.47	1.19	4.11	5.53	3.38	-0.54	3.88
rs3827760	C	-4.14	-2.14	10.03	7.91	9.50	-1.55	-6.19
rs28777	C	0.29	8.29	5.11	5.49	6.14	-0.24	-2.65
rs12896399	T	1.82	1.93	-0.11	-0.46	-0.88	-0.37	-1.12
rs10777129	G	2.33	3.00	-6.25	-6.46	-5.66	2.75	-6.00
rs1800414	G	2.50	-3.72	-4.40	-8.98	-6.52	-1.77	6.44
rs6119471	G	5.95	4.65	-4.38	-5.38	-10.46	8.86	1.04
rs4959270	A	0.55	0.09	2.06	0.83	2.50	-0.94	1.68
rs2378249	A	-1.07	-1.05	1.59	3.72	1.94	0.69	-0.58
rs12821256	C	1.17	1.36	12.92	12.58	13.23	0.23	-12.53
rs1426654	G	-10.82	-4.87	-3.20	0.63	-0.87	-14.30	-9.51

For determining the optimal model, a more conservative evaluation where a higher number of inconclusive predictions with a smaller number of incorrect predictions was considered the more appropriate model. In a forensic context, an inconclusive result would be better than an incorrect classification. For eye color, the AUC values for blue

and brown colors were similar across all models, however, the pigment + ancestry BN model had a higher AUC for the intermediate colors and although slightly more inconclusive predictions, fewer incorrect predictions (Table 2.9). Comparing the models across all possible categories of each trait, the pigment + ancestry BN model was considered optimal with similar AUC values for all pigmentation traits. Similar trends were seen with hair color, where the AUC values were relatively the same across all models (Table 2.9). For skin color, the light and dark predictions have similar AUC values across all models, however, for the intermediate category, the AUC is the highest in the pigment + ancestry BN model with 0.79 indicating prediction capability of the model, whereas all the other models have AUC values below 0.50 indicating the other models are not informative for predicting intermediate skin (Table 2.9).

2.3.4 Pigmentation Prediction Model Likelihood Ratio Evaluation

As Bayesian networks utilize Bayes' rule to determine the posterior probabilities, likelihood ratios (LR) can also be calculated and evaluated. For all pigmentation models, a 70% threshold was applied; a single category with a 70% probability equated approximately to a likelihood of 2.33 that the single highest predicted category was the classification versus the other remaining possible categories. For example, it is 2.33 times more likely an individual has blue eye color over brown or intermediate eye color given the genotypes observed. This highest single probability over the remaining categories is labeled as the single LR in Table 2.9. In other words, it is 500 times likely to not have brown eye color), but for the sake of brevity, we are demonstrating a single method.

Table 2.9 BN and MLR prediction model parameters for the pigmentation traits. The model that best predicts most accurately across all traits is highlighted in green.

		EYE COLOR			HAIR COLOR				SKIN COLOR		
Model	Parameter	BLU (n=17)	BRN (n=24)	INT (n=9)	BLD (n=9)	BLK (n=9)	BRN (n=24)	RED (n=4)	DRK (n=3)	INT (n=2)	LIT (n=45)
Pigment + Ancestry	AUC	0.92	0.89	0.72	0.81	0.92	0.66	0.73	0.96	0.79	0.97
	CORRECT(%)	82	71	11	22	89	50	0	67	50	80
	INCORRECT(%)	6	4	33	33	0	33	75	0	50	13
	INCONCLUSIVE(%)	12	33	56	44	11	17	25	33	0	7
Pigment SNPs only	AUC	0.91	0.90	0.66	0.81	0.91	0.64	0.76	0.95	0.47	0.96
	CORRECT(%)	82	67	11	22	100	46	0	100	0	73
	INCORRECT(%)	6	0	44	33	0	29	75	0	50	0
	INCONCLUSIVE(%)	12	33	44	44	0	25	25	0	50	18
Trait SNPs only	AUC	0.92	0.88	0.63	0.87	0.92	0.71	0.73	0.98	0.31	0.92
	CORRECT(%)	82	67	0	11	89	33	0	100	0	84
	INCORRECT(%)	6	0	33	11	11	21	25	0	100	2
	INCONCLUSIVE(%)	12	33	67	78	9	46	75	0	0	13
All SNPs	AUC	0.92	0.89	0.71	0.83	0.92	0.70	0.77	0.96	0.43	0.95
	CORRECT(%)	82	67	11	33	89	50	0	100	0	78
	INCORRECT(%)	6	8	44	44	0	29	75	0	50	20
	INCONCLUSIVE(%)	12	25	44	22	11	21	25	0	50	2
MLR (all SNPs)	AUC	0.90	0.85	0.73	0.83	0.88	0.61	0.56	0.97	0.31	0.87
	CORRECT(%)	65	71	11	11	67	54	0	100	0	96
	INCORRECT(%)	12	13	44	44	33	25	100	0	100	2
	INCONCLUSIVE(%)	24	17	44	44	0	21	0	0	0	2

Legend: BLU= blue, BRN= brown, INT= intermediate, BLD= blonde, BLK= black, RED=red, DRK= dark, LIT= light, AUC= area under the ROC curve, MLR= multinomial logistic regression

By combining 2 possible categories, generating an exclusion LR, this decreases the amount of classification error, although there is not a definitive assignment of a category to the individual. For example, it is 300 times more likely that the individual does not have brown eyes. Only categories with possible confusion were combined which included brown and intermediate and blue and intermediate; blue and brown against intermediate were not considered as it is not perceptually likely for an individual to confuse blue and brown color; they are on opposite ends of the melanin content spectrum. For eye color in the pigment + ancestry BN model, there were 7 more correct classifications, 42/50 correct vs 36/50 correct, when considering a two category LR over the single LR, respectively (Table 2.10). Of the 8 samples that were incorrect for both sets of LR calculations, 6 were confusions of blue and brown (5 of which were predicted blue when they were actually brown), and 2 were predicted intermediate where the consensus was either brown or blue. Looking at the strongest linked SNP predictor of blue and brown eye color, rs12913832 (*HERC2*), for the 6 individuals that were predicted as blue but visually rated as brown, 3 were homozygous TT and 3 were heterozygous CT. These genotypes suggest inaccuracy in the model for brown predictions at another SNP in the panel (or more) or due to the genotypes of the samples used in the training set, as TT at rs12913832 is strong evidence for brown eye color expression whereas heterozygous individuals have a lower predictability and can be either brown or blue. Overall, while considering the other misclassified samples, there is a higher margin of error in the prediction model and not the classification method, as 7 of the 8 had incorrect prediction classifications and only one was misclassified according to the consensus rating. Additionally, as an exclusion probability is being considered with the probabilities of multiple categories, the LR values

are higher (Tables 2.10-2.12). In a forensic context, this is still useful information as individuals with the denominator category can be eliminated from being the possible contributor.

For hair color, with 4 possible categories, different combinations can be considered for the LRs. For the pigment + ancestry BN model, 2 categories against 2 and 3 categories against 1 were considered (Table 2.11). The categories again were grouped as those possibly to be confused perceptually. For the 2 against 2 categories, blonde and red, brown and red, and brown and black were calculated. For the 3 categories, it would only seem likely to eliminate the extreme categories of blonde and black as all the others can be perceptually confused. When compared to the 3 categories, which had the higher LRs, there were no misclassifications. The single LR had 24/46 correct predictions while the 2-category LR improved accuracy with 36/46 correct predictions (Table 2.11). Exclusion LR calculations are especially useful for the hair color predictions as it was the most confused trait overall. There were 9 samples that were incorrect for the 2-category classification between brown and blonde, and 1 sample between black and red (sample 9451, Table 2.11). Sample 9451 was misclassified as red by the consensus rating. The remaining 8 samples were misclassified as either brown or blonde by the model. Again, this is in part due to method for the hair collection as only single strands of hair were analyzed, not the overall shade of head. Individuals with brown or blonde hair may have red mixed in, and individuals with black may have some brown. Furthermore, blonde hair has already been known to be difficult to predict due to age-dependent changes especially during adolescence [45].

For skin color, the 2 category combinations considered were dark and intermediate and light and intermediate. Light and dark was not considered as those categories would unlikely be confused. When comparing single LRs with the exclusion LR, there were 41/50 and 45/50 correct predictions, respectively. There were 5 samples that were incorrect for both LR calculations. For these samples, 2 of them were incorrect in their consensus rating, not the predictions (sample 6329 was self-reported and predicted dark but the consensus rating was light; sample 7181 was self-reported and predicted intermediate but the consensus rating was light, Table 2.12). The other 3 samples were predicted as intermediate or dark and were actually light (samples 1654, 9717, and 7814, Table 2.12). Of these samples, 2 of the individuals were East Asian, and one was a light-skinned northern African. These samples indicate how ancestry's influence on pigmentation may not always be the most accurate prediction, and that further ancestry or skin SNPs would need to be included to avoid this classification confusion in the model.

2.3.5 Quantitative Color Pigmentation Prediction Models

In addition to discrete, qualitative category prediction as discussed above, it was a goal of this work to develop a BN that would be able to predict continuous, or, quantitative color coordinates (RGB values). This eliminates the subjectivity in binning color coordinates into a discrete color classification. A BN that has both discrete and continuous variables is a hybrid BN. A recently developed R package, *HydeNet*, was used to design similar BN structures as the discrete BNs. All 4 of the models were tested here: all SNPs, pigment SNPs only, trait SNPs only, and pigment + ancestry. The difference is that R, G, and B nodes were created in place of the single phenotype trait node; an example of the networks used can be seen in Figure 2.11.

Table 2.10 LR comparison for eye color in the pigment + ancestry BN model. The probabilities for each category (shaded by strength of probability in green) and the likelihood ratio (single LR) of the highest predicted category are shown, and also the 2-category exclusion LR (LR values are also shaded by strength of LR in blue). The misclassifications (N=No, Y=Yes) are highlighted in red.

Sample	BLU	BRN	INT	Predicted	Consensus	Single LR	BLU+INT/ BRN LR	BRN+INT/ BLU LR	Single LR correct?	2 category LR correct?
1370	0.521	0.068	0.412	BLU	BRN	1.09	13.75	0.92	N	N
1654	0.000	0.997	0.003	BRN	BRN	314.78	0.00	4775.65	Y	Y
1736	0.000	1.000	0.000	BRN	BRN	928797.45	0.00	5259532.10	Y	Y
1784	0.000	1.000	0.000	BRN	BRN	310986.32	0.00	1173639.90	Y	Y
1892	0.805	0.000	0.194	BLU	BLU	4.13	2462.58	0.24	Y	Y
1902	0.000	1.000	0.000	BRN	BRN	1098686.53	0.00	2220987.68	Y	Y
1905	0.905	0.002	0.093	BLU	BLU	9.48	487.39	0.11	Y	Y
2079	0.134	0.736	0.130	BRN	BRN	2.79	0.36	6.44	Y	Y
2093	0.869	0.012	0.119	BLU	BLU	6.66	82.98	0.15	Y	Y
2435	0.726	0.000	0.274	BLU	BLU	2.64	6707.11	0.38	Y	Y
3187	0.786	0.016	0.198	BLU	BLU	3.68	62.89	0.27	Y	Y
3471	0.922	0.002	0.077	BLU	BLU	11.74	664.99	0.09	Y	Y
3542	0.393	0.446	0.161	BRN	INT	0.81	1.24	1.54	N	Y
4063	0.000	1.000	0.000	BRN	INT	16210.92	0.00	76848.49	N	Y
4069	0.467	0.119	0.413	BLU	BRN	0.88	7.38	1.14	N	N
4258	0.590	0.019	0.391	BLU	BLU	1.44	50.58	0.70	Y	Y
4389	0.222	0.002	0.776	INT	BRN	3.47	453.56	3.51	N	N
4635	0.961	0.000	0.039	BLU	BLU	24.62	2689.50	0.04	Y	Y
4710	0.000	1.000	0.000	BRN	BRN	263248.42	0.00	1124157.56	Y	Y

Table 2.10 continued

4819	0.907	0.001	0.092	BLU	BLU	9.70	987.12	0.10	Y	Y
5168	0.521	0.011	0.468	BLU	BRN	1.09	91.06	0.92	N	N
5230	0.000	0.999	0.001	BRN	BLU	1299.17	0.00	8163.06	N	N
6084	0.000	0.997	0.002	BRN	BRN	362.83	0.00	2845.85	Y	Y
6149	0.819	0.001	0.181	BLU	INT	4.52	1564.23	0.22	N	Y
6305	0.053	0.835	0.112	BRN	BRN	5.05	0.20	17.93	Y	Y
6329	0.000	1.000	0.000	BRN	BRN	29846.23	0.00	135334.19	Y	Y
6347	0.942	0.000	0.058	BLU	BLU	16.23	5973.22	0.06	Y	Y
6789	0.116	0.021	0.863	INT	INT	6.29	46.28	7.62	Y	Y
7181	0.000	1.000	0.000	BRN	BRN	8155.13	0.00	88055.64	Y	Y
7263	0.335	0.149	0.516	INT	INT	1.07	5.71	1.99	Y	Y
7280	0.001	0.960	0.039	BRN	BRN	23.69	0.04	925.43	Y	Y
7294	0.365	0.517	0.118	BRN	INT	1.07	0.93	1.74	N	Y
7482	0.863	0.007	0.130	BLU	BLU	6.31	137.15	0.16	Y	Y
7632	0.789	0.000	0.211	BLU	BLU	3.74	6833.49	0.27	Y	Y
7659	0.379	0.238	0.382	INT	BRN	0.62	3.20	1.64	N	N
7814	0.000	1.000	0.000	BRN	BRN	3704309.86	0.00	9653528.54	Y	Y
7890	0.004	0.991	0.005	BRN	BRN	109.87	0.01	231.50	Y	Y
8395	0.000	1.000	0.000	BRN	BRN	928696.71	0.00	1654763.16	Y	Y
8539	0.000	0.999	0.001	BRN	BRN	936.39	0.00	74122.85	Y	Y
8709	0.798	0.000	0.202	BLU	INT	3.95	61810.78	0.25	N	Y
8730	0.557	0.261	0.182	BLU	INT	1.26	2.83	0.80	N	Y
8934	0.954	0.000	0.046	BLU	BLU	20.74	2729.38	0.05	Y	Y

Table 2.10 continued

8972	0.963	0.000	0.037	BLU	BLU	25.82	4149.62	0.04	Y	Y
9167	0.200	0.207	0.592	INT	BLU	1.45	3.82	3.99	N	N
9345	0.471	0.265	0.265	BLU	BRN	0.89	2.78	1.12	N	N
9451	0.000	0.999	0.000	BRN	BRN	1380.37	0.00	2628.49	Y	Y
9628	0.593	0.077	0.330	BLU	BRN	1.46	12.06	0.69	N	N
9717	0.000	1.000	0.000	BRN	BRN	1664447.76	0.00	4818491.00	Y	Y
9785	0.918	0.004	0.079	BLU	BLU	11.19	284.32	0.09	Y	Y
9981	0.572	0.010	0.419	BLU	INT	1.34	101.87	0.75	N	Y
TOTALS									36/50 correct	42/50 correct

Table 2.11 LR comparison for hair color in the pigment + ancestry BN model. The probabilities for each category (shaded by strength of probability in green) and the likelihood ratio (single LR) of the highest predicted category are shown, and a 2- and 3- category exclusion LR (LR values also shaded by strength of LR in blue). The misclassifications (N=No, Y=Yes) are highlighted in red.

Sample	BLD	BLK	BRN	RED	Predicted	Consensus	Single LR	BLK+BRN/ BLD+RED LR	BLD+RED/ BLK+BRN LR	BRN+RED/ BLD+BLK LR	BLK+BRN +RED/ BLD LR	BLD+RED +BRN/ BLK LR	Single LR correct?	3 category LR correct?	2 category LR correct?
1370	0.142	0.000	0.849	0.008	BRN	BRN	5.64	5.64	0.18	6.03	6.03	93456.94	Y	Y	Y
1654	0.000	0.998	0.002	0.000	BLK	BRN	429.87	33771.37	0.00	0.00	709218.86	0.00	N	Y	Y
1736	0.000	0.996	0.003	0.000	BLK	BLK	264.91	3293.34	0.00	0.00	90908.10	0.00	Y	Y	Y
1784	0.000	0.998	0.001	0.000	BLK	BLK	615.3	8115.88	0.00	0.00	24212.07	0.00	Y	Y	Y
1892	0.451	0.000	0.549	0.001	BRN	BLD	1.21	1.21	0.82	1.22	1.22	146626.57	N	Y	N
1902	0.000	1.000	0.000	0.000	BLK	BRN	22023.35	22169.08	0.00	0.00	171232875.34	0.00	N	Y	Y
1905	0.206	0.000	0.793	0.001	BRN	BRN	3.84	3.84	0.26	3.85	3.86	8646.90	Y	Y	Y

Table 2.11 continued

2079	0.640	0.005	0.354	0.002	BLD	BRN	1.78	0.56	1.79	0.55	0.56	208.75	N	Y	N
2093	0.277	0.000	0.720	0.003	BRN	RED	2.57	2.57	0.39	2.61	2.61	2286.68	N	Y	Y
2435	0.241	0.000	0.699	0.060	BRN	BLD	2.32	2.32	0.43	3.15	3.15	225224.23	N	Y	N
3187	0.639	0.001	0.359	0.001	BLD	BLD	1.77	0.56	1.78	0.56	0.56	1047.76	N	Y	Y
3471	0.263	0.000	0.734	0.003	BRN	BRN	2.76	2.76	0.36	2.80	2.80	19568.47	Y	Y	Y
4063	0.000	1.000	0.000	0.000	BLK	BLK	5241.13	35004.43	0.00	0.00	1763667.41	0.00	Y	Y	Y
4069	0.170	0.000	0.829	0.001	BRN	BRN	4.84	4.84	0.21	4.88	4.89	5560.27	Y	Y	Y
4258	0.015	0.004	0.901	0.080	BRN	BRN	9.13	9.56	0.10	52.32	66.58	251.70	Y	Y	Y
4635	0.254	0.000	0.729	0.017	BRN	BRN	2.69	2.69	0.37	2.94	2.94	31644.57	Y	Y	Y
4710	0.000	1.000	0.000	0.000	BLK	BLK	8784.52	8930.06	0.00	0.00	92592591.11	0.00	Y	Y	Y
4819	0.330	0.000	0.666	0.004	BRN	BRN	2	2.00	0.50	2.03	2.03	13421.82	Y	Y	Y
5168	0.000	0.988	0.010	0.002	BLK	BRN	81.47	562.58	0.00	0.01	65788.47	0.01	N	Y	Y
5230	0.001	0.747	0.234	0.018	BLK	BRN	2.95	50.21	0.02	0.34	732.09	0.34	N	Y	Y
6084	0.041	0.370	0.587	0.002	BRN	BLK	1.42	22.34	0.04	1.43	23.29	1.70	N	Y	Y
6149	0.572	0.001	0.336	0.092	BLD	RED	1.34	0.51	1.97	0.75	0.75	1876.12	N	Y	Y
6305	0.136	0.004	0.851	0.010	BRN	RED	5.69	5.86	0.17	6.19	6.37	281.36	N	Y	Y
6329	0.000	1.000	0.000	0.000	BLK	BLK	3788.14	4023.51	0.00	0.00	29154516.94	0.00	Y	Y	Y
6347	0.874	0.000	0.125	0.000	BLD	BLD	6.96	0.14	6.98	0.14	0.14	72462.77	Y	Y	Y
6789	0.076	0.000	0.913	0.011	BRN	BLD	10.46	10.46	0.10	12.15	12.16	33556.05	N	Y	N
7181	0.000	1.000	0.000	0.000	BLK	BLK	10050.41	74598.03	0.00	0.00	9523808.68	0.00	Y	Y	Y
7263	0.156	0.000	0.844	0.000	BRN	BLD	5.4	5.40	0.19	5.40	5.41	13868.63	N	Y	N
7280	0.001	0.247	0.747	0.005	BRN	BRN	2.95	183.18	0.01	3.03	1632.75	3.04	Y	Y	Y
7294	0.140	0.000	0.859	0.001	BRN	BRN	6.08	6.10	0.16	6.10	6.13	2348.68	Y	Y	Y
7482	0.388	0.000	0.596	0.016	BRN	BLD	1.47	1.48	0.68	1.57	1.58	4376.73	N	Y	N
7632	0.729	0.000	0.267	0.004	BLD	BLD	2.69	0.36	2.75	0.37	0.37	36230.88	N	Y	Y
7659	0.048	0.000	0.946	0.005	BRN	BRN	17.61	17.65	0.06	19.61	19.66	8862.76	Y	Y	Y

Table 2.11 continued

7814	0.000	1.000	0.000	0.000	BLK	BRN	44429.08	78130.10	0.00	0.00	10101008.92	0.00	N	Y	Y
8395	0.000	1.000	0.000	0.000	BLK	BLK	164612	174992.74	0.00	0.00	222717147.66	0.00	Y	Y	Y
8539	0.000	1.000	0.000	0.000	BLK	BRN	27933.09	42362.39	0.00	0.00	189035917.01	0.00	N	Y	Y
8709	0.829	0.000	0.170	0.001	BLD	BRN	4.85	0.20	4.88	0.21	0.21	56178.78	N	Y	N
8730	0.336	0.000	0.659	0.004	BRN	BRN	1.93	1.94	0.52	1.97	1.98	3626.12	Y	Y	Y
8934	0.352	0.000	0.645	0.002	BRN	BRN	1.82	1.82	0.55	1.84	1.84	245699.25	Y	Y	Y
8972	0.933	0.000	0.065	0.002	BLD	BRN	13.98	0.07	14.34	0.07	0.07	59522.81	N	Y	N
9167	0.056	0.000	0.937	0.007	BRN	BRN	14.96	14.98	0.07	16.92	16.93	21550.72	Y	Y	Y
9451	0.000	0.970	0.029	0.001	BLK	RED	32.47	1570.84	0.00	0.03	95237.10	0.03	N	Y	N
9628	0.268	0.000	0.728	0.003	BRN	BRN	2.68	2.68	0.37	2.73	2.73	101831.99	Y	Y	Y
9717	0.000	1.000	0.000	0.000	BLK	BLK	39275.34	117480.20	0.00	0.00	3424656.37	0.00	Y	Y	Y
9785	0.304	0.000	0.694	0.002	BRN	BLD	2.27	2.27	0.44	2.29	2.29	26384.22	N	Y	N
9981	0.164	0.000	0.832	0.003	BRN	BRN	4.97	4.98	0.20	5.07	5.08	2156.59	Y	Y	Y
TOTALS													24/46 correct	46/46 correct	36/46 correct

Table 2.12 LR comparison for skin color in the pigment + ancestry BN model. The probabilities for each category (shaded by strength of probability in green) and the likelihood ratio (single LR) of the highest predicted category are shown, and a 2-category exclusion LR (LR values also shaded by strength of LR in blue). The misclassifications (N=No, Y=Yes) are highlighted in red.

Sample	DRK	INT	LIT	Predicted	Consensus	Single LR	DRK+INT/ LIT LR	LIT+INT/ DRK LR	Single LR correct?	2 category LR correct?
1370	0.000	0.000	1.000	LIT	LIT	41894.73	0.00	35102641.87	Y	Y
1654	0.358	0.617	0.026	INT	LIT	1.607942	38.09	1.79	N	N
1736	0.060	0.742	0.198	INT	INT	2.876198	4.04	15.76	Y	Y
1784	0.107	0.702	0.191	INT	LIT	2.356587	4.25	8.31	N	Y
1892	0.000	0.000	1.000	LIT	LIT	299003	0.00	110827758.79	Y	Y

Table 2.12 continued

1902	0.997	0.003	0.000	DRK	INT	381.275	9728819.27	0.00	N	Y
1905	0.000	0.000	1.000	LIT	LIT	36703.95	0.00	5075376.97	Y	Y
2079	0.000	0.000	1.000	LIT	LIT	3270.189	0.00	76665.74	Y	Y
2093	0.000	0.000	1.000	LIT	LIT	16795.8	0.00	266737.47	Y	Y
2435	0.000	0.000	1.000	LIT	LIT	17904.27	0.00	13980847.10	Y	Y
3187	0.000	0.000	1.000	LIT	LIT	32070.61	0.00	282815.64	Y	Y
3471	0.000	0.000	1.000	LIT	LIT	34594.33	0.00	112733939.19	Y	Y
3542	0.000	0.000	1.000	LIT	LIT	28539.13	0.00	2709170.90	Y	Y
4063	0.315	0.673	0.012	INT	LIT	2.06009	81.52	2.18	N	Y
4069	0.000	0.000	1.000	LIT	LIT	5227.807	0.00	996047.13	Y	Y
4258	0.000	0.003	0.997	LIT	LIT	286.2847	0.00	1034014.10	Y	Y
4389	0.000	0.000	1.000	LIT	LIT	12314.71	0.00	26104012.07	Y	Y
4635	0.000	0.000	1.000	LIT	LIT	83681.08	0.00	228656472.46	Y	Y
4710	0.994	0.006	0.000	DRK	DRK	174.7681	327418.79	0.01	Y	Y
4819	0.000	0.000	1.000	LIT	LIT	37094.39	0.00	82038145.13	Y	Y
5168	0.015	0.748	0.237	INT	LIT	2.961556	3.22	64.63	N	Y
5230	0.006	0.189	0.805	LIT	LIT	4.116254	0.24	164.94	Y	Y
6084	0.000	0.019	0.981	LIT	LIT	51.03596	0.02	2301.33	Y	Y
6149	0.000	0.001	0.999	LIT	LIT	961.6165	0.00	5232.87	Y	Y
6305	0.000	0.003	0.997	LIT	LIT	377.7958	0.00	84081.56	Y	Y
6329	0.854	0.146	0.000	DRK	LIT	5.849223	22249.06	0.17	N	N
6347	0.000	0.000	1.000	LIT	LIT	74314.97	0.00	983159.98	Y	Y
6789	0.000	0.000	1.000	LIT	LIT	9321.736	0.00	6688300.05	Y	Y
7181	0.129	0.863	0.008	INT	LIT	6.285268	127.33	6.72	N	N
7263	0.000	0.000	1.000	LIT	LIT	118225.2	0.00	2841668.30	Y	Y
7280	0.000	0.100	0.900	LIT	LIT	8.969565	0.11	2068.94	Y	Y

Table 2.12 continued

7294	0.000	0.000	1.000	LIT	LIT	44822.28	0.00	19126769.26	Y	Y
7482	0.000	0.000	1.000	LIT	LIT	4819.091	0.00	3494296.88	Y	Y
7632	0.000	0.000	1.000	LIT	LIT	131797.3	0.00	1600379.37	Y	Y
7659	0.000	0.000	1.000	LIT	LIT	5047.092	0.00	41150620.49	Y	Y
7814	0.975	0.025	0.000	DRK	LIT	38.84789	2077.96	0.03	N	N
7890	0.001	0.084	0.915	LIT	LIT	10.7108	0.09	1008.70	Y	Y
8395	0.975	0.025	0.000	DRK	DRK	39.52512	1125150.70	0.03	Y	Y
8539	0.436	0.564	0.000	INT	DRK	1.292076	932436.39	1.29	N	Y
8709	0.000	0.000	1.000	LIT	LIT	76289.17	0.00	38908532.70	Y	Y
8730	0.000	0.000	1.000	LIT	LIT	41327.64	0.00	1045479.93	Y	Y
8934	0.000	0.000	1.000	LIT	LIT	109375.1	0.00	394892169.19	Y	Y
8972	0.000	0.000	1.000	LIT	LIT	31617.57	0.00	1023395.30	Y	Y
9167	0.000	0.000	1.000	LIT	LIT	15829.61	0.00	53311166.25	Y	Y
9345	0.000	0.000	1.000	LIT	LIT	57933.48	0.00	24935776.32	Y	Y
9451	0.001	0.477	0.522	LIT	LIT	1.089973	0.92	746.39	Y	Y
9628	0.000	0.000	1.000	LIT	LIT	128532.8	0.00	313671781.77	Y	Y
9717	0.959	0.040	0.001	DRK	LIT	23.29524	1127.20	0.04	N	N
9785	0.000	0.000	1.000	LIT	LIT	220011.2	0.00	20086150.74	Y	Y
9981	0.000	0.000	1.000	LIT	LIT	6445.191	0.00	3664266.00	Y	Y
TOTALS									41/50 correct	45/50 correct

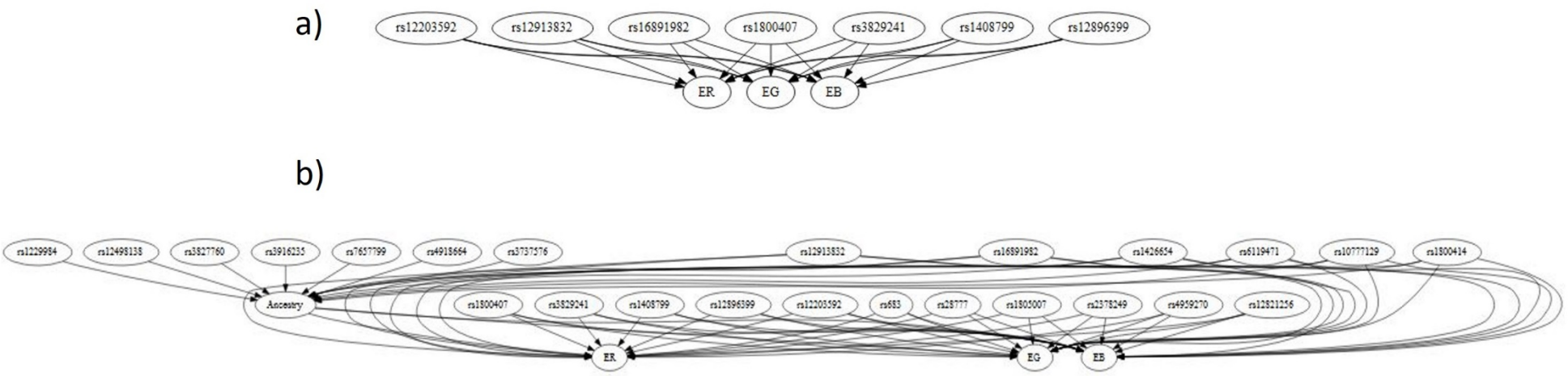


Figure 2.11 *HydeNet* built BN models. a) The trait SNPs only hybrid BN model (eye color). b) The pigment + ancestry BN model (skin color).

The following equations are examples of the *HydeNet* BN model output to calculate the predicted RGB values. Each genotype possibility is coded and inputted as 0, 1, or 2 at each SNP for each color value equation. The examples shown were used to calculate the eye color R, G, and B predictions for the trait SNPs only model (coded as ER, EG, and EB):

ER = 52.94984 + 8.03904 * rs12913832 + -12.47638 *rs16891982 + 4.98515 *rs1800407 + -3.9789 * rs3829241 + 1.30806 *rs1408799 + -0.56048 * rs12896399 + 3.39914 * rs12203952 Equation 2.4

EG = 32.31058 + 14.1561 *rs12913832 + -11.51611 *rs16891982 + 4.79901 *rs1800407 + -3.66421 *rs3829241 + 1.82756 *rs1408799 + 0.46773 *rs12896399 + 2.34781 * rs12203592 Equation 2.5

EB = 28.57741 + 15.21519 *rs12913832 + -10.65358 * rs16891982 + 3.28254 * rs1800407 + -2.53158 *rs3829241 + 2.09347 *rs1408799 + 0.8793 *rs12896399 + 1.06664 *rs12203592 Equation 2.6

As the Lu'v' color model was used for classification in the discrete BN models, it was applied for the hybrid models as well; however, to be able to visualize the RGB output as a color on a computer, the Lu'v' values were transformed to RGB values. It was performed for eye color in one model and there was no significant difference in error measurements between the transformed Lu'v' values and those where RGB was used as the input. Therefore, to eliminate the need of computational effort for color coordinate transformations, RGB values were used as input for these models. Table 2.13 shows the average error for each model for all 3 pigmentation traits. All RGB value comparisons can be seen in Appendix D, but Table 2.14 shows a subset as an example.

For eye color, the trait SNPs model had the least amount of error overall and for blue and intermediate eye colors; the all SNPs model had a slightly lower error value for

brown eye color. The pigment SNPs model had the least amount of error for hair color overall and for black and red hair color (Table 2.13). The all SNPs model had slightly lower error for brown hair, while the least amount of error for blonde was with the trait SNPs only model. For skin color, the pigment + ancestry model had the least amount of error overall and for all three skin color phenotypes, with high accuracy especially for light and dark skin with only a 1.1% error (Table 2.13). Hair color was the trait with the highest error. Again, this could be due to collection and analysis method of the hair, also as hair can be a mix of more than one shade of color and may not be accurately represented as a single RGB coordinate. Overall, the predicted RGB coordinates are in higher ranges than the measured values. Although the average error values were low for some phenotypes (1.1% for dark and light skin color), it was as high as 25.6% for blonde hair color (Table 2.13).

Table 2.13 Error rates of the RGB value predictions.

	Average Absolute Error (Average Error %)			
	Trait SNPs only	**Pigment SNPS only**	**All SNPs**	**Pigment+ Ancestry**
Eye Color- Overall	**28.66 (6.5%)**	**30.61 (6.9%)**	**31.05 (7.0%)**	**30.96 (7.0%)**
BLU (n=17)	6.34 (1.4%)	8.19 (1.9%)	7.56 (1.7%)	7.97 (1.8%)
BRN (n=24)	15.66 (3.6%)	19.27 (4.4%)	13.84 (3.1%)	37.63 (8.5%)
INT (n=9)	26.47 (6.0%)	27.34 (6.2%)	32.90 (7.5%)	52.73 (11.9%)
Hair Color -Overall	**52.08 (11.8%)**	**50.03 (11.3%)**	**52.34 (11.9%)**	**55.50 (12.6%)**
BLD (n=9)	108.60 (24.6%)	111.15 (25.1%)	110.49 (25.0%)	113.03 (25.6%)
BLK (n=9)	41.61 (9.4%)	38.93 (8.8%)	43.63 (9.9%)	35.46 (8.03%)
BRN (n=24)	8.54 (1.9%)	7.96 (1.8%)	7.45 (1.7%)	8.05 (1.82%)
RED (n=4)	45.56 (10.3%)	32.14 (7.28%)	37.99 (8.6%)	30.69 (7.0%)
Skin Color - Overall	**35.24 (8.0%)**	**33.29 (7.5%)**	**33.23 (7.5%)**	**33.57 (7.6%)**
DRK (n=3)	9.15 (2.1%)	5.43 (1.2%)	11.99 (2.7%)	4.80 (1.1%)
INT (n=2)	71.65 (16.2%)	49.81 (11.3%)	50.48 (11.4%)	47.33 (10.7%)
LIT (n=45)	7.65 (1.7%)	5.12 (1.2%)	7.60 (1.7%)	4.86 (1.1%)

Table 2.14 Subset of RGB value comparison from *HydeNet* BN models. The highlighted columns are the RGB measurements extracted from the digital skin photo collection. Only the pigmentation SNPs only model predicted RGB values are fully shown, and the percent error shown for all models tested.

Sample	SR	SG	SB	Actual	Predicted SRpig	Predicted SGpig	Predicted SBpig	Pigmentation SNPs only, Error %	All SNPs, Error %	Trait SNPs only, Error %	Ancestry+ Pigmentation, Error %
1370	200	156	125		223	186	151	10.47%	9.60%	8.47%	10.56%
1654	212	174	133		208	163	121	3.76%	7.96%	2.37%	3.85%
1736	187	134	83		208	165	124	12.56%	14.51%	17.81%	11.54%
1784	200	161	123		207	157	104	4.84%	5.07%	6.69%	5.39%
1892	228	183	128		220	181	144	4.07%	4.02%	4.95%	3.95%
1902	172	121	68		185	142	104	10.06%	8.86%	14.78%	10.04%
1905	231	201	170		213	172	137	10.65%	8.57%	7.45%	10.62%
2079	224	175	120		209	171	136	5.01%	5.22%	4.98%	5.23%
2093	208	168	127		207	170	142	3.44%	3.56%	4.75%	3.55%
2435	207	178	159		225	192	166	5.45%	4.59%	3.77%	5.37%
3187	213	155	131		212	174	139	4.71%	5.69%	6.08%	4.77%
3471	217	178	146		209	171	136	3.40%	4.21%	0.82%	3.23%
3542	216	164	113		218	181	145	8.19%	8.06%	6.89%	8.48%
4063	194	154	110		234	191	142	14.24%	18.06%	23.30%	14.56%
4069	200	175	140		218	179	145	4.36%	4.41%	4.62%	4.39%
4258	225	192	156		212	171	130	8.24%	8.96%	4.66%	7.52%
4389	207	165	118		217	179	145	7.43%	4.15%	6.66%	7.60%
4635	196	158	124		217	181	146	8.62%	11.23%	7.98%	8.67%
4710	148	87	53		185	135	90	16.09%	14.31%	24.58%	16.41%
4819	212	174	164		213	176	142	4.95%	5.07%	5.05%	4.93%
5168	203	155	115		196	151	112	1.79%	2.00%	2.44%	1.74%

2.3.6 Ancestry Prediction Model Evaluation

Ancestry was evaluated similarly to the discrete pigmentation phenotypes. There were 2 BN models considered, one with all SNPs, and one with the ancestry informative SNPs alone which consisted of 13 SNPs (Figure 2.12).

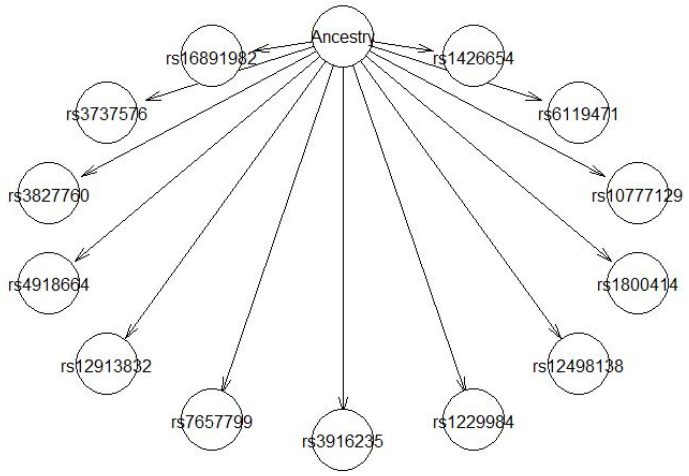


Figure 2.12 Ancestry trait SNPs BN model.

The known ancestry of the sample population samples is based on the self-reported information collected from the volunteer survey. To simplify classification, only single categories by continental group was considered. For the individuals who listed more than one ancestry, and as the majority of the sample population was of European descent, the minority group was used (e.g., if reported as African/European, African was used). There were 7 individuals in the training set that reported unknown as their ancestry. LDA was used to group these samples into one of the 5 possible ancestral categories: European, East Asian (E. Asian), Other Asian (O. Asian), Hispanic, and African. E. Asian encompassed individuals from the eastern regions of Asia such as China, Japan, Korea, and Taiwan whereas O. Asian included countries such as India and Pakistan. The

prediction model performance parameters can be seen in Table 2.15, a 70% threshold was applied similarly as the pigmentation models.

Table 2.15 Prediction Model parameters for ancestry. The model with best overall accuracy is highlighted in green.

Model	Parameter	ANCESTRY African (n=6)	E. Asian (n=4)	European (n=33)	Hispanic (n=4)	O. Asian (n=3)
All SNPs	AUC	0.99	0.96	1.0	0.99	0.89
	CORRECT(%)	83	50	97	75	67
	INCORRECT(%)	17	50	0	0	33
	INCONCLUSIVE(%)	0	0	3	25	0
Trait SNPs only	AUC	0.84	0.83	0.88	0.99	0.91
	CORRECT(%)	83	75	100	50	33
	INCORRECT(%)	17	25	0	0	33
	INCONCLUSIVE(%)	0	0	0	50	33

The BN models classified the ancestry groups with a high rate of correct predictions. The all SNPs model performed better overall for all ancestry groups than the trait SNPs only model (Table 2.15). Most confusion of classifications happened for O. Asian individuals, as well as between Hispanic and E. Asian groups. This is not surprising as E. Asian and Hispanic ancestry are most similarly related when considering the out of Africa theory of human migration [93]; the SNPs chosen to classify E. Asian and Hispanic had similar minor allele frequencies when compared to the other ancestral groups. There was one individual classified as O. Asian who was actually from North Africa (Egypt). One E. Asian, one O. Asian, and one Hispanic were also misclassified between those three ancestry groups, but none were classified as European or African. Further SNPs should be included to discriminate the Asian ancestral groups from Hispanic.

2.4 Conclusions

The goal of this work was to develop a SNP assay that can be informative for eye color, hair color, skin color, and ancestry simultaneously not only based on discriminant color categories but on a quantitative approach, while also developing an accurate model to perform the predictions. The SNP assay itself being based on allele-specific PCR is a simpler protocol than single-based extension methods, eliminating the need for multiple PCR and clean-up steps and reducing human error in pipetting and transfer between tubes. Overall, the selected SNPs predict eye and skin color fairly accurately with an ancestry influenced pigmentation model; with the exception of intermediate eye color at 11% correct prediction rate, all other phenotypes of these traits had correct prediction rates ranging from 50-82%, with blue eye color as the most accurate trait with an 82% correct prediction rate. Hair color predictions were the least accurate in all the models tested, however, the collection method retrospectively was not the optimal way to measure hair color. Although calibrated color measurement of digital photos was a reliable tool, a more representative portion of hair needs to be evaluated as individual strands have a lot of variation. Furthermore, inclusion of more known hair informative SNPs would improve prediction accuracies, especially for red hair.

Quantitative color measurement was successful. This work shows that a consensus of multiple individuals in rating pigmentation colors is more accurate and reliable than a single individual (self-reported rating), likely due to bias in self-reporting, and also as the use of consensus ratings incorporates differences in human perception of the same color. The LUV color model should be considered for quantitative color measurement as it measured color similarly to the photo-measured RGB color, is a more objective color space with a perceptually uniform color space, and is device independent.

BN models could be developed to predict quantitative color coordinates. This eliminates the need to bin colors into discriminant categories and the color can be assessed as an objective numerical value. One limitation of the color measurements in this study was the use of the average color – the whole iris was extracted and measured as a whole. Because of this, the quantitative RGB representation of the color does not necessarily portray the true shade of the visually-percepted iris. One way to work around this would be to extract and measure only a small portion of the iris. Conversely, for hair color, only a small section was measured, it might be beneficial to measure the entire strand or a photo of the entire head of hair to be able to assess the overall shade more accurately. Although error rates were determined for the color coordinate predictions, further analysis into what range of error would actually have an effect on the human perception of the color is necessary to further evaluate these models.

For ancestry, there were 97% correct predictions, with 83% for African, 75% for Hispanic, 67% for Other Asian, and 50% for East Asian when using the entire SNP panel. These are relatively accurate predictions, especially when considering they are only based on 24 SNPs. There were no well-established and highly informative SNPs for Other Asian populations found in literature at the time of the assay SNP selection, although Eurasiaplex includes 23 SNPs found to have high likelihoods of discrimination between European and South Asian populations [58]. For improvement in the Other Asian ancestries in this FPP panel, perhaps some of those SNPs should be considered for inclusion. The collected sample population was lacking in many of the ancestry groups other than European, especially in East Asian samples (with 7/200), and therefore, the model may not reliably reflect the power of prediction for the more minor ancestry

groups assessed. Furthermore, admixture of biracial individuals was not fully addressed. Individuals were categorized into a single ancestry group for ease of model building, however, further development of the models should be done to be able to consider levels of admixture that would have a conflicting result when trying to categorize as a single ancestry.

For eye color, there was 82% correct predictions for blue, 71% for brown, and 11% for intermediate. For skin color, there was 80% correct predictions for light skin, 67% correct predictions for dark, and 50% for intermediate. Not unexpected, as it is seen with eye color, the intermediate skin had the lowest percent of correct predictions within the skin color phenotypes. For hair color, there was 89% correct predictions for black hair, 50% correct predictions for brown, 22% for blonde, and 0% for red. More work needs to be done to improve this assay. This was clearly needed for the red hair phenotype. Red hair was the first trait identified for prediction from DNA [35] and yet had very poor prediction power in the developed BN models. This is in part due to selecting only one of the *MC1R* SNPs. As discussed in section 2.1.2, there is a compound effect between the several mutations within the *MC1R* gene for red hair expression; some with a stronger or weaker contribution to that expression. Not every red-haired individual will have the same haplotype of *MC1R* mutations and selecting only one of the SNPs with a strong effect has shown to be insufficient. There was also only a small set of red haired individuals in the sample population (8/200), so similarly to the East Asian ancestry, more individuals representing this phenotype would be ideal to include. Including a few more *MC1R* SNPs and optimizing the hair collection method would likely improve the accuracy of these predictions.

In terms of the prediction modeling accuracy, the BN model was shown to perform more accurately than MLR. This is not to say the MLR model was not accurate, it had comparable AUC values with the BN models (Table 2.9). However, the BN models had overall less incorrect and more inconclusive predictions. This is the best conservative approach to be used in the forensic context, as it would be better to have an inconclusive result than an incorrect one. The pigment + ancestry BN model was selected as the optimal model for this reason. However, this may not be the most optimal model for all datasets; a larger sample population with higher frequencies of the lesser represented traits should be evaluated. The likelihood ratio statistic that can also be calculated is favorable for reporting purposes. In fact, even if a probability is not strong enough in one category, a broader conclusion can still be drawn if you group categories together to still have an accurate prediction statement. For example, one can state that it is 100 times more likely that an individual has blue eyes. But stating more broadly that it is 100 times more likely that an individual has blue or intermediate eyes can still eliminate brown-eyed individuals. The phenotypic information from these predictions, as of now, is to be used to help gather intelligence for an investigation. Any reduction in possible suspects to help solve a case is a desired goal.

CHAPTER 3. METHYL-RADSEQ: A NOVEL METHOD FOR THE DISCOVERY OF CANDIDATE DNA MARKERS FOR AGE PREDICTION

3.1 Introduction

Pigmentation traits are not the only phenotypes that would aid the intelligence for an investigation. Another important phenotype is age. Age can also relate to pigmentation; for example, an individual may be predicted to have brown hair, but after a certain age, hair color can change, and thus investigators may not be looking for an individual with brown hair, but with white or grey hair. In addition, knowing the age range of an unknown DNA sample found at a crime scene would certainly reduce the number of possible contributors. One mechanism in place for aging phenomena is due to epigenetic changes to the genome over time.

Epigenetics refers to the inherited patterns of gene expression without changes in the DNA sequence [94]. Epigenetic regulation plays a role in animal and plant development and regulates the activation or repression of genes within specific cells that may be tissue specific and occur during different stages of development [94]. The best known types of epigenetic modifications include: methylation, phosphorylation, acetylation, and ubiquitination [95]. Each of these mechanisms produce changes in gene expression, thus effecting the production of specific gene products or proteins.

DNA methylation is an epigenetic factor that has potential use in forensic investigations. DNA methylation involves the addition of a methyl group (CH_3) to the aromatic ring of a DNA base. The most common methylation is the methyl group addition to the 5'-carbon structure of cytosine bases (5-mC) [95]. In particular, these

methylated cytosines are found in dinucleotide patterns with guanines and are termed CpG sites or islands, if they occur in long stretches of > 500bp [96], and are typically located in gene promoter regions. There are roughly 20 million CpG sites in the human genome, and at present there is no defined set of CpG markers with predictive relevance [97]. However, age-associated differentially methylated regions (DMRs) have been found to be conserved in several tissues, indicating age-dependent methylation is not random [97]. The association of these sites to age is based on either a positive (hypermethylated) or negative (hypomethylated) correlation as age increases.

One of the proposed panels to be included in the FPP assay is a panel of epigenetic methylation markers (each CpG site as a C/T 'SNP') that will be informative for chronological age prediction and thus represent an age phenotype profile.

3.1.1 Predicting Age

There have been some developments in which a few studies have identified potential useful age-associated CpG markers, most of which have been with microarray platforms, or bead chip arrays, which can analyze hundreds of thousands of SNPs at once. A challenging aspect is that regulation of methylation, and thus rate of methylation at specific sites, has been found to be mostly tissue-specific. However, markers that correlate in common across all tissues are known [98]. Additionally, with the use of a prediction model for specific tissue markers, a correction can be applied when comparing samples from different tissues [99].

Bocklandt et al. [100] found significant correlation between 88 CpG sites with age from saliva samples and describe a predictive model using CpG sites within the promoter regions of three genes: *NPTX2*, *TOM1L1*, and *EDARADD*. This model was

accurate for an age range of 18-70 within 5.2 years and explained 73% of age variance. However, the study is limited in that the CpG sites were found correlated in saliva samples only, therefore not broadly applicable in forensic samples. Furthermore, sex was an influencing factor, as *NPTX2* sites were not significant in the analysis of females [100]. Sex was previously found to influence methylation states between males and females, especially X-chromosome sites [101]. The correction of any gender-related influences should be included in the final selection of age-related sites used for age predictions. Florath et al. [102] used a model with 17 CpG sites that explained 71% of the variance of age with an average accuracy of 2.6 years from blood samples. The most statistically significant CpG sites (with known associated genes) found were: cg06784991 (*ZYG11A*), cg06639320 (*FHL2*), cg04875128 (*OTUD7A*), cg19283806 (*CCDC102B*), cg17110586, and cg07547549 (*SLC12A5*), cg09809672 (*EDARADD*), cg16867657 (*ELOVL2*) [102]. It was also reported that cg16867657 (*ELOVL2*) alone describes 47% of the variance of age [102]. More recently, *ELOVL2* was further analyzed on its own for age determination, where 7 CpG sites within *ELOVL2* in blood samples was able to predict age within 7 years with a 60-78% correct prediction rate [18]. Hannum et al. [99] developed a model that predicted age at 91% with an error of 5 years, adding that cg16867657 and cg23606718, both associated to *ELOVL2*, were a common significant age correlating marker for multiple tissue models [99]. Zbieć-Piekarska et al. [103] further investigated the top 8 candidate loci from Hannum et al. [99] via pyrosequencing and developed a prediction model with a standard error of 4.5 years, where 86.7% of samples had correct predictions within 5 years for ages 2-19, which gradually decreased as age increased to 50% for 60-75 years.

There are a few issues to be aware of when analyzing methylation status. There are environmental factors that influence the methylation state at certain CpG sites [101]. Factors such as sex and lifestyle choices such as the use of tobacco, nutrition, and fitness [94, 104] are correlated and could thus result in an inaccurate age prediction. Another issue to take caution of in the current studies is the reliance on selection of markers which are based on microarray analysis. As it is a more recently explored area, coverage of as many loci as possible initially is desirable. Bead chip arrays allow for that, and most published studies have used the Illumina HumanMethylation27 or HumanMethylation450 arrays, with coverage of 27,578 or 485,577 sites, respectively [96, 105]. However, the arrays are limited by the available probes as defined by the company that developed the chip array, targeting CpG islands, shores, or shelves (approximately 85% of CpG content in HumanMethylation450 array) [106]. Shores are genomic regions between 0-2kb from a CpG island, and shelves are those within 2-4kb from the island [96]. These arrays are important for discovering methylation patterns related to human diseases, especially cancer, as many cancer-related genes and their promoter regions are preferentially targeted and therefore there exists some selection bias in the panel of array markers [105].

The objective of this work is to develop a novel genome-wide sequencing method to find candidate CpG sites that can be potentially informative for age prediction for a broad range of forensically relevant tissues: saliva, blood, and semen.

3.1.2 Methyl RAD Sequencing

To discover further candidate age informative CpG sites, we propose to use a novel method termed methylation restriction-site associated DNA sequencing, or,

methyl-RADseq. This method is based on the analysis of restriction-site associated DNA (RAD) at high resolution using a sequencing platform. Traditional RADseq combines, typically, Illumina sequencing with the use of restriction enzymes and the ability to use barcodes to associate sequencing reads to a particular individual or sample [107]. It allows for genome-wide, high density SNP genotyping and genetic mapping, but is less expensive and faster than whole genome re-sequencing efforts [108, 109]. Other methods, including the Infinium arrays, utilize bisulfite conversion of the DNA (conversion of unmethylated cytosines to uracil and ultimately to thymine during PCR [110]), which can be difficult for mapping. The method developed here differs from conventional RADseq method in the use of the isoschizomeric methylation sensitive and insensitive enzymes, *Hpa*II and *Msp*I, respectively. Therefore potential methylated CpG sites are differentiated by fragments produced where the methylated cytosines are cut by *Msp*I, and not cut by *Hpa*II (Figure 3.1). The reads are mapped to a reference genome, and the methylation state at CpG sites across the genome are computed. The first step is the restriction enzyme digest of the DNA sample to generate the target DNA fragments (Figure 3.2 shows a schematic of methyl-RADseq). For this method, a double digest is performed with *Eco*RI and either *Hpa*II or *Msp*I. Following digestion, the fragments are ligated with two custom-designed adaptor oligonucleotides containing complementary sequences to the enzyme cut sites on each fragment end, with a 5bp barcode adaptor and a common adaptor (Figure 3.2). Following ligation, PCR is performed to amplify the adaptor-ligated restriction digest fragments. The PCR primers were designed with phosphothiolate-modified bases on the first two bases on the 5’ end that inhibit endonuclease and exonuclease from acting on the ligated DNA fragments during

amplification (Figure 3.2). The PCR primers also have Illumina sequences incorporated to be recognized during sequencing. Following PCR, the fragments are purified, size selected, and quantified for sequencing.

3.2 Methods

3.2.1 Sample Collection

Three samples (semen, buccal (saliva), and blood) were collected from 5 male volunteers aged 25, 33, 41, 45, and 69 years old (Indiana University IRB Approved Protocol #1402819847). Buccal swabs were taken from each volunteer for the saliva samples. Blood was collected by using single-use blood sampling devices (Unistik 2 Extra, Owen Mumford, Oxford, UK) on a finger of the volunteer's choice. Volunteers were provided a sterile falcon tube to take home and return with a semen sample. All samples were stored at -20°C upon collection or receipt.

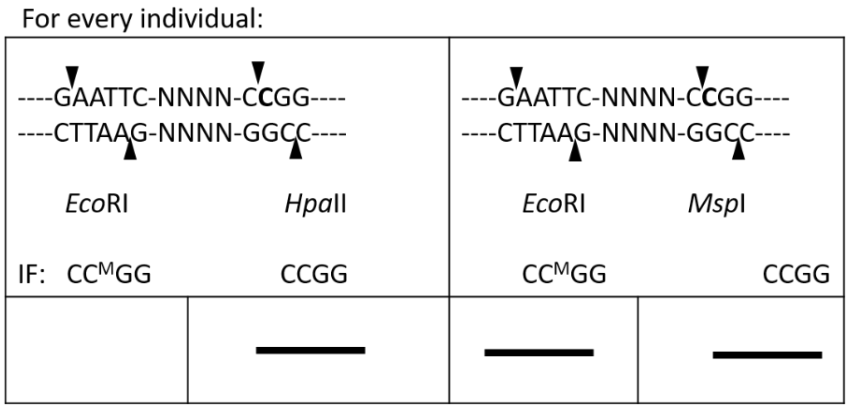


Figure 3.1 Schematic diagram illustrating the principle of methylation status using methyl-RADseq. *Msp*I will cut at the recognition site and generate fragments regardless of the methylation status of the cytosine (right) whereas *Hpa*II will not cut at the recognition site and no fragment will be generated if the cytosine is methylated (left).

Figure 3.2 Methyl-RADseq schematic of primer binding for sequencing.

3.2.2 DNA Extraction and Quantitation

The organic DNA extraction protocol as described in section 2.2.2 was used to extract the DNA from the buccal samples. For the blood samples, 50 μL of blood was added to a 1.5mL tube with 50 μL of proteinase K and 200μL of ChargeSwitch lysis buffer (Thermo Fisher Scientific) and incubated for only 4 hours at 56°C and extracted as described in section 2.2.2. For the semen samples, 50 μL of semen was added to a 1.5mL tube with 200 μL of sperm lysis buffer which was prepared as follows: 350 μL ChargeSwitch lysis buffer (Thermo Fisher Scientific), 25 μL proteinase K, and 40 μL of 390mM dithiothreitol (DTT) (Sigma-Aldrich Corp., St. Louis, MO), and incubated for 8-12 hours at 56°C and extracted as described in section 2.2.2. Quantitation of all the extracted DNA samples was performed in duplicate with the Quantifiler® Human DNA quantification kit (Thermo Fisher Scientific) following the manufacturer's protocol.

3.2.3 Methyl-RADseq Sample Preparation

The methyl-RADseq sequencing protocol was a modification of an existing RAD protocol from the University of Wyoming [111]. The semen, saliva, and blood samples for each individual were pooled in equal amounts for approximately 400 ng of DNA in 6 μL for digestion. Sample 6972 (69 year old) did not have sufficient DNA extraction from the semen sample, and therefore only the blood and saliva were used. For each individual, two digest reactions were set up (Table 3.1). The 10X T4 buffer (New England Biolabs Inc., Ipswich, MA), BSA (Promega Corp., Madison, WI), *Eco*RI (Promega Corp.), and NaCl (Acros Organics, Thermo Fisher Scientific) all have the same input amount, the only difference is the use of *Msp*I (New England Biolabs Inc.) or *Hpa*II (New England Biolabs Inc.) (Table 3.1).

Table 3.1 Double digestion protocol setup f or each sample.

Reaction Component	*Eco*RI + *Msp*I reaction (μL)	*Eco*RI +*Hpa*II reaction (μL)
DNA (398 ng)	6.0	6.0
10X T4 buffer	1.0	1.0
1M NaCl	0.5	0.5
BSA (10 mg/mL)	0.05	0.05
*Eco*RI (5,000 U)	0.7	0.7
*Msp*I (20,000 U) or *Hpa*II (10,000 U)	0.2	0.4
H_2O	0.2	0.0
Total Volume	8.65	8.65

Following digestion, the *Eco*RI and *Msp*I*/Hpa*II double stranded adaptors were ligated to the digested fragments. For the adaptor preparation (Table 3.2), 10μM of each adaptor was produced using an Eppendorf Mastercycler Pro (Eppendorf) thermal cycler to heat the oligonucleotide to 95°C for 5 minutes, then slowly cooled to room temperature by decreasing the temperature by 2°C every 50 seconds. Ligation was performed using 10X T4 buffer (New England Biolabs Inc.), 1M NaCl (Acros, Thermo Fisher Scientific), BSA (Promega Corp.), and T4 DNA ligase (400,000 U/mL, New England Biolabs, Inc.) using an Eppendorf Mastercycler Pro thermal cycler (Eppendorf) at 16°C for 2 hours with heated lid at 20°C followed by a 4°C hold. The ligated products were diluted by adding 90 μL of TE buffer (Thermo Fisher Scientific). PCR (Table 3.3 for primer sequences) was performed using iProof™ High-Fidelity PCR kit (Bio-Rad Laboratories, Hercules, CA) on an Eppendorf Mastercycler Pro thermal cycler (Eppendorf) with the following conditions: 98°C for 30 seconds, followed by 30 cycles of 98°C for 20 seconds, 60°C for 30 seconds, and 72°C for 40 seconds, a final extension at 72°C for 10 minutes, followed by a 4°C hold.

Table 3.2 Methyl-RADseq Adaptor Sequences

Adaptor	Adaptor Sequence
ComAd1	CGGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
ComAd2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
Bar1a	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**TGCGA**
Bar1c	AATTTCGCAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
Bar2a	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**CGCTT**
Bar2b	AATTAAGCGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
Bar3a	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**TCACC**
Bar3b	AATTGGTGAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
Bar4a	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**CTAGC**
Bar4b	AATTGCTAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
Bar5a	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**ACAAA**
Bar5b	AATTTTTGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Bold = barcode identifier

Table 3.3 Methyl-RADseq PCR Primer Sequences

PCR Primer	Primer Sequence
Ill-pcr1b	A*T*TGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
Ill-pcr2b	C*A*AGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

* = phosphothiolated base

The PCR products of the digested, ligated DNA were pooled into a single library and purified using Agencourt AMPure XP beads (Beckham Coulter Inc., Brea, CA), QC analysis on an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA) and size selection on a BluePippin (Sage Science, Inc., Beverly, MA) (Genomics and Bioinformatics Core Facility at the University of Notre Dame).

3.2.4 Methyl-RADseq MiSeq Preparation

The MiSeq sequencing was performed in duplicate. To prepare the library for massively parallel sequencing, the Illumina MiSeq System Denature and Dilute Libraries Guide was followed, available on the Illumina website [112]. Briefly, the library was diluted to a final concentration of 2pM. The library was then spiked with Illumina Phi-X sequencing control v3 (Illumina, San Diego, CA) at 35% for run 1 and 20% for run 2. The spiked library was then loaded onto an Illumina MiSeq Reagent Kit v2 cartridge (Illumina) and loaded onto an Illumina MiSeq FGx System (Illumina) following the manufacturer's protocol and set to collect FastQ only. There were 2 paired-end runs performed, producing 4 FASTQ files, 2 forward read files and 2 reverse read files. However, the reverse run failed during run 1 and was not further analyzed.

3.2.5 Methyl-RADseq Computational Analysis

Several programs were used to extract the potential candidate CpG sites from the sequencing data. The first phase of analysis was to convert the MiSeq generated FASTQ files to BAM files. This was performed using Galaxy (https://usegalaxy.org), an open source, web-based platform with many bioinformatics tools for data analysis [113]. Within Galaxy, the FASTQ files of run 1 forward reads and run 2 forward and reverse reads were uploaded. The forward and reverse reads from run 2 were merged using

PEAR [114] with the following parameters: a minimum overlap of 5 bases, maximum length of 302 bases, a minimum length of 150 bases, and an upper bound quality score of 30. This reduced the data from 3 files to 2: forward reads for run 1 and run 2. Barcode Splitter, a component of the FASTX-Toolkit, was then used to parse out the two FASTQ files into the different aged individuals based on the barcodes. The resulting files were then trimmed and filtered for quality using the Filter FASTQ tool [113] with the following parameters: minimum length of 50 bases, minimum quality score of 30, and maximum number of bases outside of quality range of 15 bases (90%). The quality trimmed and filtered reads were then concatenated between the 2 runs into one FASTQ file for each age and enzyme, creating 10 files: 25Hpa, 25Msp, 33Hpa, 33Msp, 41Hpa, 41Msp, 45Hpa, 45Msp, 69Hpa, and 69Msp, corresponding to the age and restriction enzyme. Each FASTQ file was then mapped against the human genome (hg38) using Bowtie2 [115], generating BAM files for each of the 10 files. The BAM files were uploaded to the Mason cluster at Indiana University. Within Mason, BEDTools v2.26.0 [116] was used to convert the BAM files to BED files, as well as compute genome coverages and read counts. BEDOPS v2.4.26 [117] was then used to find the *Msp*I-cut regions that were in common between all 5 ages; these are considered the control regions. The *Msp*I regions where *Hpa*II fragments were either found in common or found to be unique (where *Hpa*II did not map) from the control regions were extracted. The *Msp*I control regions were used to compare against 35 known age informative CpG sites that were found in multiple studies or between multiple tissues that used Infinium HumanMethylation27 or HumanMethylation450 arrays (Table 3.4). The *Msp*I regions of interest were combined and read counts were normalized between the *Msp*I and *Hpa*II

regions for each age. The regions were filtered using Excel to find those where at least one age had at least 10 reads (10X coverage). The regions were also filtered to those that either decreased or increased in methylation when comparing the *Hpa*II to *Msp*I read counts from the youngest to oldest ages and the final list were those which had an r^2 value of at least 0.7 (70% correlation). The 0.7 threshold was chosen as it is a similar to the 70% threshold used in the pigmentation prediction modeling, however, a more stringent threshold can be considered during validation. The final list of regions were then viewed via the BAM files on the UCSC genome browser (http://genome.ucsc.edu) [118], and annotated to the NCBI RefSeq database within the browser. Gene ontology (GO) terms of the biological processes were identified for the genes using the PANTHER database, as well as GO term enrichment analysis [119].

3.3 Results and Discussion

The number of reads mapped to the human genome, following quality filter and trimming, and genome coverages can be seen in Table 3.4 along with the normalization factors between the *Msp*I and *Hpa*II reads for each age. After filtering by the 0.7 correlation threshold and exclusion of the sex chromosomes, there were 491 candidate CpG sites (Appendix F); an example of the correlation seen with 5 sites can be seen in Figure 3.3. When comparing the 35 known CpG sites (Table 3.5) to the regions of the sample data, there was one found in common: cg08097417 within the *KLF14* gene. There are two reasons as to why more known sites may not have been found. The first is that most of the known CpG sites are within CpG islands (Table 3.5), which again are generally defined as genomic regions with long stretches (> 500 bp) that have a GC content of greater than 55% and excludes *Alu* elements [120]. The Infinium

HumanMethylation arrays were designed to specifically emphasize genomic regions that included genes and CpG islands [96]. The methyl RADseq method does not target any specific feature within the human genome, however it is limited to regions where there is enzyme cut site recognition, of which only one end of the fragment is a CpG recognition site (*Msp*I or *Hpa*II). Secondly, many of the known sites are specific for only one type of tissue analyzed, mostly blood, and the data generated here was between 3 combined tissue types: blood, saliva, and semen. Many sites have been found to be tissue specific and therefore may not correlate beyond one tissue type. The CpG site in common, designated by the Infinium array as cg08097417, showed an increase in methylation with an r^2 value of 0.43. Again, previous studies have only showed this to be significant in blood [99, 121] and the signal for each individual fluid may not be optimally expressed as they are pooled together in this data; individual fluid correlations cannot be identified.

Table 3.4 Mapped read count and normalization factors

Age	Enzyme	Number of mapped reads	Average Genome Read Coverage	Median Genome Read Coverage	Normalization Factor (*Msp*I/*Hpa*II)
25	*Hpa*II	562862	6.3	2	1.5
25	*Msp*I	858317	5.7	3	
33	*Hpa*II	415208	5.9	2	1.9
33	*Msp*I	777150	5.4	3	
41	*Hpa*II	503236	6.0	2	1.4
41	*Msp*I	683160	4.8	3	
45	*Hpa*II	428555	6.0	2	1.4
45	*Msp*I	599591	4.5	2	
69	*Hpa*II	467287	5.9	2	1.7
69	*Msp*I	787406	5.0	3	

Table 3.5 Known age-associated CpG sites from previous studies.

CpG Site	Gene	Chr	Start-End	Human Methylation Chip	CpG Pattern	Methylation Pattern (if reported)	Previous Studies	Tissue Type
cg07533148	*TRIM58*	1	247857510-247857511	27	ISLAND	hyper	[100, 122]	Saliva and Blood
cg09809672	*EDARADD*	1	236394382-236394383	27	none	hypo	[99, 100, 123]	Saliva and Blood
cg19945840	*B3GALT6*	1	1232656-1232657	27	ISLAND	hyper	[100, 124]	Saliva and Blood
cg06639320	*FHL2*	2	105399282-105399283	450	ISLAND	hyper	[99, 102, 121, 125]	Blood
cg11176990	*LOC375196*	2	38960392-38960393	450	ISLAND	hyper	[102, 121]	Blood
cg16232126	*SLC5A7*	2	107986549-107986550	27	ISLAND	hyper	[100, 124]	Saliva and Blood
cg22158769	*LOC375196*	2	38960398-38960399	450	ISLAND	hyper	[102, 121]	Blood
cg22454769	*FHL2*	2	105399310-105399311	450	ISLAND	hyper	[99, 121, 125]	Blood
cg24079702	*FHL2*	2	105399314-105399315	450	ISLAND	hyper	[99, 121, 125]	Blood
cg27320127	*KCNK12*	2	47571257-47571258	27	ISLAND	hyper	[100, 124]	Saliva and Blood
cg07553761	*TRIM59*	3	160450189-160450190	450	ISLAND	hyper	[99, 102, 121]	Blood
cg25148589	*GRIA2*	4	157220784-157220785	27	ISLAND	hyper	[100, 122]	Saliva and Blood
cg00059225	*GLRA1*	5	151924796-151924797	27	ISLAND	hyper	[100, 121, 124]	Saliva and Blood
cg19885761	*CPLX2*	5	175796643-175796644	27	ISLAND	hyper	[100, 124]	Saliva and Blood

Table 3.5 continued

cg16867657	*ELOVL2*	6	11044644-11044645	450	ISLAND	hyper	[18, 99, 121, 125, 126]	Blood
cg215572722	*ELOVL2*	6	11044661-11044662	450	ISLAND	hyper	[121],[18, 125]	Blood
cg24724428	*ELOVL2*	6	11044655-11044656	450	ISLAND	hyper	[18, 121, 125]	Blood
cg22736354	*NHLRC1*	6	18122488-18122488	27	ISLAND	hyper	[123, 127]	Blood
cg08097417	*KLF14*	7	130734372-130734373	450	ISLAND	hyper	[99, 121]	Blood
cg12799895	*NPTX2*	7	98617340-98617341	27	ISLAND	hyper	[100, 122]	Saliva and Blood
cg12837463	*ZC3H12A*	7	35260617-35260618	450	N_SHORE	hypo	[128]	Semen
cg19594666	*LEP*	7	128241227-128241228	27	ISLAND	hyper	[124]	Saliva and Blood
cg23571857	*BIRC4BP*	7	6593466-6593467	27	none	hypo	[122]	Saliva and Blood
cg15747595	*TSPYL5*	8	97277652-97277653	27	ISLAND	hyper	[124]	Saliva and Blood
cg16219603	*PENK*	8	56448027-56448028	450	ISLAND	hyper	[121, 125]	Blood
cg18898125	*NEFM*	8	24912868-24912869	450	N_SHORE	hyper	[102]	Blood
cg01530101	*KCNQ1DN*	11	2869868-2869869	27	ISLAND	hyper	[122]	Saliva and Blood
cg06979108	*NOX4*	11	89589683-89589684	450	ISLAND	hyper	[128]	Semen
cg01820374	*LAG3*	12	6772917-6772918	450	N_SHORE	hypo	[121, 123]	Blood

Table 3.5 continued

cg18236477	*ATP8A2*	13	25468928-25468929	27	ISLAND	hyper	[100, 124, 129]	Saliva and Blood
cg06304190	*TTC7B*	14	90817262-90817263	450	S_SHORE	hypo	[128]	Semen
cg04875128	*OTUD7A*	15	31483692-31483693	450	ISLAND		[99, 102]	Blood
cg21801378	*BRUNOL6*	15	72319784-72319785	27	ISLAND	hyper	[100, 124, 129]	Saliva and Blood
cg19283806	*CCDC102B*	18	68722183-68722184	450	none		[99, 102]	Blood
cg07547549	*SLC12A5*	20	46029586-46029587	450	ISLAND		[99]	Blood

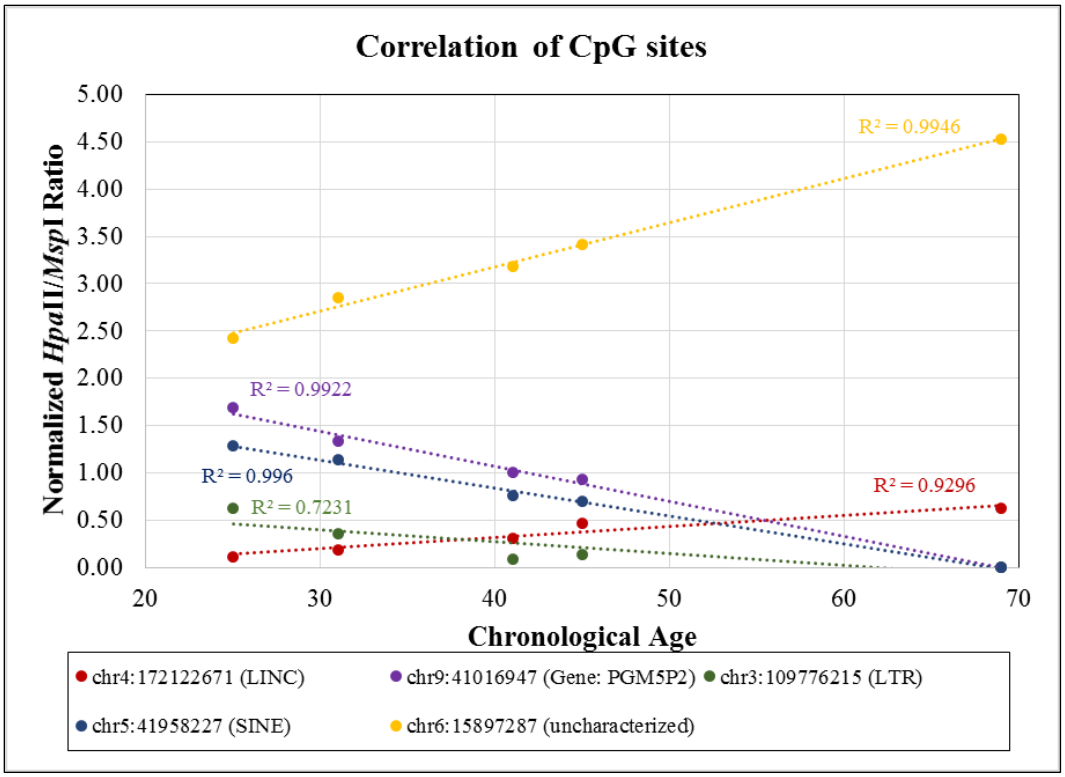


Figure 3.3 Example of individual CpG site age correlations. The variation in methylation correlation ($r^2 \geq 0.70$) and type of methylation (hyper or hypo) can be seen in the 5 sites shown, as well as the linear regression correlation values (R^2).

One of the most highly significant age-correlated CpG sites found in blood is cg16867657 in the *ELOVL2* gene, and our data did have reads that mapped to this gene, however, they were not located at this specific site and were not found to correlate in a linear pattern with age. Distribution of the candidate CpG sites by chromosome and genomic features can be seen in Figure 3.4. Interestingly, only chromosome 10 did not have hypermethylated sites, only hypomethylated sites were found correlated to age.

There were 222 genes found with at least one candidate CpG site located near (flanking region of 1000bp or less) or within the gene (45% of sites, Figure 3.4b). Thirteen of these genes were uncharacterized, and three genes had 2 CpG sites found: *CDH11*, *NLGN1*, and *PPARGC1B*.

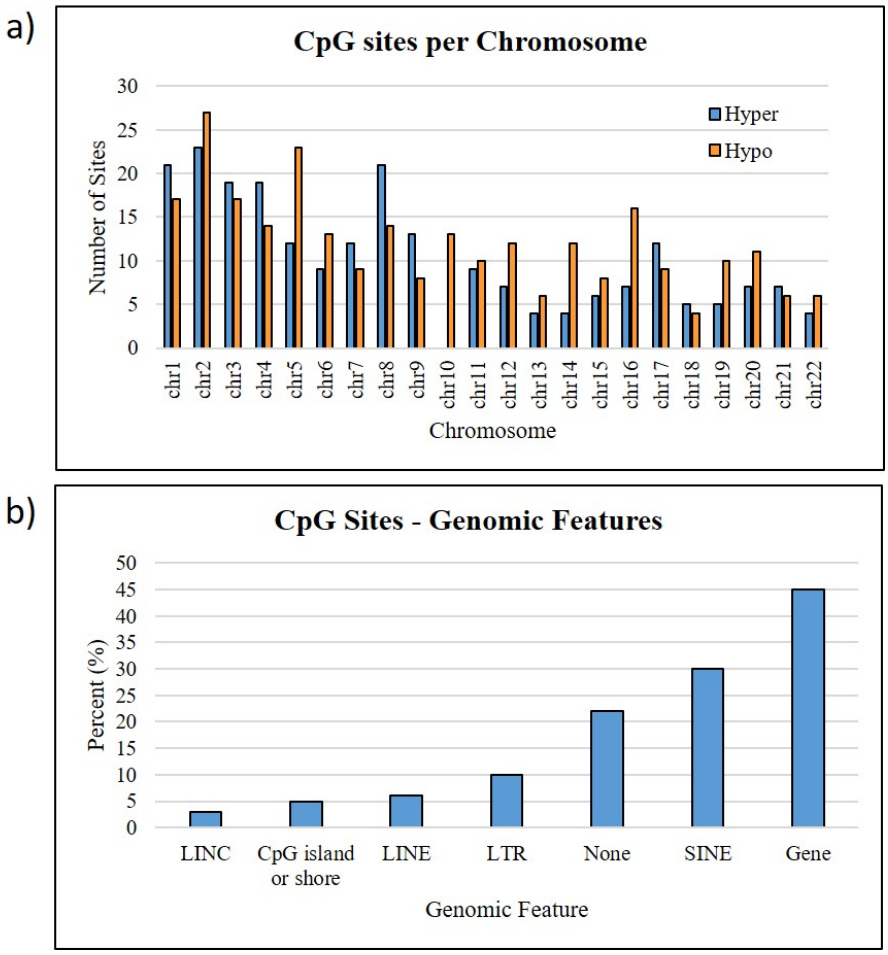


Figure 3.4 Candidate CpG site distribution. a) The number of hypermethylated and hypomethylated CpG sites by chromosome, and b) genomic features represented by CpG sites.

Gene ontology (GO) analysis resulted in 197 genes clustered into biological process GO terms, a breakdown of these categories can be seen in Figure 3.5.The largest group of clustered genes related to cellular processes (88 genes). Cell communication was the largest subsection (43 genes) which includes cell-cell signaling (8 genes) and signal transduction (37 genes). Further breaking down the signal transduction genes shows an almost equal split between two categories: cell surface receptor signaling pathways (17 genes) and intracellular signal transduction (19 genes), which includes G-protein coupled receptor signaling, transmembrane receptor serine/threonine kinase and tyrosine kinase,

and cytokine-mediated signaling. Aging is considered a time-dependent functional decline mainly caused by the accumulation of cellular damage [130]. Many of these highlighted genes associated with intercellular communication, which is thought to be one of the hallmarks of aging [130].

Metabolic processes was the next largest category (56 genes) with most related to primary metabolic processes (42 genes), followed by responses to a stimulus (29 genes) which includes responses to stress, cellular defense, external stimulus, and immune response. There were 10 genes relating to immune system processes. One of these genes, *MAP3K13*, was found related not only to immune response, but stress response, cell death, cell morphogenesis, phosphate-containing metabolic process, and NF-kappaβ cascade. Immune response decreases with age as there is a decline in both T- and B-cell function and the development of a chronic inflammatory state, referred to as "inflammaging" [130, 131]. The NF-kappaβ signaling pathway regulates developmental processes, host defense (innate immunity), and cell survival functions e.g., inhibition of apoptosis [132]. The methylation of the CpG site at this gene was found to increase, suggesting inactivation of the pathway which would lead to more cell apoptosis, and downregulation of immune response, both which are expected due to aging.

Other cellular processes important in aging can be related to transcription regulation, nuclear trafficking and organization, protein translation, proteostasis, autophagy, mitochondrial dysfunction, and cytoskeleton and membrane integrity [133]. There were 14 genes were found associated to cellular component organization, e.g., chromatin and cytoskeleton organization and protein complex assembly. There were 22 genes associated to both multicellular organismal processes and biological regulation which includes processes such as blood circulation, muscle contraction, sensory and visual

perception, calcium ion homeostasis, catalysis, and DNA/RNA binding transcription activity. These categories are logical when considering the physiological changes that occur with aging, many different factors are controlled by different cell signaling pathways, ion channel transporters, and transcriptional regulators. GO enrichment analysis of these categories showed that only the cell communication category had significant enrichment ($p < 0.05$) when compared to the human genome. As discussed above, altered intercellular communication, which can include deregulation of signaling and increase in inflammation, is considered a hallmark of aging [130]. Furthermore, it has been shown that senescent cells can induce senescence in neighboring cells via gap junction-mediated cell-cell contact, which may contribute to the overall aging process [134].

Approximately half (45%) of the candidate CpG sites were associated with a gene. However, there were other genomic features abundantly represented within the candidate sites including: short interspersed nuclear elements (SINEs), long interspersed nuclear

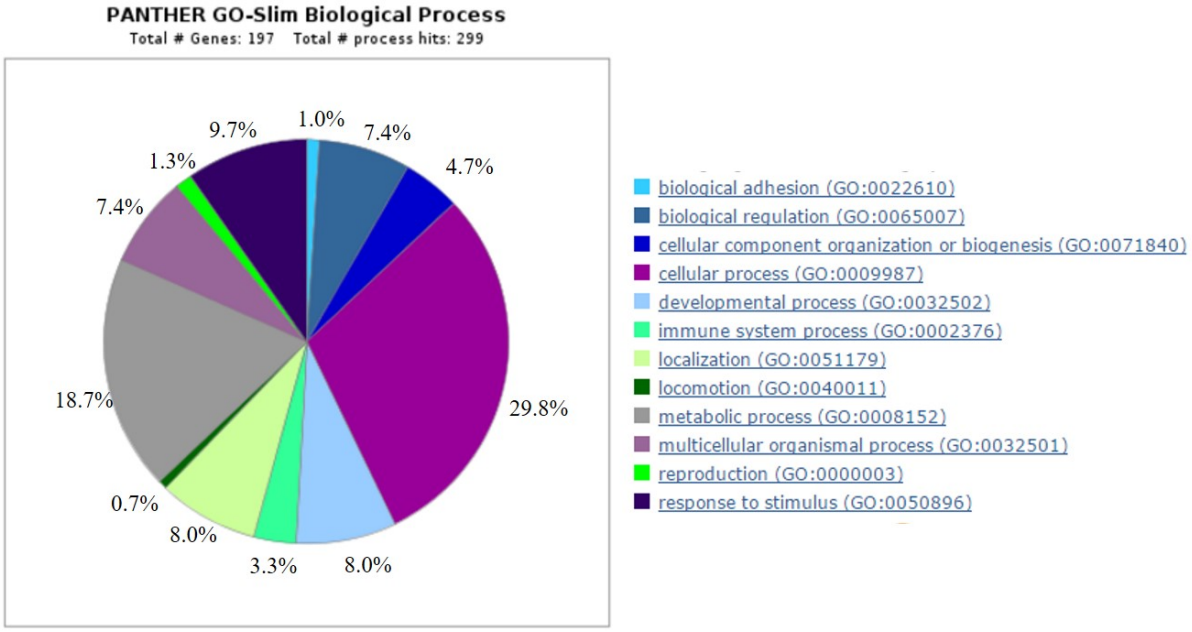


Figure 3.5 GO terms of the biological processes of the genes associated with the candidate CpG sites.

elements (LINEs), long terminal repeats (LTRs), and long intergenic non-protein coding RNA (LINCs). Of the regions in which the candidate CpG sites were located, 31% were SINEs, 6% LINEs, 10% LTRs, and 2% LINCs. SINEs, LINEs, and LTRs are the three major families of transposable elements that are abundant in mammalian genomes [135].

The most common SINE in the human genome and also found in the candidate CpG sites is the *Alu* family of elements. Transposable elements are a significant source of regulating signals for transcription [136] and therefore it is not surprising they would have an impact on gene regulation which may be important during the aging process. It has also been shown that DNA methylation of the heterochromatin of major retrotransposon elements decreases during senescence, and the chromatin structure becomes more open, leading to an increase in transcription and ultimately transposition [137]. While these heterochromatic regions, which includes transposable elements, become more open in gene-poor regions, it is also found that euchromatin in gene-rich regions become more closed [137]. In further support of this, the expression of LINEs and SINEs has been found to increase with normal aging in humans and mice [135]. As for LINCs, they have been found to be influential in the molecular processes of age-associated phenotypes by impacting cellular processes including proliferation, differentiation, senescence, and response to stress and immune agents [138]. LINCs can serve as recruiters of chromatin-modification factors, assemblers of transcriptional activators or repressors, maintenance of nuclear compartmentalization, and regulators of post-transcriptional gene expression [138]. One well known example is *Xist*, which is responsible for inactivating the X chromosome [139]. There was a candidate CpG site found within the LINC-PINT region, which has been found responsible for histone modifications and eliciting the TGF-β, MAPK, and p53 pathways which are associated with senescence and aging [138, 140].

Many LINCs have been identified but not yet characterized and this can also be seen in the data.

One last feature of known importance for aging are CpG islands. Within the candidate CpG sites, 25 (5%) sites were near or within CpG islands. Again, many of the known age-associated CpG sites are found in CpG islands, shores, or shelves as those are the targets of interest for the Infinium arrays typically used. CpG islands are typically located in promoter regions near the transcription start sites (TSSs) and can regulate the transcriptional activity of the gene by increasing in methylation to inhibit transcription of the gene as most CpG islands are inherently non-methylated regions [141, 142]. CpG islands found within the gene body may enhance transcription elongation or play a role in alternative splicing and inhibiting transposable elements [142]. Several studies show that methylation patterns typically increases within CpG islands, but decreases at non-island CpG sites with age [142, 143]. This pattern is not reflected in the data, however, as only 8 CpG sites (32%) of all the sites found within CpG islands or shores were hypermethylated with age.

3.4 Conclusions

The methyl RADseq method developed here has generated data locating potential candidate CpG sites that may be correlated with age-related changes and therefore age prediction. Many of the genes associated with the candidate CpG sites have been found related to cellular processes and functions that influence aging. Other genomic features include CpG sites in sites of repetitive elements, such as SINEs and LINEs, which has also been found to regulate processes important in the aging process. Unlike the Infinium array studies, only a small proportion of the data was found within or near CpG islands.

The Infinium arrays target CpG islands and promoter regions, however, it has been shown that methylation levels that correlate to gene expression levels are concentrated in non-CpG island regions [142]; therefore hypomethylated sites outside of CpG islands and promoter regions should be further investigated.

Methylation changes are critical in normal human development and aging, but as previously discussed, variation in methylation may also indicate diseases, especially cancer, as well as be influenced by environmental factors. In this study, 5 healthy male individuals were used to provide proof of concept and generate this preliminary data. To be able to control the results more with environmental factors, additional information regarding the volunteers' lifestyle should be collected, and therefore this data is promising, but further testing of the candidate CpG sites found with a larger cohort of individuals will be necessary. This method is not without limitations; the candidate CpGs found are restricted to those recognized by the enzyme recognition sites, and some regions did show incomplete enzyme digestion. Furthermore, as the 3 biological fluids were initially pooled, individual fluid correlations cannot be identified. However, this method is versatile and may be adapted for single tissue assessment if that is the desired goal. Furthermore, it is an alternative to the selection bias of the Infinium HumanMethylation arrays to explore and discover other potential age-correlated CpG sites. Recently, the next level of Infinium bead chip array has been developed, MethylationEPIC, which covers over 850,000 CpG sites and extends coverage to a larger proportion of non-CpG island regions to CpG islands and therefore reduces the selection bias of the earlier arrays [144]. For those laboratories without the resources to invest in microarray methods, or if they have interest in investigating methylomes of species other than human, methyl-RADseq is an alternative sequencing method for discovering methylated differences in genomes.

CHAPTER 4. EFFECTS OF MICROBIAL DNA ON HUMAN DNA PROFILES

**This work has been submitted for publication and is currently under review to the Journal of Forensic and Legal Medicine.*

4.1 Abstract

Most crime scenes are not sterile and therefore may be contaminated with environmental DNA, especially if a decomposing body is found. Collecting biological evidence from this individual will yield DNA samples mixed with microbial DNA. This also becomes important if postmortem swabs are collected from sexually assaulted victims. Although genotyping kits undergo validation tests, including bacterial screens, they do not account for the diverse microbial load during decomposition. We investigated the effect of spiking human DNA samples with known concentrations of DNA from 17 microbe species associated with decomposition on DNA profiles produced using the Promega PowerPlex® 16 HS system. Two species, *Bacillus subtilis* and *Mycobacterium smegmatis*, produced an extraneous allele at the TPOX locus. When repeated with the PowerPlex® Fusion kit, the extra allele no longer amplified with these two species. This experiment demonstrates that caution should be exhibited if microbial load is high and the PowerPlex® 16 HS system is used.

4.2 Background

As the sensitivity of human DNA profiling has increased, the types of samples being successfully profiled has also increased, including samples with little human DNA ("touch" DNA) [145], genotyping as little as single cells [1], or samples which may be

degraded [146]. Similarly, decomposed tissues often produce low template and degraded DNA samples in addition to the increased microbial presence. However, little attention has been paid to the effects of the high microbial load on the resulting DNA profiles. Increased sensitivity of DNA genotyping, along with the increased ability to amplify even single cells of human DNA, may also increase the possibility of amplifying environmental DNA sources, such as bacteria, that may have been co-extracted with the evidentiary sample. This is especially pertinent for swab collection when sexual assaults are suspected. When DNA is extracted from a substrate, all DNA is extracted; it is not limited to only human DNA. Before DNA amplification, the amount of human DNA must be determined. Since forensic DNA quantitation kits are human specific, it remains unknown how much non-human DNA is present in a forensic sample [147]. This exogenous DNA is not typically a problem because human forensic identification tests are also designed to be human-specific [147]. Furthermore, an important step in validating new forensic DNA profiling kits is performing a developmental validation, which includes amplification of other common species' DNA (such as domestic animals), and microbial pools of extracted DNA [148-151]. With such an obvious abundant source of external DNA present in forensic samples, it is surprising how little research has been done. The majority of scholarly articles associated with this type of DNA contamination are related to the identification of the pathogen rather than its effect on forensic DNA profiles [152-154]. The proliferation of microbes during decomposition should be of interest to the forensic community if the presence can possibly affect the outcome of human identification analysis, especially in sexual assault cases. Most sexual assault samples are collected from orifices such as the mouth and genital regions [155], and many microbial communities already thrive in these areas while the body is alive.

There are many species of bacteria present on a living human body; in fact the ratio of microbial to animal cells is 10:1 [11, 156]. In the mouth alone, 200 species have been found [156]. Other systems in humans, such as the intestinal tract, naturally contain 300-500 species of bacteria [157], which can translate to 100 trillion bacterial cells [156]. Some types of bacteria are still found in tissue and bone samples following death [158]. These living populations will grow rapidly at the onset of decomposition, which starts occurring 4 minutes after death [159]. This does not account for bacteria found in external environments, such as soil, that may also contribute to DNA collected from a decomposing body. In some cases, 50,000 microbial species can be found in one gram of soil [156].

In one study, the effect of 30 different microbial species on human DNA profiles (based on single locus amplifications) was examined [160]. Three of the 10 loci produced no artifacts (TPOX, TH01 and CSF1PO), and two produced non-specific artifacts (HLA-DQA1 and PM). However, at the D1S80 locus, fragments were produced within the range of true D1S80 alleles with six of the tested microbial species [160]. The artifact was present when only 100 pg of microbial DNA was amplified. However, this study does not offer much value today as D1S80 is no longer a locus of interest in current human DNA typing kits. Furthermore, the current STR multiplex kits are not tested against a comprehensive set of microbial species, although most validations include some microbial species specificity testing. For example, in the developmental validation study for the PowerPlex® 16 HS system, a small panel of microbial species is included to test for species cross-reactivity, however, it is a small subset: only 5 species (2 fungi, 3 bacteria) were assessed [149]. This small number of species is insufficient for testing a sample if collected from a decomposed individual.

A possible solution to handling the co-isolation of bacterial DNA with human DNA is simply cataloging the bacterial DNA that shows up in human profiles, which was the basis of this work. The ability to recognize and categorize any artifact will allow for reasonable and justified identification of extraneous peaks that may be present in DNA profiles. This becomes increasingly important as samples with low DNA concentrations are amplified, or more importantly, with mixtures. In this work, we intentionally spiked human DNA samples with microbial DNA and amplified them to analyze any effects the microbial DNA may have on the interpretation of the human DNA profile.

4.3 Methods

A total of 17 microbial species were analyzed as part of this pilot study (Table 4.1). Eleven species were cultured (American Type Culture Collection, ATCC, Manassas, VA). Lyophilized cultures (ATCC) were re-suspended in 5 mL of Tryptic Soy broth (DOT Scientific Inc., Burton, MI) and incubated overnight at 37°C. Cultures were streaked on Tryptic Soy plates (Tryptic Soy broth with 2% agar) to grow isolated colonies and incubated at 37°C. Cultures for DNA extraction were started by inoculating 5 mL Tryptic Soy broth (DOT Scientific Inc.), followed by overnight incubation at 37°C, then centrifuged and measured by weight (70 mg -130 mg). For each culture, an organic extraction was performed to extract the DNA. Briefly, the pellets were re-suspended in 200 µL of water, and incubated at 37°C for 1 hour with 5 µL of lysozyme (50 mg/mL, Thermo Fisher Scientific Inc., Waltham, MA). Following cell lysis, 25 µL of proteinase K (20mg/mL, Thermo Fisher Scientific Inc.) in 500 µL of lysis buffer (Invitrogen Corp., Carlsbad, CA) was added and the samples were incubated at 56°C overnight. Then, 500 µL of phenol/chloroform/isoamyl alcohol (25:24:1) (Thermo Fisher Scientific Inc.) was

added, gently inverted, and then centrifuged at 13,000 rpm for 1 minute. The aqueous layer was removed to a new tube and 500 µL of chloroform:isoamyl alcohol (24:1) (Thermo Fisher Scientific Inc.) was added, inverted, and centrifuged at 13,000 rpm for 1 minute. The aqueous layer was removed to a new tube and 25 µL of 0.2 M NaCl and 500 µL of 95% ethanol were added. The samples were centrifuged at 4°C for 15 minutes at 15,000 rpm. The supernatant was removed, and 500 µL of cold 70% ethanol was added and centrifuged at 4°C at 15,000 rpm for 5 minutes. The liquid was removed and the pellet was allowed to air dry before being re-suspended in 25 µL of TE buffer (Thermo Fisher Scientific Inc.).

The remaining six species were obtained as purified, freeze-dried, extracted DNA (ATCC), and were prepared by incubating the dried DNA overnight at 4ºC, then for 1 hour at 65ºC before being rehydrated with either 100 µL or 200 µL deionized water (depending on the initial starting concentration). All microbial DNA samples were quantified using the Qubit® 2.0 fluorometer (Thermo Fisher Scientific Inc.) following the manufacturer's protocol.

Buccal swabs were collected from three human volunteers (Indiana University IRB Approval Protocol #1507469161) and extracted using a phenol-chloroform extraction (as above, except lysozyme was not added) and quantified in duplicate following the manufacturer's protocol using the Quantifiler® Human DNA Quantification kit (Thermo Fisher Scientific Inc.) on a 7500 Real Time PCR System (Thermo Fisher Scientific Inc.). In order to examine the analytical effects of microbial DNA on human DNA profiles, 1.0 ng of one sample of human DNA was mixed with 10.0 ng, 50.0 ng, and 100.0 ng of each of the 17 microbial DNA samples (Table 4.1).

Table 4.1 List of microbial species used to spike human DNA samples, chosen because of their association in human decomposition.

Species
Bacillus subtilis
*Bacteroides fragilis**
Candida albicans
*Clostridium perfringens**
Proteus mirabilis
*Helicobacter pylori**
*Clostridium difficile**
Enterococcus faecalis
Escherichia coli
*Lactobacillus acidophilus**
Lactobacillus casei
Mycobacterium smegmatis
*Neisseria flava**
Pseudomonas aeruginosa
Staphylococcus aureus
Staphylococcus epidermidis
Streptococcus mutans

* = Purchased DNA from ATCC, not extracted from lab culture

The resulting mixtures (human DNA + microbial DNA) were re-quantified, in duplicate, using Quantifiler® Human DNA Quantification kit (Thermo Fisher Scientific Inc.). T-tests were performed (α=0.05) comparing the average quantitation values among each tested human + microbial DNA mixtures tested, and the average of a human DNA sample alone, tested in duplicate. DNA profiles were amplified using the PowerPlex® 16 HS system (Promega Corp., Madison, WI) following the manufacturer's protocol with 1 ng of human DNA from 2 of the human samples amplified alone, and also mixed with either 10, 50, or 100 ng of microbial DNA on a Veriti thermal cycler (Thermo Fisher Scientific Inc.) using the 9600 emulation mode. The fragments were separated and detected using capillary electrophoresis with 1 µL PCR product, 9.0 µL Hi-Di™ Formamide (Thermo Fisher Scientific Inc.), and 0.5 µL ILS 600 (Promega Corp.) on a 3500 Genetic Analyzer (Thermo Fisher Scientific Inc.) following the manufacturer's protocol. Data analysis was

performed using GeneMarker HID v2.6.0 genotyping software (SoftGenetics, State College, PA) with an analytical threshold of 100 RFUs. Any microbial species that produced artifacts were further tested by amplifying with the PowerPlex® Fusion system (Promega Corp.) at the same ratios following the manufacturer's protocol. The amplified fragments were then separated and detected as described above. Average peak heights were determined across all loci for each profile and statistical t-tests with Bonferroni correction were performed to look for deviations from the average RFU values of the human standard alone. Average heterozygosity values between species and amount of microbial DNA were also calculated among all loci for the combined human samples. This was done using Microsoft Excel by dividing the RFU values for the heterozygote alleles across all loci, the smaller peak height divided by the larger peak height. Each locus had at least one heterozygous genotype between the two human samples tested, and if both samples were heterozygous at the same locus, the average value was reported. In cases where an artifact was observed, the microbial DNA sample was amplified alone (as above) and in the third human sample to verify the artifact's presence. Buccal swabs were also collected and extracted from the two individuals working on this project as elimination standards.

4.4 Results and Discussion

The microbial species tested were chosen based on either their abundance in main locations on the human body such as the skin, mouth, gut, genital region, or anus [13, 161, 162], predominance at certain stages of decomposition [163], or if they are common environmental species, such as those found in soil [164]. The capability to culture some of the species was also a factor.

Human DNA quantitation was not significantly affected by the presence of microbial DNA at any concentration for any species (Table 4.2). Generally, addition of any microbial DNA produced profiles with similar average peak heights as seen in the human standard alone ($p > 0.05$ in all measurements after Bonferroni correction) (Figure 4.1). However, two microbial DNA species (with no overlap between the two samples in species or quantity of microbial DNA input) had statistically different average peak heights across the entire profile without any discernible pattern (Figure 4.1). For the average heterozygosity values calculated with the addition of microbial DNA to a human DNA sample, there were no significant differences. This indicates that the addition of any quantity of microbial DNA tested of any of the microbial species tested did not have an effect on heterozygosity balance of a human DNA profile (Figure 4.3).

Table 4.2 Average Quantifiler values showed no significant difference between the human DNA input values and the human + microbial DNA mixtures.

DNA Sample (Human input (ng):Bacteria input (ng))	**Average concentration (ng/µL) (± S.D.)**
Human only (N = 2)	0.959 (± 0.013)
1:10 (N = 26)	0.941 (± 0.049)
1:50 (N = 26)	0.918 (± 0.066)
1:100 (N = 26)	0.919 (± 0.037)

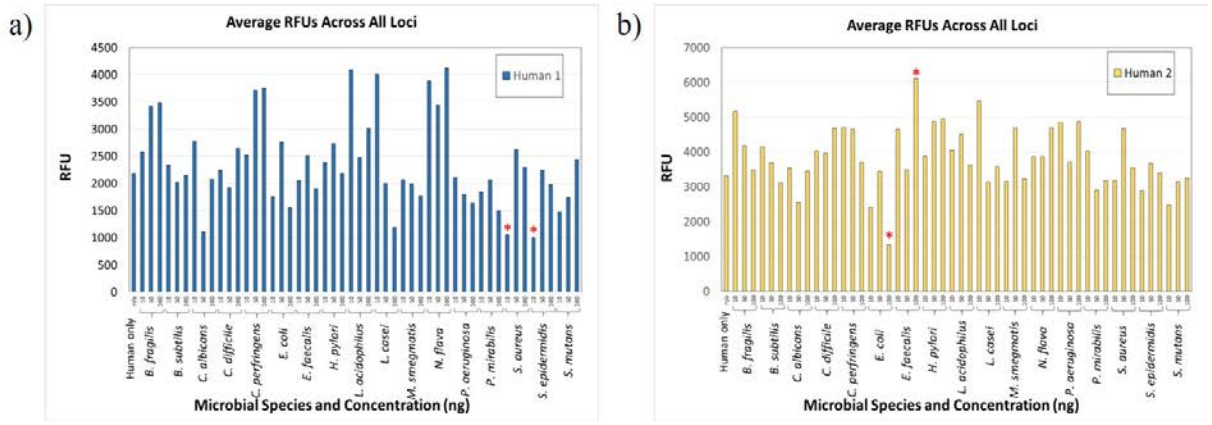


Figure 4.1 Average RFU values. The peak heights between the 2 human samples, a) Human 1 (s.d±791) and b) Human 2 (s.d.±856),and all tested microbial species and ratios (N=51). Only 4 samples showed a significant difference in average peak heights as compared to the human only standards. None of these significant differences were indicative of a trend among the species nor concentration of microbe DNA added to the sample. * = significant difference

More importantly, two of the 17 microbial species produced discordant DNA profiles (Figure 4.2). *Bacillus subtilis* and *Mycobacterium smegmatis* produced a significant artifact at all three microbial DNA input amounts in all three human samples tested (3rd sample was tested as confirmation): a drop-in of allele 5 of the TPOX locus (neither elimination standards contain this allele). The peak height of the artifact increased in intensity as the concentration of microbial DNA increased (Figure 4.2). The same peak was observed when the microbial DNA was amplified in the absence of human DNA (Figure 4.4). The two bacteria samples alone were also quantified with the Quantifiler® Human kit (Thermo Fisher Scientific, Inc.) to ensure there was no human DNA contamination in the samples. To do this, six species were tested, the two that amplified the artifact and four randomly chosen species. All species were quantified in duplicate at 100 ng bacterial DNA input and resulted in either undetermined, or a small quantity result, with an average of 0.0017 ng/µL (highest value of 0.002 ng/µL).

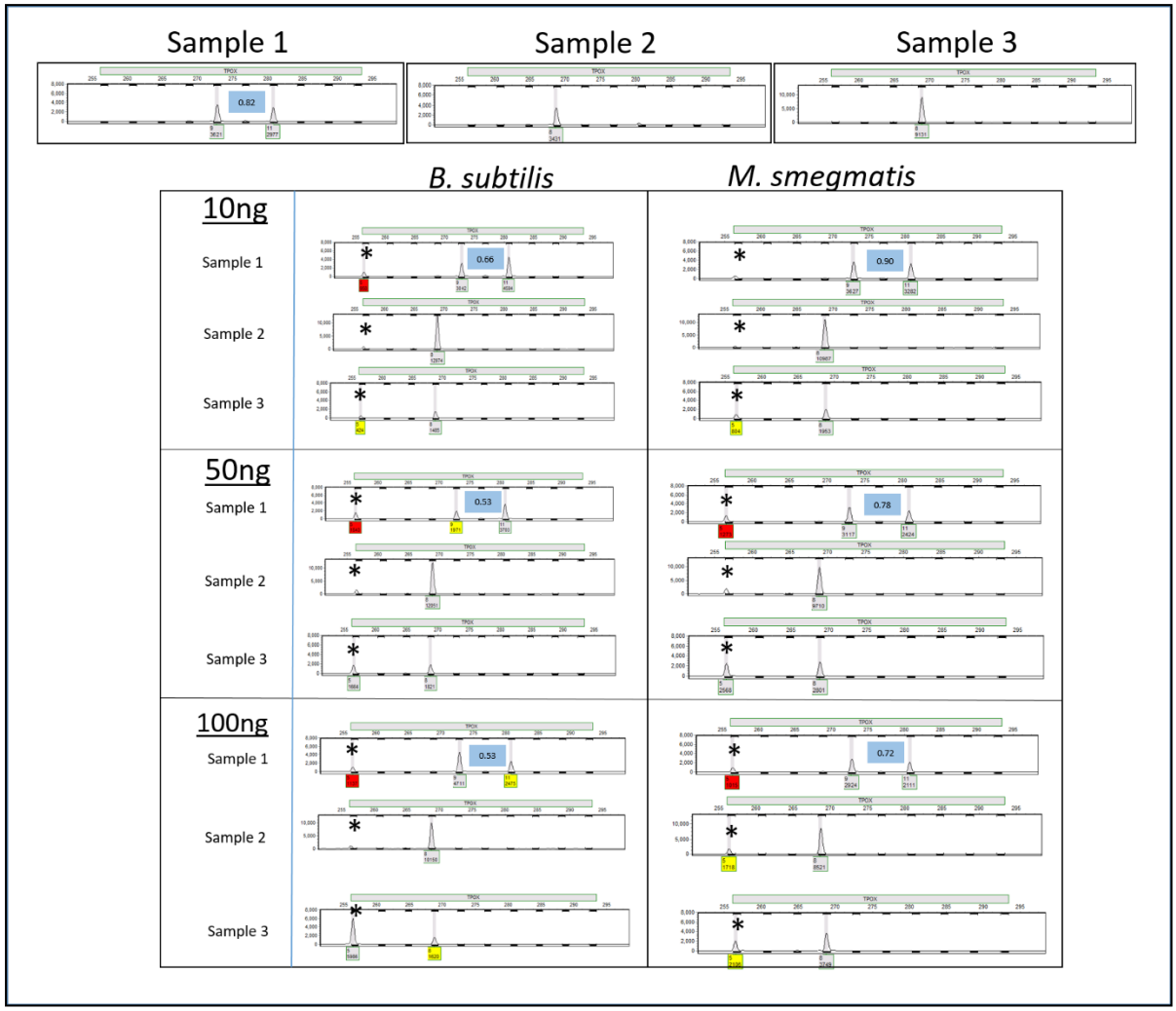


Figure 4.2 Microbial artifacts (*) at TPOX locus in three human DNA profile samples. Three human samples (1ng) were amplified alone (top panel), and with DNA from *B. subtilis* and *M. smegmatis* at 10ng, 50ng, and 100 ng. Allelic drop-in was produced at the same size as the '5' allele at the TPOX locus. In one case (Sample 3), the artifact allele of *B. subtilis* at 100ng is higher (RFU) than that of the allele of the actual human genotype at the locus. The blue boxes indicate reduced (but not significant) peak height ratios of the heterozygote genotype when the microbial DNA was present. Red boxes are produced by the genotyping software to indicate the detection of an extra allele at the locus when only a maximum of two are expected for a single source profile.

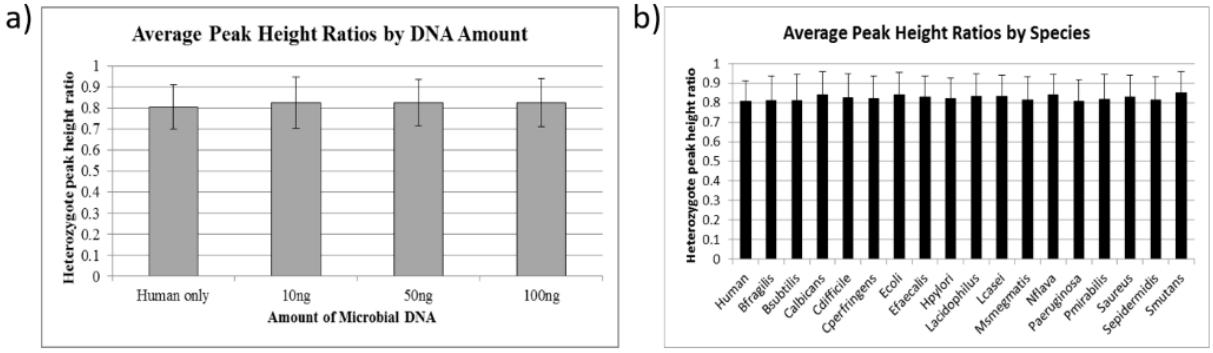


Figure 4.3 Average heterozygosity values of spiked human samples among all microbial DNA species and all microbial DNA quantities, tested in duplicate. a) Heterozygosity by microbial DNA amounts (N=34, N=2 for human only) and b) heterozygosity by microbial species (N=6, N=2 for human). There were no significant differences (all $p > 0.05$).

To test whether this amount of input DNA would produce the allele observed, we amplified a human DNA sample, in duplicate, at 0.002ng with the PowerPlex® 16 HS kit (Promega Corp.) to ensure no amplification would occur at this quantity to the level seen with the artifact. One human DNA replicate had no amplification, which was the same with the negative control. The second human DNA replicate had a drop-in allele at one locus only 13 RFU above the set analytical threshold (D8S1179, 14 allele, 113 RFU). The 14 allele at D8S1179 is not found in the tested individual's known profile (data not shown) and therefore attributed as a low template drop-in artifact. This demonstrates that this low level of human DNA input would not cause the level of amplification of the TPOX "5" allele seen with the microbial species. The TPOX "5" allele amplified by the two microbial species in the absence of human DNA was seen at 564 RFU in *B. subtilis* and 744 RFU in *M. smegmatis* (Figure 4.4).

We replicated the same experiment for these two species with a different kit, the PowerPlex® Fusion system (Promega Corp.), using one of the human samples (Figure 4.5). We note that this artifact (the TPOX '5' allele) was only found in the PowerPlex® 16 HS system, no artifact was produced using the Fusion kit (Figure 4.5).

Furthermore, no artifact was produced in the absence of human DNA for these two species using the Fusion kit (data not shown). This difference, however, is likely due to the fact that there is a different primer set used for the TPOX locus between the two systems (personal correspondence, Promega Corp.).

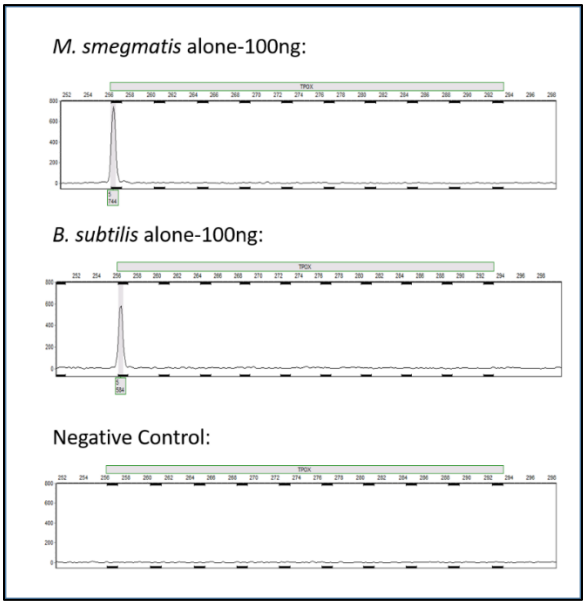


Figure 4.4 TPOX artifact produced with bacteria samples in the absence of human DNA using the PowerPlex ® 16 HS system. *B. subtilis* and *M. smegmatis* were amplified with 100 ng, along with a negative control, demonstrating the same artifact is produced in the absence of human DNA**.**

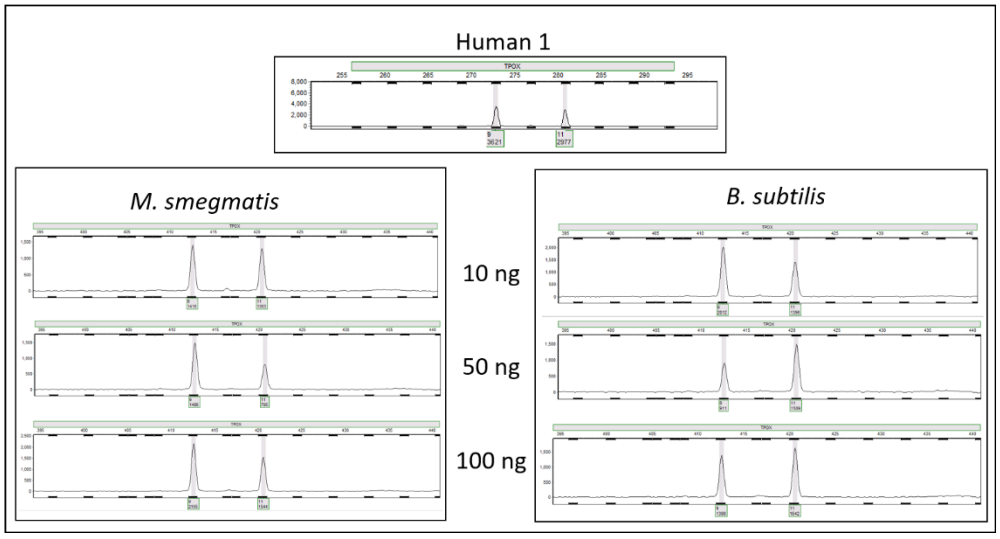


Figure 4.5 The human sample tested with the PowerPlex® Fusion system at the TPOX locus with 1:10, 1:50, and 1:100 ratios with *B. subtilis* and *M. smegmatis*. No other alleles other than the human profile is produced.

4.5 Conclusions

The possibility of observing exogenous DNA from the environment, specifically microbial DNA, affecting human DNA evidence has been shown here for 1 ng of input DNA. STR genotyping kits are designed to be human-specific, but the presence of microbes in forensic samples ensures microbial DNA will be co-extracted with the human DNA and possibly affect human DNA profile interpretations. There was little to no effect on the DNA quantitation, and little effect on the human DNA profiles with the majority of the microbial species tested here. However, two bacterial species, *B. subtilis* and *M. smegmatis*, produced a peak the same size as the "5" allele of the human TPOX locus, verified in three individual profiles. *Bacillus subtilis* is found in the human gastrointestinal tract [165, 166], and *Mycobacterium smegmatis* is found in the genital regions of both men and women [167]. Both species are found in soils, water and plants [164, 168] and both phyla of these species (Actinobacteria and Firmicutes) have also be found during decomposition [163]. Only 17 species of possibly thousands of species were tested herein. With the creation of the Human Microbiome Project (HMP) and more recent studies on identifying the microbiome of human decomposition [11, 13, 169], it would be important to further test for effects of additional decomposition bacteria on human DNA profiles. One other consideration of possible microbial species is the stage of decomposition of the body from which evidence is collected, as it has been found that the microbial profile shifts throughout the process, generally from aerobic to anaerobic bacterial species [163]. Additionally, low DNA concentrations were not tested here; it remains unknown whether their quantitation and DNA profile interpretations would be impacted by the presence of environmental DNA, and this should be further investigated. Furthermore, as additional loci are incorporated into DNA genotyping kits, other

commercially available kits should be evaluated – this study simply highlights how a single kit produced an artifact that could be used in the interpretation of a DNA profile.

Worst case scenario in this situation might lead to the possible exclusion of a suspect due to their lack of the '5' allele at the TPOX locus. As the most extreme example: a forensic sample is recovered, microbial DNA is co-extracted, and the sample is genotyped to produce a heterozygote locus (5,8, as seen in Figure 4.2 for sample #3 at 50ng input microbial DNA). The true contributor, or suspect, in this scenario (sample #3) is homozygous at the TPOX locus (8,8). If all alleles fall above stochastic threshold for the remainder of the profile, which may be rare occurrence but cannot be ruled out, then, an analyst would need to exclude the suspect as a contributor due to the inconsistencies in the TPOX locus genotypes (Carl Sobieralski, Indiana State Police, personal communication). Thus, we urge caution in DNA profile interpretation in any instances where human decomposition has occurred and the possibility of co-extracted microbial DNA exists, especially when using PowerPlex® 16 HS amplification kits.

A better knowledge of microbial genomes that have a propensity to interfere with the interpretation of human DNA can help overcome these quality effects. Information regarding the effects of foreign DNA on human sources suggests that STR genotyping kits that are designed to be human specific may be negatively affected by the presence of bacterial DNA. We can speculate that with new technologies, such as the use of massively parallel sequencing for DNA profiling, these artifacts would be designated as such (presumably the artifacts would have different sequences), however, we do not have any data to support this. Having knowledge of these effects provides the analyst with a system of categorizing DNA profiles from samples with possible extraneous microbial DNA contamination, such as with decomposition. It is not likely possible to eliminate

microbes (and therefore microbial DNA) from human DNA samples, but recognizing the effects they cause allows additional information for an analyst to assess and explain artifacts found in a sample, especially those involved in decomposition cases.

CHAPTER 5. ESTIMATING NUMBER OF CONTRIBUTORS IN THEORETICALLY GENERATED MIXTURE PROFILES

5.1 Introduction

Mixtures are a common challenge in DNA STR profile interpretation. DNA mixtures are more frequently encountered in forensic casework than in the earlier years of STR typing. This is mainly because of increased kit sensitivity and increased cycle number, and also because there are more requests for 'touch' or low copy number (LCN) DNA samples to be tested (Carl Sobieralski, Indiana State Police, personal communication). One laboratory published a study where they retroactively reviewed 1547 cases for the 4 years (1997-2000). Out of 2424 samples reviewed in the study, 163 (6.7%) showed a mixture profile, where only 8 of the 163 (0.3%) samples were mixtures of more than two contributors [170]. A decade later, a survey study initiated by SWGDAM in 2008 collected case data from 14 laboratories on 4541 samples, where 45.2% showed a mixture profile, and 33.6% of those samples had 2 contributors and 526 (11.6%) samples were mixtures of more than two contributors [171]. This survey was the basis of the 2010 SWGDAM DNA interpretation guidelines to focus on single source and two person mixture samples [171], although updated SWGDAM guidelines do include criteria for more than 2 contributors [172].

The first step in interpreting a DNA profile is identifying the presence of a mixture, or, a profile with more than one individual. This is typically determined by analyzing the number of allelic peaks, and peak height ratios, while considering stochastic effects, especially stutter products. According to SWGDAM guidelines, if one or more loci have 3 or more alleles present, excluding tri-allelic loci, then the sample is assumed to

be a mixture [172]. The next logical step is determining the number of contributors in that mixture. This is a key step to begin deconvolution of the mixture to assign genotypes to each individual present and providing statistical weight to the evidence. The most common approach for estimating number of contributors in a mixture profile is maximum allele count [171]. Maximum allele count is used to estimate the number of minimum contributors to the mixed sample by evaluating the locus that has the greatest number of allelic peaks [172]. This is because a single individual will only have a maximum of 2 alleles at a locus. For example, if a locus in a profile has 5 allele peaks (all above the stochastic threshold), there has to be a minimum of 3 contributors because for a two person mixture, you would expect the maximum number of alleles to be seen at a locus to be 4.

There have been some previous studies to characterize the number of contributors according to maximum allele count. Paoletti et al. [173] generated conceptual 3 and 4 person mixtures from an FBI database which contained genotypes from the 13 common CODIS STR loci from 959 individuals. Based on maximum allele count, they found that 3% of the 146,536,159 three-person mixtures could be mischaracterized as two-person mixtures, and that 76% of the 57,211,376 four-person mixtures could be mischaracterized as two- or three-person mixtures [173]. Haned et al. [174] also conducted simulations from published genotypes of individuals with 15 STR loci (13 of which are the core CODIS loci) by generating 1000 mixtures comprised of 2-5 contributors to compare maximum allele count with maximum likelihood [174]. They concluded that mixtures of 2 or 3 contributors was greater than 90% for both methods, but with mixtures of 4 or 5 contributors, maximum likelihood yielded greater success rates (2-15 fold higher) [174].

To the authors' knowledge, there have not been any published studies to evaluate how well the number of contributors can be determined for mixture profiles using the expanded core STR loci, which has increased to 20 loci. The PowerPlex® Fusion 6C system (Promega Corp., Madison, WI) incorporates 27 loci, which includes the expanded 20 CODIS core loci [175]. This kit was recently internally validated for use by the Indiana State Police (ISP) laboratory in 2016. It was the objective of this work to evaluate how the maximum allele count method would determine number of contributors for theoretically generated combined 2-, 3-, 4-, 5-, and 6-person mixtures (4,976,355 total mixture profiles) using the PowerPlex® Fusion 6C kit (Promega Corp.).

5.2 Material and Methods

Single source reference DNA profiles were amplified from non-related anonymous volunteers collected by the Indiana State Police Laboratory. The automated DNA sample processing using the BioMek NX and BioMek 3000 Automated Workstations standard casework operating procedures of the Indiana State Police Laboratory were used to generate the DNA profiles [176]. For amplification, the PowerPlex® Fusion 6C System (Promega Corp.) was used following standard casework operating procedures which follow the manufacturer's protocol. The genotypes, each designated with a random number identifier, were entered into an electronic database using Microsoft Excel for theoretical mixture generation and analysis. There was a total of 236 profiles used to generate the 4,976,355 mixture combinations.

A macro using Visual Basic in Microsoft Excel was used to generate all possible combinations of 2- and 3-person mixtures. The macro was also used to generate combinations for the 4-, 5-, and 6-person mixtures, however, due to the large number of

possible combinations and the limitation in number of rows possible in Excel (1,048,576 rows), only a random subset of all possible combinations were analyzed (see Results). Two separate sample sets were generated for the 4 person combinations to ensure allele count distributions were representative of the whole set (see Results). For analysis of all the mixture combinations, all statistics were performed in Microsoft Excel. As the kit has 3 Y-STR markers, a separate analysis of the Y-STRs was performed whereas the generated mixture combinations were filtered to analyze those between males only.

5.3 Results and Discussion

The number of profiles generated can be seen in Table 5.1. The distribution of the two separate sets of 4-person mixtures can be seen in Table 5.2. For all the mixture combinations, the profiles were considered under assumed ideal conditions where contributors are in equal input ratio (1:1, 1:1:1, etc.), there was no stutter or artifacts, and all alleles were above the stochastic threshold. Therefore, allele count analysis was calculated based on the assumed presence of all possible allelic peaks from all individuals in the mixture. This does not reflect the possibilities of mixed ratios, stutter, or allele-dropout, which are not unexpected in casework mixtures [177, 178]. Each set of mixtures were analyzed with the following defined parameters. The minimum allele count is the count across all loci per profile that had the lowest number of alleles observed in at least one locus. The maximum allele count is the count across all loci per profile that had the highest number of alleles observed in at least one locus. The overall count is the frequency distribution of all allele counts across all loci of all possible *n* person profile combinations.

Table 5.1 Number of mixture combinations.

Number of Contributors	Number of mixtures generated (number of database samples used)
2 person	27,730 (236)
3 person	2,162,940 (236)
4 person	916,895 (70)*
5 person	962,598 (43)*
6 person	906,192 (32)*
TOTAL	4,976,355

*= subset of total possible combinations

Table 5.2 Comparison of allele distributions between two separately generated combinations of 4-person mixtures (N=916,895). There was no significant difference ($p > 0.5$).

4-person Mixture Combinations: Allelic Distribution		
Allele	**Group 1**	**Group 2**
1	5555	11204
2	533410	623641
3	3467181	3671196
4	6601424	6289826
5	5805377	5596867
6	3236480	3307998
7	1194533	1302849
8	244625	285004
TOTAL	21088585	21088585

5.3.1 Two-Person Mixtures

For the 27,730 two-person mixtures, the minimum allele count was 2 alleles in 70% of mixtures, and 1 allele in the remaining 30% of profiles (Figure 5.1a). Although it was still possible to see loci with only 1 allele, which is more typical in single source profiles, when you consider maximum allele count (Figure 5.1b), 99.99% of the time there is at least one locus that has 4 alleles. There were 4 (0.01%) profiles that had a 3 allele maximum in at least one locus. In either case of a 3 or 4 maximum allele count, a minimum of 2 contributors would still be indicated (Figure 5.1b). Based on maximum

allele count, 2-person mixtures could accurately be determined as having 2 contributors in all cases. SE33, D1S1656, and Penta E are among the loci that have the highest frequency of the highest allele count of 4 (Figure 5.2).

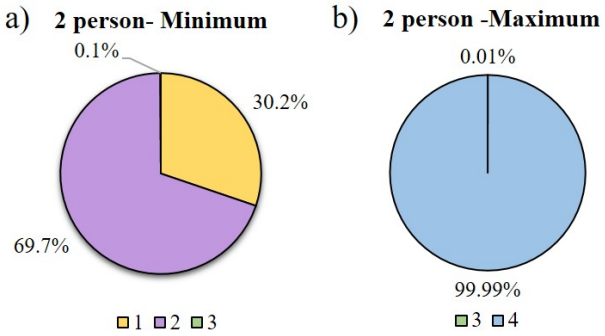


Figure 5.1 Two-person mixture allele counts. a) The minimum allele count distribution; and b) the maximum allele count distribution.

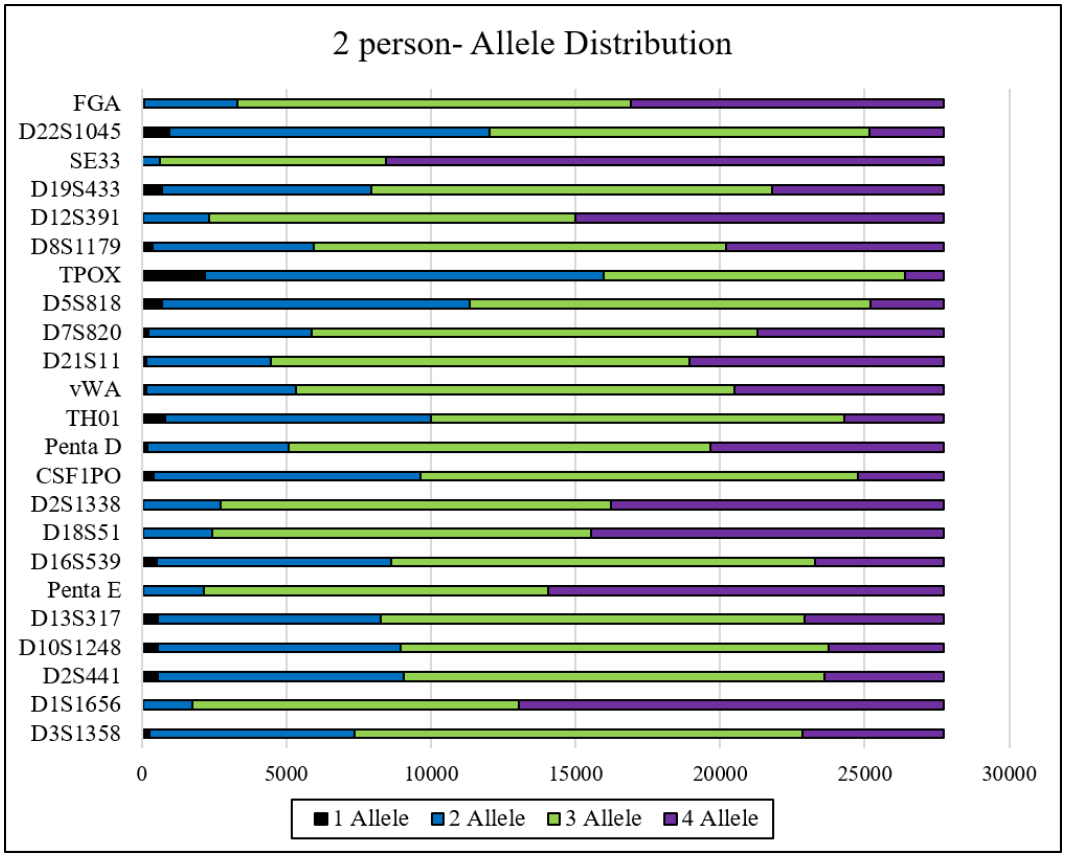


Figure 5.2 Frequency of allele counts by locus of the 2-person mixtures (N= 27,730).

5.3.2 Three-Person Mixtures

For the 2,162,940 three-person mixtures, the majority (80%) exhibited at least one locus with a minimum of 2 alleles (Figure 5.3a). There were still 100,831 (4.7%) profiles where there was at least one locus with a minimum allele count of 1 allele (Figure 5.3a). One 3-person mixture (0.0005%) was shown to have a minimum of 4 alleles, meaning no other loci had less than 4 alleles across the whole profile. Over 2 million profiles were generated and only one profile exhibited this pattern indicating how rare it is for three individuals to have at least 4 unique alleles between them at every locus. The maximum allele count method is accurate for 3-person mixtures in that 99.99% of the time (21.5% with 5 alleles, 78.5% with 6 alleles) there was at least one locus in the mixture profile that had 5 or 6 alleles, both of which indicate a minimum of 3 contributors (Figure 5.3b); there were 277 (0.01%) profiles which may have been confused as a mixture with only 2 contributors. SE33, D1S1656, and Penta E remain the loci with the highest frequencies of the highest allele count of 6, whereas TPOX had the highest frequency of only 2 alleles (Figure 5.4).

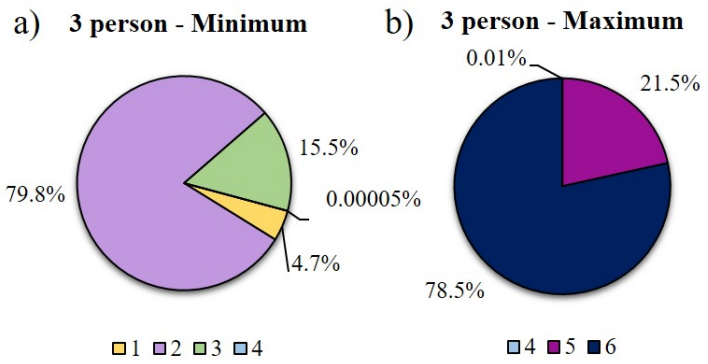


Figure 5.1 Three-person mixture allele counts. a) The minimum allele count distribution; and b) the maximum allele count distribution.

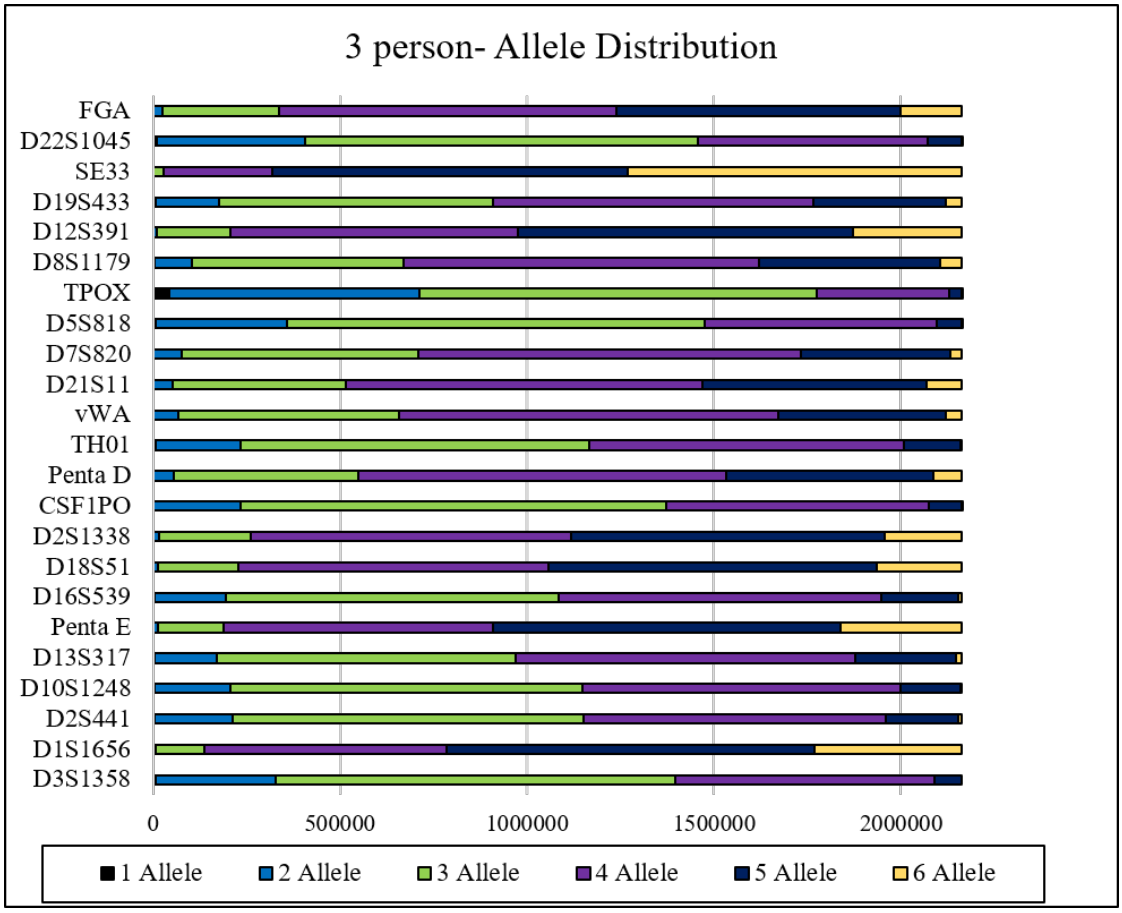


Figure 5.2 Frequency of allele counts by locus of the 3-person mixtures (N= 2,162,940).

5.3.3 Four-Person Mixtures

Minimum allele counts in the 916, 895 four-person mixture profiles were nearly evenly split between 2 and 3 alleles, 52% and 46.7%, respectively (Figure 5.5a). There were 11,174 (1.2%) of profiles that still showed at least one locus with only 1 allele within the profile (Figure 5.5a). At least one locus exhibited 7 or 8 alleles in approximately 90% (61% and 28.7%, respectively) of observed 4-person mixtures (Figure 5.5b). However, there were 94,880 profiles (10.3%) that would be confused as a 3-person mixture as there was only a maximum count of 5 or 6 alleles in at least one locus across the profile. The loci that have higher allele count frequencies for the highest possible allele count of 8 remain SE33, D1S1656, and Penta E as seen for the two- and three- person mixtures,

although the frequencies are relatively lower here, indicating there are fewer profiles for 4-person mixtures that exhibit the maximum number of unique alleles at these loci (Figure 5.6).

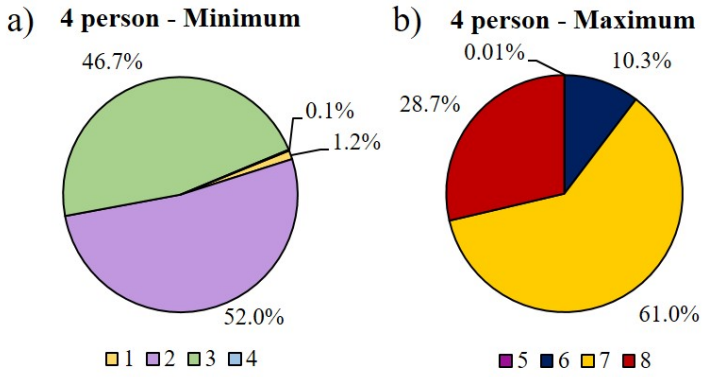


Figure 5.3 Four-person mixture allele counts. a) The minimum allele count distribution; and b) maximum allele count distribution.

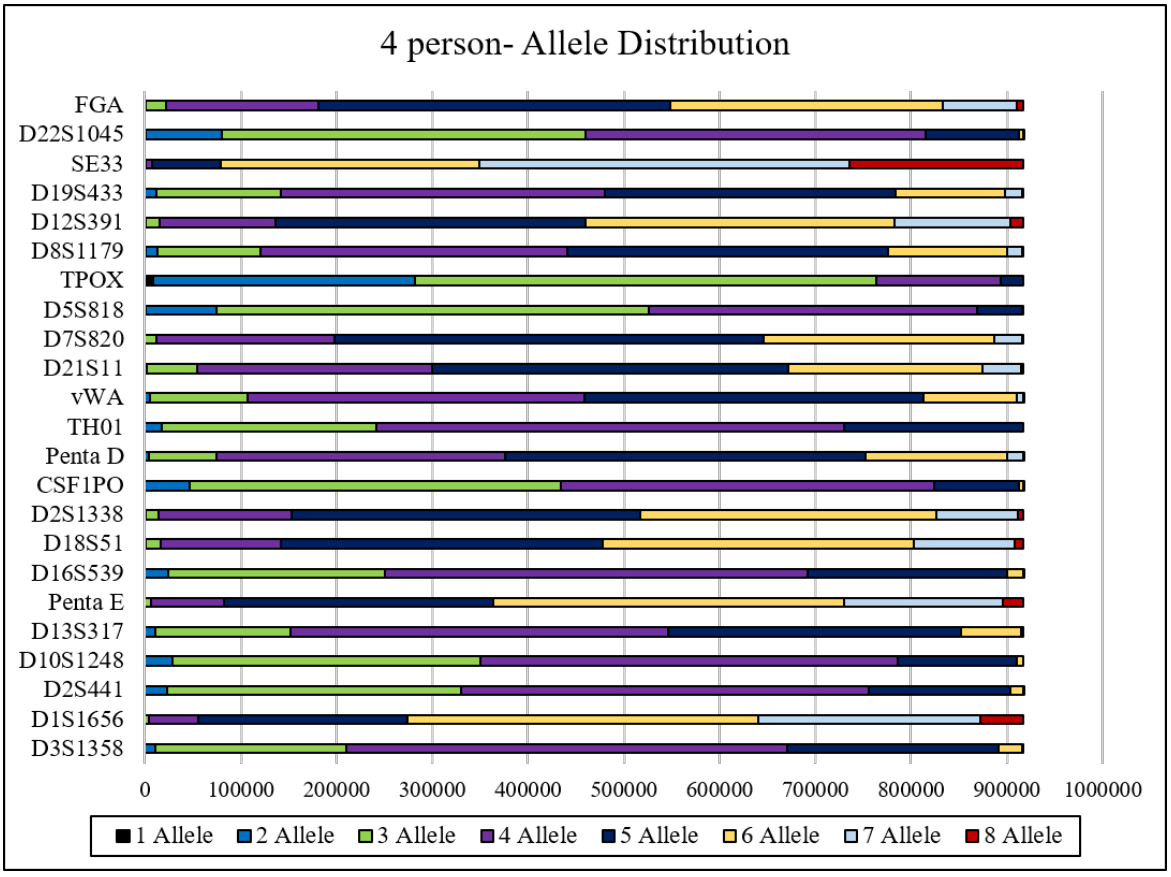


Figure 5.4 Frequency of allele counts by locus of the 4-person mixtures (N= 916,895).

5.3.4 Five-Person Mixtures

It was shown to be more difficult to discriminate a minimum of 5 contributors than mixtures with fewer contributors. The majority (72.5%) of the 962,598 profiles had at least one locus with a minimum of 3 alleles, followed by 24.5% with 2 alleles (Figure 5.7a). There were 3216 (0.3%) profiles that had at least one locus with only 1 allele. In terms of maximum allele count, 42.9% of profiles had at least one locus with 9 alleles, which is indicative of a 5-person mixture (Figure 5.7b). However, the next highest distribution of profiles (37.4%) were observed to have a maximum allele count of 8, which would indicate a 4-person mixture (Figure 5.7b). Approximately 57% of observed 5-person mixtures had at least one locus with 9 or 10 alleles, thereby making estimations of the number of contributors inaccurate for nearly half of the profiles generated here (Figure 5.7b). Per locus, SE33 still had the highest distribution exhibiting a maximum of 10 alleles with 204,446 (21%) profiles (Figure 5.8). The second and third highest loci with a maximum allele count of 10 were at a much lower distribution: D12S391 with 26,974 (2.8%) profiles, and Penta E with 16,964 (1.8%) profiles (Figure 5.8).

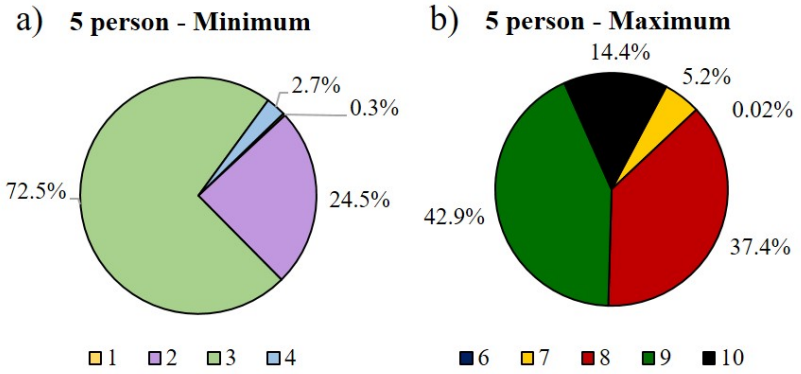


Figure 5.5 Five-person mixture allele counts. a) The minimum allele count distribution; and b) maximum allele count distribution.

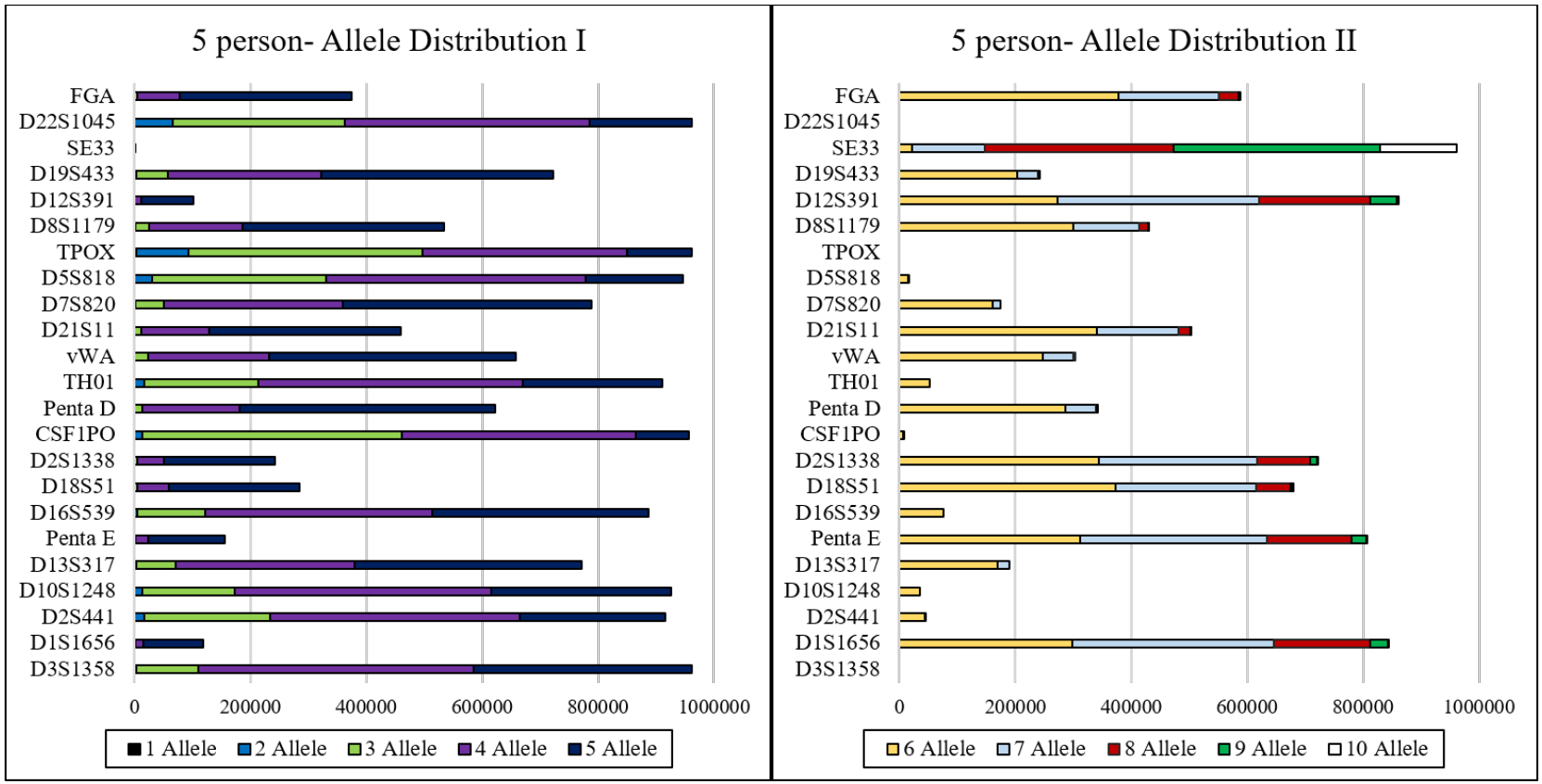
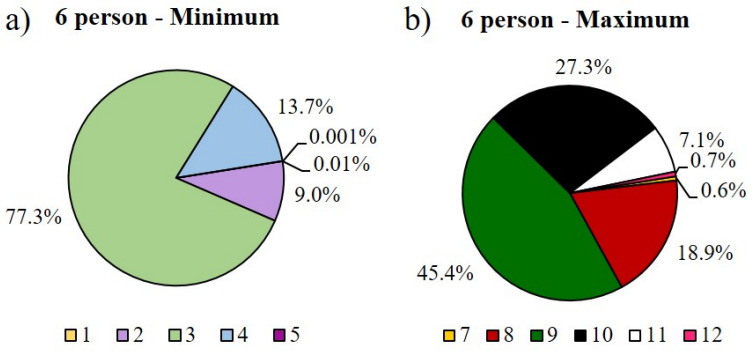


Figure 5.6 Frequency of allele counts by locus of the 5-person mixtures (N= 962,598)

5.3.5 Six-Person Mixtures

Six-person mixtures, as expected, were the most difficult to discriminate by allele count as compared to the 2-, 3-, 4-, and 5-person mixtures. The minimum allele count was similar to the 5- person mixtures with only approximately a 5% increase in profiles with at least one locus with a minimum of 3 alleles (72.5% to 77.3%, Figure 5.9a). There were 98 (0.01%) profiles where at least one locus only showed 1 allele (Figure 5.9a); conversely, there were 6.



int

distribution; and b) maximum allele count distribution.

The maximum count that would distinguish a 6-person mixture from a 5-person mixture would be 11 or 12 alleles, and only 71,176 (7.8%) of profiles had at least one loci with either 11 or 12 alleles (7.1% and 0.7%, respectively, Figure 5.9b). The most common maximum allele count was 9 with 411,324 (45.4%) profiles (Figure 5.9b). Based on these low frequencies, there is a low probability of accurately estimating the correct number of contributors from a 6-person mixture, they are most likely to be assumed 5-person mixtures. In terms of specific loci, only 3 were observed to produce profiles with 12 alleles, all with low frequency: SE33 with 6,332 profiles (0.07%), D12S91 with 68 (0.0008%) profiles, and D18S51 with 22 (0.0002%) profiles (Figure 5.10).

The two loci with the highest profile counts with 11 alleles were SE33 with 60,676 (0.7%) profiles and D12S391 with 2,464 (0.03%) profiles (Figure 5.10).

5.3.6 Overall by locus

As previously discussed, SE33 was the most polymorphic locus, with the greatest number of profiles with the maximum allele counts for mixtures with all number of contributors (Figure 5.11). This is not surprising considering SE33 has 58 distinguishable alleles, which is twice the number of unique alleles compared to FGA, which is the next most variable STR [179]. If SE33 is considered as the only locus, 97.7% of profiles have 3 or 4 alleles for 2-person mixtures, and with 3-person mixtures, 85.3% of profiles have 5 or 6 alleles which would indicate minimally 2 and 3 contributors, respectively (Table 5.2). Looking at allele count distribution across number of contributors, 70% of profiles with 4 alleles is observed at SE33 is a 2-person mixture, 44% of profiles with 5 alleles is a 3-person mixture, with much smaller percentages of 5 alleles seen in other mixtures (Table 5.2). The locus is less informative with the higher count of alleles, where 59% of profiles with 8 alleles could be a 5- or 6-person mixture, and if this is the maximum allele count, it would be incorrectly assumed a 4-person mixture (Table 5.2). For 5-person mixtures, 58% of profiles showed 9 or 10 alleles. For 6-person mixtures, only 7% of profiles at SE33 had 11 or 12 alleles observed. The majority of allele distribution for 6-person mixtures was 8 or 9 alleles (Table 5.2). All of this from a single locus, however, SE33 is not one of the expanded core loci. In fact, it is not included in many of the current amplification kits [180]. An issue also arises in successfully genotyping SE33, as it is susceptible to allele drop-out and mobility shifts due to mutations in the primer binding sites, causing discordant genotyping results between different kits [181].

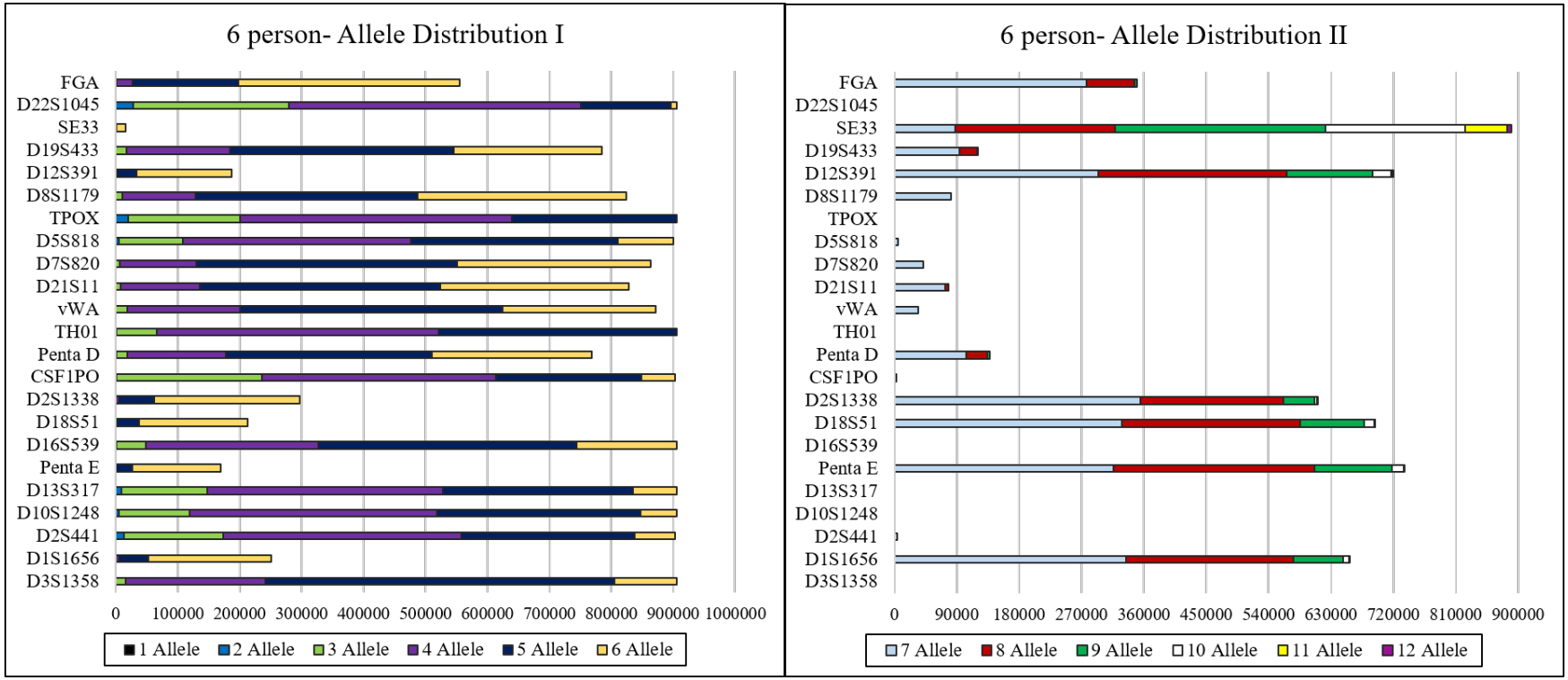


Figure 5.8 Frequency of allele counts by locus of the 6-person mixtures (N= 906,192).

When considering other loci, D12S391, which is one of the newly expanded core loci, and D18S51 were the only other loci to have a maximum of 12 alleles for the 6-person mixtures (Figure 5.11). Conversely, TPOX had the lowest allele maximum counts especially for 4-, 5-, and 6- person mixtures; there were only a maximum count of 5 alleles (Figure 5.11). This is not surprising as TPOX is considered the least polymorphic of 24 of the most commonly used STR loci [182] and has one of the highest allele frequencies of the autosomal loci (0.54 for allele 8 in the Caucasian population) [183]. All loci exhibited the possibility of maximum allele count of 4 for 2-person mixtures, and all loci but D3S1358 exhibited the maximum possibility of 6 alleles for 3-person mixtures (Figure 5.11). The maximum possible counts for 4-, 5-, and 6- person mixtures were more variable between all loci, with fewer loci exhibiting the higher possible counts (Figure 5.11). This is not a surprising trend as it is inherently difficult to deconvolute higher count mixtures.

Table 5.2 Frequency of 2-, 3-, 4-, 5-, and 6-person mixtures based on number of alleles seen at the SE33 locus among all observed mixture profiles.

Alleles	2 person (%)	3 person (%)	4 person (%)	5 person (%)	6 person (%)
1	0.03	0.00005			
2	2.3	0.03			
3	28.2	1.3	0.02		
4	**69.5**	13.4	0.75	0.002	0.001
5		**44.0**	7.8	0.14	0.1
6		41.3	29.4	2.2	1.7
7			**42.2**	13.1	9.7
8			19.8	33.8	25.4
9				**37.0**	**33.4**
10				20.9	12.4
11					6.7
12					0.7

Bold = highest distribution of alleles per *n*-person mixture

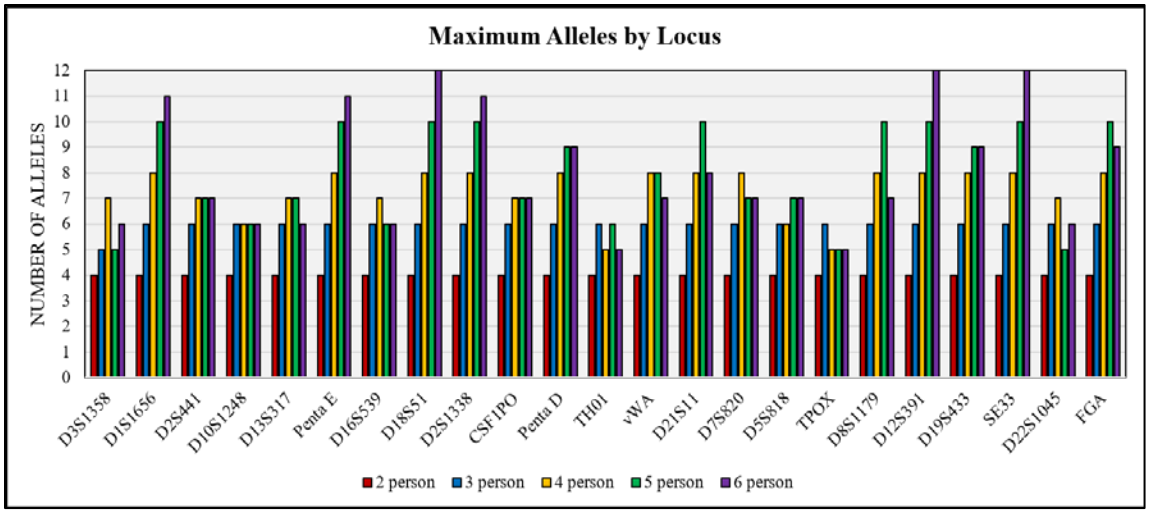


Figure 5.9 Maximum alleles per loci for all mixtures.

5.3.7 Total Allele Count Distributions

A DNA profile is considered as a whole and not just by a single locus. Another overall observation to make with all these mixture profiles is the average number of alleles seen across all 23 autosomal loci in the profile. Maximum allele count is important for assuming number of contributors, but if there are trends in number of overall allele counts and number of contributors, that could be important too. For 2-person mixtures across all loci in the profile, 49% of all alleles had 3 alleles, and 27.3% had 4 alleles (Figure 5.12a). When comparing 3-person mixtures to the 2-person mixture distributions, there is a decrease in the number of loci that exhibit 3 alleles, and an increase in those that exhibit 4 alleles (Figure 5.12a,b). Also, 7.2% and 5.8% of loci show 5 or 6 alleles, respectively (Figure 5.12b). When considering total allele count across a 4-person mixture compared to a 3-person mixture, there is a decrease in loci that exhibit 4 alleles and increase in those exhibiting 5 and 6 alleles (Figure 5.12b,c). However, when comparing between 4-, 5-, and 6-person mixtures, the distribution of loci,

other than the increase of the highest 2 allele counts between each, is very similar demonstrating that discriminating a 4-, 5-, and 6-person mixture based on allele counts from this set of 23 autosomal loci would be very difficult (Figure 5.12c,d,e).

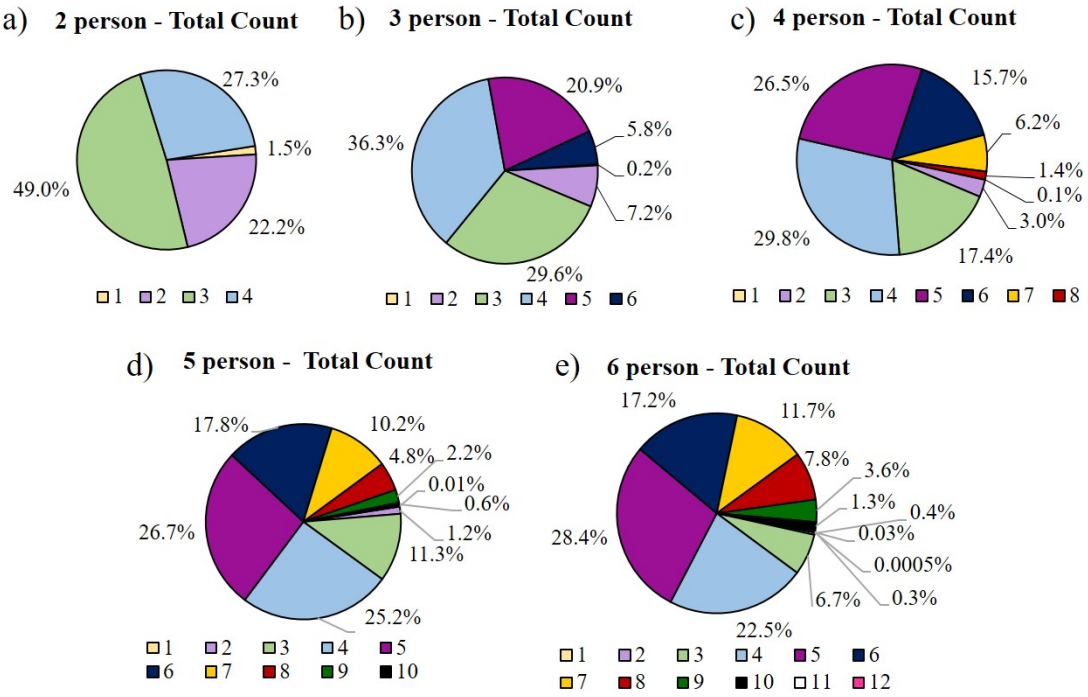


Figure 5.10 Total allele count distributions across all the autosomal loci across all observed profiles for a) 2-person mixtures, b) 3-person mixtures, c) 4-person mixtures, d) 5-person mixtures, and e) 6 person mixtures.

5.3.8 Y-STR Analysis

The PowerPlex® Fusion 6C system (Promega Corp.) includes three Y-STR markers, two of which were an addition since the PowerPlex® Fusion system (Promega Corp.). As stated previously (see Methods), the Y-STR data was analyzed separately and only for mixtures that had all male contributors (e.g., 2-person mixtures with 2 males, 3-person mixtures with 3 males, etc.). The purpose was to determine how often a 2-, 3-, 4-, 5-, and 6-person male mixtures would be correctly identified based solely on the Y-STR

loci. The trends are similar with the Y-STRs as with the autosomal loci for the 2 person mixture, where approximately 97% of 2-person -male mixtures exhibit at least one Y-STR with 2 alleles (Table 5.3). For 3-person male mixtures, approximately two-thirds of all profiles (67%) exhibited at least one locus with 3 alleles (Table 5.3). Similarly to 3-person male mixtures, most 4-person male mixtures (64%) exhibited 3 or less alleles. Considering 5- and 6-person male mixtures, approximately 50% for both sets exhibited 4 or less alleles (Table 5.3). Therefore the 3 Y-STR markers were informative for number of contributors in the 2-person male mixtures, and for the majority of 3-person male mixtures, but are not very discriminating for number of contributors in the 4-, 5-, and 6-person male mixtures. DYS576 was found to be the most diverse of the 3 YSTRs analyzed.

Table 5.3 Y-STR Distributions per *n*-person male mixtures.

Mixture	Number of Mixtures	Max Alleles	Frequency (%)
2 males	6,105	1	3.3
		At least one locus with 2	96.7
3 males	221,815	1	0.1
		2 or less	32.9
		At least one locus with 3	67.0
4 males	40,919	2 or less	11.0
		3 or less	64.0
		At least one locus with 4	25.0
5 males	4,278	2 or less	1.5
		3 or less	40.0
		4 or less	52.5
		At least one locus with 5	6.0
6 males	5,005	2 or less	0.2
		3 or less	10.8
		4 or less	51.0
		5 or less	35
		At least one locus with 6	3.0

5.4 Conclusions

This study shows that using maximum allele count with profiles generated with the PowerPlex® Fusion 6C (Promega Corp.), which includes the expanded core STR loci, is accurate for estimating number of contributors for 2- and 3-person mixtures. Maximum allele count has already been shown just as efficient as maximum likelihood for 2- and 3- person mixtures based on 15 autosomal loci [174]; it is shown here to not be much more improved when expanded to 23 autosomal loci. The capability for estimations are less accurate as number of contributors increase, even with the expanded panel of standard loci. SWGDAM mixture interpretation guidelines state that there are essentially two approaches to determining the statistical weight of inclusions: 1) binary or 2) probabilistic [172]. Binary statistical models (i.e., random match probability (RMP), likelihood ratio (LR), and combined probability of exclusion/inclusion (CPE/CPI)) are still very common in practice, however, they are limited to cases where at least one contributor can be deconvoluted from the mixture. They also require the analyst to assume the number of contributors in order to perform the statistic, with the exception of CPE/CPI [171]. However, the CPE/CPI is also limited in that it can only be calculated when all alleles are present within the profile; there cannot be any allele drop-out [177]. For this calculation, in many cases then, not all the profile information can be used and the statistical result is not as informative. Furthermore, according to the recently released PCAST report, the CPE/CPI statistic was deemed inadequate and subjective [184]. The PCAST report also points to the fact that probabilistic genotyping methods are an improvement, but further testing should be done to ensure the scientific validity on reliability and on the algorithms being implemented [184].

While the future of mixture interpretation is heading towards these computer-based probabilistic genotyping methods, current practices in many laboratories still implement maximum allele count in their standard operating procedures. This study highlights that caution still needs to be given when assuming the number of contributors in suspected mixtures of greater than 3 individuals as even with the expanded core STR loci, discrimination for 4-. 5-, and 6-person mixtures is complex and difficult. Highly variable STRs, such as SE33, can be more useful for these higher order mixtures as seen here, however, there are still limitations in the successful typing of this locus and it is not available in all genotyping kits. Other genotype factors are taken into consideration when interpreting a profile besides the presence or absence of alleles, such as the peak height ratios and stutter etc. The analyses reported here were based on theoretical 1:1 mixtures under ideal conditions and unrelated individuals, and therefore further empirical testing with different ratios of contributors and possibly related individuals (which would have alleles in common than unrelated individuals) would be valuable to perform as it would simulate more realistic conditions of casework mixture samples.

CHAPTER 6. CONCLUSIONS AND FUTURE DIRECTIONS

The overall scope of the work presented was to address some of the limitations of current forensic DNA-based identification technologies. For example, one of those limitations was when a conventional STR profile from biological evidence does not provide a match to a suspect or database hit. Additional information can be determined from the evidence that will be useful to gain intelligence for the investigation by developing a SNP assay to provide a phenotypic profile of the contributor. Pigmentation contributes to many externally visible traits including eye, hair, and skin color, and since many pigmentation traits are related (they all stem from the melanogenesis pathway), there is a relationship between the elucidation of many of the components of melanogenesis and the SNPs found responsible for the expression of different pigmentation traits. This is especially true for some characteristic phenotypes such as red hair, blue and brown eyes, and light skin in Europeans, as these traits have at least one SNP/gene (*MC1R*, rs12913832, rs1426654, respectively) that contributes much of the variance in their expression. However, intermediate color categories (e.g., green and hazel eye colors) are known to be difficult to predict from previous studies [20, 40], and were not improved by the selected SNPs in this work. Therefore, optimization of prediction models that combine the possible traits into groups (e.g., blue or intermediate eye colors) results in likelihood calculations that reduce inaccuracy by providing information that is still reliable for use in an investigation, albeit a less specific prediction. Further studies into completing elucidation of all components related to the melanogenesis pathway and GWAS with larger cohorts of individuals with these traits will provide more guidance on identification of additional genes and SNPs that may

contribute to the expression of these traits. Furthermore, as more studies are heading towards quantitative color measurements as opposed to a single individual's visual determination for color assignment, intermediate colors can be more objectively defined and consistently identified for more accurate prediction analyses. The quantitative color prediction model developed here is a step in that direction.

Skin color pigmentation can also implicate ancestry due to the dispersal of melanin is correlated to geography; there are higher levels of melanin expression (i.e., darker skin, darker hair) in regions closer to the equator because of higher levels of UV (sun) exposure [185, 186]. Skin color is not a sole indication of ancestry, and therefore many of the additional ancestry SNPs selected for the panel developed in this work were not related to melanin expression, but were SNPs with alleles found to be in high frequencies in a specific population. The ancestry predictions in this work of the 5 main continental groups seen in the U.S. had greater than 50% correct predictions. A limitation of this work was that the sample size of each reported ancestry was very small, especially for East Asians, and therefore predictions were not as accurate as those with higher reported numbers of individuals, such as Europeans. Additionally, further development of prediction models to be able to account for admixed or biracial individuals is necessary, as the models built in this work could only assign a single ancestry category.

DNA phenotyping is not limited to pigmentation and ancestry. Age is one trait that has gained much interest in prediction studies recently. The main technology used for these studies has been bead chip microarrays, for example, the Infinium HumanMethylation arrays. However, these arrays suffer from bias in selection of region

coverage, typically targeting CpG islands in promoter regions and disease-causing genes [142]. The novel method developed here reduced the association to disease, included three forensically relevant fluids, and resulted in genome-wide coverage. Further analysis of the genes and features identified in the candidate CpG sites shows that these sites are likely involved in the regulation of many of the biological processes affected by aging, especially cell communication, metabolism, and immune response. Further testing of these sites with a larger sample population of healthy individuals, including females to eliminate and/or correct any influence of sex, would be necessary for validation. However, once validated, they can be incorporated into the FPP assay to add age to the predicted phenotype of an individual, furthering the information that can be gained for an investigation from a DNA sample that otherwise was a dead-end from its STR profile.

The other major consequence of current DNA technologies is the increased probability of mixture profiles due to increased sensitivity of amplification kits. Mixtures are inherent for certain crimes (e.g., sexual assaults); but the ability to deconvolute a mixture to individual genotypes of each contributor is essential to identifying the contributors correctly. Typically, mixture separation can accurately be determined for most 2-person mixtures and some 3-person mixtures by allele counting. It was also hypothesized that increasing the number of loci in forensic DNA profiles would improve the accuracy of the determination of the number of contributors. However, the millions of 2-, 3-, 4-, 5-, and 6-person mixture profiles generated in this work suggests that with the expanded core STR loci does not offer much improvement in accuracy. One locus, SE33, improved the determination of the number of contributors, however, it is a large locus, and suffers from a greater probability of drop out in forensic samples that have

degraded [171]. Furthermore, SE33 is not in the newly expanded core STR loci and therefore not all amplification kits include it in their panel of loci. It is also difficult to successfully amplify. When using Y-STRs for determining the number of male contributors, discrimination greater than a 3-person mixture was difficult. Furthermore, the mixture data generated here was assumed under ideal conditions: 1:1 ratios, no stutter, and all alleles above the stochastic threshold. Additional work can be done to analyze more realistic conditions such as differing ratios of each contributor with differing amount of input DNA. Practicing laboratories have mixture protocols to follow, and it should be cautioned that allele counting is not the best method for all mixture cases. More recently, computer-based probabilistic genotyping has gained recent attention as a more accurate method, considering all aspects of the genotype (e.g., peak height ratios, probability of drop-out, stutter, etc.). Although this is the future of mixture interpretation, some laboratories do not have the resources to implement these methods and therefore reliance on the allele-counting method is not yet obsolete. Another potential option is generating DNA profiles using NGS methods, as they have the capability of producing additional variant information that can discriminate individuals by mutations in flanking regions [187]; however, this technology is not cost effective for many laboratories yet and requires additional validation work to implement for common practice.

Besides a higher probability of developing human DNA mixtures, issues from non-human, extraneous DNA may also arise from crime scene evidence. Amplification and quantitation kits for forensic DNA analysis are designed to be human-specific to purposely avoid amplifying extraneous sources of DNA from a crime scene. However,

the amount of non-human DNA that may be present is not known in an evidence sample. When validating current forensic DNA genotyping kits, sources of common extraneous DNA such as domestic animals and common bacteria are included to ensure little or no artifactual amplifications. However, there are many more species of bacteria present on a living human body than is currently included during validation, and the proportion increases once decomposition sets in. Furthermore, if a body is found in an outdoor setting, there are many more microbial species that are found in soil and water that may be collected. This work sought to determine the effect of common microbial DNA on DNA profiles generated using the PowerPlex® 16 HS kit (Promega Corp.). Two species produced an allelic artifact at the TPOX locus, mimicking a true allele of the locus, and thereby interfering with the interpretation of the DNA profile. In one case, the microbial artifact amplified at equal ratio to the human allele. In a worst case scenario, this could lead to the possible exclusion of the actual contributor due to an inconsistent genotype at TPOX. Expanding the microbial species tested could be beneficial to extend this catalog of possible interfering bacterial sources. Again, the artifact allele was recognized with one kit, and not reproduced when repeated with another (personal communications with Promega revealed primer differences at the TPOX locus between kits). Additional testing of currently used amplification kits will help gauge how often this interference occurs. Furthermore, there was only three amounts of DNA reported here: 10ng, 50ng, and 100ng. An additional study into how much microbial DNA is actually collected from a decomposing body in different environments would be able to further guide the species and ratios tested during developmental validation of future amplification kits.

Advancements in technology have rapidly developed the area of DNA-based identification, this is evident especially in the last 30 years since the development of DNA fingerprinting and PCR. Overall, the work presented here addresses aspects of modern DNA-based identification. It is the hope that the work here provides proof of some of the limitations and consequences of these modern methods. Furthermore, this work provides guidance for ways these limitations can be overcome in future research as technologies continue to advance.

REFERENCES

[1] Findlay, I., Taylor, A., Quirke, P., Frazier, R., Urquhart, A. DNA fingerprinting from single cells. Nature. 1997;389:555-6.

[2] Jeffreys, A.J., Wilson, V., Thein, S.L. Hypervariable 'minisatellite' regions in human DNA. Nature. 1985;317.

[3] Jeffreys, A.J., Brookfield, J.F., Someoneoff, R. Positive ID of an immigration test-case using human DNA fingerprints. Nature. 1985;317.

[4] Aronson, J.D. Genetic Witness: science, law, and controversy in the making of DNA profiling. New Brunswick, NJ: Rutgers University Press; 2007.

[5] Lee, H.C., Ladd, C., Bourke, M.T., Pagliaro, E., Tirnady, F. DNA typing in forensic science: I. Theory and Background. American Journal of Forensic Medicine and Pathology. 1994;15:269-82.

[6] Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., Erlich, H. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harbor Symposia on Quantitative Biology. 1986;LI.

[7] Maeda, M., Murayama, N., Ishii, H., Uryu, N., Ota, M., Tsuji, K., Inoko, H. A simple and rapid method for HLA-DQA1 genotyping by digestion of PCR-amplified DNA with allele specific restriction endonucleases. Tissue Antigens. 1989;34.

[8] Helmuth, R., Fildes, N., Blake, E., Luce, M.C., Chimera, J., Madej, R., Gorodezky, C., Stoneking, M., Schmill, N., Klitz, W., Higuchi, R., Erlich, H.A. HLA-DQa allele and genotype frequencies in various human populations, determined by using enzymatic amplification and oligonucleotide probes. Am J Hum Genet. 1990;47.

[9] Li, R. Forensic Biology. Boca Raton; London; New York: CRC Press; 2008.

[10] van Oorschot, R.A.H., Jones, M.J. DNA fingerprints from fingerprints. Nature. 1997;387.

[11] Group, N.H.W., Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J.A., Bonazzi, V., McEwen, J.E., Wetterstrand, K.A., Deal, C., Baker, C.C., Di Francesco, V., Howcroft, T.K., Karp, R.W., Lunsford, R.D., Wellington, C.R., Belachew, T., Wright, M., Giblin, C., David, H., Mills, M., Salomon, R., Mullins, C., Akolkar, B., Begg, L., Davis, C., Grandison, L., Humble, M., Khalsa, J., Little, A.R., Peavy, H., Pontzer, C., Portnoy, M., Sayre, M.H., Starke-Reed, P., Zakhari, S., Read, J., Watson, B., Guyer, M. The NIH Human Microbiome Project. Genome Res. 2009;19:2317-23.

[12] Fierer, N., Lauber, C.L., Zhou, N., McDonald, D., Costello, E.K., Knight, R. Forensic identification using skin bacterial communities. Proc Natl Acad Sci U S A. 2010;107:6477-81.

[13] Finley, S.J., Benbow, M.E., Javan, G.T. Microbial communities associated with human decomposition and their potential use as postmortem clocks. Int J Legal Med. 2015;129:623-32.

[14] Hares, D.R. Selection and implementation of expanded CODIS core loci in the United States. Forensic Sci Int Genet. 2015;17:33-4.

[15] Butler, J.M., Coble, M.D., Vallone, P.M. STRs vs. SNPs: thoughts on the future of forensic DNA testing. Forensic Sci Med Pathol. 2007;3:200-5.

[16] Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., Rothberg, J.M. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437:376-80.

[17] Rogers, Y.-H., Venter, J.C. Massively parallel sequencing. Nature. 2005;309:326-7.

[18] Zbiec-Piekarska, R., Spolnicka, M., Kupiec, T., Makowska, Z., Spas, A., Parys-Proszek, A., Kucharczyk, K., Ploski, R., Branicki, W. Examination of DNA methylation status of the ELOVL2 marker may be useful for human age prediction in forensic science. Forensic Sci Int Genet. 2015;14:161-7.

[19] Kayser, M., Schneider, P. DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations. Forensic Science International: Genetics. 2009;3:154-61.

[20] Walsh, S., Liu, F., Ballantyne, K.N., van Oven, M., Lao, O., Kayser, M. IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. Forensic Science International: Genetics. 2011;5:170-80.

[21] D'Mello, S.A., Finlay, G.J., Baguley, B.C., Askarian-Amiri, M.E. Signaling Pathways in Melanogenesis. Int J Mol Sci. 2016;17.

[22] Park, H.Y., Kosmadaki, M., Yaar, M., Gilchrest, B.A. Cellular mechanisms regulating human melanogenesis. Cell Mol Life Sci. 2009;66:1493-506.

[23] Thody, A.J., Higgins, E.M., Wakamatsu, K., Ito, S., Burchill, S.A., Marks, J.M. Pheomelanin as well as Eumelanin is Present in Human Epidermis. Journal of Investigative Dermatology. 1991;97:340-4.

[24] Prota, G., Hu, D.N., Vincensi, M.R., McCormick, S.A., Napolitano, A. Characterization of melanins in human irides and cultured uveal melanocytes from eyes of different colors. Exp Eye Res. 1998;67:293-9.

[25] Tully, G. Genotype versus phenotype: human pigmentation. Forensic Science International: Genetics. 2007;1:105-10.

[26] Imesch, P.D., Wallow, I.H.L., Albert, D.M. The color of the human eye: a review of morphologic correlates and of some conditions that affect iridal pigmentation. Surv Ophthalmol. 1997;41:S117-S23.

[27] Cichorek, M., Wachulska, M., Stasiewicz, A., Tymińska, A. Skin melanocytes: biology and development. Advances in Dermatology and Allergology. 2013;1:30-41.

[28] Kayser, M., Liu, F., Janssens, C.J.W., Rivadeneira, F., Lao, O., van Duijn, K., Vermeulen, M., Arp, P., Jhamai, M.M., van IJcken, W.F.J., den Dunnen, J.T., Heath, S., Zelenika, D., Despriet, D.D.G., Klaver, C.C.W., Vingerling, J.R., de Jong, P.T.V.M., Hofman, A., Aulchenko, Y.S., Uitterlinden, A.G., Oostra, B.A., van Duijn, C.M. Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene. The American Journal of Human Genetics. 2008;82:411-23.

[29] Sulem, P., Gudbjartsson, D.F., Stacey, S.N., Helgason, A., Rafnar, T., Magnusson, K.P., Manolescu, A., Karason, A., Palsson, A., Thorleifsson, G., Jakobsdottir, M., Steinberg, S., Pálsson, S., Jonasson, F., Sigurgeirsson, B.T., K., Ragnarsson, R., Benediktsdottir, K.R., Aben, K.K., Kiemeney, L.A., Olafsson, J.H., Gulcher, J., Kong, A., Thorsteinsdottir, U., Stefansson, K. Genetic determinants of hair, eye and skin pigmentation in Europeans. Nat Genet. 2007;39:1443-52.

[30] Sulem, P., Gudbjartsson, D.F., Stacey, S.N., Helgason, A., Rafnar, T., Jakobsdottir, M., Steinberg, S., Gudjonsson, S.A., Palsson, A., Thorleifsson, G., Palsson, S., Sigurgeirsson, B., Thorisdottir, K., Ragnarsson, R., Benediktsdottir, K.R., Aben, K.K., Vermeulen, S.H., Goldstein, A.M., Tucker, M.A., Kiemeney, L.A., Olafsson, J.H., Gulcher, J., Kong, A., Thorsteinsdottir, U., Stefansson, K. Two newly identified genetic determinants of pigmentation in Europeans. Nat Genet. 2008;40:835-7.

[31] Han, J.L., Kraft, P., Nan, H., Guo, Q., Chen, C., Qureshi, A., Hankinson, S.E., Hu, F.B., Duffy, D.L., Zhao, Z.Z., Martin, N.G., Montgomery, G.W., Hayward, N.K., Thomas, G., Hoover, R.N., Chanock, S., Hunter, D.J. A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. PLoS Genetics. 2008;4:e1000074.

[32] Larsson, M., Duffy, D.L., Zhu, G., Liu, J.Z., Macgregor, S., McRae, A.F., Wright, M.J., Sturm, R.A., Mackey, D.A., Montgomery, G.W., Martin, N.G., Medland, S.E. GWAS findings for human iris patterns: associations with variants in genes that influence normal neuronal pattern development. Am J Hum Genet. 2011;89:334-43.

[33] Candille, S.I., Absher, D.M., Beleza, S., Bauchet, M., McEvoy, B., Garrison, N.A., Li, J.Z., Myers, R.M., Barsh, G.S., Tang, H., Shriver, M.D. Genome-wide association studies of quantitatively measured skin, hair, and eye pigmentation in four European populations. PLoS One. 2012;7:e48294.

[34] Beleza, S., Johnson, N.A., Candille, S.I., Absher, D.M., Coram, M.A., Lopes, J., Campos, J., Araujo, I.I., Anderson, T.M., Vilhjalmsson, B.J., Nordborg, M., Correia, E., Silva, A., Shriver, M.D., Rocha, J., Barsh, G.S., Tang, H. Genetic architecture of skin and eye color in an African-European admixed population. PLoS Genet. 2013;9:e1003372.

[35] Grimes, E.A., Noake, P.J., Dixon, L., Urquhart, A. Sequence polymorphism in the human melanocortin 1 receptor gene as an indicator of red hair phenotype. Forensic Sci Int. 2001;1:124-9.

[36] Valverde, P., Healy, E., Jackson, I., Rees, J.L., Thody, A.J. Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans. Nature Genetics. 1995;11.

[37] Eiberg, H., Troelsen, J., Nielsen, M., Mikkelsen, A., Mengel-From, J., Kjaer, K.W., Hansen, L. Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. Human Genetics. 2008;123:177-87.

[38] Sturm, R.A., Duffy, D.L., Zhao, Z.Z., Leite, F.P.N., Stark, M.S., Hayward, N.K., Martin, N.G., Montgomery, G.W. A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. The American Journal of Human Genetics. 2008;82:424-31.

[39] Liu, F., van Duijn, K., Vingerling, J., Hofman, A., Uitterlinden, A., Janssens, A., Kayser, M. Eye color and the prediction of complex phenotypes from genotypes. Current Biology. 2009;19:R192-R3.

[40] Dembinski, G.M., Picard, C.J. Evaluation of the IrisPlex DNA-based eye color prediction assay in a United States population. Forensic Sci Int Genet. 2014;9:111-7.

[41] Dembinski, G.M., Picard, C.J. Corrigendum to "Evaluation of the IrisPlex DNA-based eye color prediction assay in a United States population" [Forensic Sci. Int. Genet. (2014) 111-117]. Forensic Sci Int Genet. 2016;24:211-3.

[42] S. Walsh, F.L., A. Wollstein, L. Kovatsi, A. Ralf, A. Kosiniak-Kamysz, W. Branicki, M. Kayser. The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. Forensic Sci Int Genet. 2013;7:98-115.

[43] Pośpiech, E., Draus-Barini, J., Kupiec, T., Wojas-Pelc, A., Branicki, W. Gene-gene interactions contribute to eye colour variation in humans. J Hum Genet. 2011;56:447-55.

[44] Branicki, W., Liu, F., van Duijn, K., Draus-Barini, J., Pospiech, E., Walsh, S., Kupiec, T., Wojas-Pelc, A., Kayser, M. Model-based prediction of human hair color using DNA variants. Hum Genet. 2011;129:443-54.

[45] Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., Branicki, W., Kayser, M. The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. Forensic Sci Int Genet. 2013;7:98-115.

[46] Lamason, R.L., Mohideen, M.P.K., Mest, J.R., Wong, A.C., Norton, H.L., Aros, M.C., Jurynec, M.J., Mao, X., Humphreville, V.R., Humbert, J.E., Sinha, S., Moore, J.L., Jagadeeswaran, P., Zhao, W., Ning, G., Makalowska, I., McKeigue, P.M., O'Donnell, D., Kittles, R., Parra, E.J., Magini, N.J., Grunwald, D.J., Shriver, M.D., Canfield, V.A., Cheng, K.C. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science. 2005;310.

[47] Donnelly, M.P., Paschou, P., Grigorenko, E., Gurwitz, D., Barta, C., Lu, R.B., Zhukova, O.V., Kim, J.J., Siniscalco, M., New, M., Li, H., Kajuna, S.L., Manolopoulos, V.G., Speed, W.C., Pakstis, A.J., Kidd, J.R., Kidd, K.K. A global view of the OCA2-HERC2 region and pigmentation. Hum Genet. 2012;131:683-96.

[48] Maronas, O., Phillips, C., Söchtig, J., Gomez-Tato, A., Cruz, R., Alvarez-Dios, J., de Cal, M.C., Ruiz, Y., Fondevila, M., Carracedo, A., Lareu, M.V. Development of a forensic skin colour predictive test. Forensic Sci Int Genet. 2014;13C:34-44.

[49] Beleza, S., Santos, A.M., McEvoy, B., Alves, I., Martinho, C., Cameron, E., Shriver, M.D., Parra, E.J., Rocha, J. The timing of pigmentation lightening in Europeans. Mol Biol Evol. 2013;30:24-35.

[50] Spichenok, O., Budimlija, Z.M., Mitchell, A.A., Jenny, A., Kovacevic, L., Marjanovic, D., Caragine, T., Prinz, M., Wurmbach, E. Prediction of eye and skin color in diverse populations using seven SNPs. Forensic Science International: Genetics. 2011;5:472-8.

[51] Hart, K.L., Kimura, S.L., Mushailov, V., Budimlija, Z.M., Prinz, M., Wurmbach, E. Improved eye- and skin-color prediction based on 8 SNPs. Croatian Medical Journal. 2013;54:248-56.

[52] Lao, O., van Duijn, K., Kersbergen, P., de Knijff, P., Kayser, M. Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. Am J Hum Genet. 2006;78:680-90.

[53] Kersbergen, P., van Duijn, K., Kloosterman, A.D., den Dunnen, J.T., Kayser, M., de Knijff, P. Developing a set of ancestry-sensitive DNA markers reflecting continental origins of humans. BMC Genet. 2009;10:69.

[54] Phillips, C., Lareu, M., Salas, A., Fondevila, M., Berniell Lee, G., Carracedo, Á., Morling, N., Schneider, P., Syndercombe Court, D. Population specific single nucleotide polymorphisms. International Congress Series. 2004;1261:233-5.

[55] Nievergelt, C.M., Maihofer, A.X., Shekhtman, T., Libiger, O., Wang, X., Kidd, K.K., Kidd, J.R. Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. Investigative Genetics. 2013;4.

[56] Phillips, C., Salas, A., Sánchez, J.J., Fondevila, M., Gómez-Tato, A., Álvarez-Dios, J., Calaza, M., Casares de Cal, M., Ballard, D., Lareu, M.V., Carracedo, Á., Consortium, S. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci Int Genet. 2007;1:273-80.

[57] Kidd, J.R., Friedlaender, F.R., Speed, W.C., Pakstis, A.J., De La Vega, F.M., Kidd, K.K. Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. Investig Genet. 2011;2:1.

[58] Phillips, C., Freire Aradas, A., Kriegel, A.K., Fondevila, M., Bulbul, O., Santos, C., Serrulla Rech, F., Perez Carceles, M.D., Carracedo, Á., Schneider, P.M., Lareu, M.V. Eurasiaplex: a forensic SNP assay for differentiating European and South Asian ancestries. Forensic Sci Int Genet. 2013;7:359-66.

[59] Fondevila, M., Phillips, C., Santos, C., Freire Aradas, A., Vallone, P.M., Butler, J.M., Lareu, M.V., Carracedo, Á. Revision of the SNPforID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended population studies. Forensic Sci Int Genet. 2013;7:63-74.

[60] Rajeevan, H., Osier, M.V., Cheung, K.-H., Deng, H., Druskin, L., Heinzen, R., Kidd, J.R., Stein, S., Pakstis, A.J., Tosches, N.P., Yeh, C.-C., Miller, P.L., Kidd, K.K. ALFRED: the ALelle FREquency Database. Update. Nucleic Acids Research. 2003;31.

[61] Kidd, K. ALFRED: The Allele Frequency Database. Yale University2012.

[62] Halder, I., Shriver, M., Thomas, M., Fernandez, J.R., Frudakis, T. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. Hum Mutat. 2008;29:648-58.

[63] Nassir, R., Kosoy, R., Tian, C., White, P.A., Butler, L.M., Silva, G., Kittles, R., Alarcon-Riquelme, M.E., Gregersen, P.K., Belmont, J.W., De La Vega, F.M., Seldin, M.F. An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. BMC Genet. 2009;10:39.

[64] Kosoy, R., Nassir, R., Tian, C., White, P.A., Butler, L.M., Silva, G., Kittles, R., Alarcon-Riquelme, M.E., Gregersen, P.K., Belmont, J.W., De La Vega, F.M., Seldin, M.F. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Hum Mutat. 2009;30:69-78.

[65] Kidd, K.K., Speed, W.C., Pakstis, A.J., Furtado, M.R., Fang, R., Madbouly, A., Maiers, M., Middha, M., Friedlaender, F.R., Kidd, J.R. Progress toward an efficient panel of SNPs for ancestry inference. Forensic Sci Int Genet. 2014;10:23-32.

[66] Sturm, R.A., Box, N.F., Ramsay, M. Human pigmentation genetics: the difference is only skin deep. BioEssays. 1998;20:712-21.

[67] Hunt, R.W.G. Measuring colour. In: Pointer M, editor. 4th ed. ed. Chichester, West Sussex, U.K. :: Wiley; 2011.

[68] CIE. International Commission on Illumination. Vienna, Austria: CIE.

[69] Yam, K.L., Papadakis, S.E. A simple digital imaging method for measuring and analyzing color of food surfaces. Journal of Food Engineering. 2004;61:137-42.

[70] Core Team, R. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria2015.

[71] Branicki, W., Brudnik, U., Wojas-Pelc, A. Interactions between HERC2, OCA2 and MC1R may influence human pigmentation phenotype. Annals of Human Genetics. 2009;73:160-70.

[72] Mengel-From, J., Børsting, C., Sanchez, J.J., Eiberg, H., Morling, N. Human eye colour and HERC2, OCA2, and MATP. Forensic Sci Int Genet. 2010;4:323-8.

[73] Valenzuela, R.K., Henderson, M.S., Walsh, M.H., Garrison, N.A., Kelch, J.T., Cohen-Barak, O., Erickson, D.T., Meaney, F.J., Walsh, J.B., Cheng, K.C., Ito, S., Wakamatsu, K., Frudakis, T., Thomas, M., Brilliant, M.H. Predicting phenotype from genotype: normal pigmentation. Journal of Forensic Sciences. 2010;55:315-22.

[74] Branicki, W. Studies on predicting pigmentation phenotype for forensic purposes. Problems of Forensic Science. 2009;LXXVII:29-52.

[75] Pospiech, E., Wojas-Pelc, A., Walsh, S., Liu, F., Maeda, H., Ishikawa, T., Skowron, M., Kayser, M., Branicki, W. The common occurrence of epistasis in the determination of human pigmentation and its impact on DNA-based pigmentation phenotype prediction. Forensic Sci Int Genet. 2014;11:64-72.

[76] Frudakis, T., Thomas, M., Gaskin, Z., Venkateswarlu, K., Chandra, K.S., Ginjupalli, S., Gunturi, S., Natrajan, S., Ponnuswamy, V.K., Ponnuswamy, K.N. Sequences associated with human iris pigmentation. Genetics. 2003;165:2071-83.

[77] Hoekstra, H. The secret of a natural blond. Nat Genet. 2014;46:660-1.

[78] Edwards, M., Bigham, A., Tan, J., Li, S., Gozdzik, A., Ross, K., Jin, L., Parra, E.J. Association of the *OCA2* polymorphism His615Arg with melanin content in East Asian populations: Further evidence of convergent evolution of skin pigmentation. PLoS Genet. 2010;6.

[79] Soejima, M., Koda, Y. Population differences of two coding SNPs in pigmentation-related genes SLC24A5 and SLC45A2. Int J Legal Med. 2007;121:36-9.

[80] Gettings, K.B., Lai, R., Johnson, J.L., Peck, M.A., Hart, J.A., Gordish-Dressman, H., Schanfield, M.S., Podini, D.S. A 50-SNP assay for biogeographic ancestry and phenotype prediction in the U.S. population. Forensic Sci Int Genet. 2014;8:101-8.

[81] Giardina, E., Pietrangeli, I., Martínez-Labarga, C., Martone, C., de Angelis, F., Spinella, A., De Stefano, G., Rickards, O., Novelli, G. Haplotypes in SLC24A5 as Ancestry Informative Markers in Different Populations. Current Genomics. 2008;9:110-4.

[82] Phillips, C. Forensic genetic analysis of bio-geographical ancestry. Forensic Sci Int Genet. 2015;18:49-65.

[83] Phillips, C., Fondevila, M., Vallone, P.M., Carla, S., Freire-Aradas, A., Butler, J.M., Lareu, M.V., Carracedo, Á. Characterization of U.S. population samples using a 34plex ancestry informative SNP multiplex. Forensic Science International: Genetics Supplement Series. 2011;3:e182-e3.

[84] Hosmer, D.W., Lemeshow, S. Applied Logistic Regression. 1st ed. New York: Wiley; 1989.

[85] Morgan, S.L., Bartick, E.G. Discrimination of Forensic Analytical Chemical Data Using Multivariate Statistics. In: Blackledge RD, editor. Forensic Analysis on the Cutting Edge: New Methods for Trace Evidence Analysis. Hoboken, NJ: Wiley & Sons, Inc.; 2007.

[86] Malovini, A., Nuzzo, A., Ferrazzi, F., Puca, A.A., Bellazzi, R. Phenotype forecasting with SNPs data through gene-based Bayesian networks. BMC Bioinformatics. 2009;10 Suppl 2:S7.

[87] Sebastiani, P., Perls, T.T. Complex genetic models. Bayesian Networks: a practical guide to applications. Hoboken, NJ: Wiley; 2008.

[88] Friedman, N., Geiger, D., Goldszmidt, M. Bayesian Network Classifiers. Machine Learning. 1997;29.

[89] Sebastiani, P., Ramoni, M.F., Nolan, V., Baldwin, C.T., Steinberg, M.H. Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. Nat Genet. 2005;37:435-40.

[90] Pośpiech, E., Draus-Barini, J., Kupiec, T., Wojas-Pelc, A., Branicki, W. Prediction of eye color from genetic data using Bayesian approach. Journal of Forensic Sciences. 2012;57:880-6.

[91] Lantz, B. Machine Learning with R. Birmingham, U.K.: Packt Publishing; 2013.

[92] Shachter, R.D. Bayes-Ball: rational pastime (for determining irrelvance and requisite information in belief networks and influence diagrams). Proceedings of the Fourteenth conference on Uncertainty in artifical intelligence. 1998.

[93] Liu, H., Prugnolle, F., Manica, A., Balloux, F. A geographically explicit genetic model of worldwide human-settlement history. Am J Hum Genet. 2006;79:230-7.

[94] Feil, R. Environmental and nutritional effects on the epigenetic regulation of genes. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2006;600:46-57.

[95] Calvanese, V., Lara, E., Kahn, A., Fraga, M.F. The role of epigenetics in aging and age-related diseases. Ageing Res Rev. 2009;8:268-76.

[96] Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L., Fan, J.B., Shen, R. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98:288-95.

[97] Winnefeld, M., Lyko, F. The aging epigenome: DNA methylation from the cradle to the grave. Genome Biol. 2012;13:165.

[98] Day, K., Waite, L.L., Thalacker-Mercer, A., West, A., Bamman, M.M., Brooks, J.D., Myers, R.M., Absher, D. Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. Genome Biol. 2013;14.

[99] Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.B., Gao, Y., Deconde, R., Chen, M., Rajapakse, I., Friend, S., Ideker, T., Zhang, K. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49:359-67.

[100] Bocklandt, S., Lin, W., Sehl, M., Sánchez, F., Sinsheimer, J.S., Horvath, S., Vilain, E. Epigenetic predictor of age. PLoS Genetics. 2011;6.

[101] Liu, J., Morgan, M., Hutchinson, K., Calhoun, V.D. A study of the influence of sex on genome wide methylation. PLoS One. 2010;5.

[102] Florath, I., Butterbach, K., Muller, H., Bewerunge-Hudler, M., Brenner, H. Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites. Hum Mol Genet. 2014;23:1186-201.

[103] Zbiec-Piekarska, R., Spolnicka, M., Kupiec, T., Parys-Proszek, A., Makowska, Z., Paleczka, A., Kucharczyk, K., Ploski, R., Branicki, W. Development of a forensically useful age prediction method based on DNA methylation analysis. Forensic Sci Int Genet. 2015;17:173-9.

[104] Philibert, R.A., Gunter, T.D., Beach, S.R., Brody, G.H., Madan, A. MAOA methylation is associated with nicotine and alcohol dependence in women. Am J Med Genet B Neuropsychiatr Genet. 2008;147B:565-70.

[105] Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R., Gunderson, K.L. Genome-wide DNA methylation profiling using Infinium(R) assay. Epigenomics. 2009;1:177-200.

[106] Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M.A., Bibikova, M., Esteller, M. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics. 2014;6:692-702.

[107] Davey, J.W., Blaxter, M.L. RADSeq: next-generation population genetics. Brief Funct Genomics. 2010;9:416-23.

[108] Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., Johnson, E.A. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLoS One. 2008;3.

[109] Miller, M.R., Dunham, J.P., Amores, A., Cresko, W.A., Johnson, E.A. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res. 2007;17:240-8.

[110] Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L., Paul, C.L. A genomic sequencing protocol that yields a positive display of 5-methylcysotine residues in individual DNA strands. Proc Natl Acad Sci U S A. 1992;89.

[111] Parchman, T., Gompert, Z., Buerkle, A. Amplified restriction fragments for genomic enrichment. University of Wyoming2011.

[112] Illumina. Preparing Libraries for Sequencing on the MiSeq. https://support.illumina.com/downloads/prepare_libraries_for_sequencing_miseq_15039740.html2013.

[113] Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Cech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., Gruning, B., Guerler, A., Hillman-Jackson, J., Von Kuster, G., Rasche, E., Soranzo, N., Turaga, N., Taylor, J., Nekrutenko, A., Goecks, J. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. 2016;44:W3-W10.

[114] Craig, D.W., Pearson, J.V., Szelinger, S., Sekar, A., Redman, M., Corneveaux, J.J., Pawlowski, T.L., Laub, T., Nunn, G., Stephan, D.A., Homer, N., Huentelman, M.J. Identification of genetic variants using bar-coded multiplexed sequencing. Nat Methods. 2008;5:887-93.

[115] Langmead, B., Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357-9.

[116] Quinlan, A.R., Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841-2.

[117] Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., Sandstrom, R., Humbert, R., Stamatoyannopoulos, J.A. BEDOPS: high-performance genomic feature operations. Bioinformatics. 2012;28:1919-20.

[118] Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, D. The human genome browser at UCSC. Genome Res. 2002;12:996-1006.

[119] Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., Thomas, P.D. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. 2017;45:D183-D9.

[120] Takai, D., Jones, P.A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A. 2002;99:3740-5.

[121] Kananen, L., Marttila, S., Nevalainen, T., Jylhava, J., Mononen, N., Kahonen, M., Raitakari, O.T., Lehtimaki, T., Hurme, M. Aging-associated DNA methylation changes in middle-aged individuals: the Young Finns study. BMC Genomics. 2016;17:103.

[122] Koch, C.M., Wagner, W. Epigenetic-aging-signature to determine age in different tissues. Aging. 2011;3.

[123] Horvath, S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14.

[124] Rakyan, V.K., Down, T.A., Balding, D.J., Beck, S. Epigenome-wide association studies for common human diseases. Nat Rev Genet. 2011;12:529-41.

[125] Garagnani P., B.M., Pirazzini C, Gori D, Giuliani C, Mari D, DiBlasio AM, Gentilini D, Vitale G, Collino S, Rezzi S, Castellani G, Capri M, Salvioli S, Franceschi C. Methylation of ELOVL2 gene as a new epigenetic marker of age. Aging Cell. 2012;11:1132-4.

[126] Johansson, A., Enroth, S., Gyllensten, U. Continuous aging of the human DNA methylome throughout the human lifespan. PLoS One. 2013;8:e67378.

[127] Bell, J.T., Tsai, P., Yang, T., Pidsley, R., Nisbet, J., Glass, D., Mangino, M., Zhai, G., Zhang, F., Valdes, A., Shin, S., Dempster, E.L., consortium., M., Dermitzakis, E.T., McCarthy, M.I., Mill, J., Spector, T.D., Deloukas, P. Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. PLoS Genet. 2012;8.

[128] Lee, H.Y., Jung, S.E., Oh, Y.N., Choi, A., Yang, W.I., Shin, K.J. Epigenetic age signatures in the forensically relevant body fluid of semen: a preliminary study. Forensic Sci Int Genet. 2015;19:28-34.

[129] Almén, M.S., Nilsson, E.K., Jacobsson, J.A., Kalnina, I., Klovins, J., Fredriksson, R., Schioth, H.B. Genome-wide analysis reveals DNA methylation markers that vary with both age and obesity. Gene. 2014;548:61-7.

[130] Franceschi, C., BonafÈ, M., Valensin, S., Olivieri, F., De Luca, M., Ottaviani, E., De Benedictis, G. Inflamm-aging: An Evolutionary Perspective on Immunosenescence. Annals of the New York Academy of Sciences. 2000;908:244-54.

[131] Grolleau-Julius, A., Ray, D., Yung, R.L. The role of epigenetics in aging and autoimmunity. Clin Rev Allergy Immunol. 2010;39:42-50.

[132] Salminen, A., Kaarniranta, K. Insulin/IGF-1 paradox of aging: regulation via AKT/IKK/NF-kappaB signaling. Cell Signal. 2010;22:573-7.

[133] DiLoreto, R., Murphy, C.T. The cell biology of aging. Mol Biol Cell. 2015;26:4524-31.

[134] Nelson, G., Wordsworth, J., Wang, C., Jurk, D., Lawless, C., Martin-Ruiz, C., von Zglinicki, T. A senescent cell bystander effect: senescence-induced senescence. Aging Cell. 2012;11.

[135] De Cecco, M., Criscione, S.W., Peterson, A.L., Neretti, N., Sedivy, J.M., Kreiling, J.A. Transposable elements become active and mobile in the genomes of aging mammalian somatic issues. Aging. 2013;5.

[136] Thornburg, B.G., Gotea, V., Makalowski, W. Transposable elements as a significant source of transcription regulating signals. Gene. 2006;365:104-10.

[137] De Cecco, M., Criscione, S.W., Peckham, E.J., Hillenmeyer, S., Hamm, E.A., Manivannan, J., Peterson, A.L., Kreioling, J.A., Neretti, N., Sedivy, J.M. Genomes of replicatively senescent cells undergo global epigenetic changes leading to gene silencing and activation of transposable elements. Aging Cell. 2013;12.

[138] Grammatikakis, I., Panda, A.C., Abdelmohsen, K., Gorospe, M. Long noncoding RNAs (lncRNAs) and the molecular hallmarks of aging. Aging. 2014;6.

[139] Degirmenci, U., Lei, S. Role of lncRNAs in Cellular Aging. Frontiers in Endocrinology. 2016;7.

[140] Marín-Béjar, O., Marchese, F.P., Athie, A., Sánchez, Y., González, J., Segura, V., Huang, L., Moreno, I., Navarro, A., Monzó, M., García-Foncillas, J., Rinn, J.L., Guo, S., Huarte, M. *Pint* lincRNA connects the p53 pathway with epigenetic silencing by the Polycomb repressive complex 2. Genome Biol. 2013;14.

[141] Bird, A., Taggart, M., Frommer, M., Miller, O.J., Macleod, D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell. 1985;40:91-9.

[142] Marttila, S., Kananen, L., Hayrynen, S., Jylhava, J., Nevalainen, T., Hervonen, A., Jylha, M., Nykter, M., Hurme, M. Ageing-associated changes in the human DNA methylome: genomic locations and effects on gene expression. BMC Genomics. 2015;16:179.

[143] Christensen, B.C., Houseman, E.A., Marsit, C.J., Zheng, S., Wrensch, M.R., Wiemels, J.L., Nelson, H.H., Karagas, M.R., Padbury, J.F., Bueno, R., Sugarbaker, D.J., Yeh, R.F., Wiencke, J.K., Kelsey, K.T. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. PLoS Genet. 2009;5:e1000602.

[144] Moran, S., Arribas, C., Esteller, M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics. 2016;8:389-99.

[145] Taberlet, P., Griffin, S., Goossens, B., Questiau, S., Manceau, V., Escaravage, N., Waits, L.P., Bouvet, J. Reliable genotyping of samples with very low DNA quantities using PCR. Nucleic Acids Res. 1996;24:3189-94.

[146] Dixon, L.A., Dobbins, A.E., Pulker, H.K., Butler, J.M., Vallone, P.M., Coble, M.D., Parson, W., Berger, B., Grubwieser, P., Mogensen, H.S., Morling, N., Nielsen, K., Sanchez, J.J., Petkovski, E., Carracedo, A., Sanchez-Diz, P., Ramos-Luis, E., Brion, M., Irwin, J.A., Just, R.S., Loreille, O., Parsons, T.J., Syndercombe-Court, D., Schmitter, H., Stradmann-Bellinghausen, B., Bender, K., Gill, P. Analysis of artificially degraded DNA using STRs and SNPs - results of a collaborative European (EDNAP) exercise. Forensic Sci Int. 2006;164:33-44.

[147] Nicklas, J.A., Buel, E. Quantification of DNA in forensic samples. Anal Bioanal Chem. 2003;376:1160-7.

[148] Oostdik, K., Lenz, K., Nye, J., Schelling, K., Yet, D., Bruski, S., Strong, J., Buchanan, C., Sutton, J., Linner, J., Frazier, N., Young, H., Matthies, L., Sage, A., Hahn, J., Wells, R., Williams, N., Price, M., Koehler, J., Staples, M., Swango, K.L., Hill, C., Oyerly, K., Duke, W., Katzilierakis, L., Ensenberger, M.G., Bourdeau, J.M., Sprecher, C.J., Krenke, B., Storts, D.R. Developmental validation of the PowerPlex((R)) Fusion System for analysis of casework and reference samples: A 24-locus multiplex for new database standards. Forensic Sci Int Genet. 2014;12:69-76.

[149] Ensenberger, M.G., Thompson, J., Hill, B., Homick, K., Kearney, V., Mayntz-Press, K.A., Mazur, P., McGuckian, A., Myers, J., Raley, K., Raley, S.G., Rothove, R., Wilson, J., Wieczorek, D., Fulmer, P.M., Storts, D.R., Krenke, B.E. Developmental validation of the PowerPlex 16 HS System: an improved 16-locus fluorescent STR multiplex. Forensic Sci Int Genet. 2010;4:257-64.

[150] Tucker, V.C., Hopwood, A.J., Sprecher, C.J., McLaren, R.S., Rabbach, D.R., Ensenberger, M.G., Thompson, J.M., Storts, D.R. Developmental validation of the PowerPlex((R)) ESI 16 and PowerPlex((R)) ESI 17 Systems: STR multiplexes for the new European standard. Forensic Sci Int Genet. 2011;5:436-48.

[151] Oostdik, K., French, J., Yet, D., Smalling, B., Nolde, C., Vallone, P.M., Butts, E.L., Hill, C.R., Kline, M.C., Rinta, T., Gerow, A.M., Allen, S.R., Huber, C.K., Teske, J., Krenke, B., Ensenberger, M., Fulmer, P., Sprecher, C. Developmental validation of the PowerPlex(R) 18D System, a rapid STR multiplex for analysis of reference samples. Forensic Sci Int Genet. 2013;7:129-35.

[152] Millar, B.C., Xu, J., Moore, J.E. Risk Assessment Models and Contamination Management: Implications for Broad-Range Ribosomal DNA PCR as a Diagnostic Tool in Medical Bacteriology. J Clin Microbiol. 2002;40:1575-80.

[153] Nikkari, S., McLaughlin, I.J., Bi, W., Dodge, D.E., Relman, D.A. Does blood of healthy subjects contain bacterial ribosomal DNA? J Clin Microbiol. 2001;39:1956-9.

[154] Muhl, H., Kochem, A.J., Disque, C., Sakka, S.G. Activity and DNA contamination of commercial polymerase chain reaction reagents for the universal 16S rDNA real-time polymerase chain reaction detection of bacterial pathogens in blood. Diagn Microbiol Infect Dis. 2010;66:41-9.

[155] Ingemann-Hansen, O., Charles, A.V. Forensic medical examination of adolescent and adult victims of sexual violence. Best Pract Res Clin Obstet Gynaecol. 2013;27:91-102.

[156] Milo, R., Jorgensen, P., Moran, U., Weber, G., Springer, M. BioNumbers--the database of key numbers in molecular and cell biology. Nucleic Acids Res. 2010;38:D750-3.

[157] Guarner, F., Malagelada, J.-R. Gut flora in health and disease. Lancet. 2003;360.

[158] De Ungria, M.C.A., Calacal, G.C. Fungal DNA Challenge in Human STR Typing of Bone Samples. J Forensic Sci. 2005;50:1-8.

[159] Vass, A.A., Bass, W.M., Wolt, J.D., Foss, J.E., Ammon, J.T. Time Since Death Determinations of Human Cadavers Using Soil Solution. J Forensic Sci. 1992;37.

[160] Fernandez-Rodriguez, A., Alonso, A., Albarran, C., Martin, P., Iturralde, M.J., Montesino, M., Sancho, M. Microbial DNA challenge studies of PCR-based systems used in forensic genetics. Adv Haemogenetics. 1996;6.

[161] Human Microbiome Project, C. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207-14.

[162] White, B.A., Creedon, D.J., Nelson, K.E., Wilson, B.A. The vaginal microbiome in health and disease. Trends Endocrinol Metab. 2011;22:389-93.

[163] Hyde, E.R., Haarmann, D.P., Lynne, A.M., Bucheli, S.R., Petrosino, J.F. The living dead: bacterial community structure of a cadaver at the onset and end of the bloat stage of decomposition. PLoS One. 2013;8:e77733.

[164] Tsukamura, M. Properties of *Mycobacterium smegmatis* freshly isolated from soil. Japan J Microbiology. 1976;20.

[165] Tam, N.K., Uyen, N.Q., Hong, H.A., Duc le, H., Hoa, T.T., Serra, C.R., Henriques, A.O., Cutting, S.M. The intestinal life cycle of Bacillus subtilis and close relatives. J Bacteriol. 2006;188:2692-700.

[166] Hong, H.A., Khaneja, R., Tam, N.M., Cazzato, A., Tan, S., Urdaci, M., Brisson, A., Gasbarrini, A., Barnes, I., Cutting, S.M. Bacillus subtilis isolated from the human gastrointestinal tract. Res Microbiol. 2009;160:134-43.

[167] Bissa, S., Songara, D., Bohra, A., Bohra, A. Microbes as Pathological Agents. In: Tripathi G, editor. Cellular and biochemical science. New Delhi: I.K. International Publishing House Pvt. Ltd.; 2010. p. 1047.

[168] Goodfellow, M. Ecology of actinomycetes. Annu Rev Microbiol. 1983;37:189-216.

[169] Hyde, E.R., Haarmann, D.P., Petrosino, J.F., Lynne, A.M., Bucheli, S.R. Initial insights into bacterial succession during human decomposition. Int J Legal Med. 2015;129:661-71.

[170] Torres, Y., Flores, I., Prieto, V., López-Soto, M., Farfán, M.a.J., Carracedo, A., Sanz, P. DNA mixtures in forensic casework: a 4-year retrospective study. Forensic Sci Int. 2003;134:180-6.

[171] Butler, J.M. Advanced Topics in Forensic DNA Typing: Interpretation. Oxford, England; San Diego, California: Academic Press; 2015.

[172] SWGDAM. SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories. 2010.

[173] Paoletti, D.R., Doom, T.E., Krane, C.M., Raymer, M.L., Krane, D.E. Empirical Analysis of the STR Profiles Resulting from Conceptual Mixtures. J Forensic Sci. 2005;50.

[174] Haned, H., Pene, L., Lobry, J.R., Dufour, A.B., Pontier, D. Estimating the number of contributors to forensic DNA mixtures: does maximum likelihood perform better than maximum allele count? J Forensic Sci. 2011;56:23-8.

[175] Ensenberger, M.G., Lenz, K.A., Matthies, L.K., Hadinoto, G.M., Schienman, J.E., Przech, A.J., Morganti, M.W., Renstrom, D.T., Baker, V.M., Gawrys, K.M., Hoogendoorn, M., Steffen, C.R., Martin, P., Alonso, A., Olson, H.R., Sprecher, C.J., Storts, D.R. Developmental validation of the PowerPlex((R)) Fusion 6C System. Forensic Sci Int Genet. 2016;21:134-44.

[176] ISP. Forensic Biology Section Casework Test Methods. Indiana State Police. 2016.

[177] Gill, P., Brenner, C.H., Buckleton, J.S., Carracedo, A., Krawczak, M., Mayr, W.R., Morling, N., Prinz, M., Schneider, P.M., Weir, B.S., Genetics, D.N.A.c.o.t.I.S.o.F. DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. Forensic Sci Int. 2006;160:90-101.

[178] Budowle, B., Onorato, A.J., Callaghan, T.F., Della Manna, A., Gross, A.M., Guerrieri, R.A., Luttman, J.C., McClure, D.L. Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework. J Forensic Sci. 2009;54:810-21.

[179] Butler, J.M., Hill, C.R., Kline, M.C., Duewer, D.L., Sprecher, C.J., McLaren, R.S., Rabbach, D.R., Krenke, B.E., Storts, D.R. The single most polymorphic STR Locus: SE33 performance in U.S. populations. Forensic Science International: Genetics Supplement Series. 2009;2:23-4.

[180] Butler, J.M., Hill, C.R., Coble, M.D. Variability of New STR Loci and Kits in US Population Groups. https://www.promega.com/resources/profiles-in-dna/2012/variability-of-new-str-loci-and-kits-in-us-population-groups/2012.

[181] Butler, J.M., Hill, C.R., Kline, M.C., Bastisch, I., Weirich, V., McLaren, R.S., Storts, D.R. SE33 variant alleles: Sequences and implications. Forensic Sci Int Genet Supp. 2011;3:e502-e3.

[182] Butler, J.M., Hill, C.R. Biology and Genetics of New Autosomal STR Loci Useful for Forensic DNA Analysis. Forensic Sci Rev. 2012;24.

[183] Moretti, T.R., Moreno, L.I., Smerick, J.B., Pignone, M.L., Hizon, R., Buckleton, J.S., Bright, J.A., Onorato, A.J. Population data on the expanded CODIS core STR loci for eleven populations of significance for forensic DNA analyses in the United States. Forensic Sci Int Genet. 2016;25:175-81.

[184] PCAST. Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. President's Council of Advisors on Science and Technology; 2016.

[185] Jablonski, N.G., Chaplin, G. The evolution of human skin coloration. J Hum Evol. 2000;39:57-106.

[186] Jablonski, N.G., Chaplin, G. Human skin pigmentation as an adaptation to UV radiation. Proc Natl Acad Sci U S A. 2010;107 Suppl 2:8962-8.

[187] Gettings, K.B., Kiesler, K.M., Faith, S.A., Montano, E., Baker, C.H., Young, B.A., Guerrieri, R.A., Vallone, P.M. Sequence variation of 22 autosomal STR loci detected by next generation sequencing. Forensic Sci Int Genet. 2016;21:15-21.

APPENDIX A. SELF-REPORTING SURVEY

Development of a Multiplex Forensic Phenotype Profile (FPP) Assay

Sex: MALE FEMALE

Eye Color:

BLUE BROWN GREEN HAZEL OTHER: _____

Skin Color:

LIGHT DARK INTERMEDIATE

Skin type when exposed to sun:

Type I (always burn)	Type II (rarely tan)	Type III (sometimes tan)	Type IV (likely to tan)	Type V (tan easily)	Type VI (never burn)
Very fair →	Fair →	Light →	Olive →	Brown →	Black

Do you tan regularly? YES NO If yes, how often (days/month): _____

Natural Hair Color:

BLONDE BROWN BLACK RED

Do you dye your hair? YES NO If yes, last time it was dyed (days): _____

Ancestry (check all that apply):

	Maternal			Paternal		
	Mother	Grandmother	Grandfather	Father	Grandmother	Grandfather
European						
E. Asian						
Asian (other)						
Hispanic						
African-American						
Other						
Unknown						

Age (years): _____ Height: _____ Weight (lbs): _____

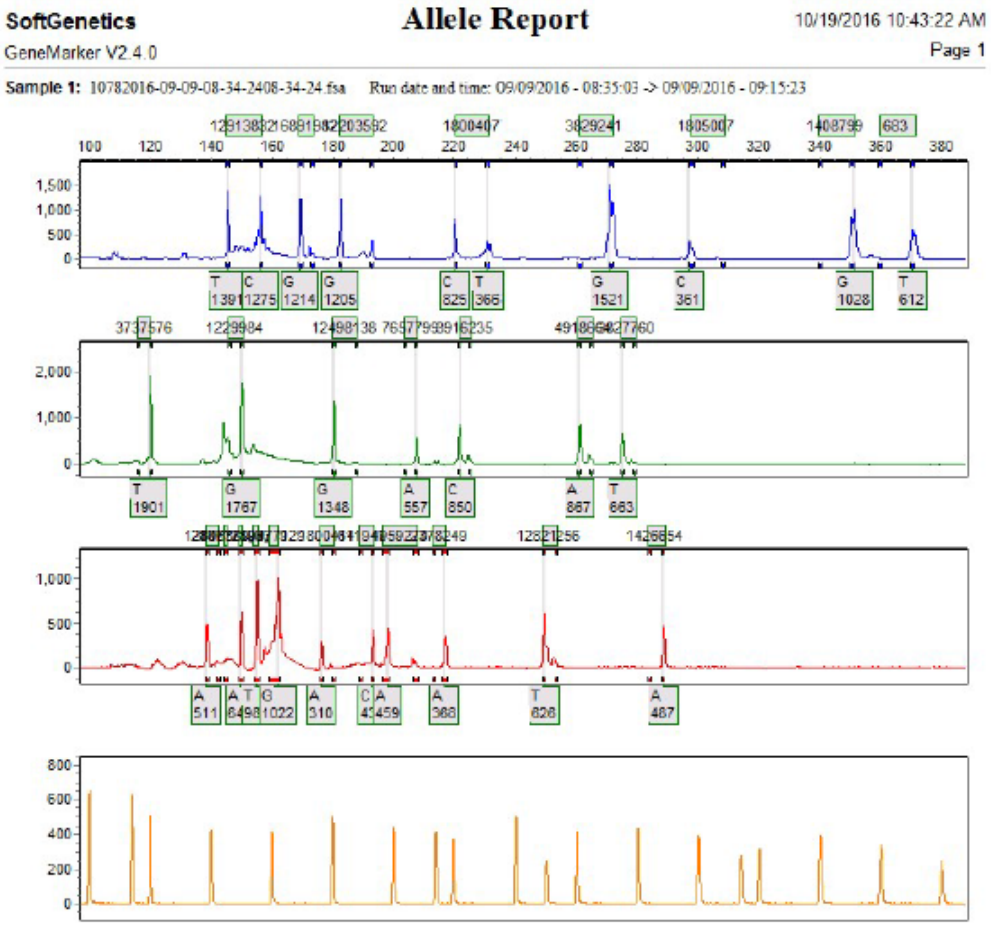
Do you smoke/use tobacco? _____ If so, how often (cigarettes/day): _____

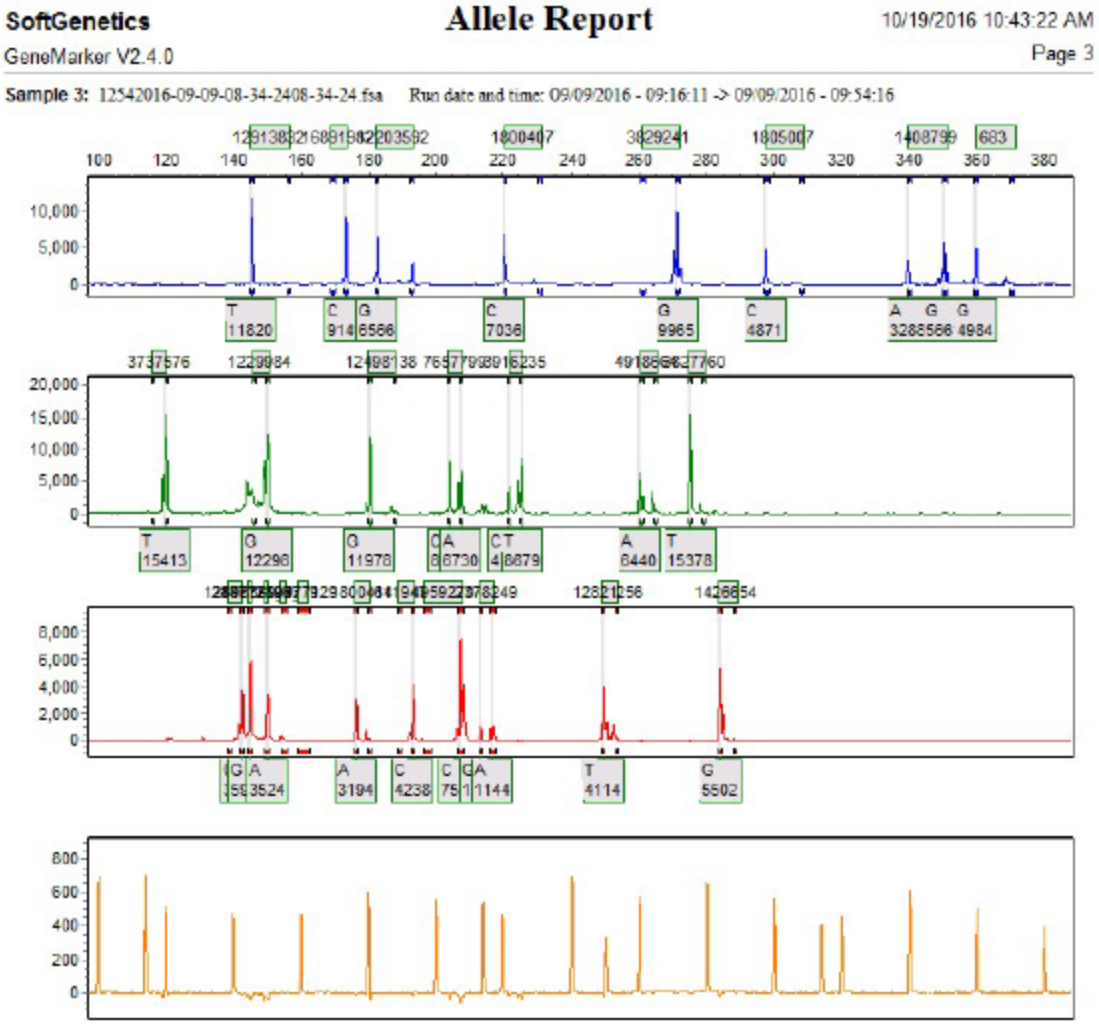
Mark box if you no longer smoke: How long did you smoke (years): _____

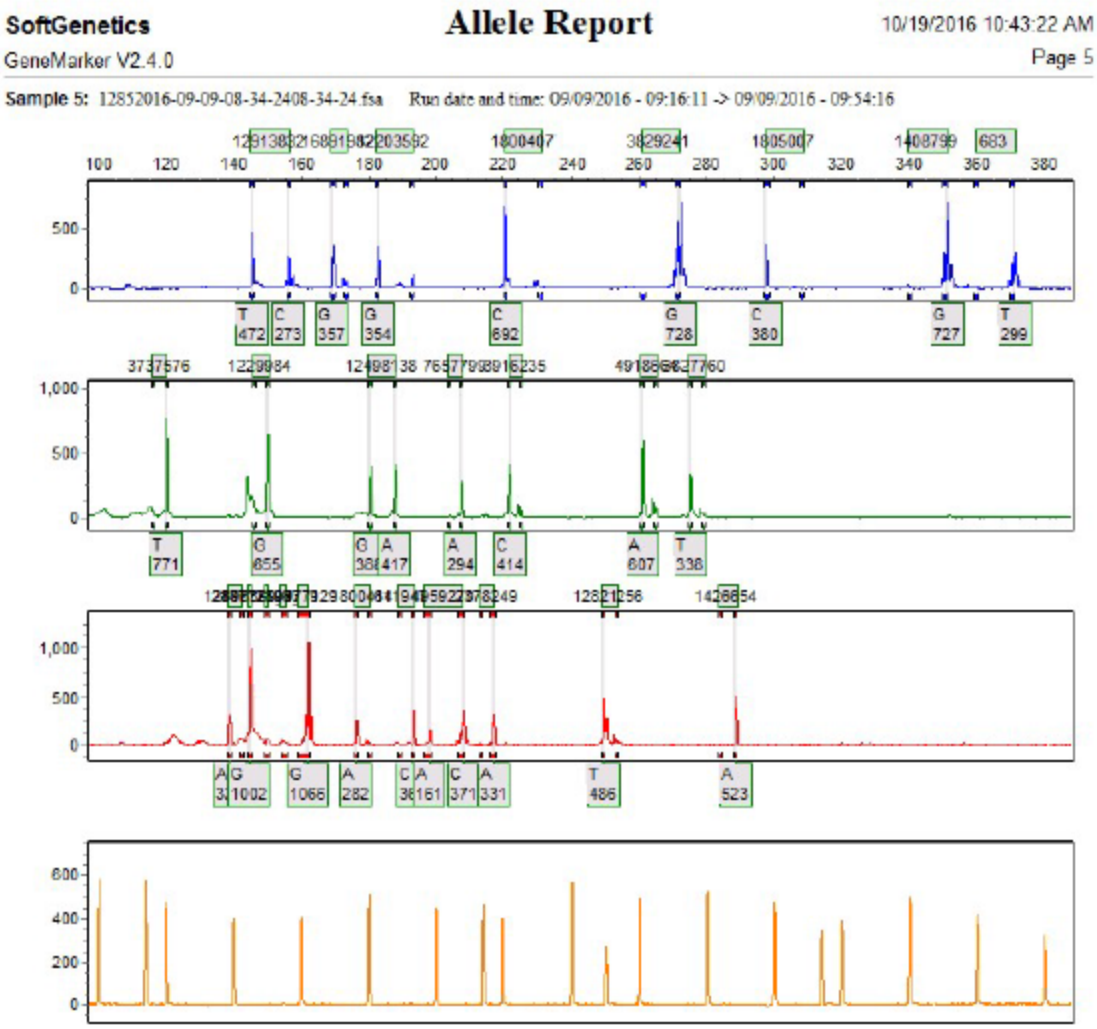
Do you exercise: YES NO How often (days/week): _____

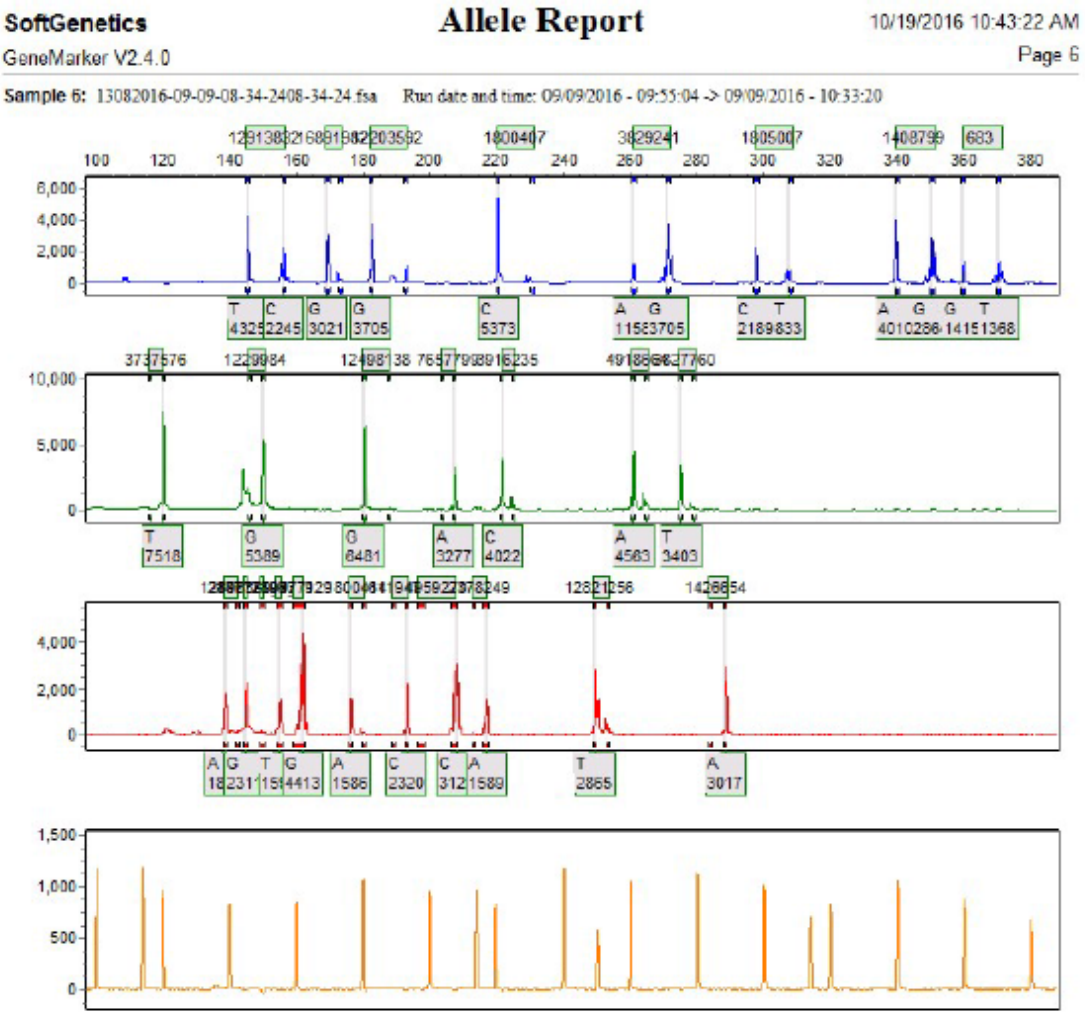
Do you drink alcohol: YES NO How often (days/month): _____

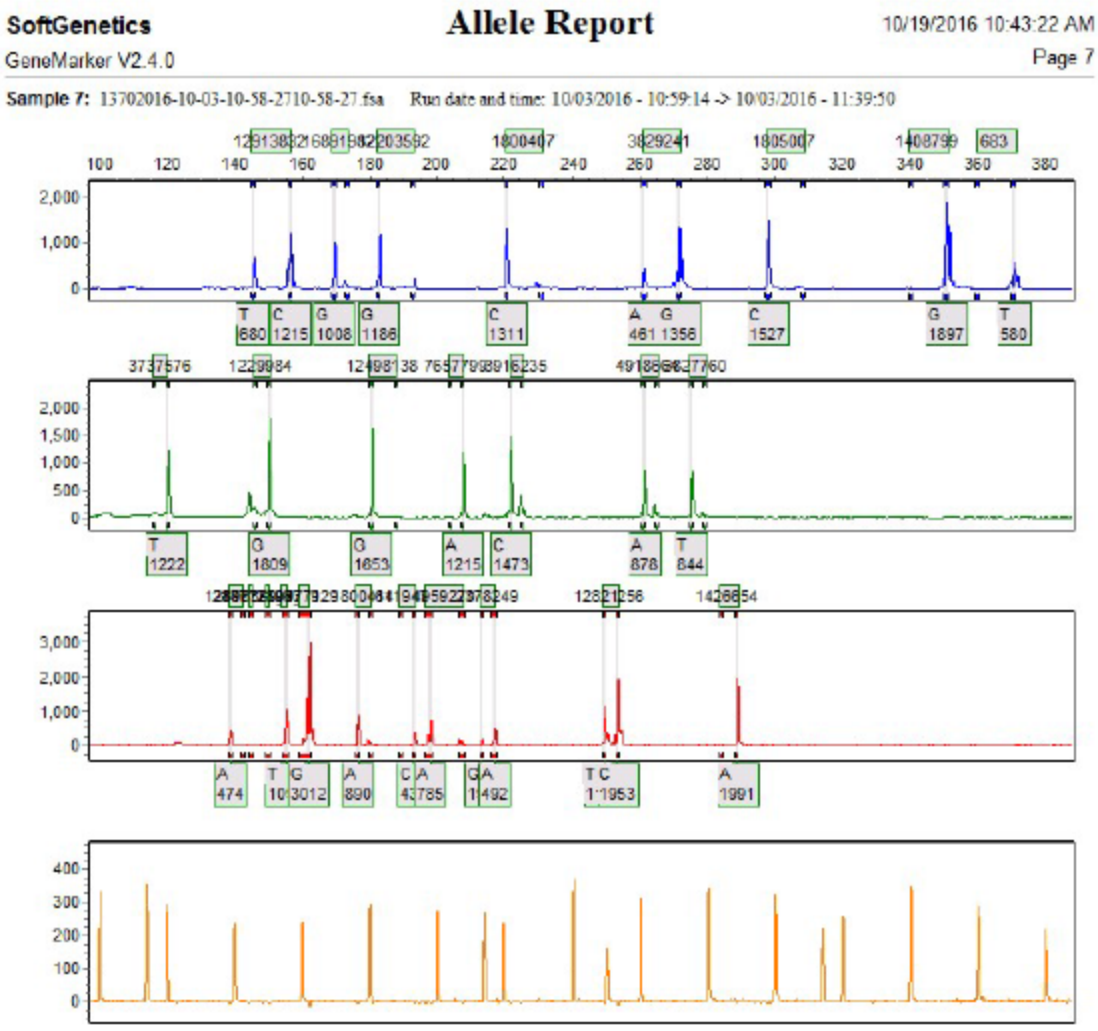
APPENDIX B. FPP PROFILES OF THE SAMPLE POPULATION (N=200)

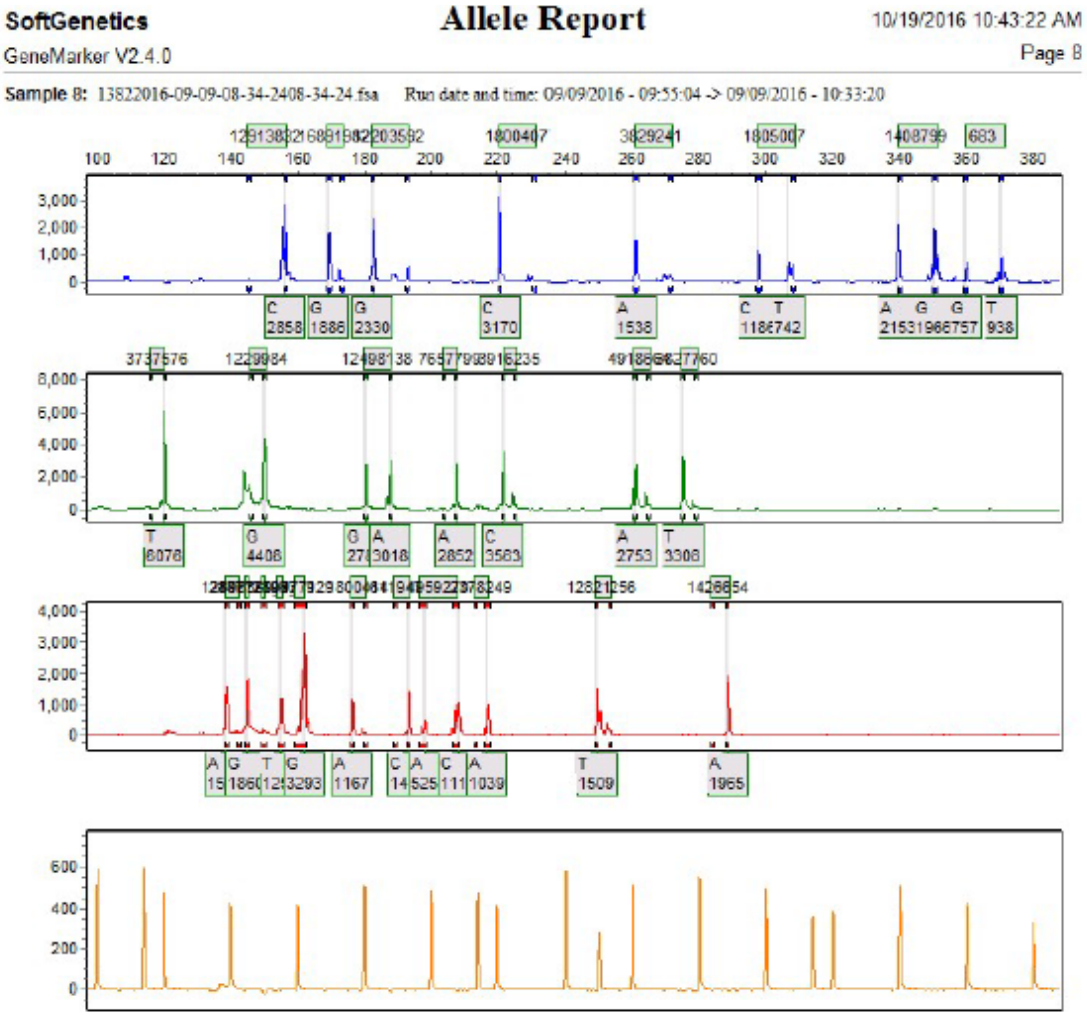



SoftGenetics
Allele Report
10/19/2016 10:43:22 AM
GeneMarker V2.4.0
Page 3
Sample 3: 12542016-09-09-08-34-2408-34-24.fsa Run date and time: 09/09/2016 - 09:16:11 -> 09/09/2016 - 09:54:16


SoftGenetics
Allele Report
10/19/2016 10:43:22 AM
GeneMarker V2.4.0
Page 5
Sample 5: 12852016-09-09-08-34-2408-34-24.fsa Run date and time: 09/09/2016 - 09:16:11 -> 09/09/2016 - 09:54:16


SoftGenetics
Allele Report
10/19/2016 10:43:22 AM
GeneMarker V2.4.0
Page 6
Sample 6: 13082016-09-09-08-34-2408-34-24.fsa Run date and time: 09/09/2016 - 09:55:04 -> 09/09/2016 - 10:33:20

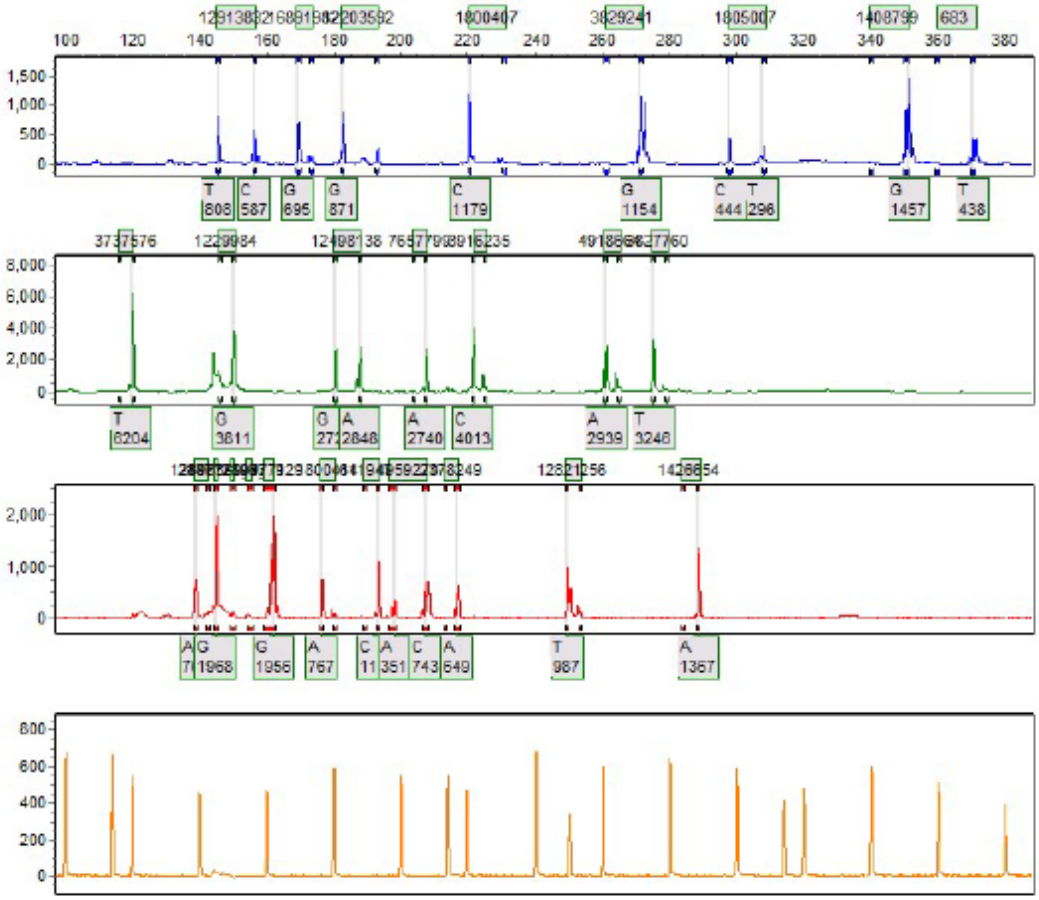

SoftGenetics
Allele Report
10/19/2016 10:43:22 AM
GeneMarker V2.4.0
Page 7
Sample 7: 13702016-10-03-10-58-2710-58-27.fsa Run date and time: 10/03/2016 - 10:59:14 -> 10/03/2016 - 11:39:50


SoftGenetics
Allele Report
10/19/2016 10:43:22 AM
GeneMarker V2.4.0
Page 8
Sample 8: 13822016-09-09-08-34-2408-34-24.fsa Run date and time: 09/09/2016 - 09:55:04 -> 09/09/2016 - 10:33:20

SoftGenetics **Allele Report** 10/19/2016 10:43:22 AM

GeneMarker V2.4.0 Page 9

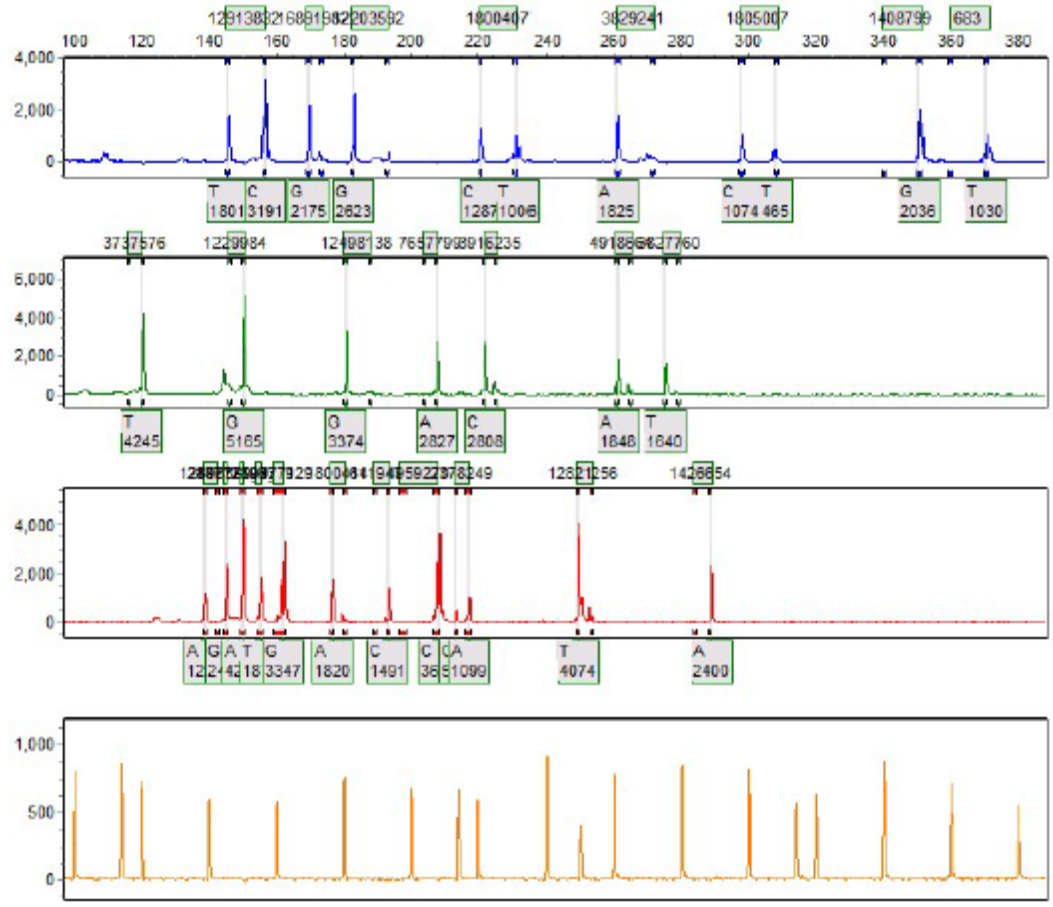
Sample 9: 14272016-09-09-08-34-2408-34-24.fsa Run date and time: 09/09/2016 - 09:55:04 -> 09/09/2016 - 10:33:20



SoftGenetics **Allele Report** 10/19/2016 10:43:23 AM

GeneMarker V2.4.0 Page 10

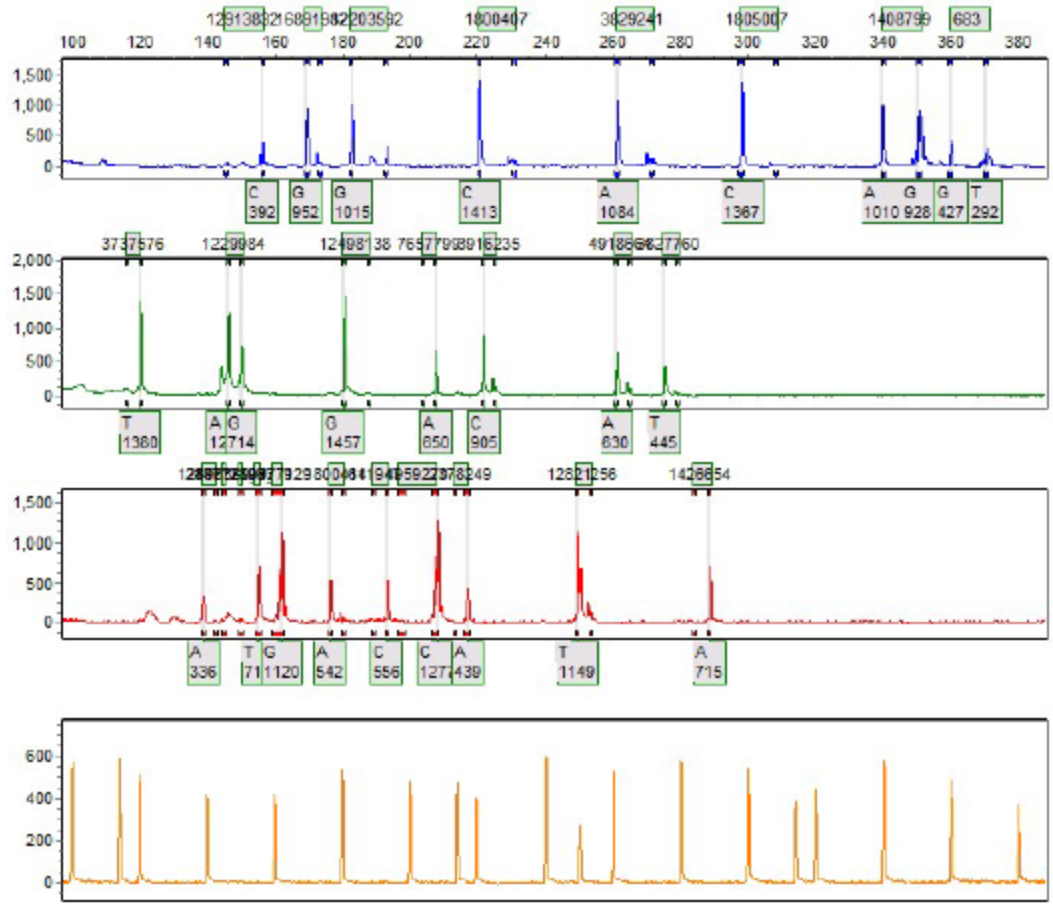
Sample 10: 15722016-10-03-10-58-2710-58-27.fsa Run date and time: 10/03/2016 - 10:59:14 -> 10/03/2016 - 11:39:50

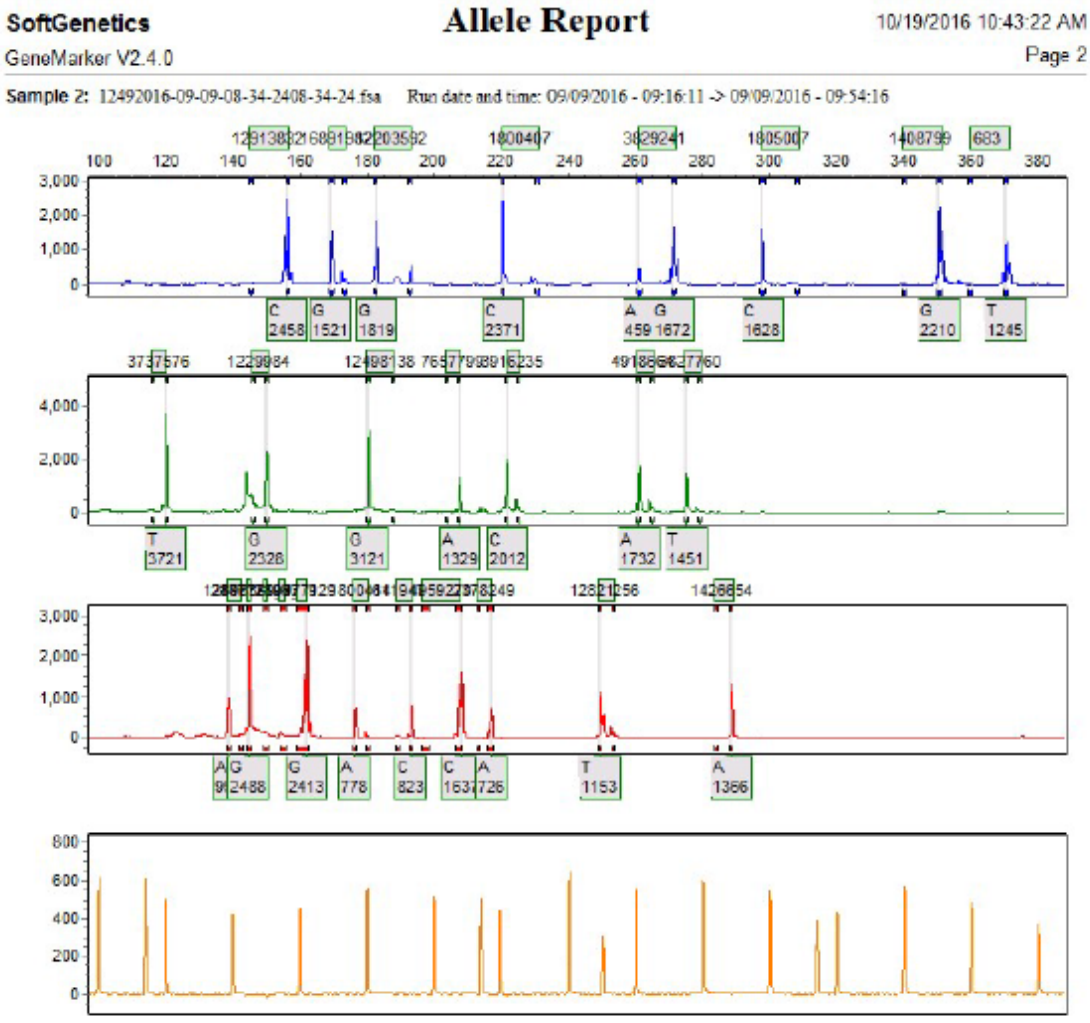


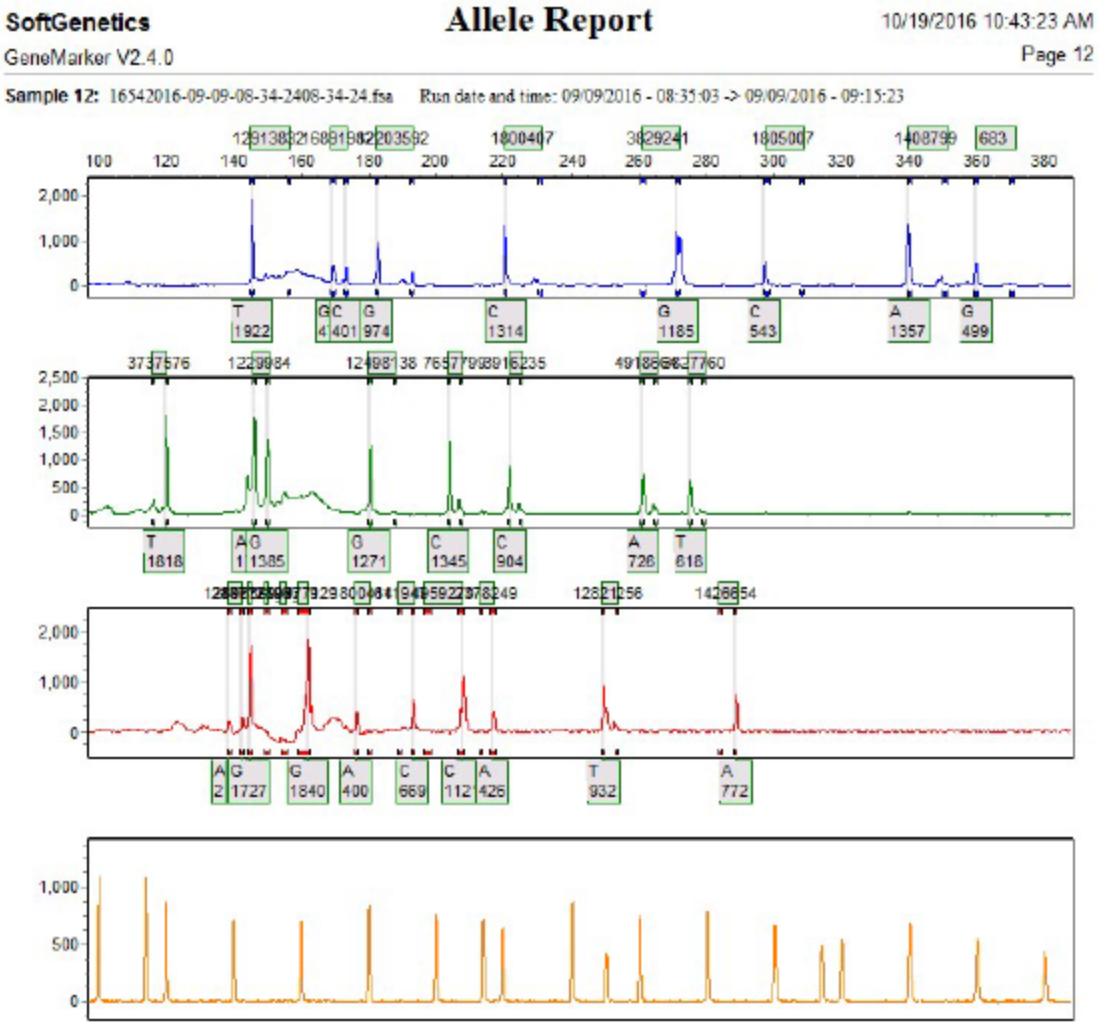
SoftGenetics **Allele Report** 10/19/2016 10:43:23 AM

GeneMarker V2.4.0 Page 11

Sample 11: 15802016-09-22-16-13-3916-13-39.fsa Run date and time: 09/22/2016 - 17:41:20 -> 09/22/2016 - 18:19:30



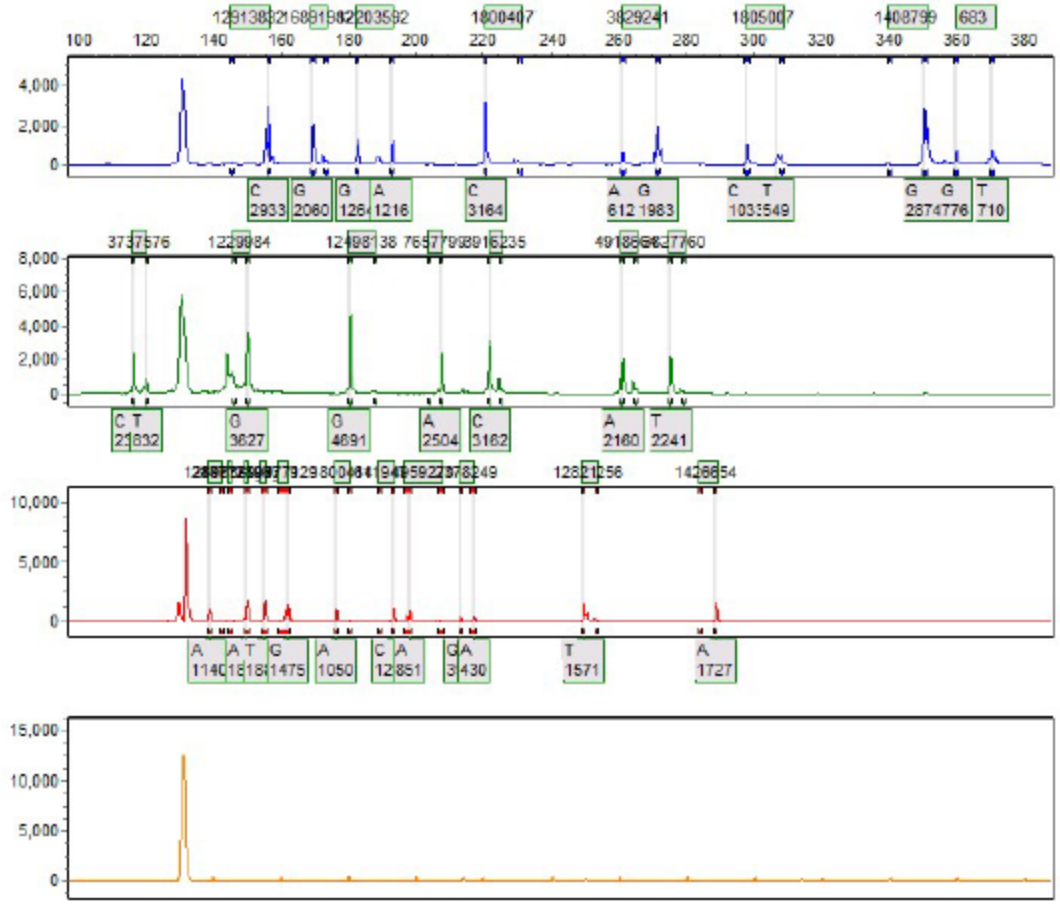

SoftGenetics
Allele Report
10/19/2016 10:43:22 AM
GeneMarker V2.4.0
Page 2
Sample 2: 12492016-09-09-08-34-2408-34-24.fsa Run date and time: 09/09/2016 - 09:16:11 -> 09/09/2016 - 09:54:16

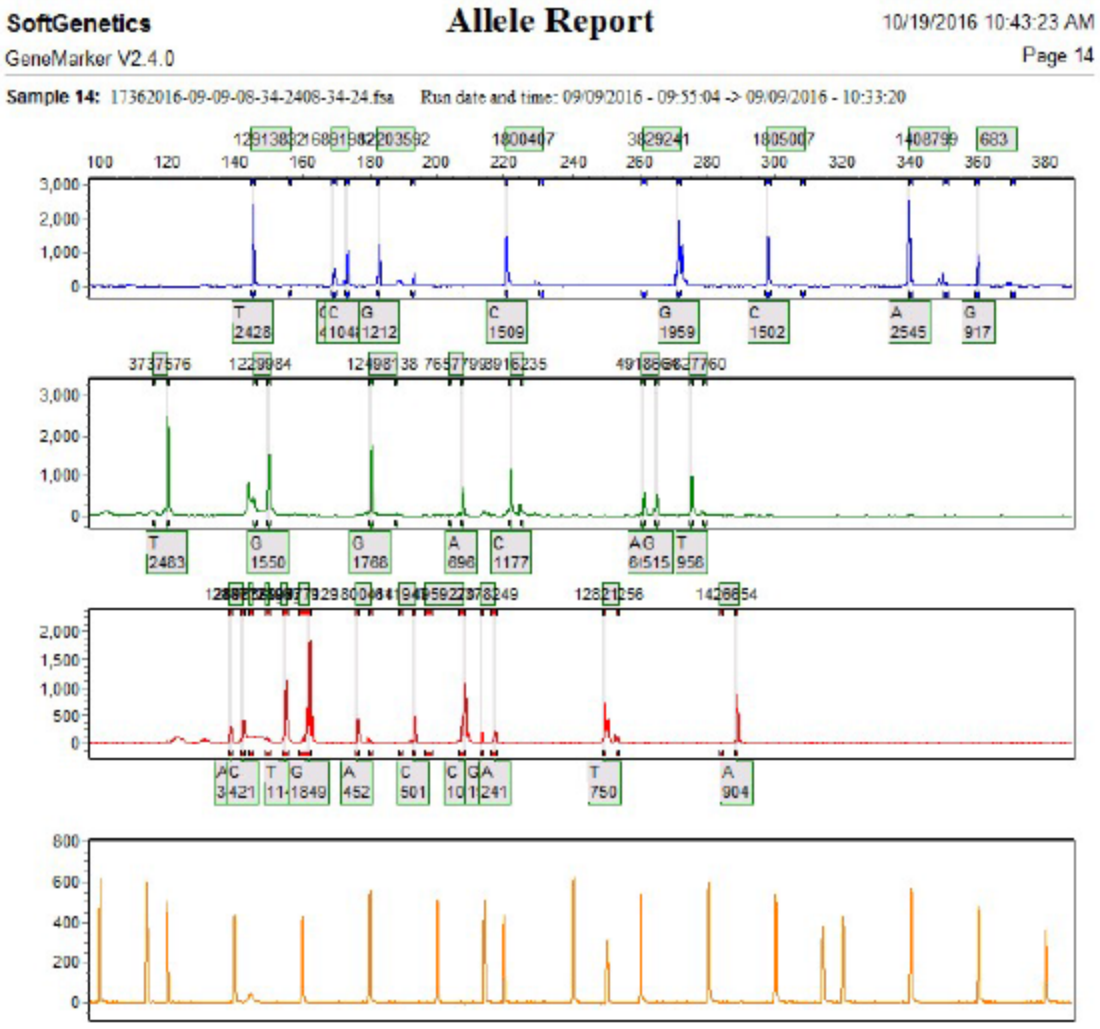

SoftGenetics
Allele Report
10/19/2016 10:43:23 AM
GeneMarker V2.4.0
Page 12
Sample 12: 16542016-09-09-08-34-2408-34-24.fsa Run date and time: 09/09/2016 - 08:35:03 -> 09/09/2016 - 09:15:23

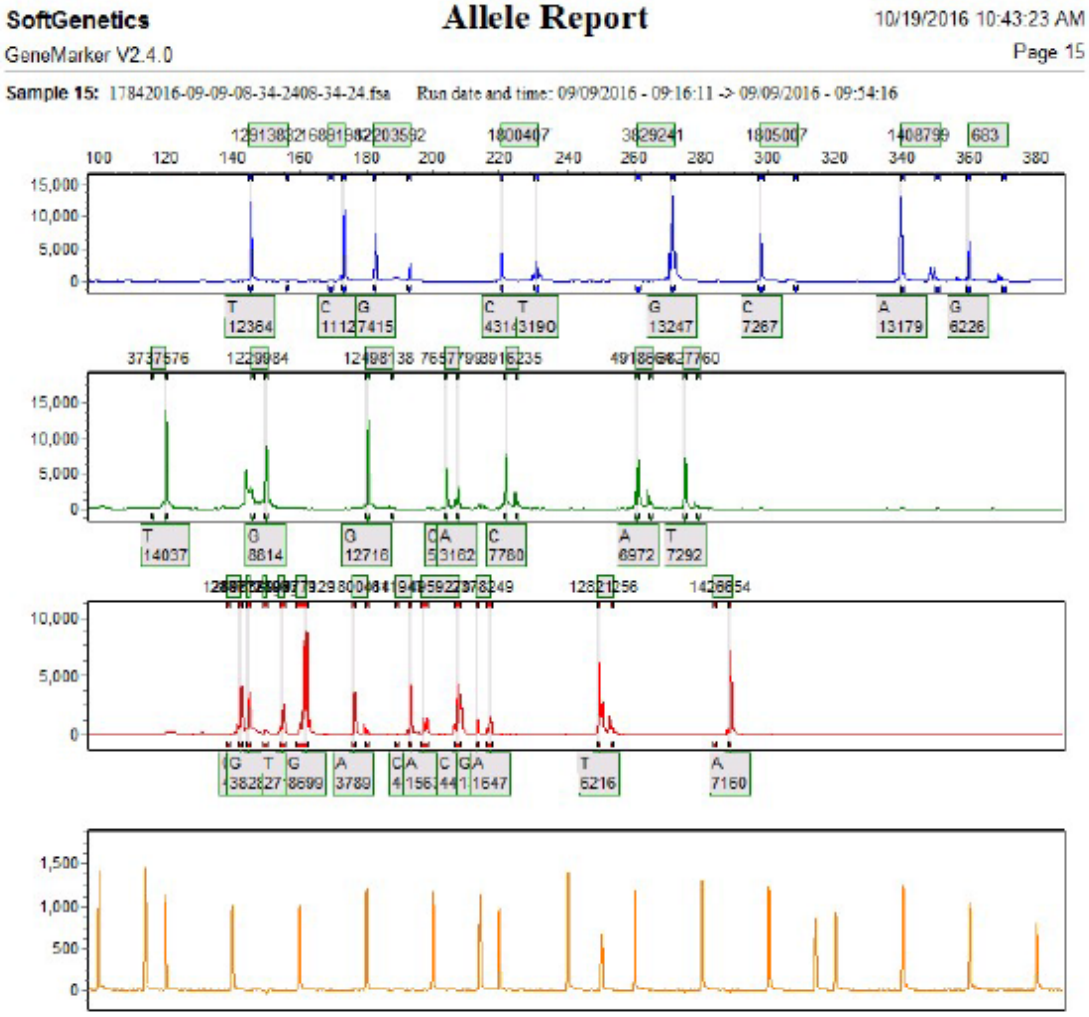
SoftGenetics **Allele Report** 10/19/2016 10:43:23 AM

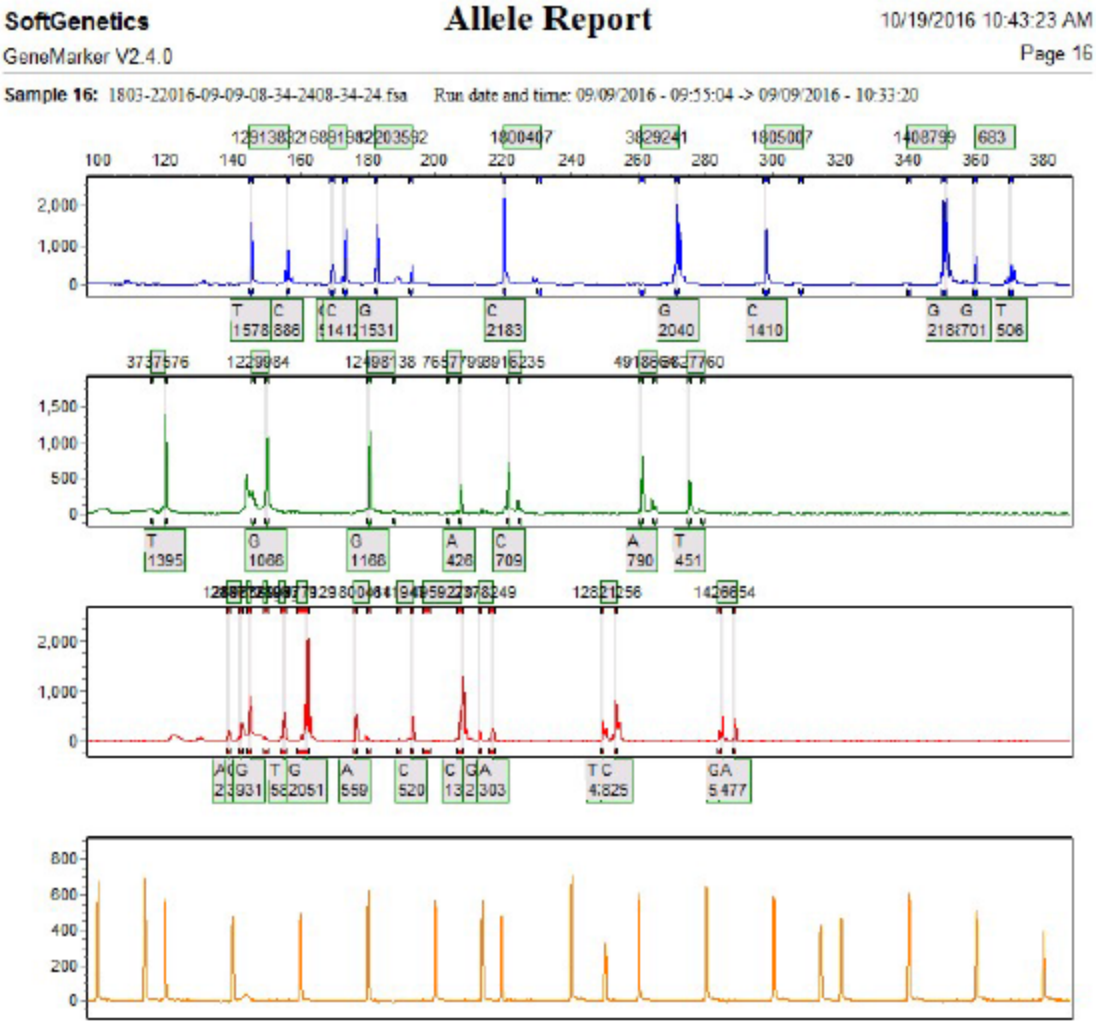
GeneMarker V2.4.0 Page 13

Sample 13: 17262016-09-09-08-34-2408-34-24.fsa Run date and time: 09/09/2016 - 09:55:04 -> 09/09/2016 - 10:33:20




SoftGenetics
Allele Report
10/19/2016 10:43:23 AM
GeneMarker V2.4.0
Page 14
Sample 14: 17362016-09-09-08-34-2408-34-24.fsa Run date and time: 09/09/2016 - 09:55:04 -> 09/09/2016 - 10:33:20

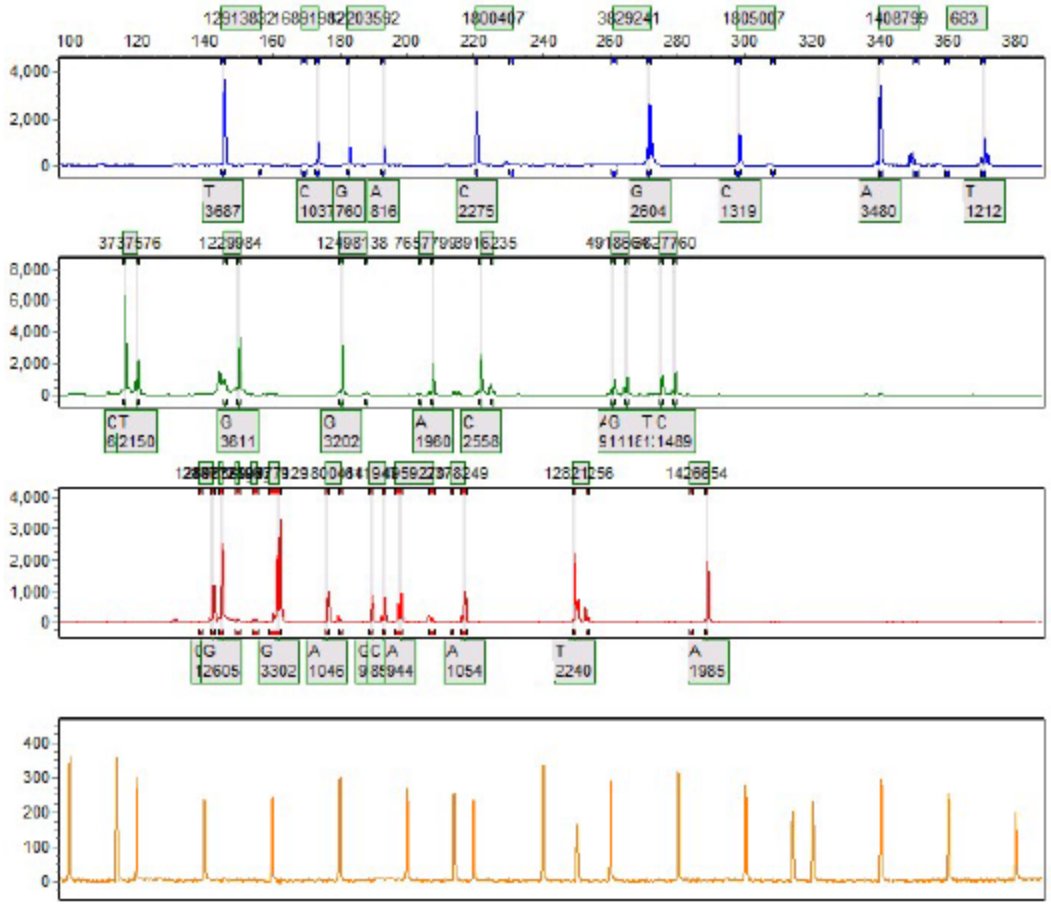

SoftGenetics
Allele Report
10/19/2016 10:43:23 AM
GeneMarker V2.4.0
Page 15
Sample 15: 17842016-09-09-08-34-2408-34-24.fsa Run date and time: 09/09/2016 - 09:16:11 -> 09/09/2016 - 09:54:16

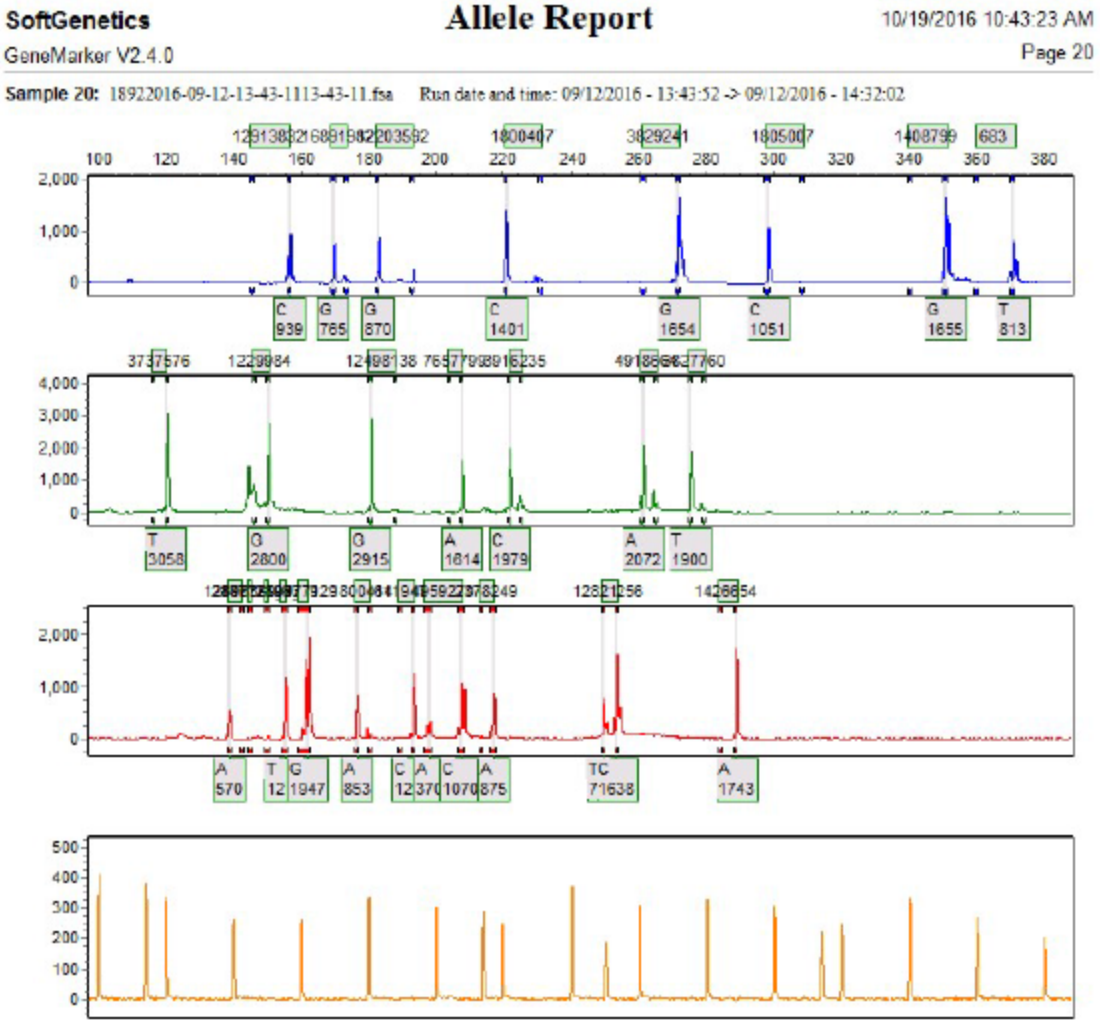

SoftGenetics
Allele Report
10/19/2016 10:43:23 AM
GeneMarker V2.4.0
Page 16
Sample 16: 1803-22016-09-09-08-34-2408-34-24.fsa Run date and time: 09/09/2016 - 09:55:04 -> 09/09/2016 - 10:33:20

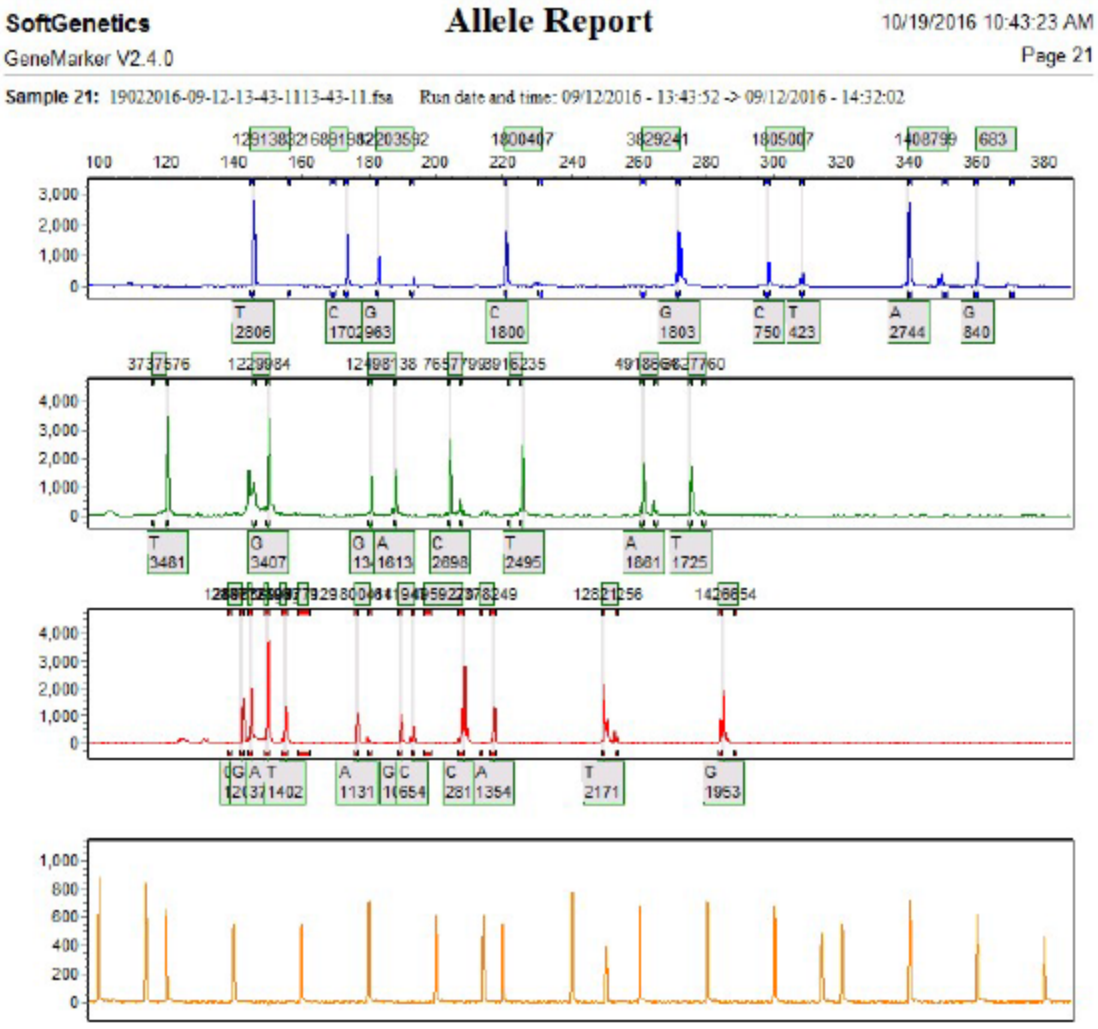
SoftGenetics **Allele Report** 10/19/2016 10:43:23 AM

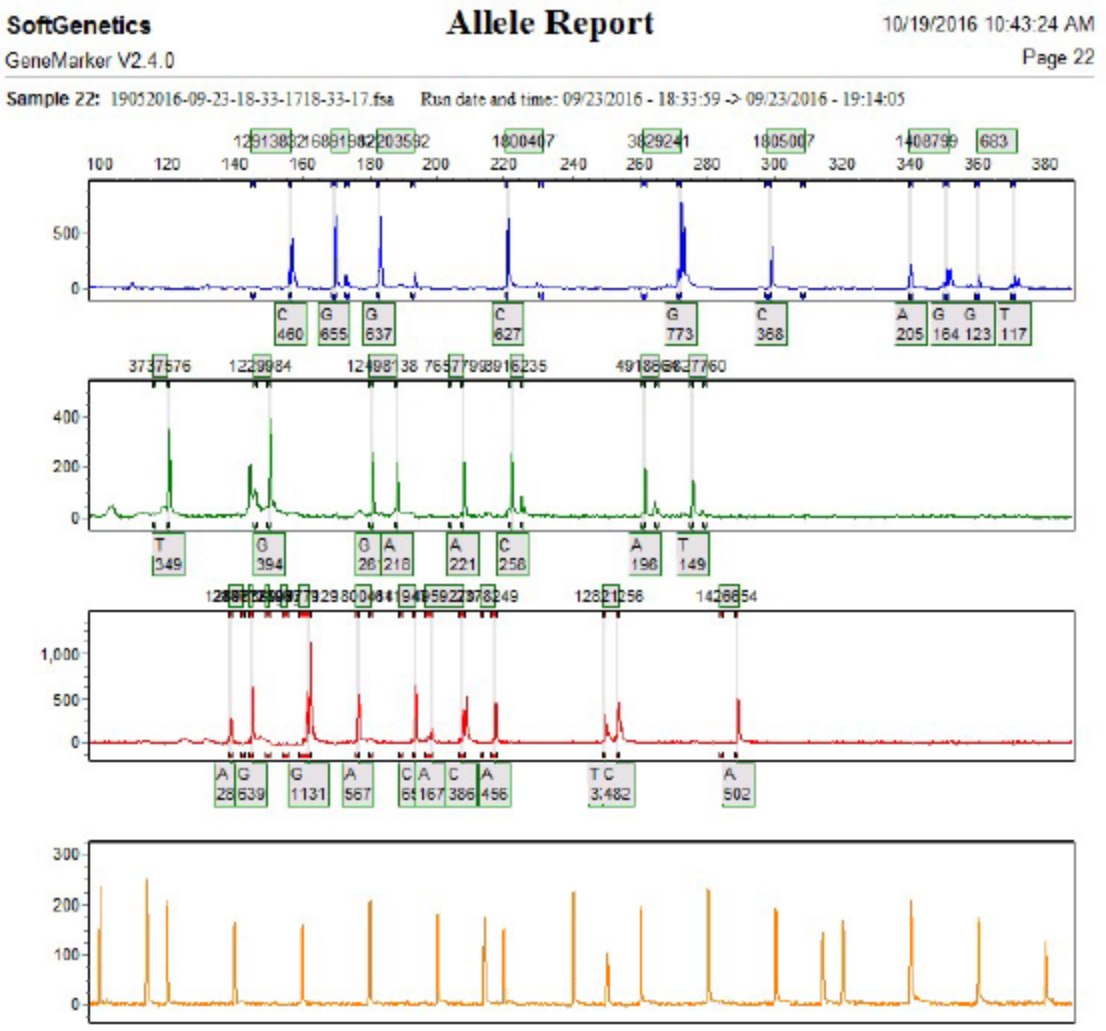
GeneMarker V2.4.0 Page 17

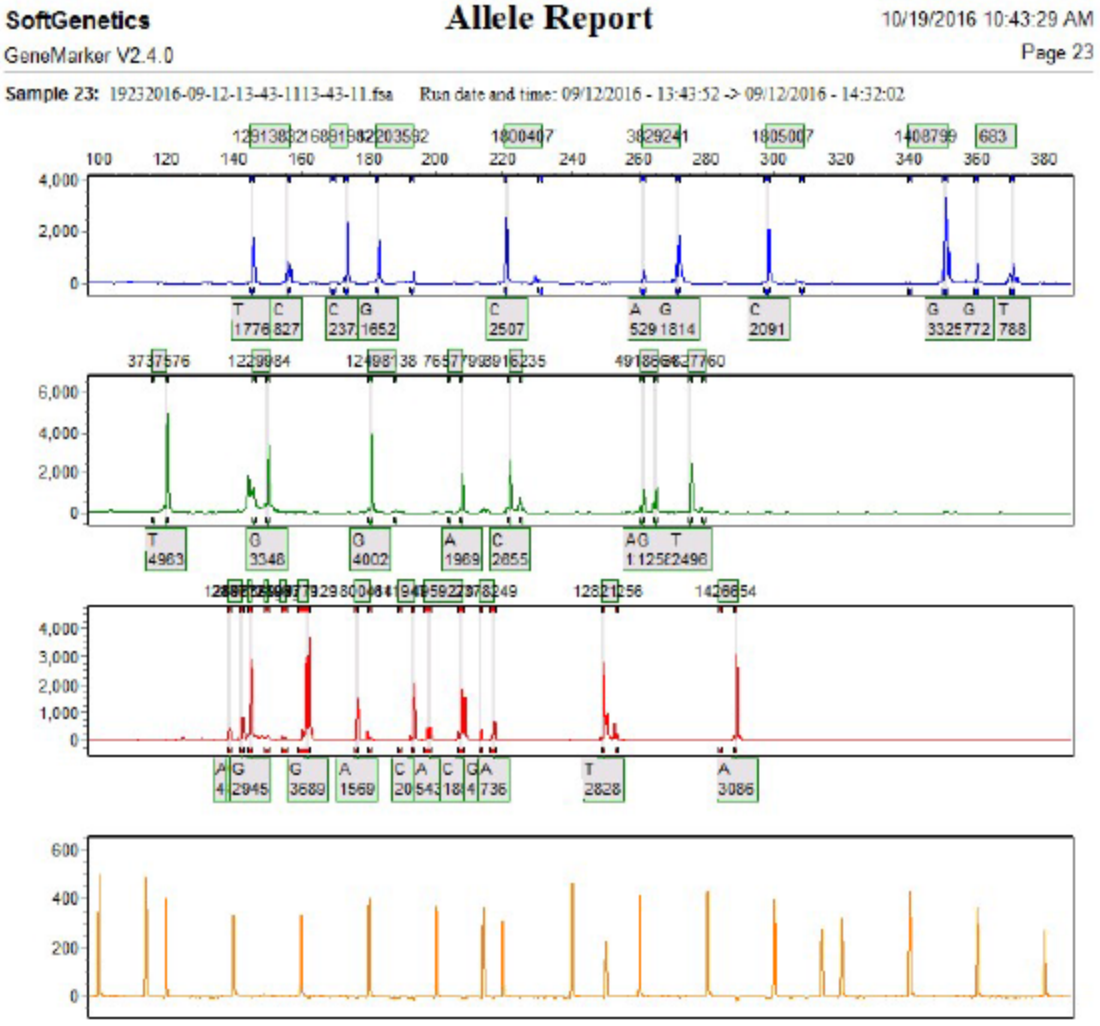
Sample 17: 18242016-09-12-13-43-1113-43-11.fsa Run date and time: 09/12/2016 - 13:43:52 -> 09/12/2016 - 14:32:02

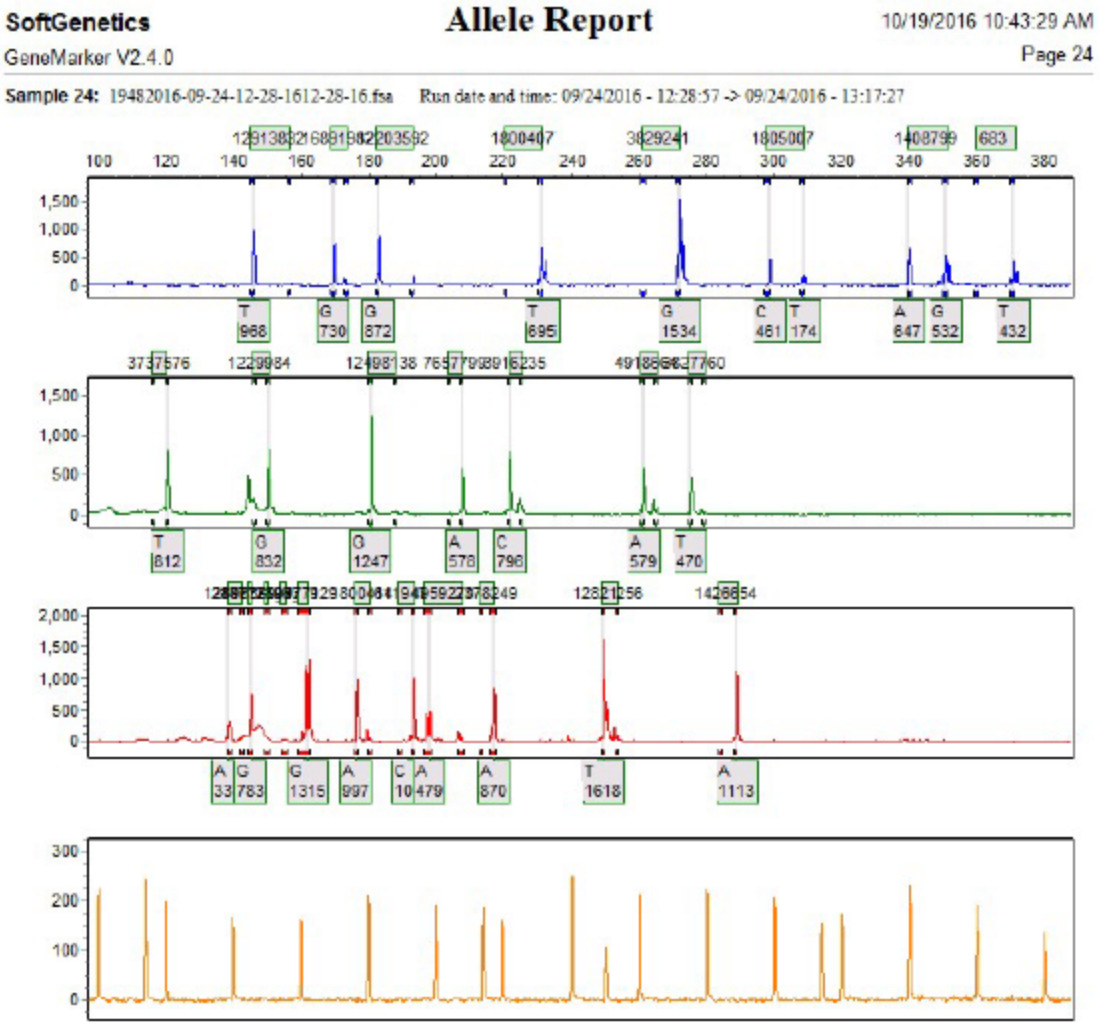



SoftGenetics
Allele Report
10/19/2016 10:43:23 AM
GeneMarker V2.4.0
Page 20
Sample 20: 18922016-09-12-13-43-1113-43-11.fsa Run date and time: 09/12/2016 - 13:43:52 -> 09/12/2016 - 14:32:02
C 939
G 765
G 870
C 1401
G 1654
C 1051
G 1655
T 813
T 3058
G 2800
G 2915
A 1614
C 1979
A 2072
T 1900
A 570
T 12
G 1947
A 853
C 12
A 370
C 1070
A 875
TC 71638
A 1743


SoftGenetics
Allele Report
10/19/2016 10:43:23 AM
GeneMarker V2.4.0
Page 21
Sample 21: 19022016-09-12-13-43-1113-43-11.fsa Run date and time: 09/12/2016 - 13:43:52 -> 09/12/2016 - 14:32:02


SoftGenetics
GeneMarker V2.4.0
Allele Report
10/19/2016 10:43:24 AM
Page 22
Sample 22: 19052016-09-23-18-33-1718-33-17.fsa Run date and time: 09/23/2016 - 18:33:59 -> 09/23/2016 - 19:14:05

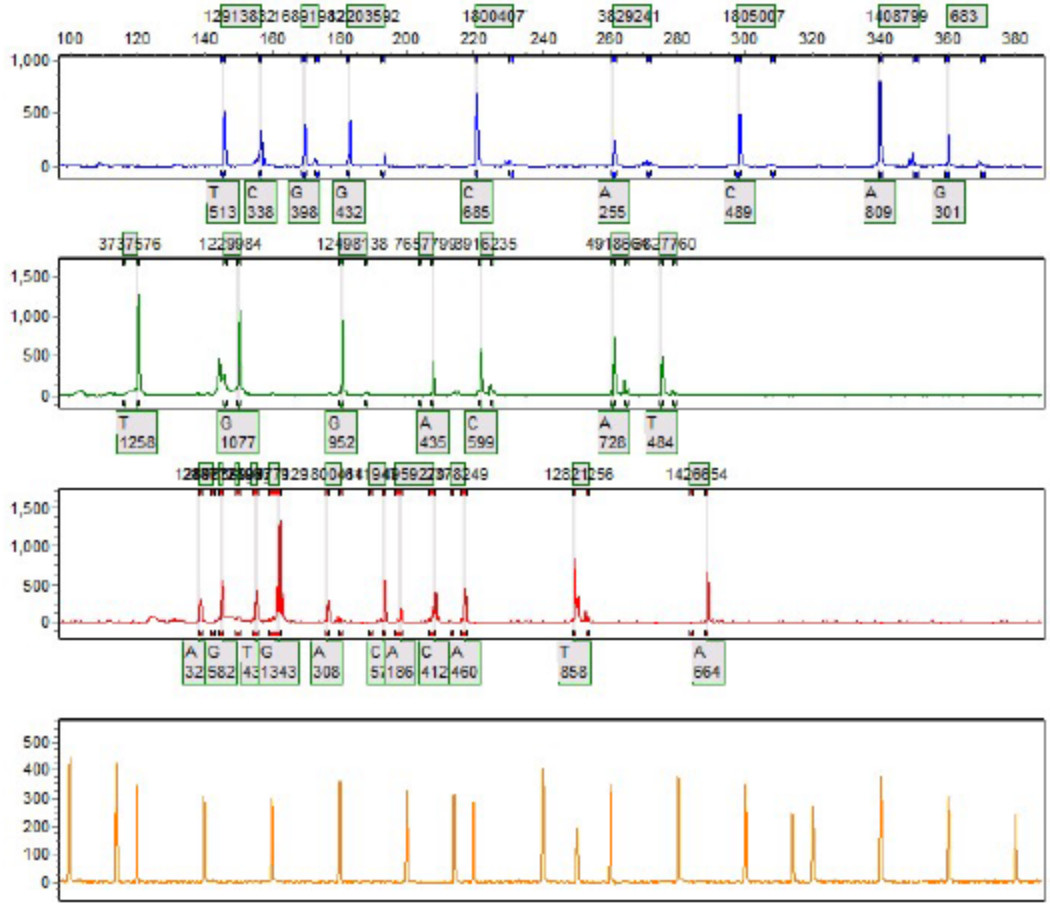

SoftGenetics
Allele Report
10/19/2016 10:43:29 AM
GeneMarker V2.4.0
Page 23
Sample 23: 19232016-09-12-13-43-1113-43-11.fsa Run date and time: 09/12/2016 - 13:43:52 -> 09/12/2016 - 14:32:02


SoftGenetics
Allele Report
10/19/2016 10:43:29 AM
GeneMarker V2.4.0
Page 24
Sample 24: 19482016-09-24-12-28-1612-28-16.fsa Run date and time: 09/24/2016 - 12:28:57 -> 09/24/2016 - 13:17:27

SoftGenetics **Allele Report** 10/19/2016 10:51:44 AM

GeneMarker V2.4.0 Page 1

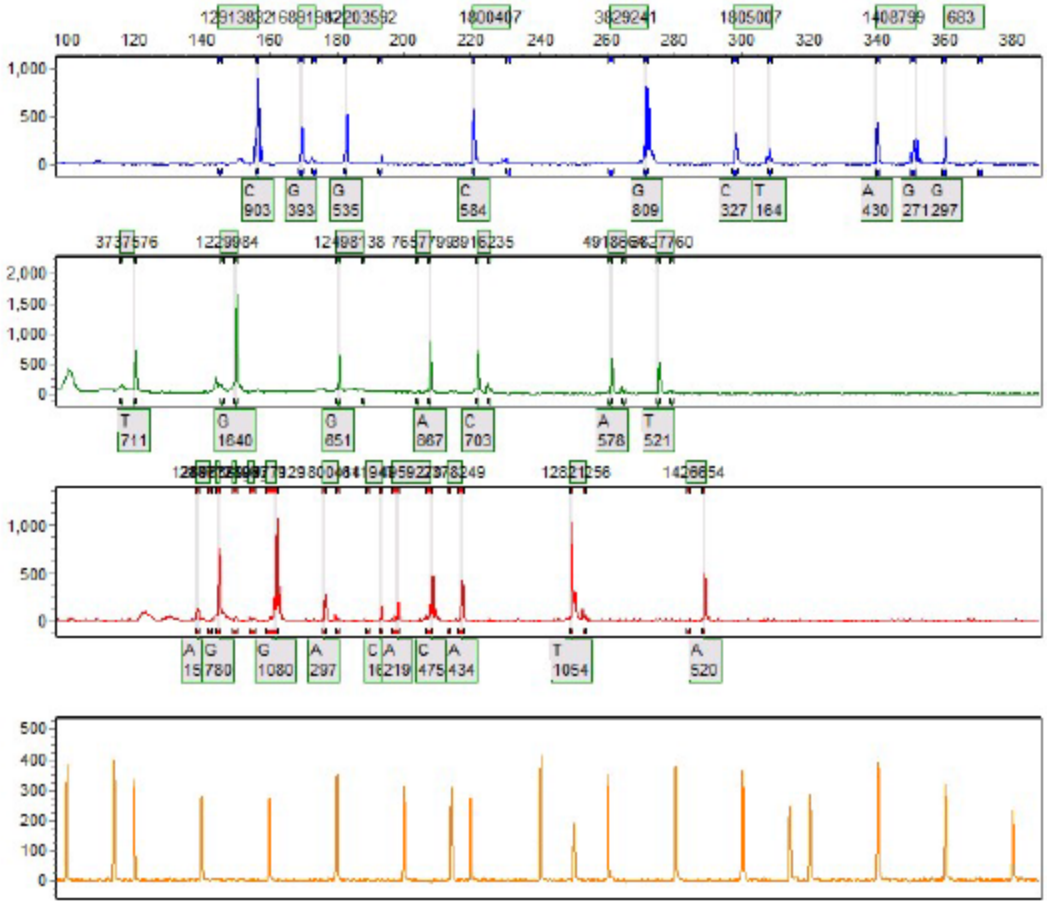
Sample 1: 20792016-09-12-13-43-1113-43-11.fsa Run date and time: 09/12/2016 - 13:43:52 -> 09/12/2016 - 14:32:02



SoftGenetics **Allele Report** 10/19/2016 10:51:44 AM

GeneMarker V2.4.0 Page 2

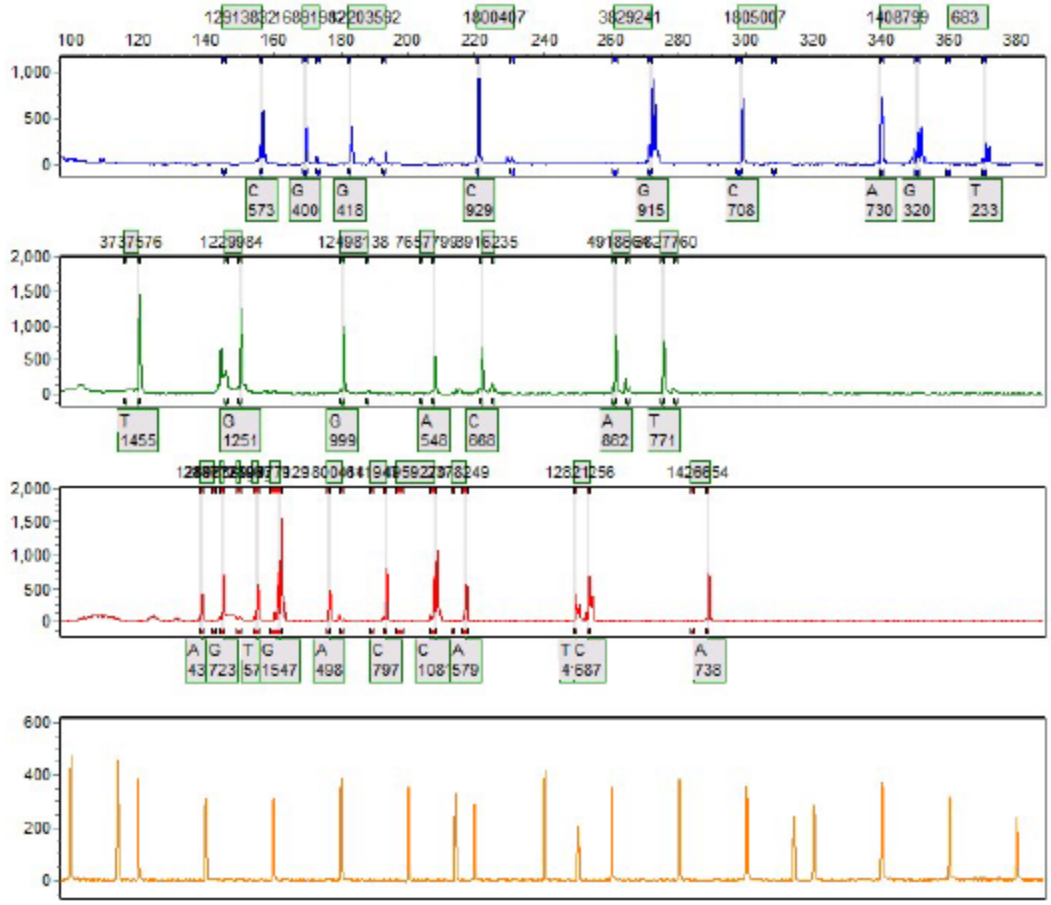
Sample 2: 20932016-10-03-10-58-2710-58-27.fsa Run date and time: 10/03/2016 - 10:59:14 -> 10/03/2016 - 11:39:50

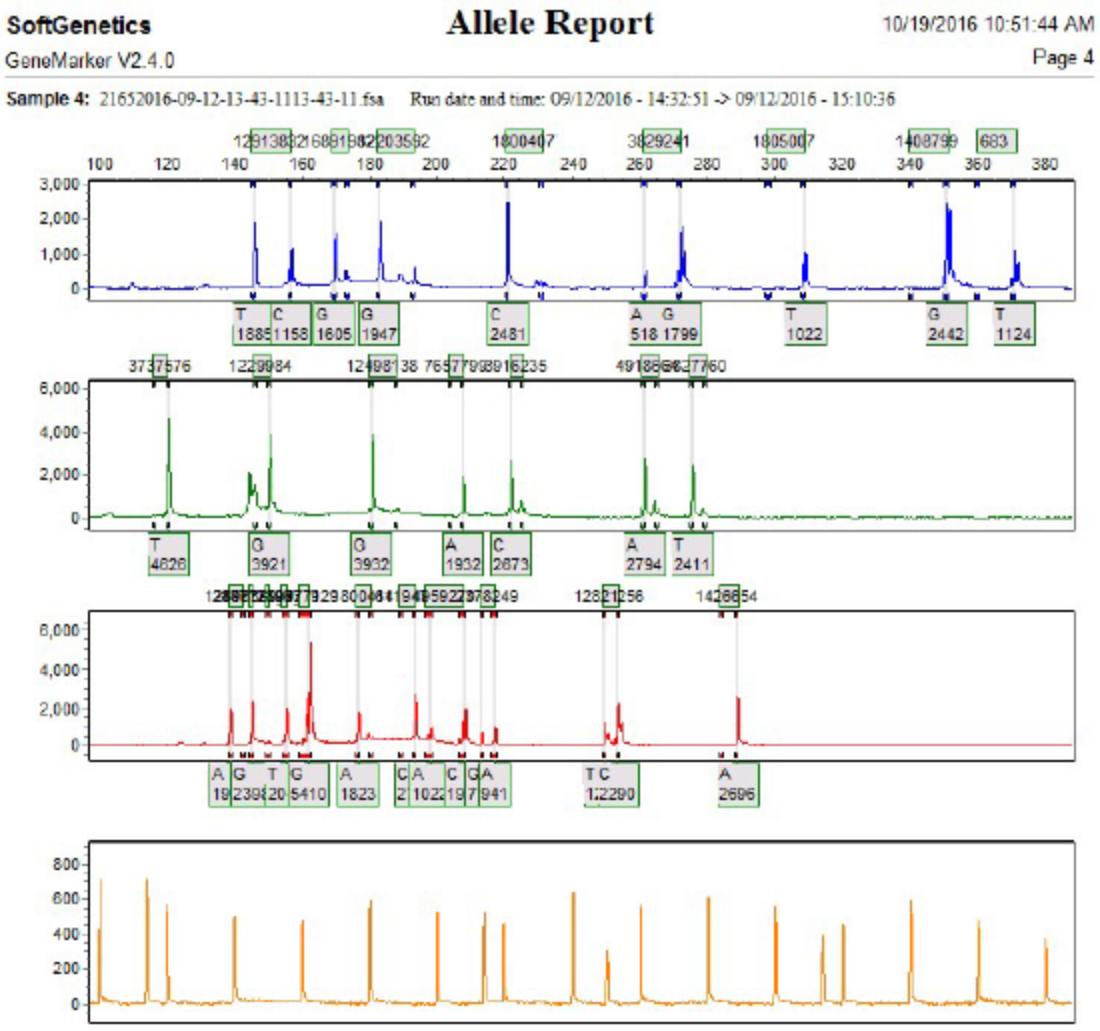


SoftGenetics **Allele Report** 10/19/2016 10:51:44 AM

GeneMarker V2.4.0 Page 3

Sample 3: 21562016-09-12-13-43-1113-43-11.fsa Run date and time: 09/12/2016 - 14:32:51 -> 09/12/2016 - 15:10:36

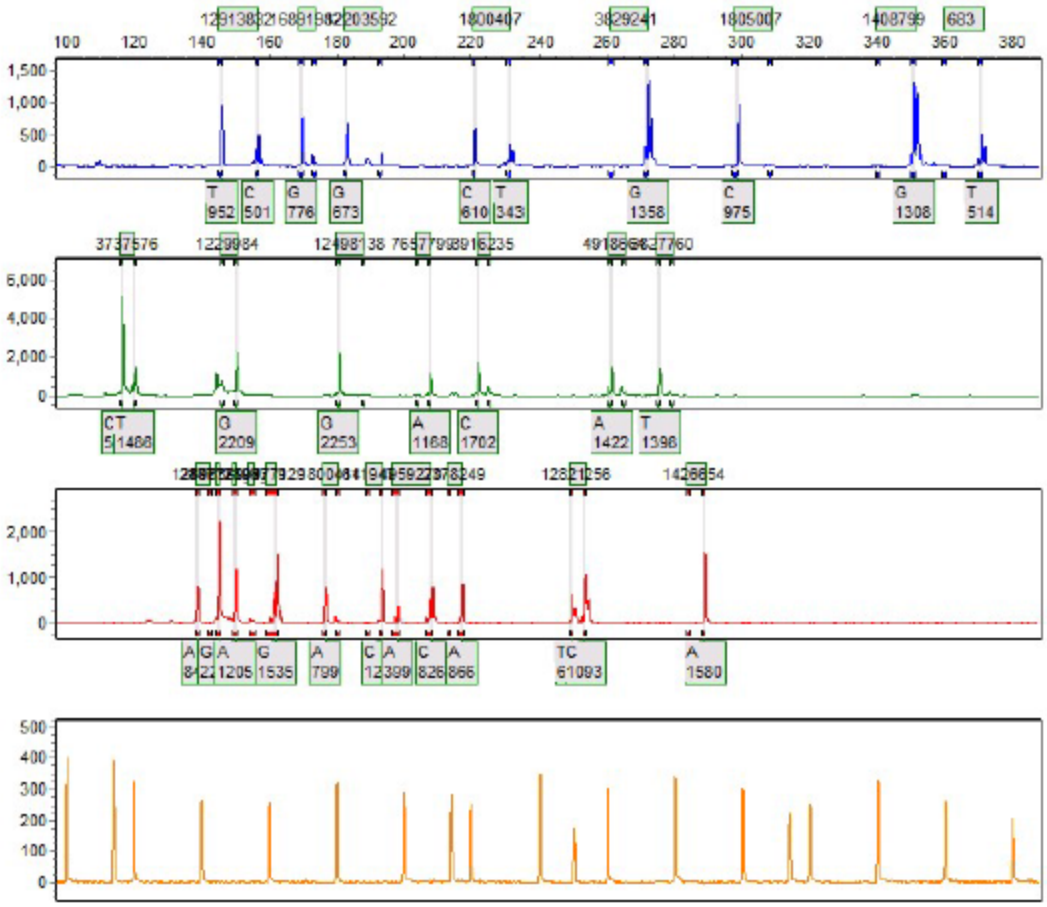


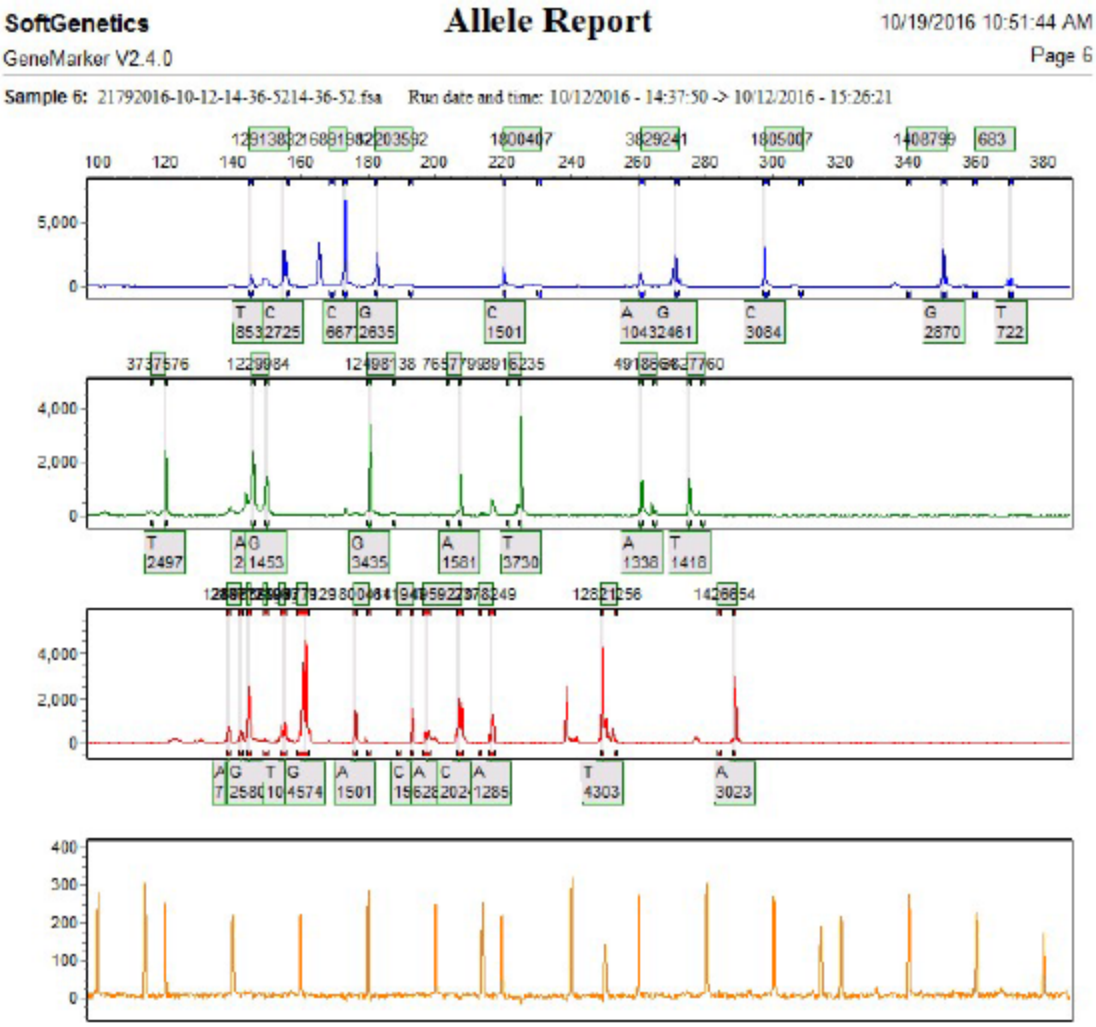

SoftGenetics
Allele Report
10/19/2016 10:51:44 AM
GeneMarker V2.4.0
Page 4
Sample 4: 21652016-09-12-13-43-1113-43-11.fsa Run date and time: 09/12/2016 - 14:32:51 -> 09/12/2016 - 15:10:36

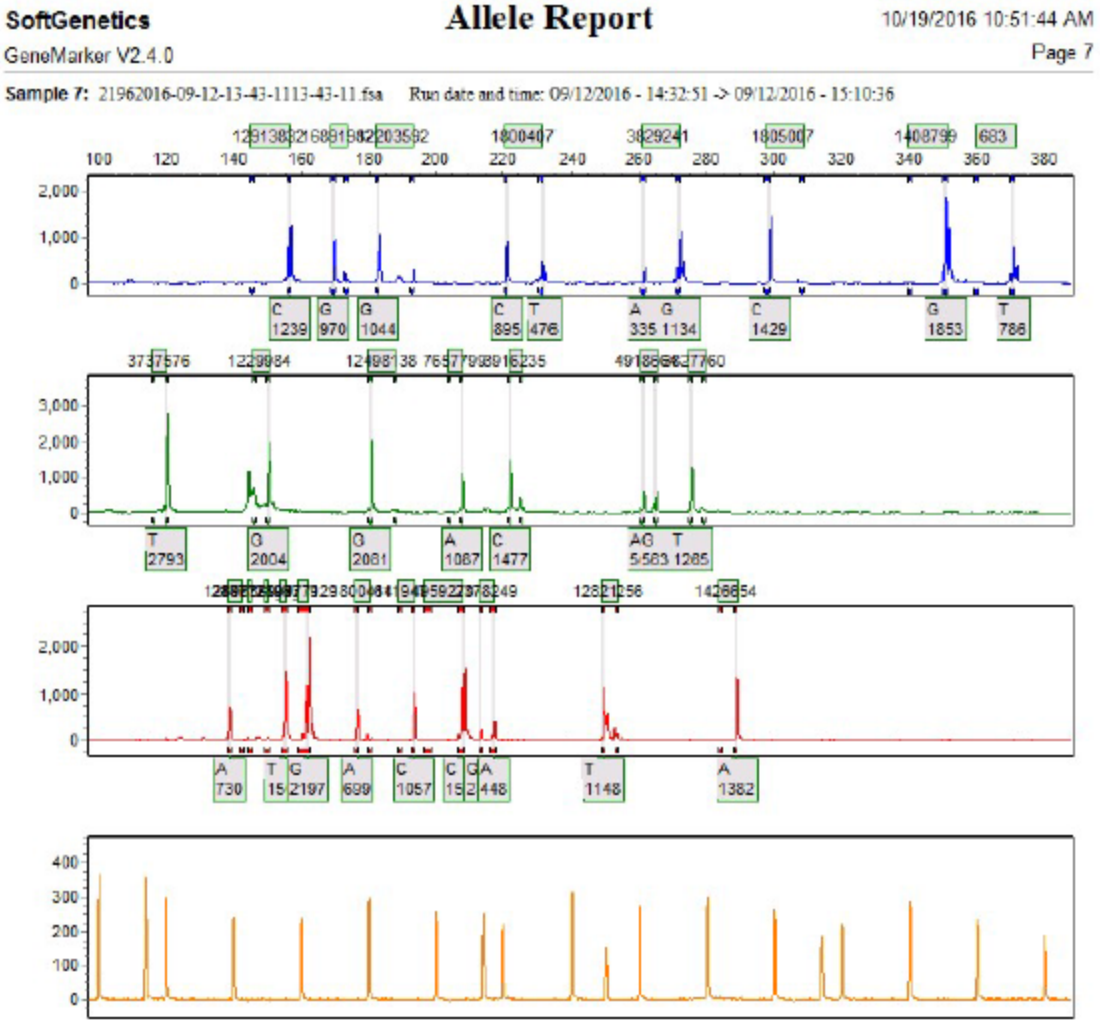
SoftGenetics **Allele Report** 10/19/2016 10:51:44 AM

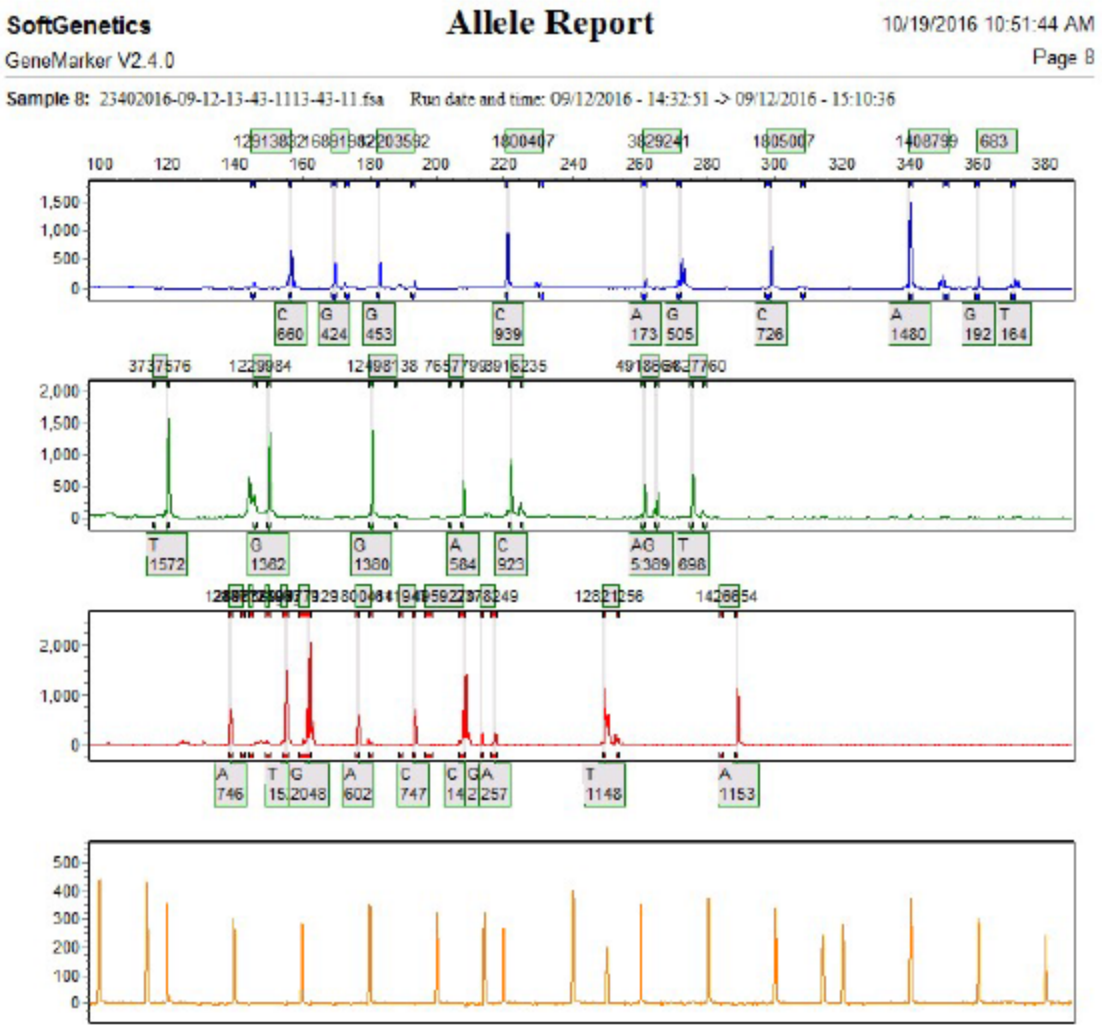
GeneMarker V2.4.0 Page 5

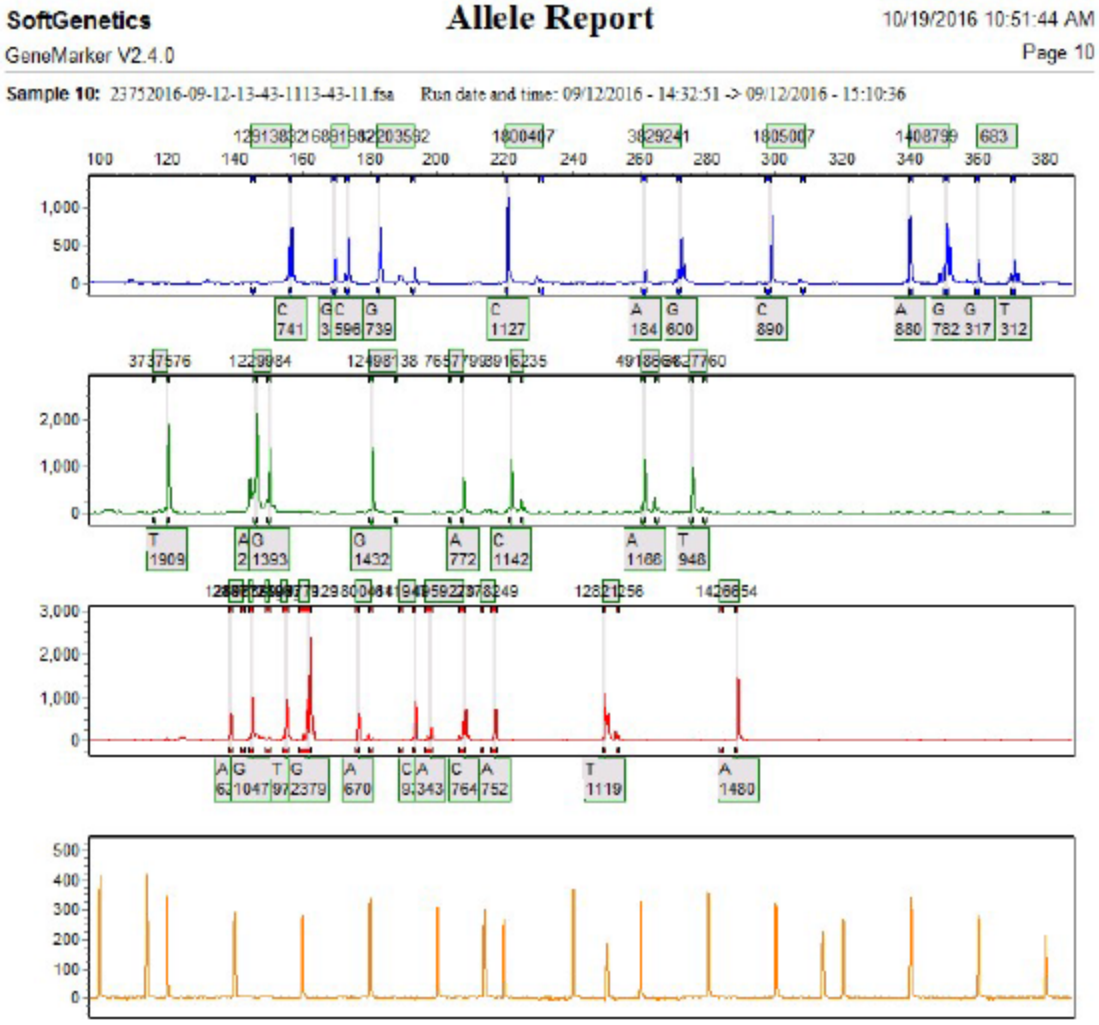
Sample 5: 21742016-09-12-13-43-1113-43-11.fsa Run date and time: 09/12/2016 - 14:32:51 -> 09/12/2016 - 15:10:36

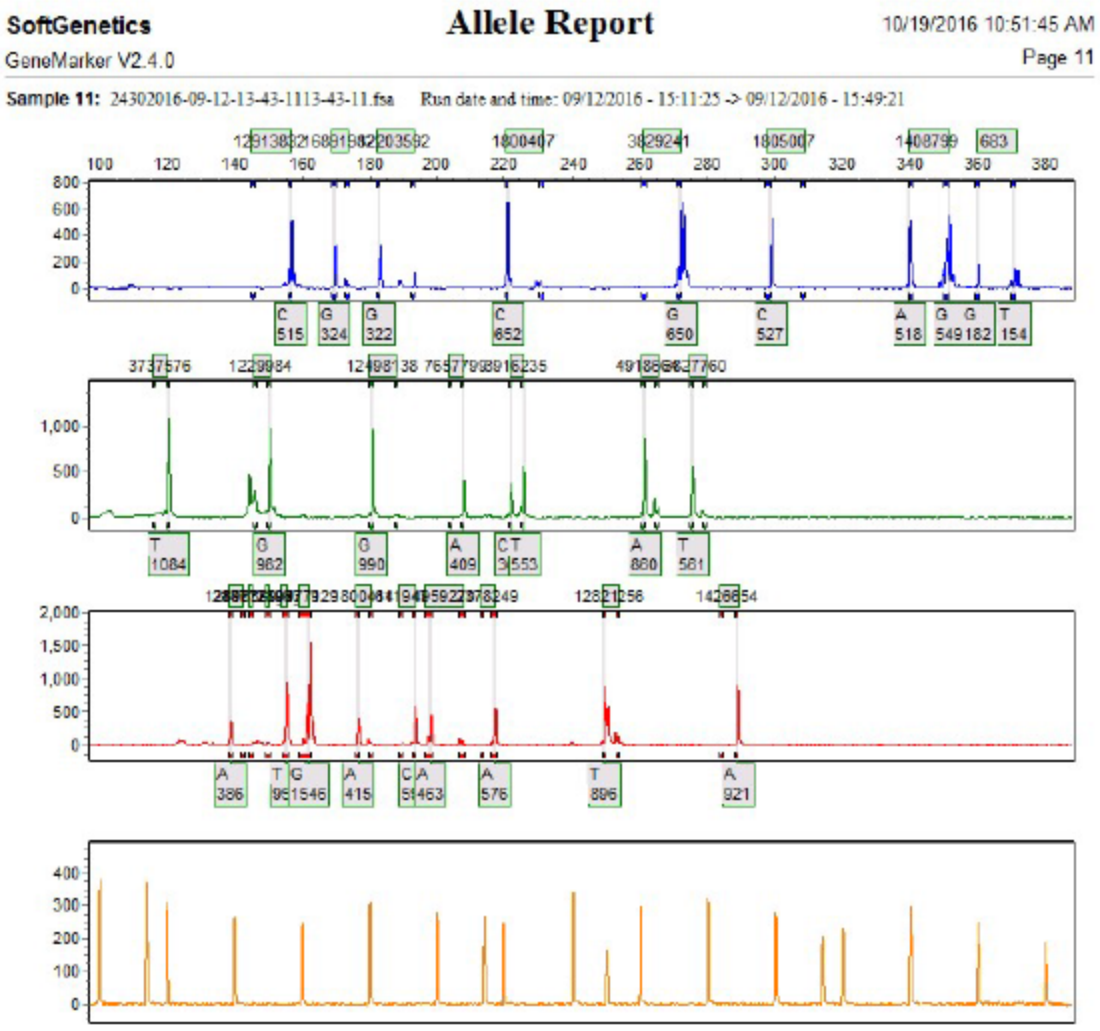



SoftGenetics
Allele Report
10/19/2016 10:51:44 AM
GeneMarker V2.4.0
Page 6
Sample 6: 21792016-10-12-14-36-5214-36-52.fsa Run date and time: 10/12/2016 - 14:37:50 -> 10/12/2016 - 15:26:21


SoftGenetics
Allele Report
10/19/2016 10:51:44 AM
GeneMarker V2.4.0
Page 7
Sample 7: 21962016-09-12-13-43-1113-43-11.fsa Run date and time: 09/12/2016 - 14:32:51 -> 09/12/2016 - 15:10:36


SoftGenetics
Allele Report
10/19/2016 10:51:44 AM
GeneMarker V2.4.0
Page 8
Sample 8: 23402016-09-12-13-43-1113-43-11.fsa Run date and time: 09/12/2016 - 14:32:51 -> 09/12/2016 - 15:10:36

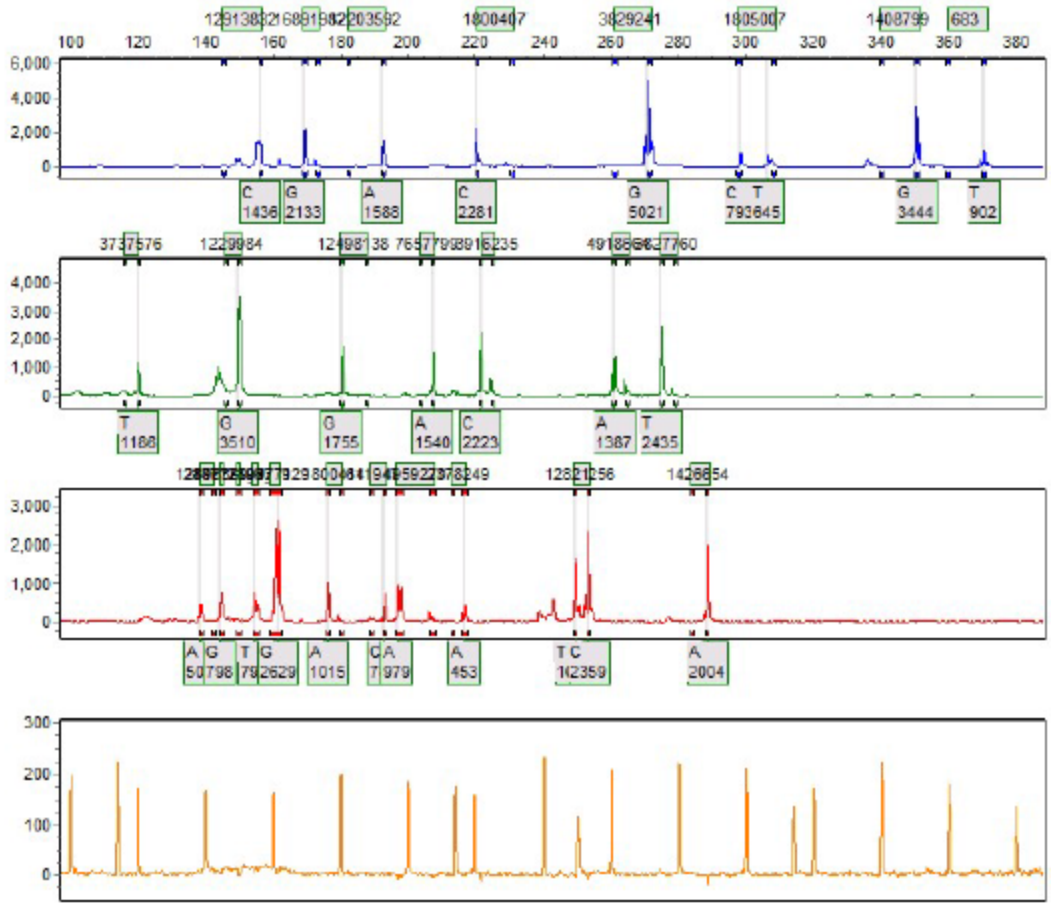

SoftGenetics
Allele Report
10/19/2016 10:51:44 AM
GeneMarker V2.4.0
Page 10
Sample 10: 23752016-09-12-13-43-1113-43-11.fsa Run date and time: 09/12/2016 - 14:32:51 -> 09/12/2016 - 15:10:36

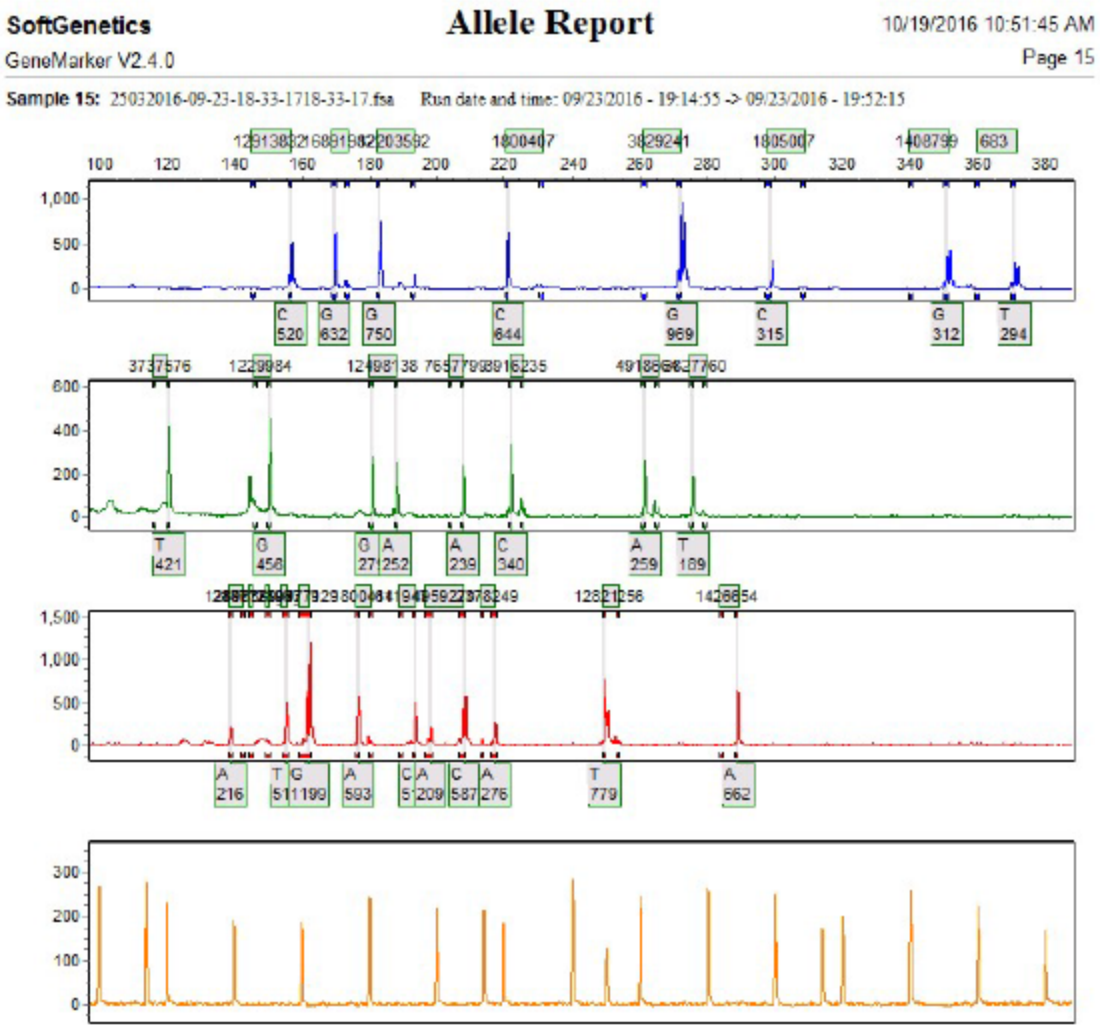

SoftGenetics
Allele Report
10/19/2016 10:51:45 AM
GeneMarker V2.4.0
Page 11
Sample 11: 24302016-09-12-13-43-1113-43-11.fsa Run date and time: 09/12/2016 - 15:11:25 -> 09/12/2016 - 15:49:21

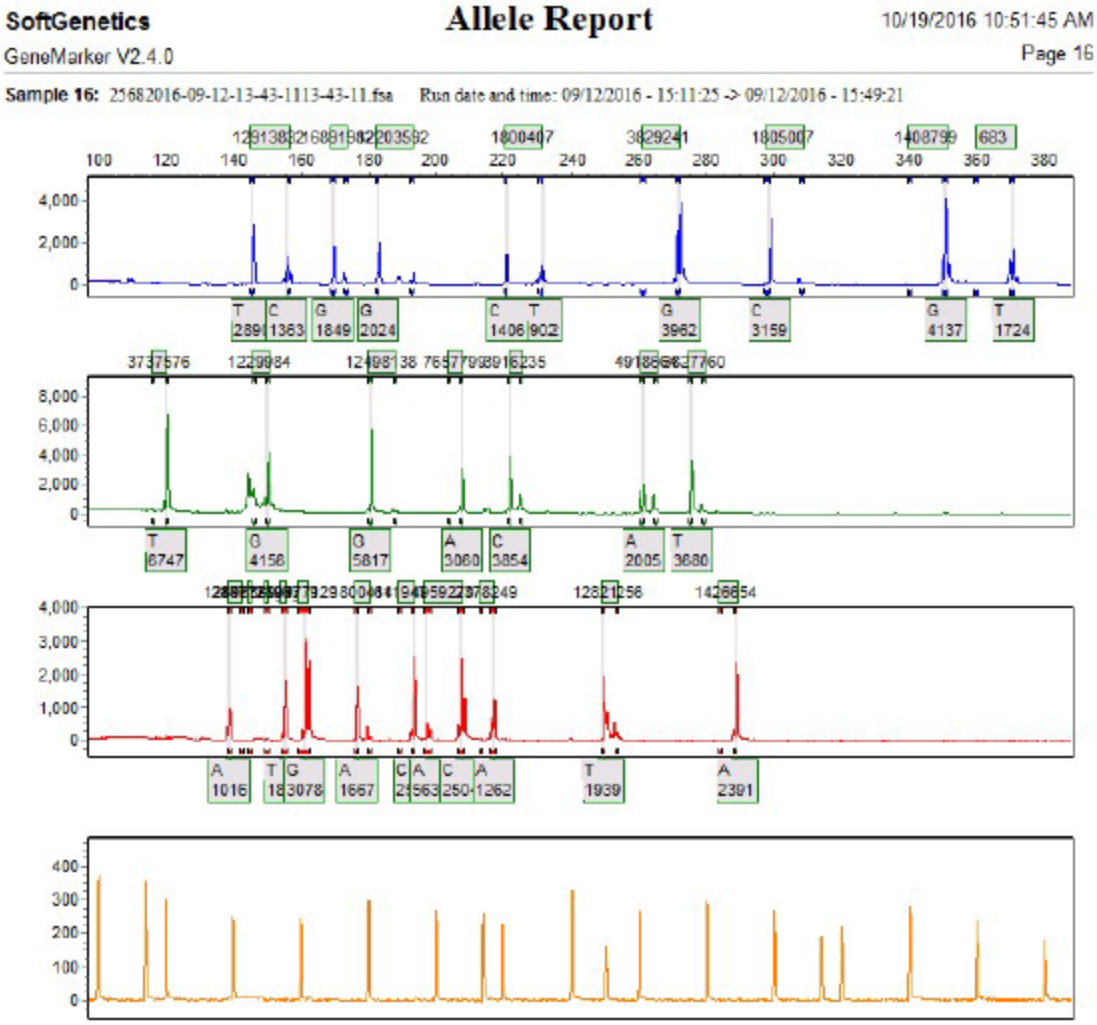
SoftGenetics **Allele Report** 10/19/2016 10:51:45 AM

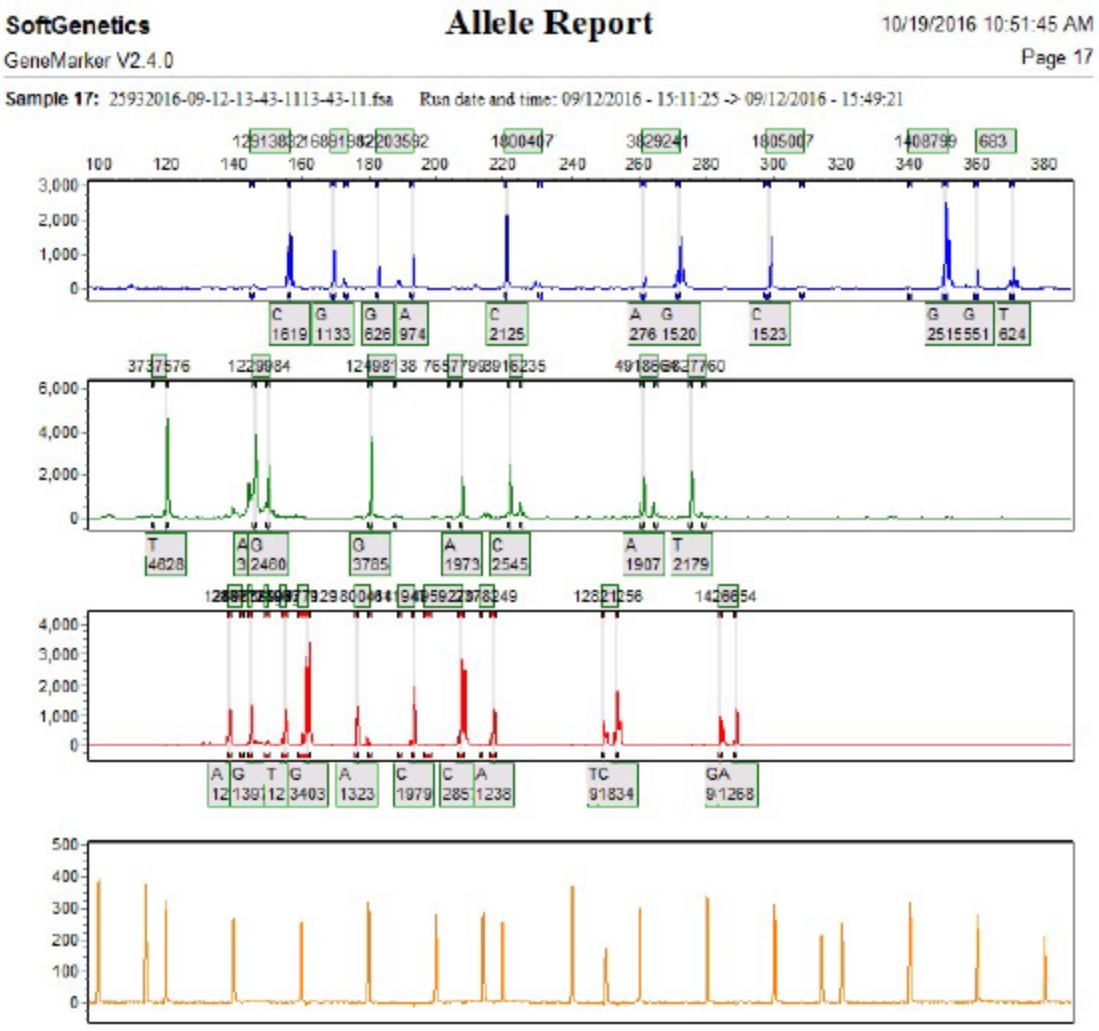
GeneMarker V2.4.0 Page 12

Sample 12: 24352016-10-13-11-00-3111-00-31.fsa Run date and time: 10/13/2016 - 11:01:15 -> 10/13/2016 - 11:50:15




SoftGenetics
Allele Report
10/19/2016 10:51:45 AM
GeneMarker V2.4.0
Page 15
Sample 15: 25032016-09-23-18-33-1718-33-17.fsa Run date and time: 09/23/2016 - 19:14:55 -> 09/23/2016 - 19:52:15

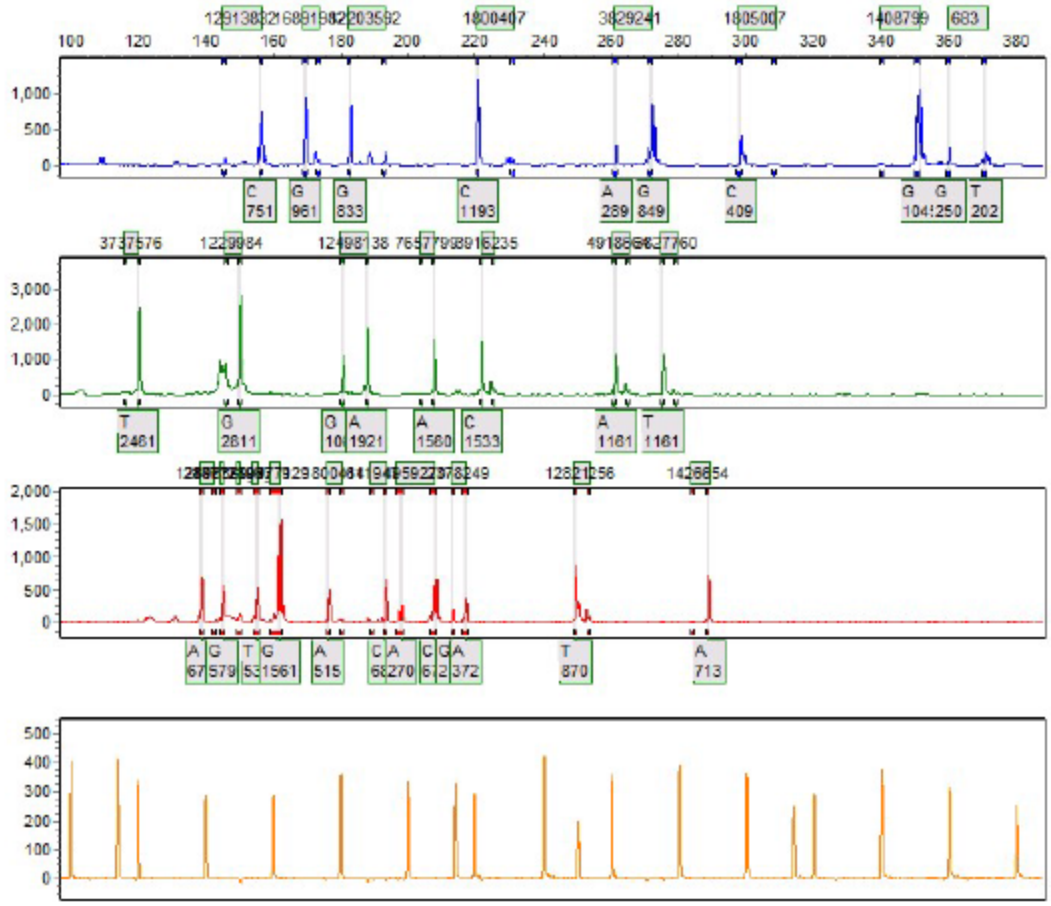

SoftGenetics
GeneMarker V2.4.0
Allele Report
10/19/2016 10:51:45 AM
Page 16
Sample 16: 25682016-09-12-13-43-1113-43-11.fsa Run date and time: 09/12/2016 - 15:11:25 -> 09/12/2016 - 15:49:21


SoftGenetics
Allele Report
10/19/2016 10:51:45 AM
GeneMarker V2.4.0
Page 17
Sample 17: 25932016-09-12-13-43-1113-43-11.fsa Run date and time: 09/12/2016 - 15:11:25 -> 09/12/2016 - 15:49:21

SoftGenetics **Allele Report** 10/19/2016 10:51:45 AM

GeneMarker V2.4.0 Page 18

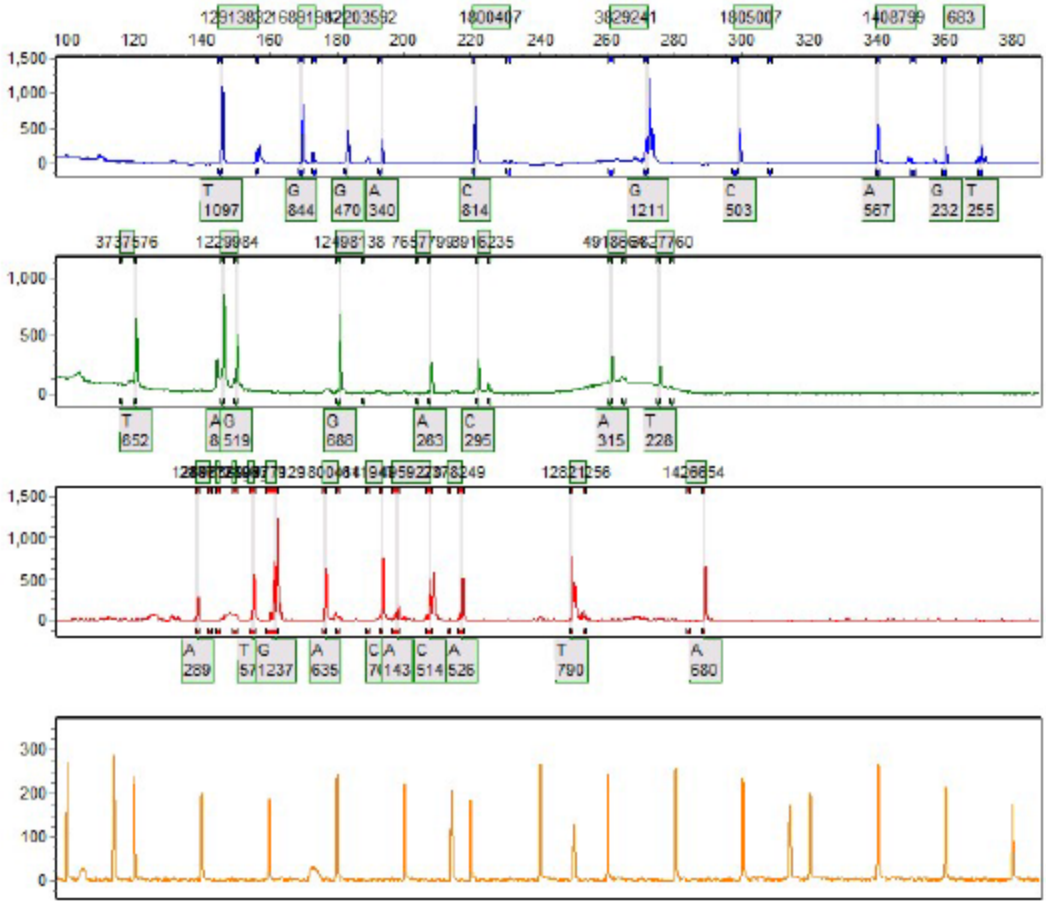
Sample 18: 26592016-09-21-17-39-3617-39-36.fsa Run date and time: 09/22/2016 - 00:53:59 -> 09/22/2016 - 01:32:04



SoftGenetics **Allele Report** 10/19/2016 10:51:45 AM

GeneMarker V2.4.0 Page 19

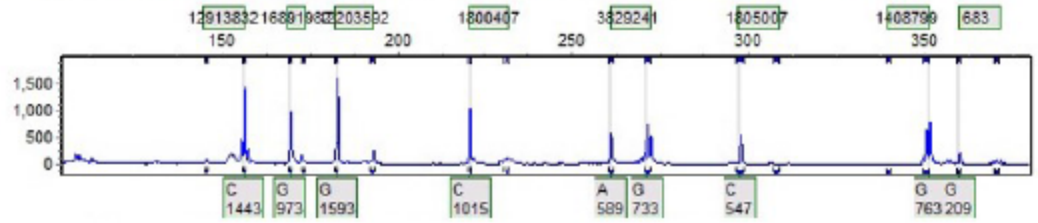
Sample 19: 29102016-09-23-18-33-1718-33-17.fsa Run date and time: 09/23/2016 - 19:14:55 -> 09/23/2016 - 19:52:15



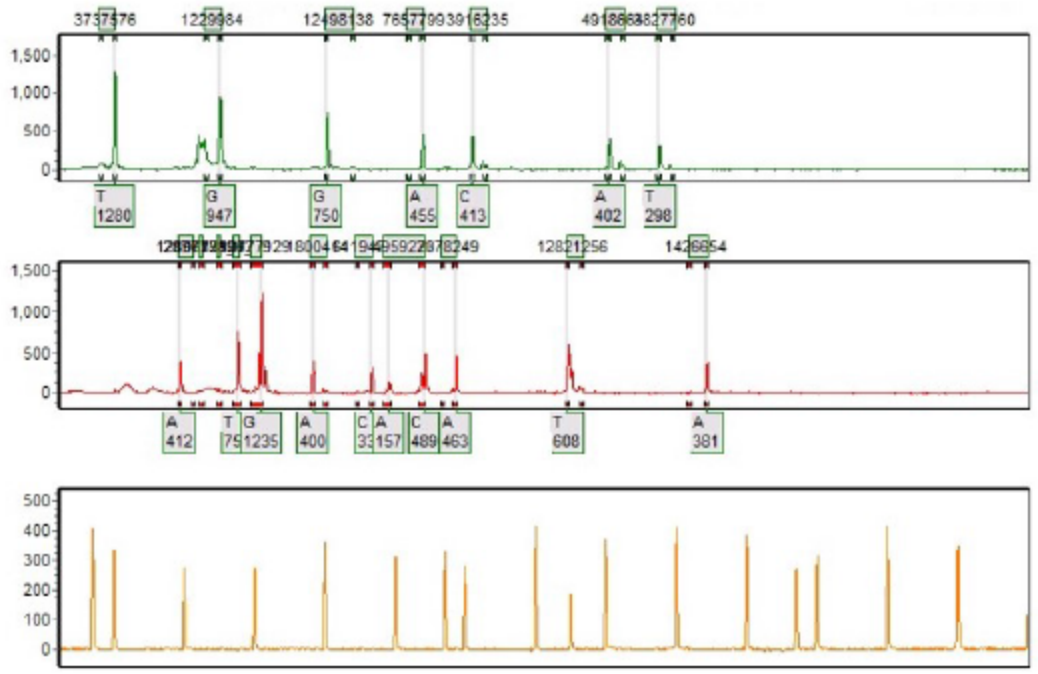
SoftGenetics **Allele Report** 10/19/2016 11:49:32 AM

GeneMarker V2.4.0 Page 4

Sample 4: 2971F2016-10-07-07-32-3307-32-33.fsa Run date and time: 10/07/2016 - 07:33:16 -> 10/07/2016 - 08:21:37



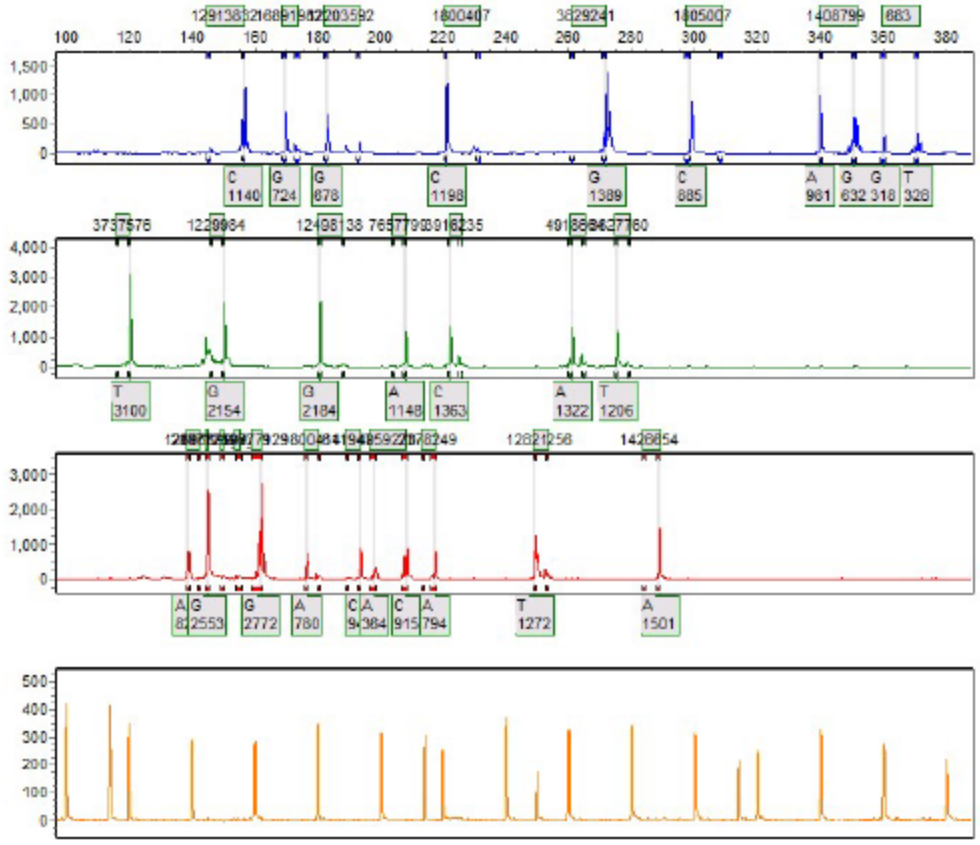
Sample 3: 29712016-09-22-16-13-3916-13-39.fsa Run date and time: 09/22/2016 - 18:20:20 -> 09/22/2016 - 18:58:21



SoftGenetics **Allele Report** 10/19/2016 10:51:45 AM

GeneMarker V2.4.0 Page 22

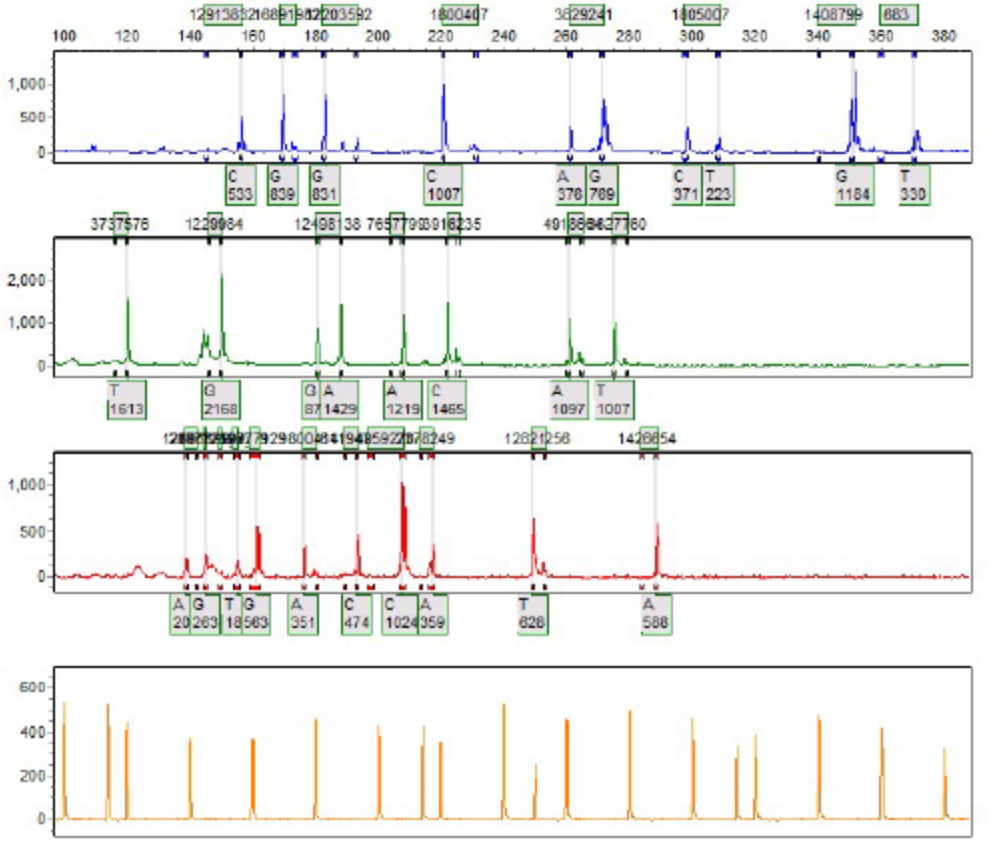
Sample 22: 29732016-09-12-13-43-1113-43-11.fsa Run date and time: 09/12/2016 - 15:11:25 -> 09/12/2016 - 15:49:21

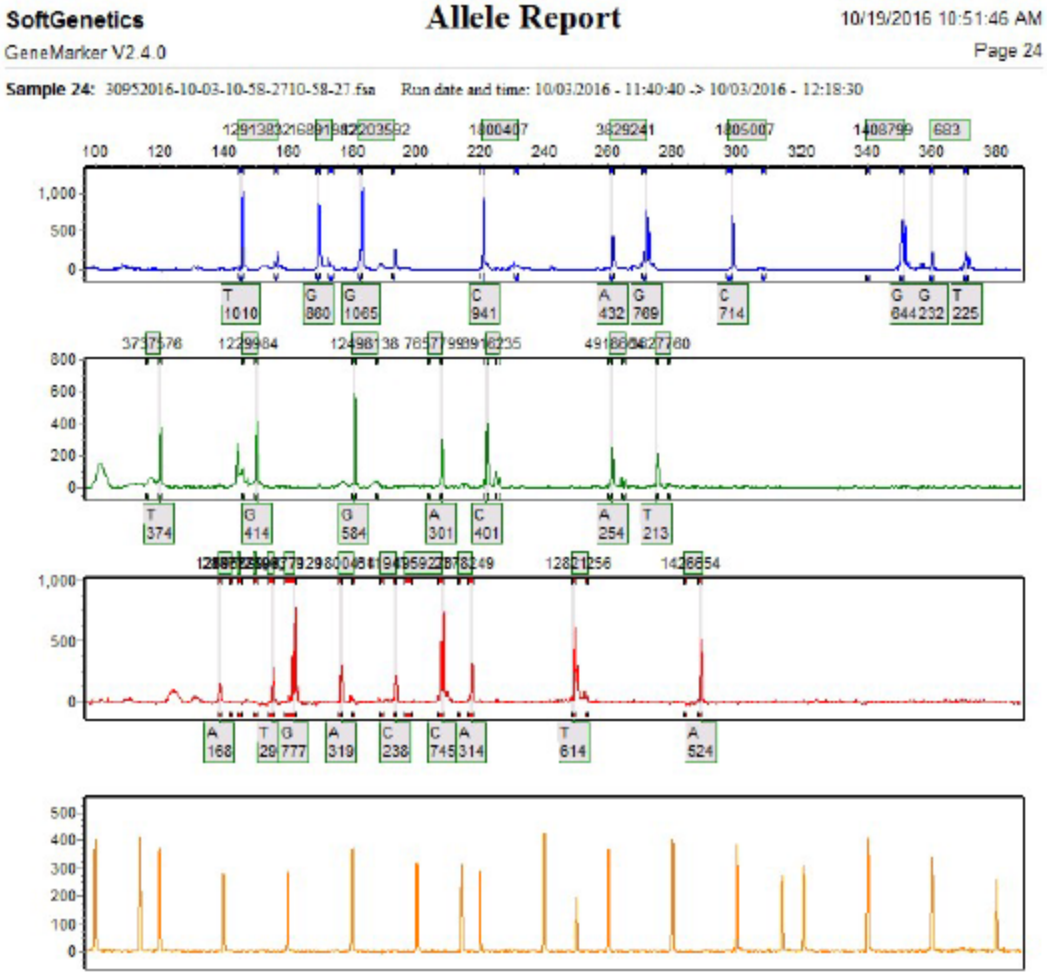


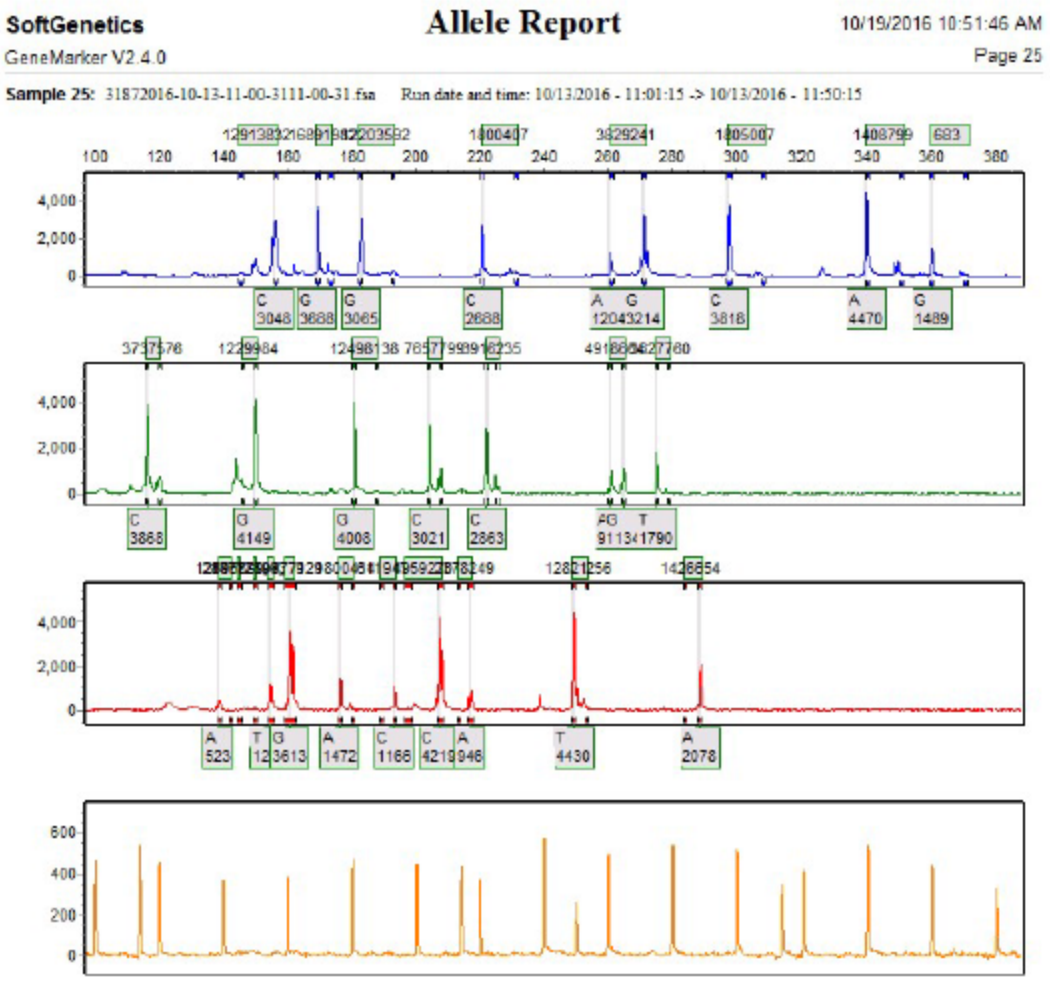
SoftGenetics **Allele Report** 10/19/2016 10:51:45 AM

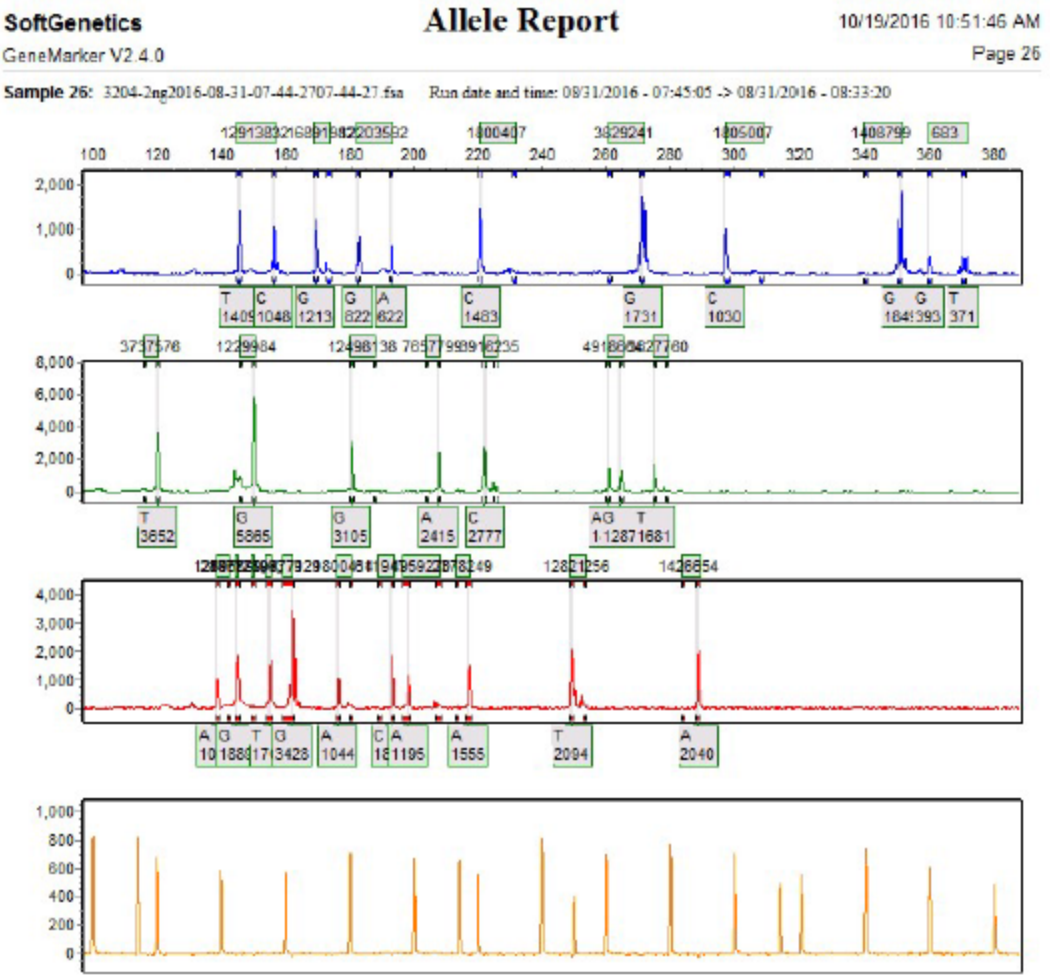
GeneMarker V2.4.0 Page 23

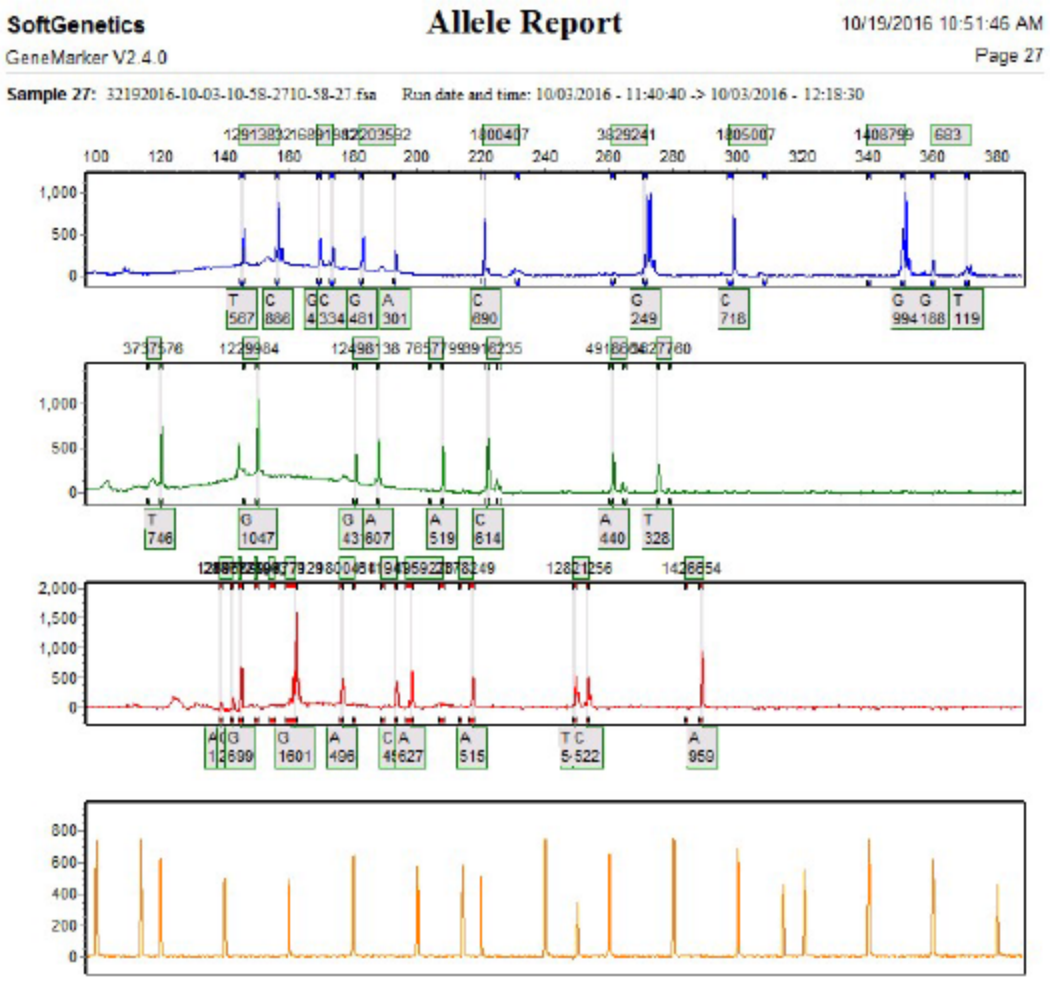
Sample 23: 30572016-09-21-17-39-3617-39-36.fsa Run date and time: 09/22/2016 - 00:15:08 -> 09/22/2016 - 00:53:08

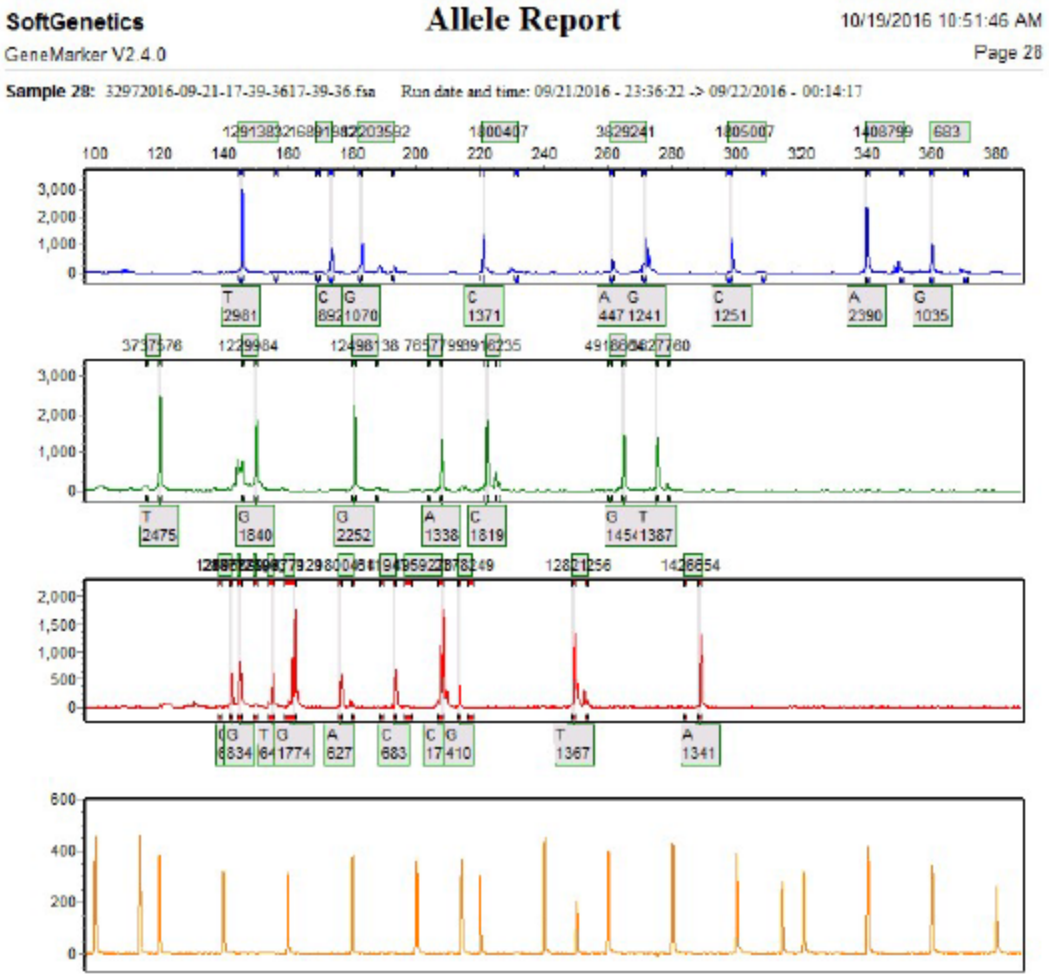



SoftGenetics
Allele Report
10/19/2016 10:51:46 AM
GeneMarker V2.4.0
Page 24
Sample 24: 30952016-10-03-10-58-2710-58-27.fsa Run date and time: 10/03/2016 - 11:40:40 -> 10/03/2016 - 12:18:30


SoftGenetics
Allele Report
10/19/2016 10:51:46 AM
GeneMarker V2.4.0
Page 25
Sample 25: 31872016-10-13-11-00-3111-00-31.fsa Run date and time: 10/13/2016 - 11:01:15 -> 10/13/2016 - 11:50:15

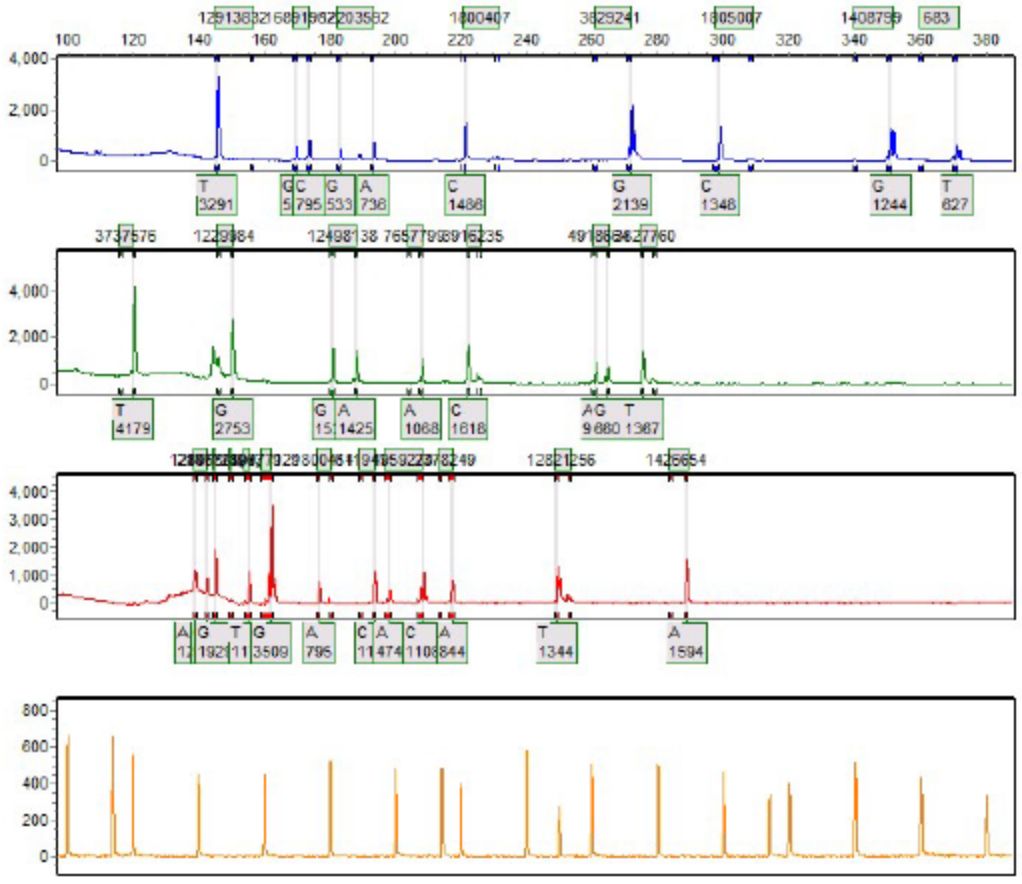

SoftGenetics
Allele Report
10/19/2016 10:51:46 AM
GeneMarker V2.4.0
Page 26
Sample 26: 3204-2ng2016-08-31-07-44-2707-44-27.fsa Run date and time: 08/31/2016 - 07:45:05 -> 08/31/2016 - 08:33:20


SoftGenetics
Allele Report
10/19/2016 10:51:46 AM
GeneMarker V2.4.0
Page 27
Sample 27: 32192016-10-03-10-58-2710-58-27.fsa Run date and time: 10/03/2016 - 11:40:40 -> 10/03/2016 - 12:18:30


SoftGenetics
Allele Report
10/19/2016 10:51:46 AM
GeneMarker V2.4.0
Page 28
Sample 28: 32972016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 23:36:22 -> 09/22/2016 - 00:14:17

Allele Report

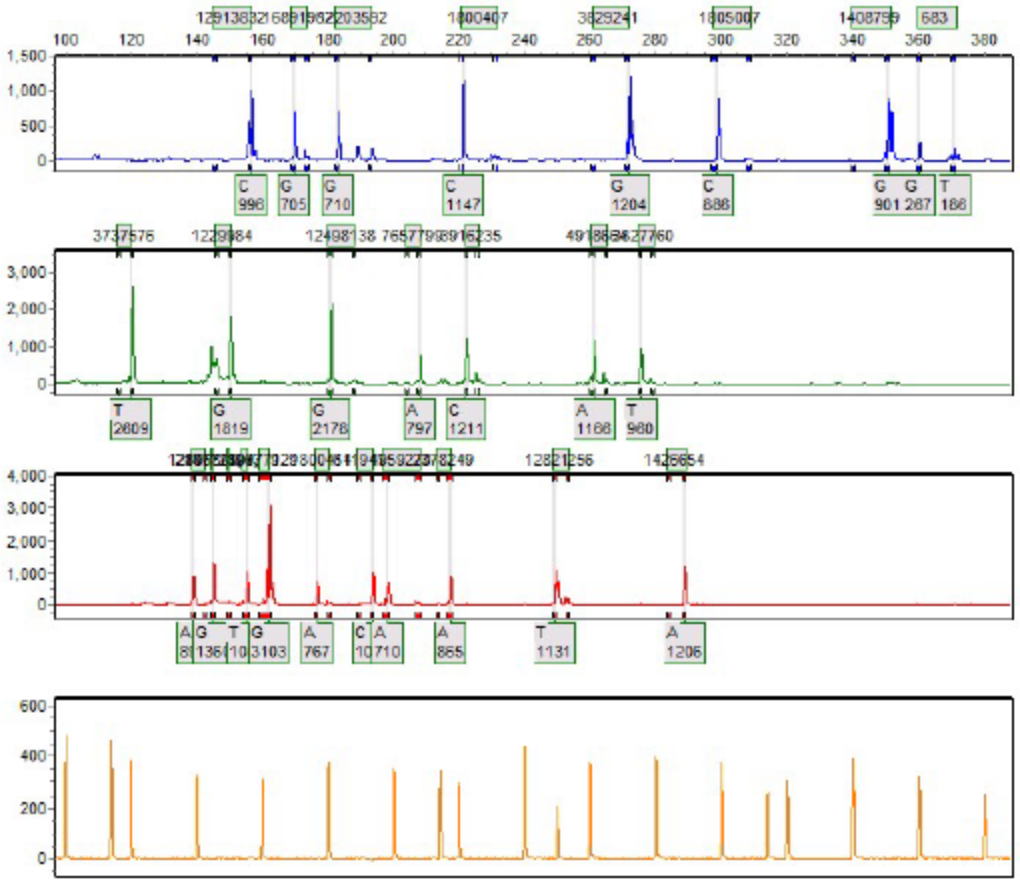
Sample 29: 34292016-09-12-13-43-1113-43-11.fsa Run date and time: 09/12/2016 - 15:50:10 -> 09/12/2016 - 16:28:05



SoftGenetics **Allele Report** 10/19/2016 10:51:46 AM

GeneMarker V2.4.0 Page 30

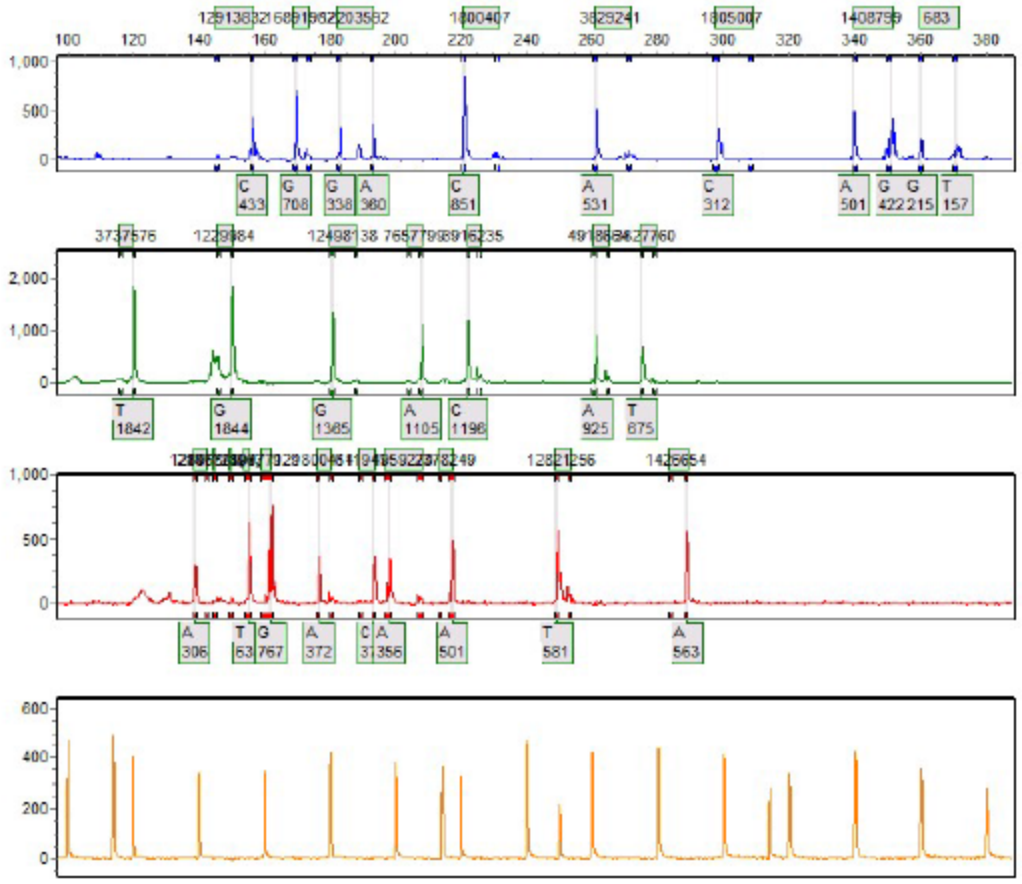
Sample 30: 34712016-09-12-13-43-1113-43-11.fsa Run date and time: 09/12/2016 - 15:50:10 -> 09/12/2016 - 16:28:05



SoftGenetics **Allele Report** 10/19/2016 10:51:46 AM

GeneMarker V2.4.0 Page 31

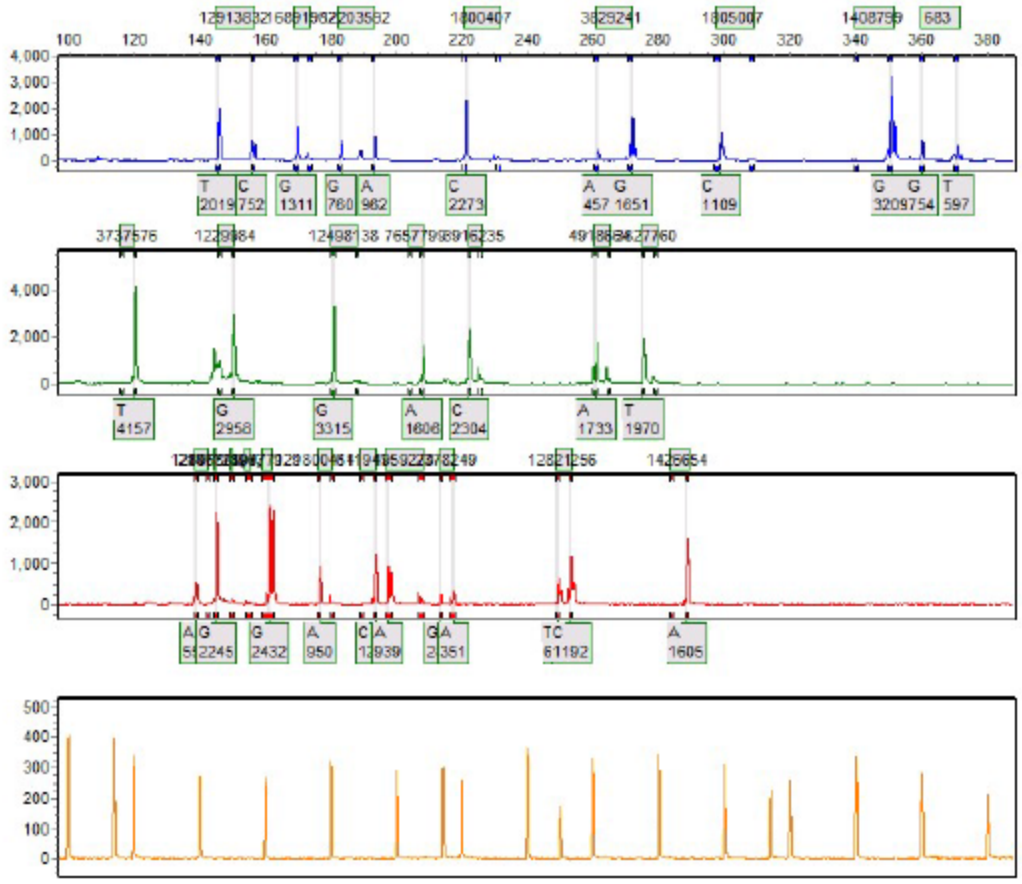
Sample 31: 35172016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 22:57:21 -> 09/21/2016 - 23:35:31



SoftGenetics **Allele Report** 10/19/2016 10:51:46 AM

GeneMarker V2.4.0 Page 32

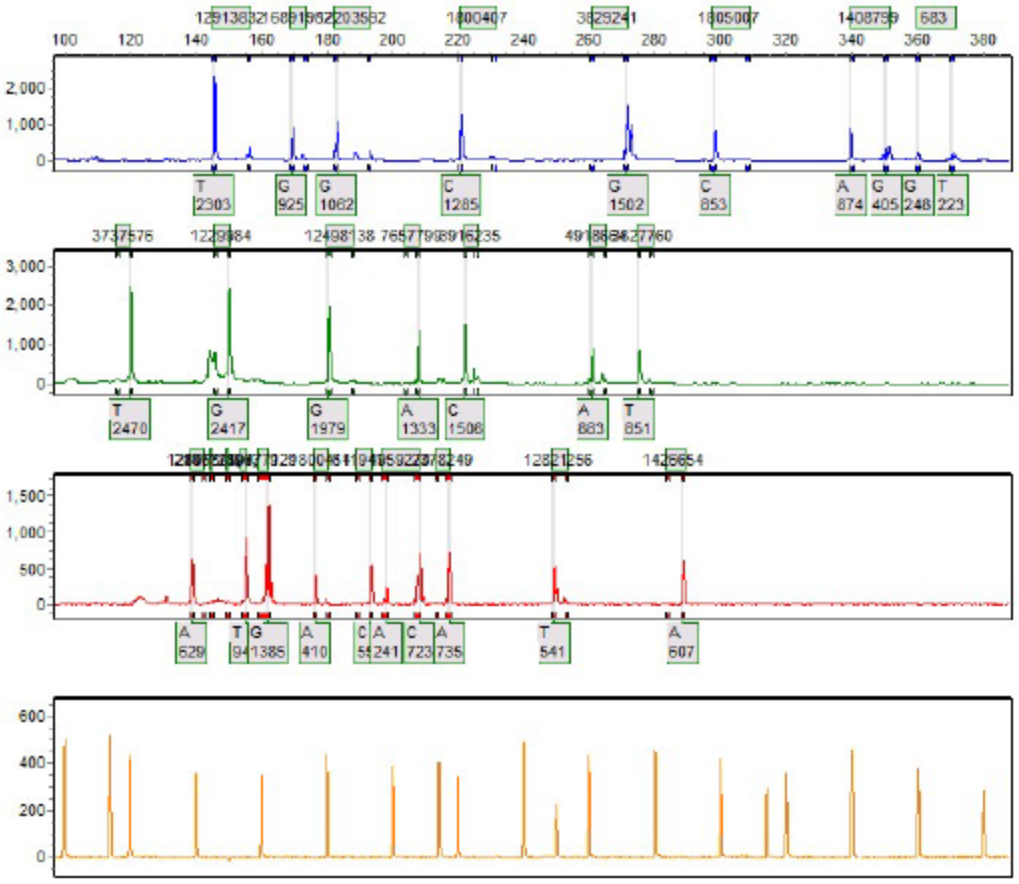
Sample 32: 35272016-09-12-13-43-1113-43-11.fsa Run date and time: 09/12/2016 - 15:50:10 -> 09/12/2016 - 16:28:05

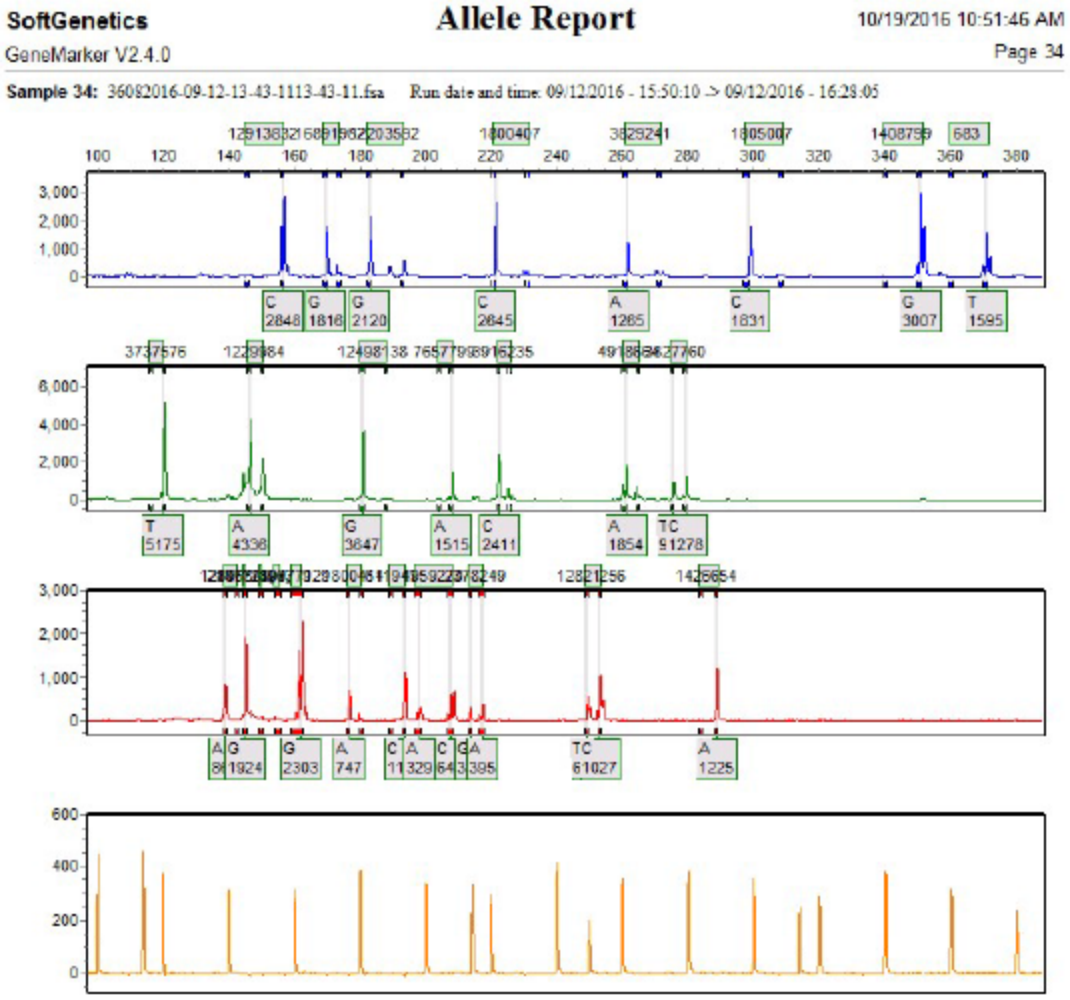


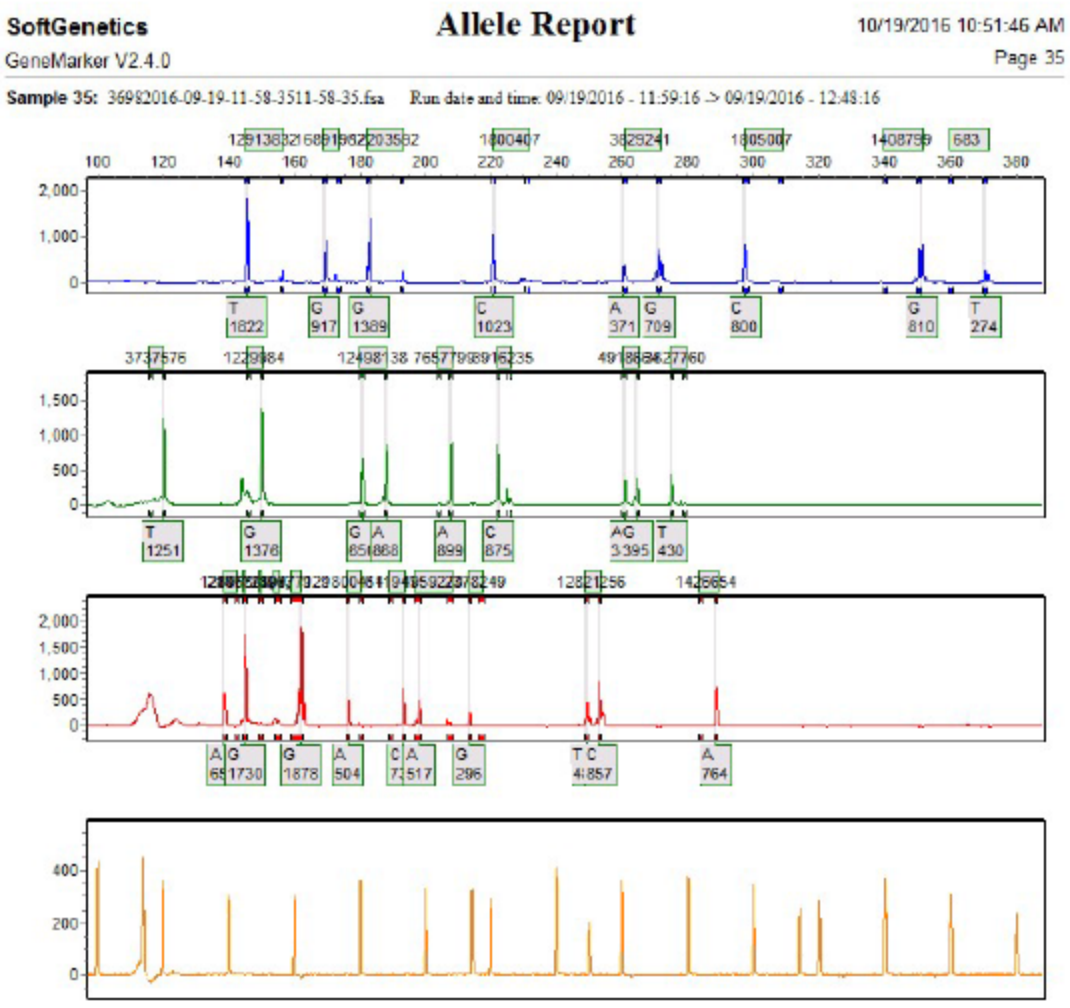
SoftGenetics **Allele Report** 10/19/2016 10:51:46 AM

GeneMarker V2.4.0 Page 33

Sample 33: 35422016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 23:36:22 -> 09/22/2016 - 00:14:17



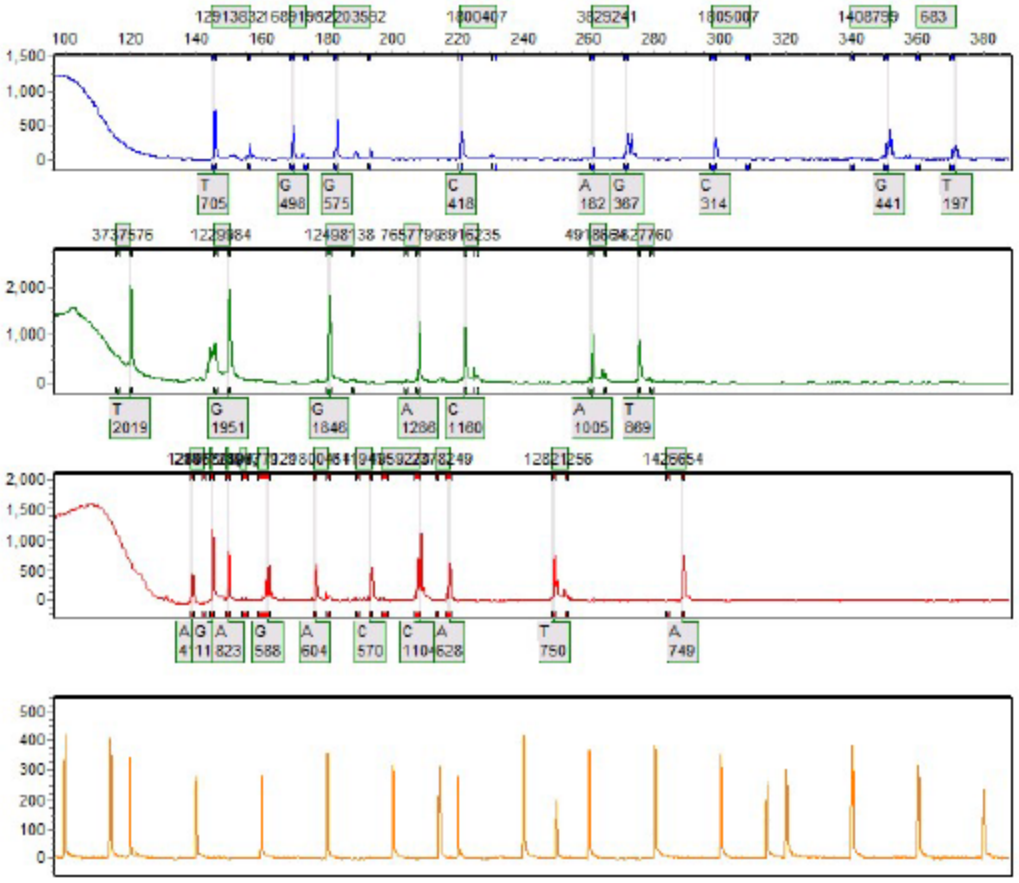

SoftGenetics
Allele Report
10/19/2016 10:51:46 AM
GeneMarker V2.4.0
Page 34
Sample 34: 36082016-09-12-13-43-1113-43-11.fsa Run date and time: 09/12/2016 - 15:50:10 -> 09/12/2016 - 16:28:05


SoftGenetics
Allele Report
10/19/2016 10:51:46 AM
GeneMarker V2.4.0
Page 35
Sample 35: 36982016-09-19-11-58-3511-58-35.fsa Run date and time: 09/19/2016 - 11:59:16 -> 09/19/2016 - 12:48:16

SoftGenetics **Allele Report** 10/19/2016 10:51:47 AM

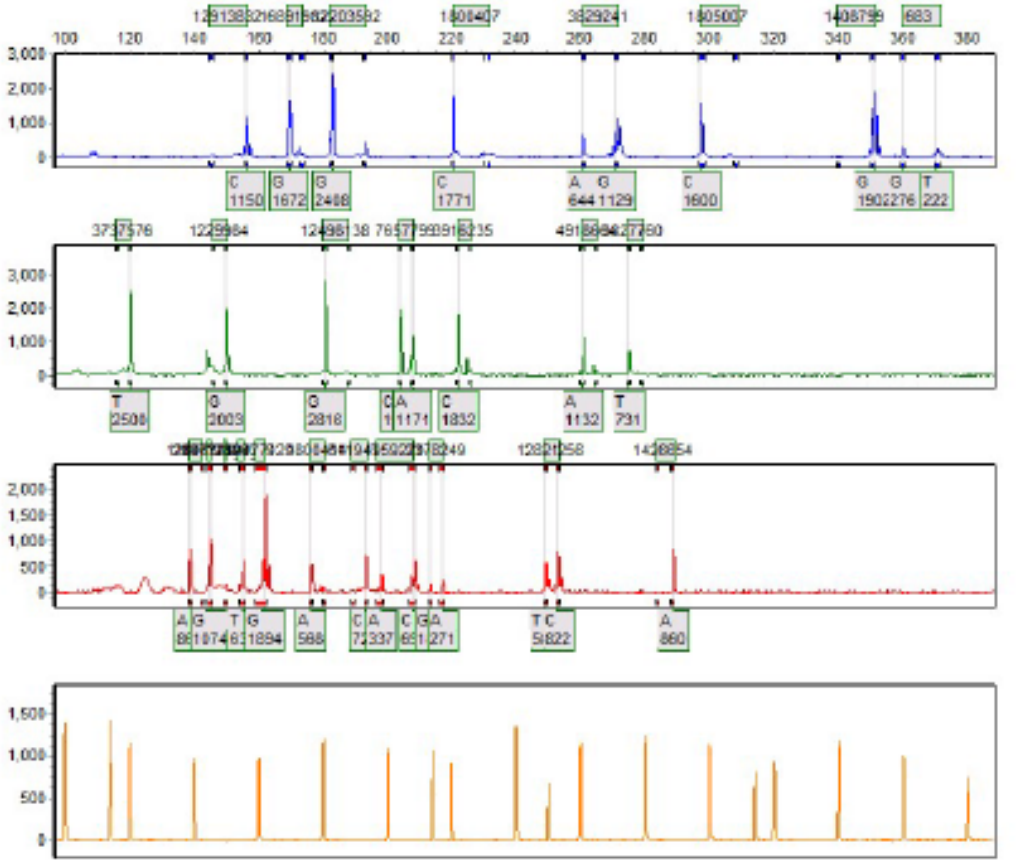
GeneMarker V2.4.0 Page 36

Sample 36: 37062016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 23:36:22 -> 09/22/2016 - 00:14:17

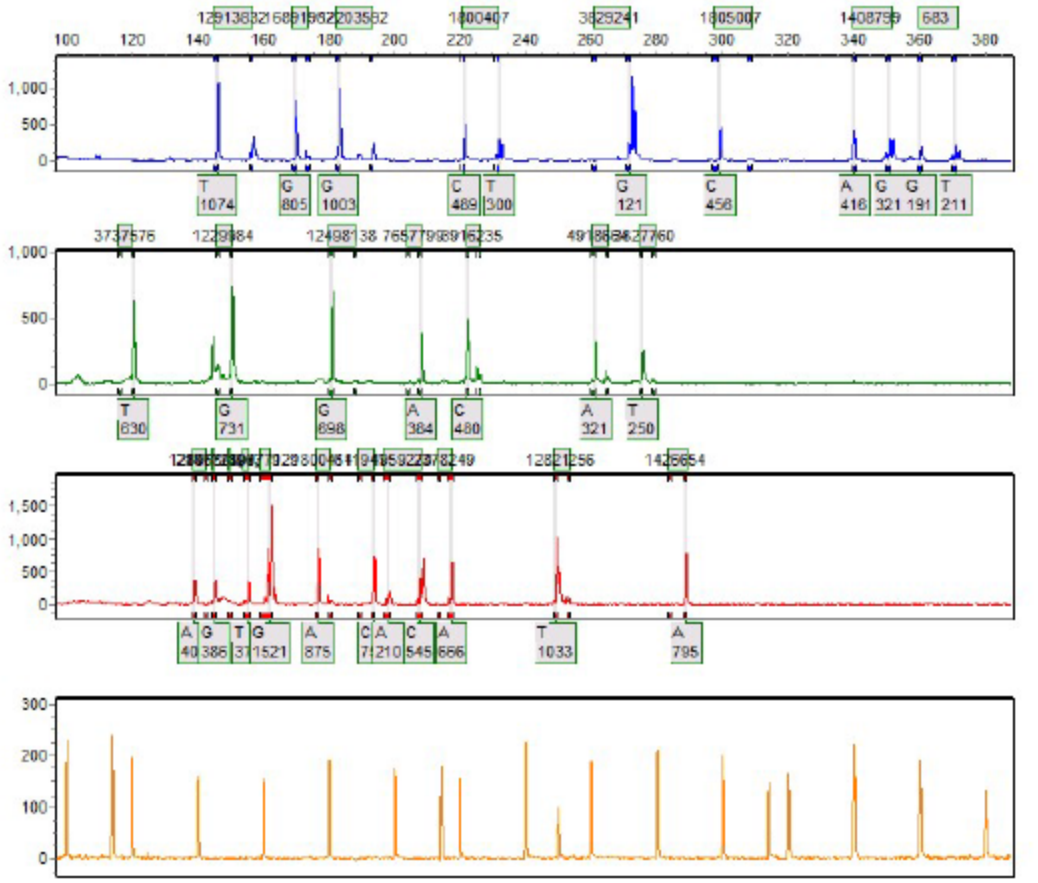


Allele Report

Sample 37: 38412016-09-19-11-38-3511-38-35.fsa Run date and time: 09/19/2016 - 11:59:16 -> 09/19/2016 - 12:48:16



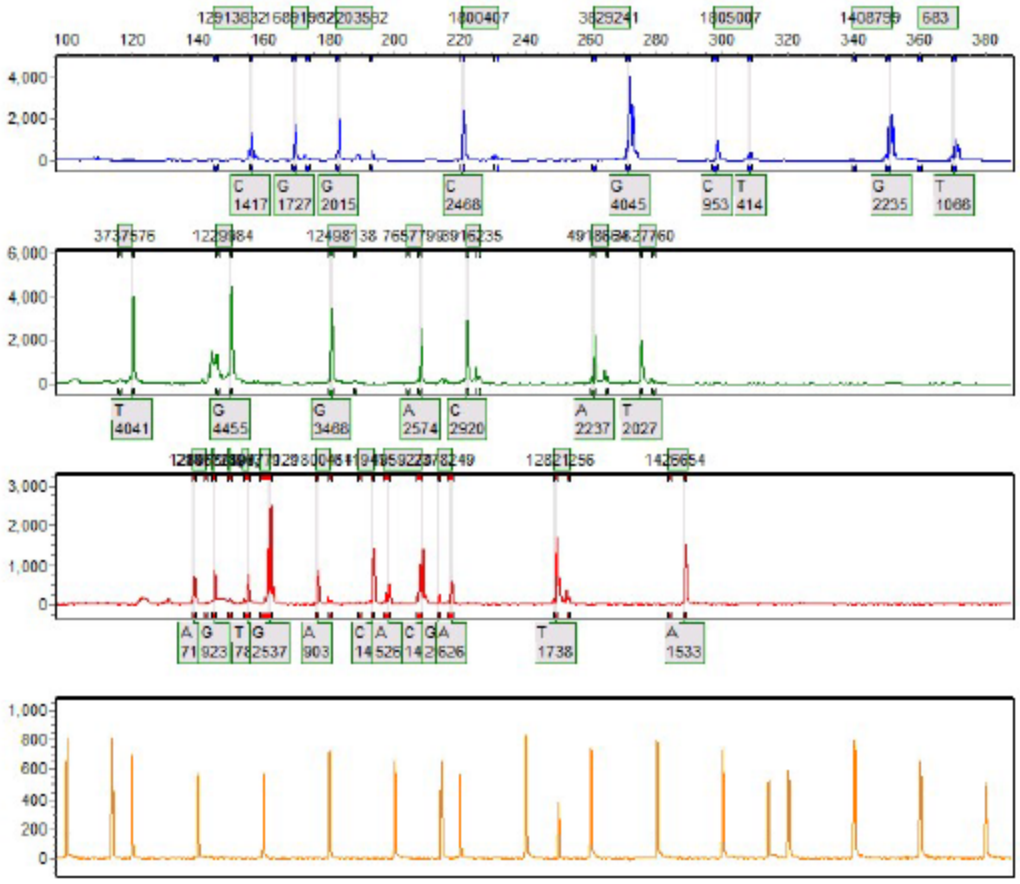
Sample 38: 38722016-09-23-18-33-1718-33-17.fsa Run date and time: 09/23/2016 - 19:14:55 -> 09/23/2016 - 19:52:15



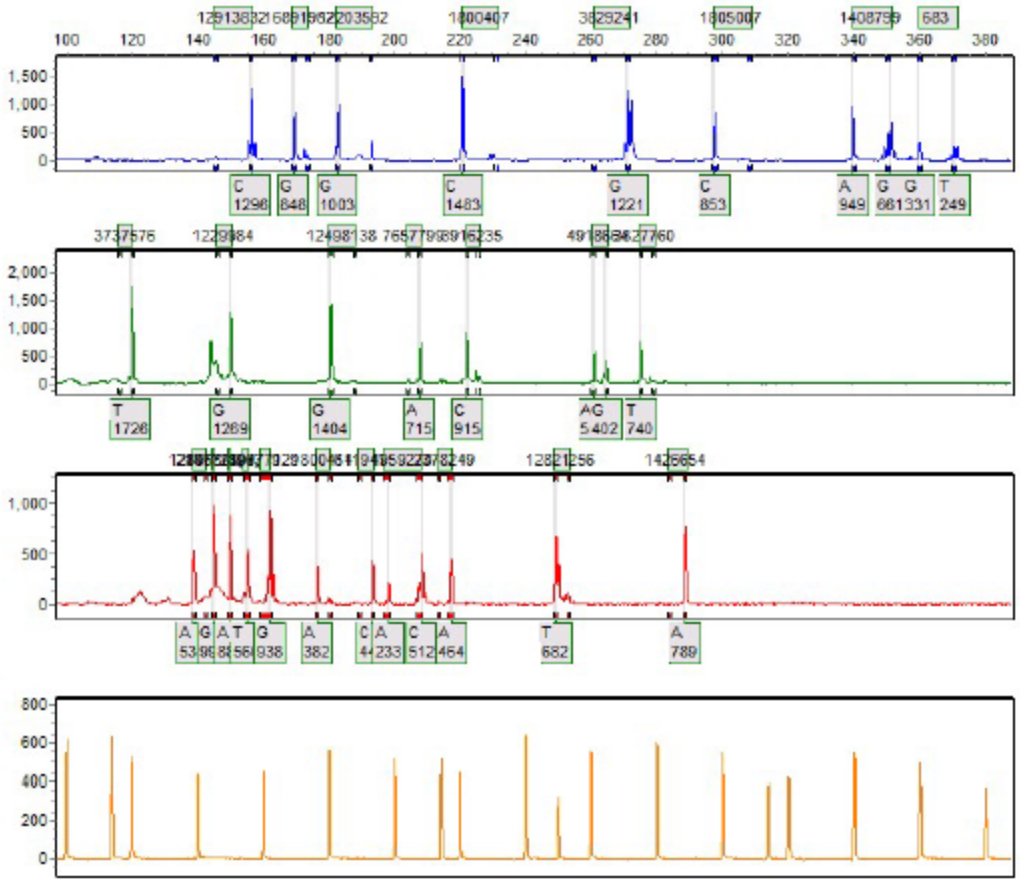
SoftGenetics **Allele Report** 10/19/2016 10:51:47 AM

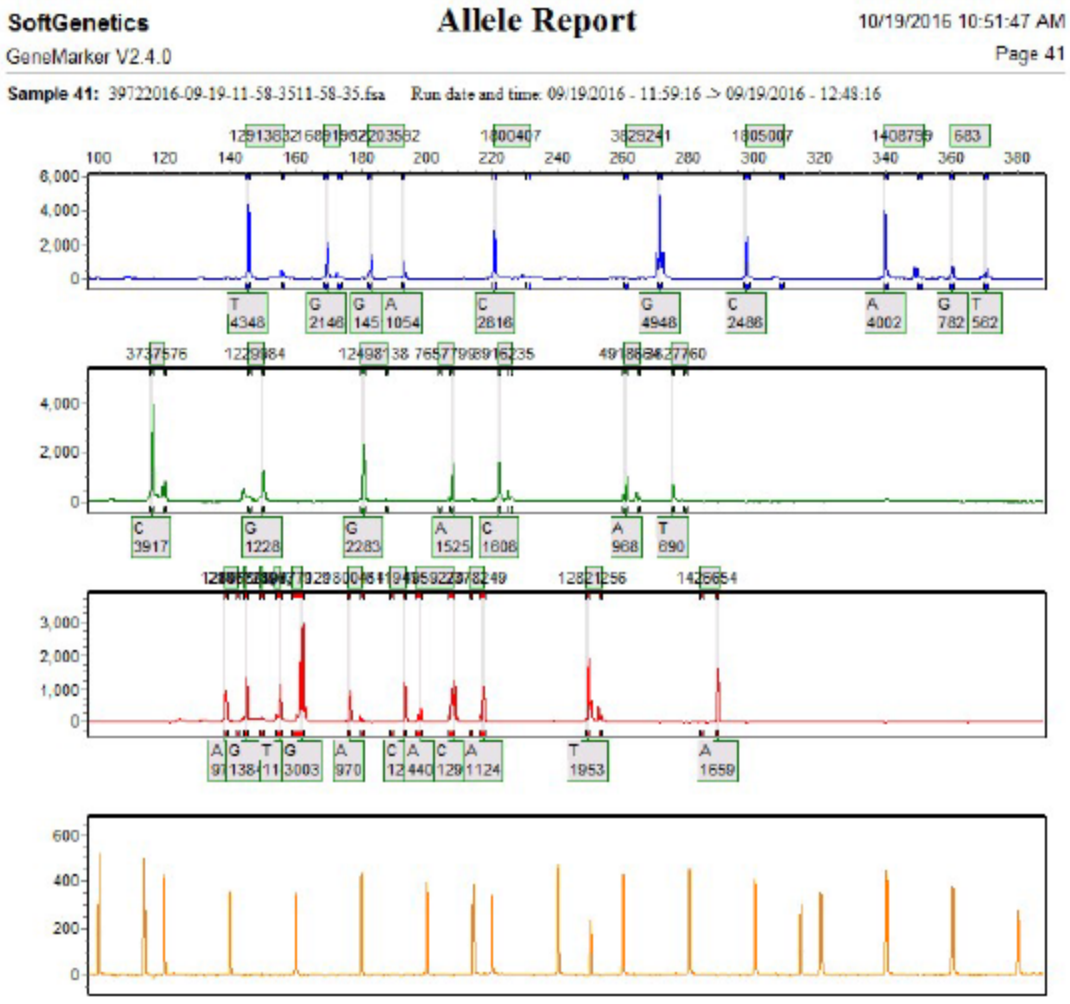
GeneMarker V2.4.0 Page 39

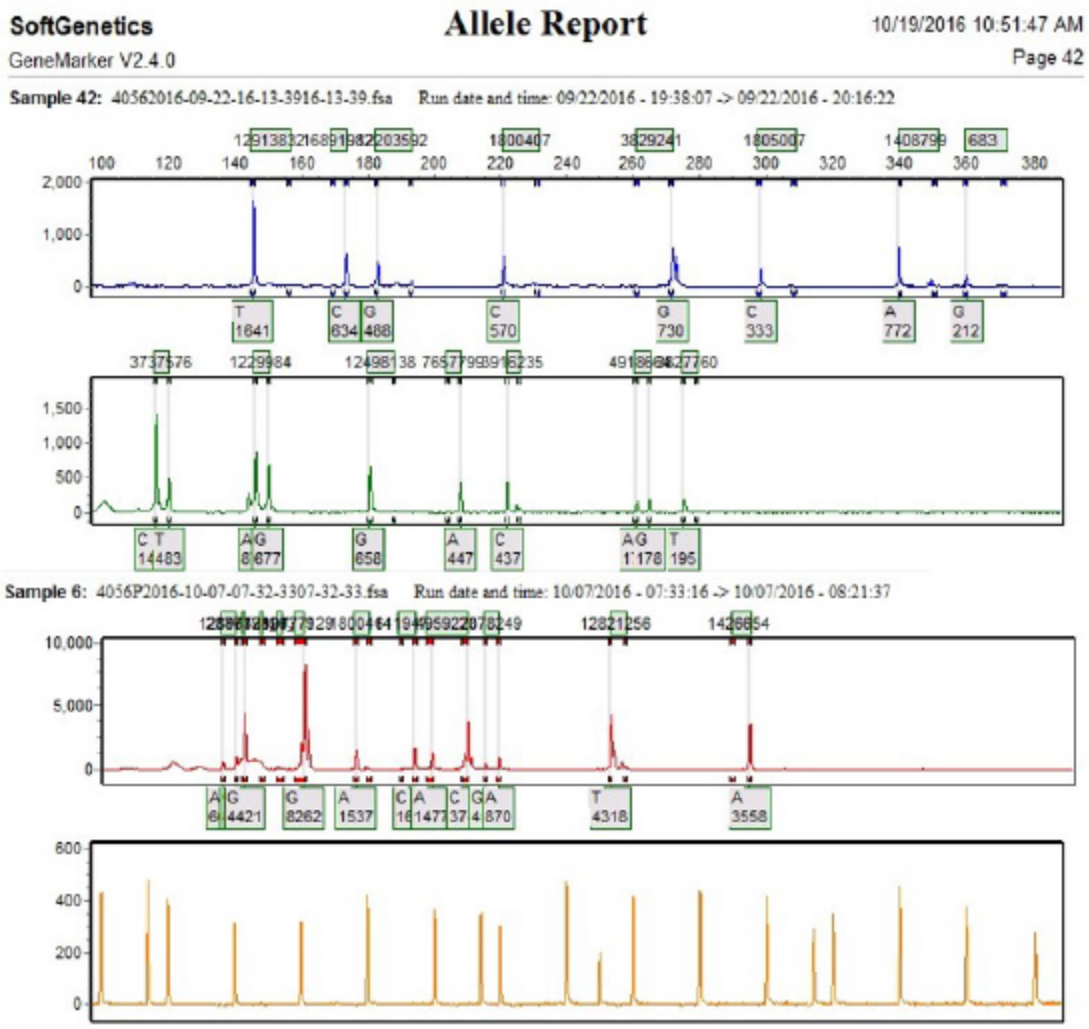
Sample 39: 38742016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 22:57:21 -> 09/21/2016 - 23:35:31

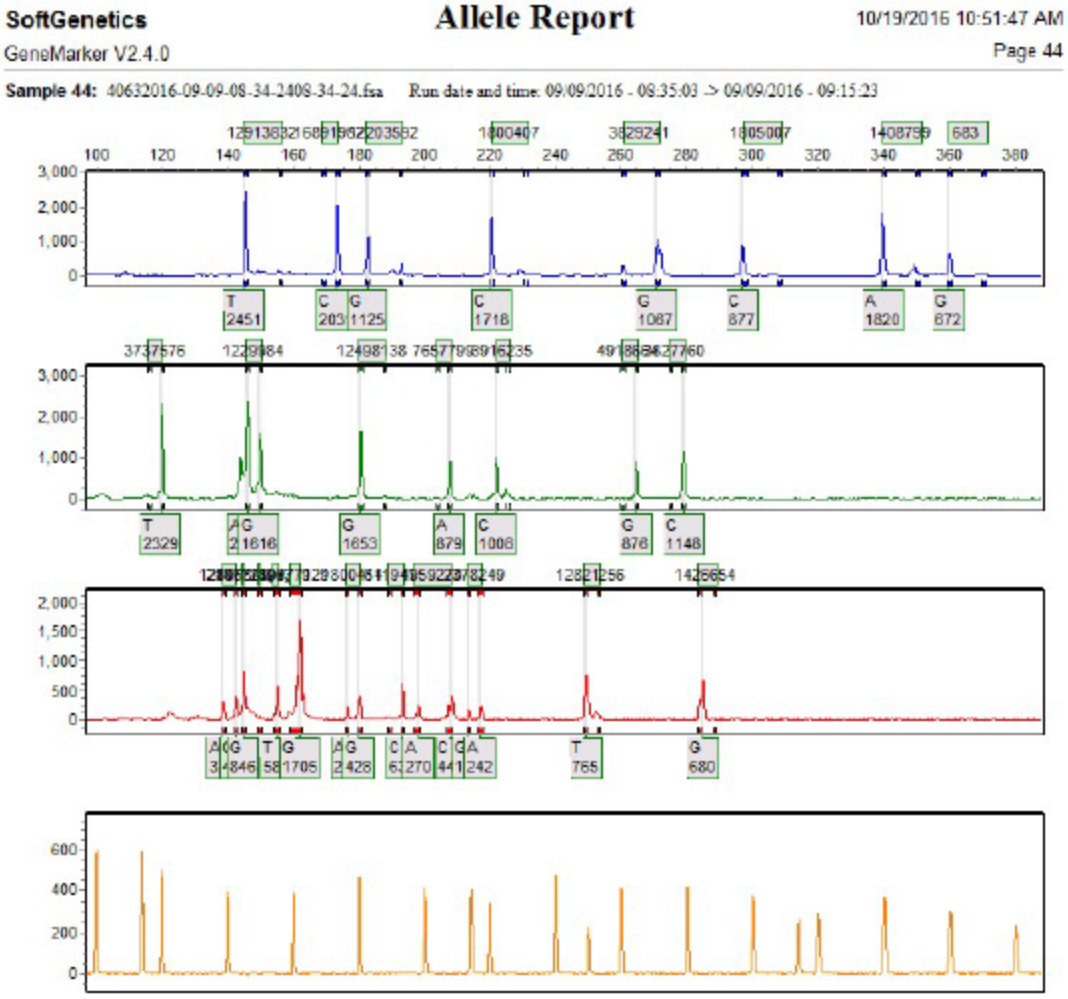


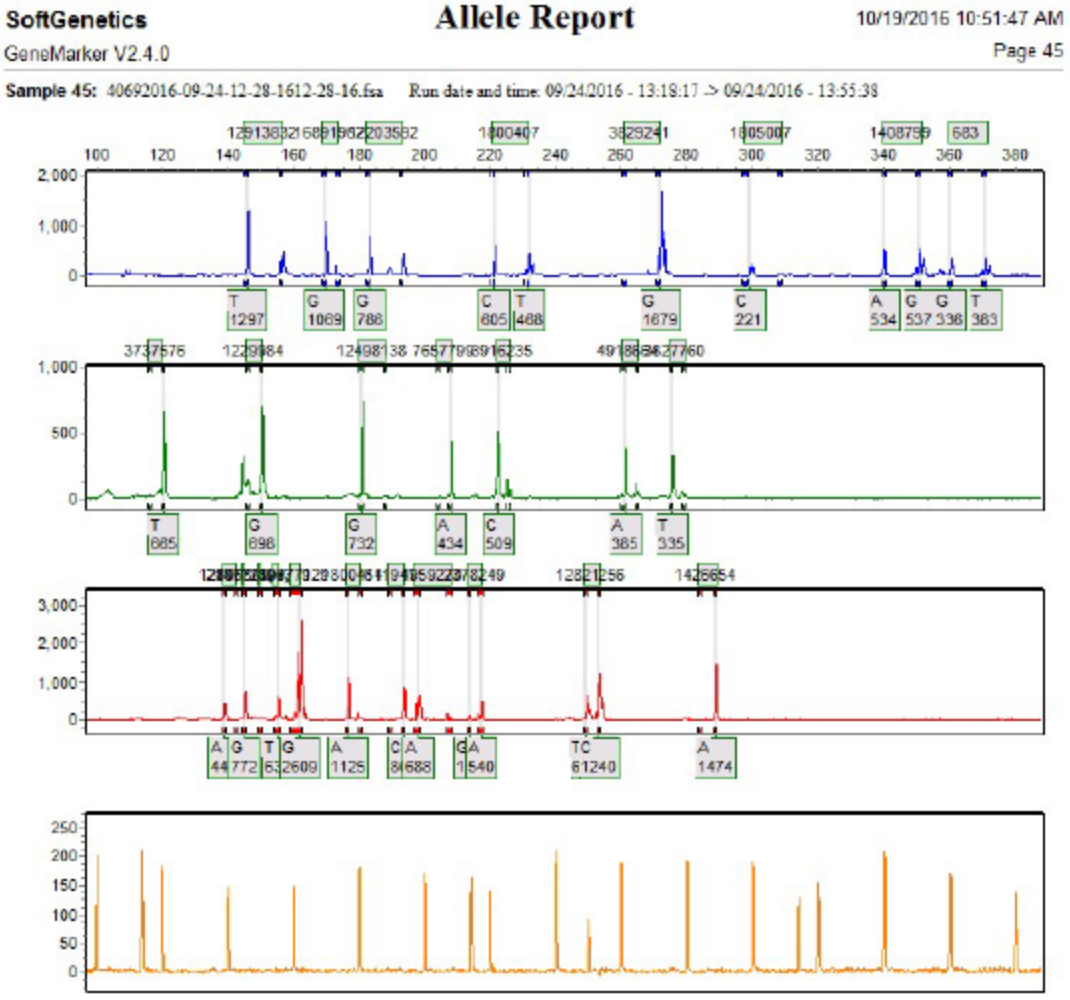
Sample 40: 39502016-09-09-08-34-2408-34-24.fsa Run date and time: 09/09/2016 - 09:16:11 -> 09/09/2016 - 09:54:16




SoftGenetics
Allele Report
10/19/2016 10:51:47 AM
GeneMarker V2.4.0
Page 41
Sample 41: 39722016-09-19-11-58-3511-58-35.fsa Run date and time: 09/19/2016 - 11:59:16 -> 09/19/2016 - 12:48:16


SoftGenetics
Allele Report
10/19/2016 10:51:47 AM
GeneMarker V2.4.0
Page 42
Sample 42: 40562016-09-22-16-13-3916-13-39.fsa Run date and time: 09/22/2016 - 19:38:07 -> 09/22/2016 - 20:16:22
Sample 6: 4056P2016-10-07-07-32-3307-32-33.fsa Run date and time: 10/07/2016 - 07:33:16 -> 10/07/2016 - 08:21:37

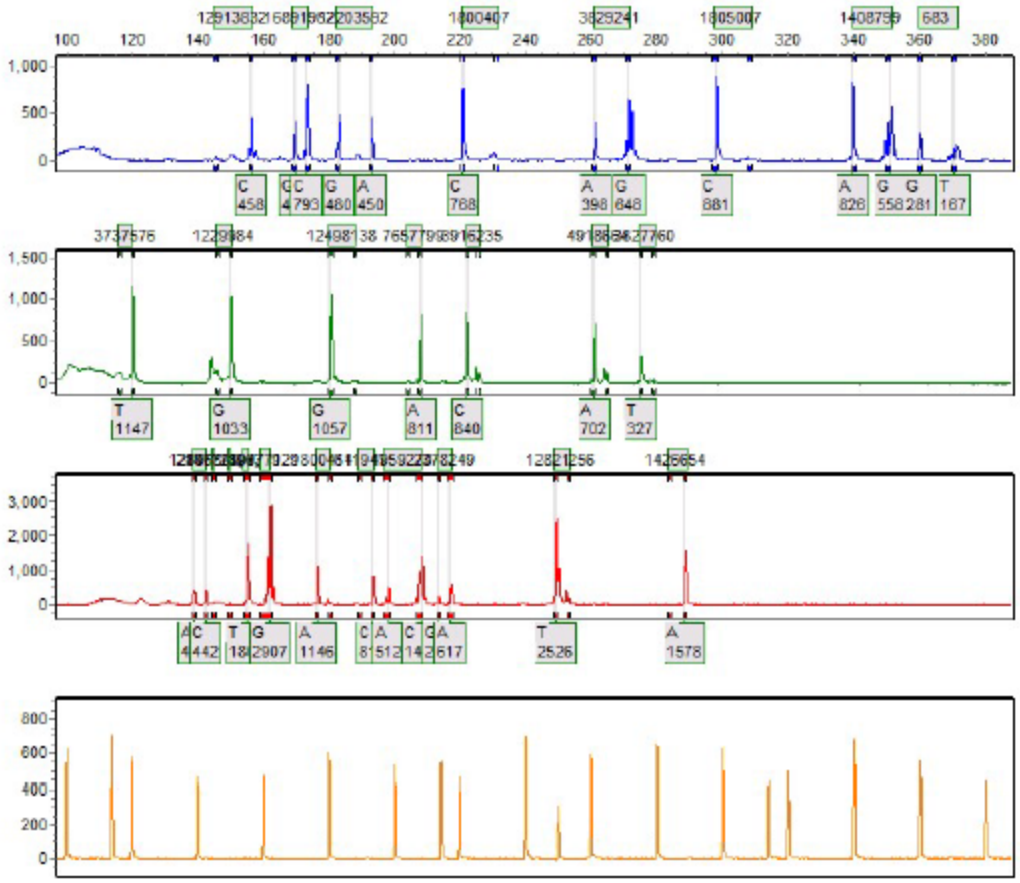

SoftGenetics
Allele Report
10/19/2016 10:51:47 AM
GeneMarker V2.4.0
Page 44
Sample 44: 40632016-09-09-08-34-2408-34-24.fsa Run date and time: 09/09/2016 - 08:35:03 -> 09/09/2016 - 09:15:23

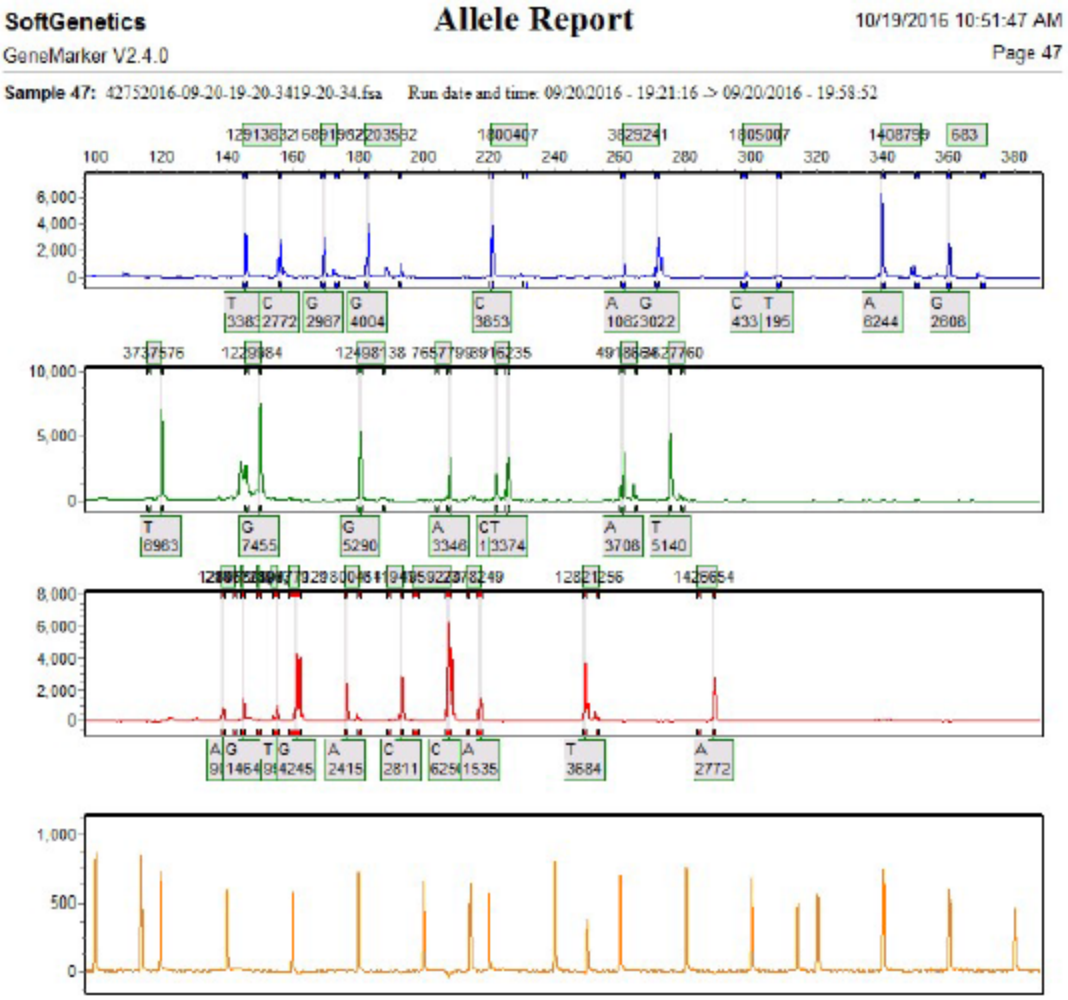

SoftGenetics
Allele Report
10/19/2016 10:51:47 AM
GeneMarker V2.4.0
Page 45
Sample 45: 40692016-09-24-12-28-1612-28-16.fsa Run date and time: 09/24/2016 - 13:18:17 -> 09/24/2016 - 13:55:38

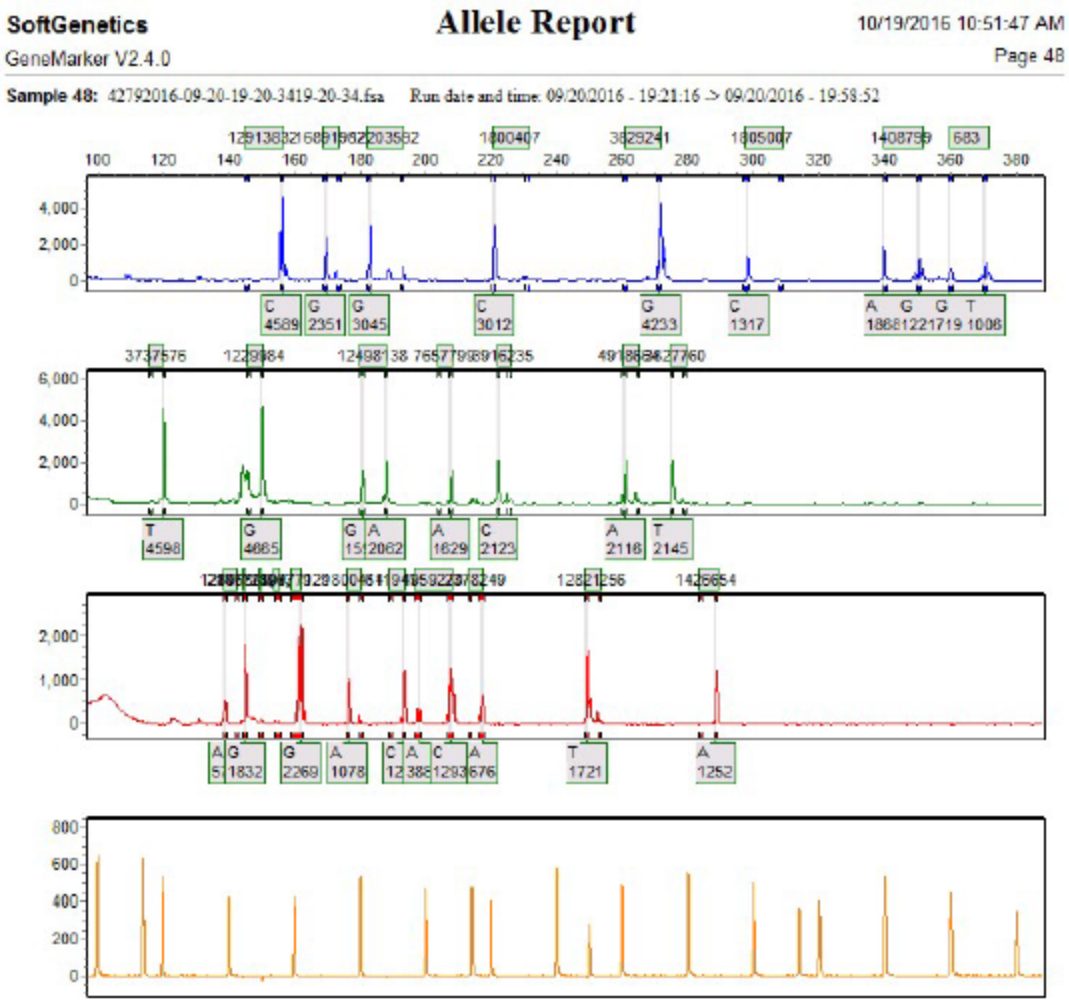
SoftGenetics **Allele Report** 10/19/2016 10:51:47 AM

GeneMarker V2.4.0 Page 46

Sample 46: 42582016-09-22-16-13-3916-13-39.fsa Run date and time: 09/22/2016 - 18:20:20 -> 09/22/2016 - 18:58:21



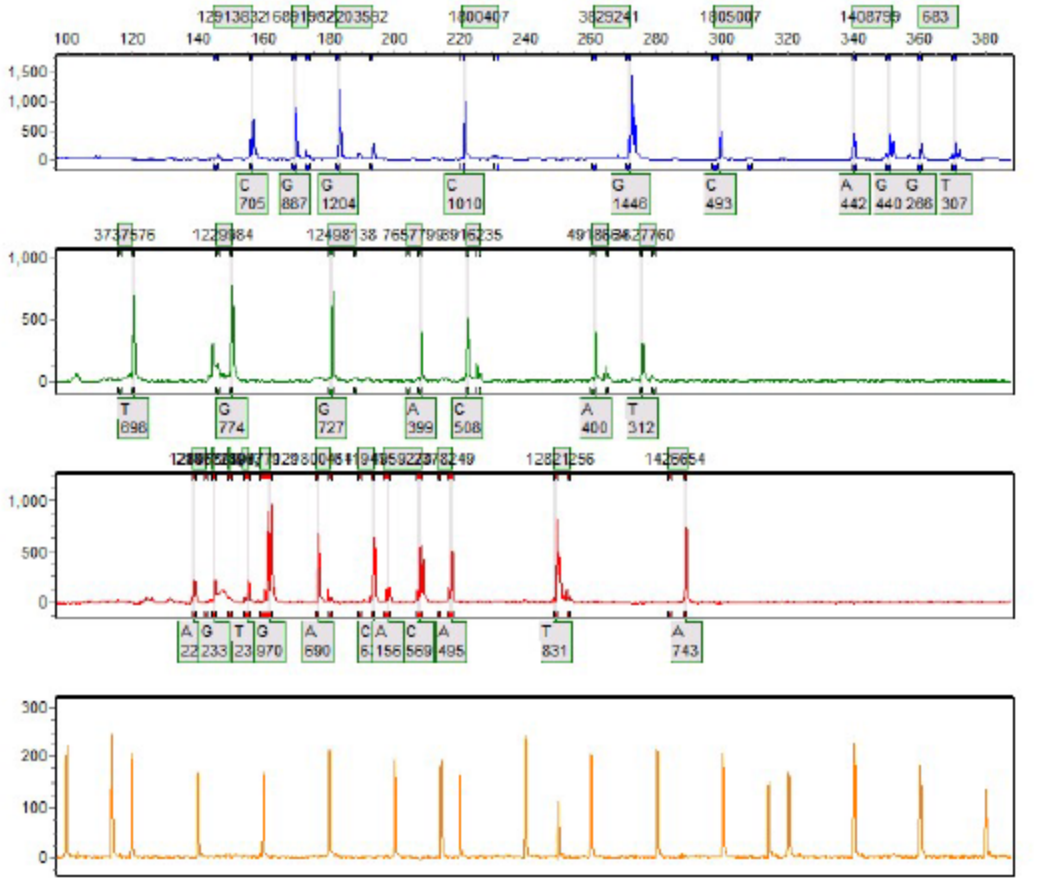

SoftGenetics
Allele Report
10/19/2016 10:51:47 AM
GeneMarker V2.4.0
Page 47
Sample 47: 42752016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 19:21:16 -> 09/20/2016 - 19:58:52

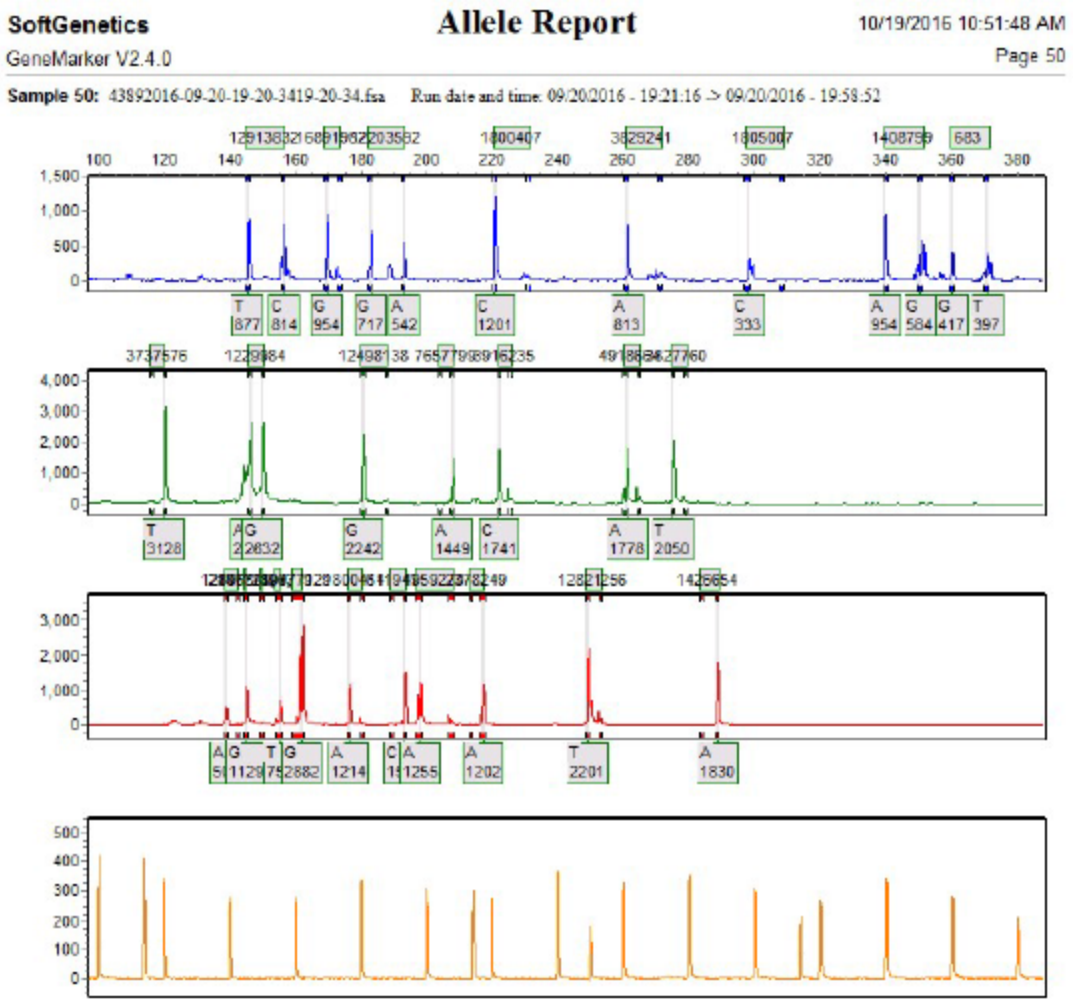

SoftGenetics
Allele Report
10/19/2016 10:51:47 AM
GeneMarker V2.4.0
Page 48
Sample 48: 42792016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 19:21:16 -> 09/20/2016 - 19:58:52

SoftGenetics **Allele Report** 10/19/2016 10:51:48 AM

GeneMarker V2.4.0 Page 49

Sample 49: 43022016-09-23-18-33-1718-33-17.fsa Run date and time: 09/23/2016 - 19:53:05 -> 09/23/2016 - 20:30:41

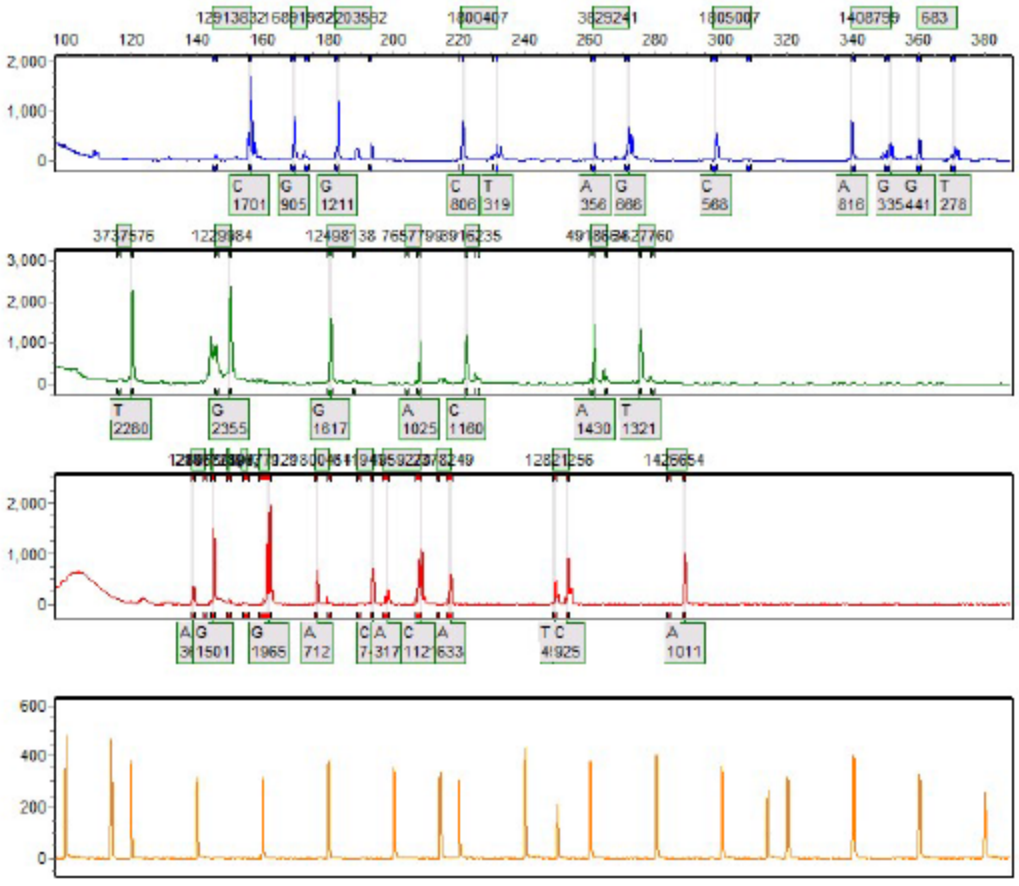


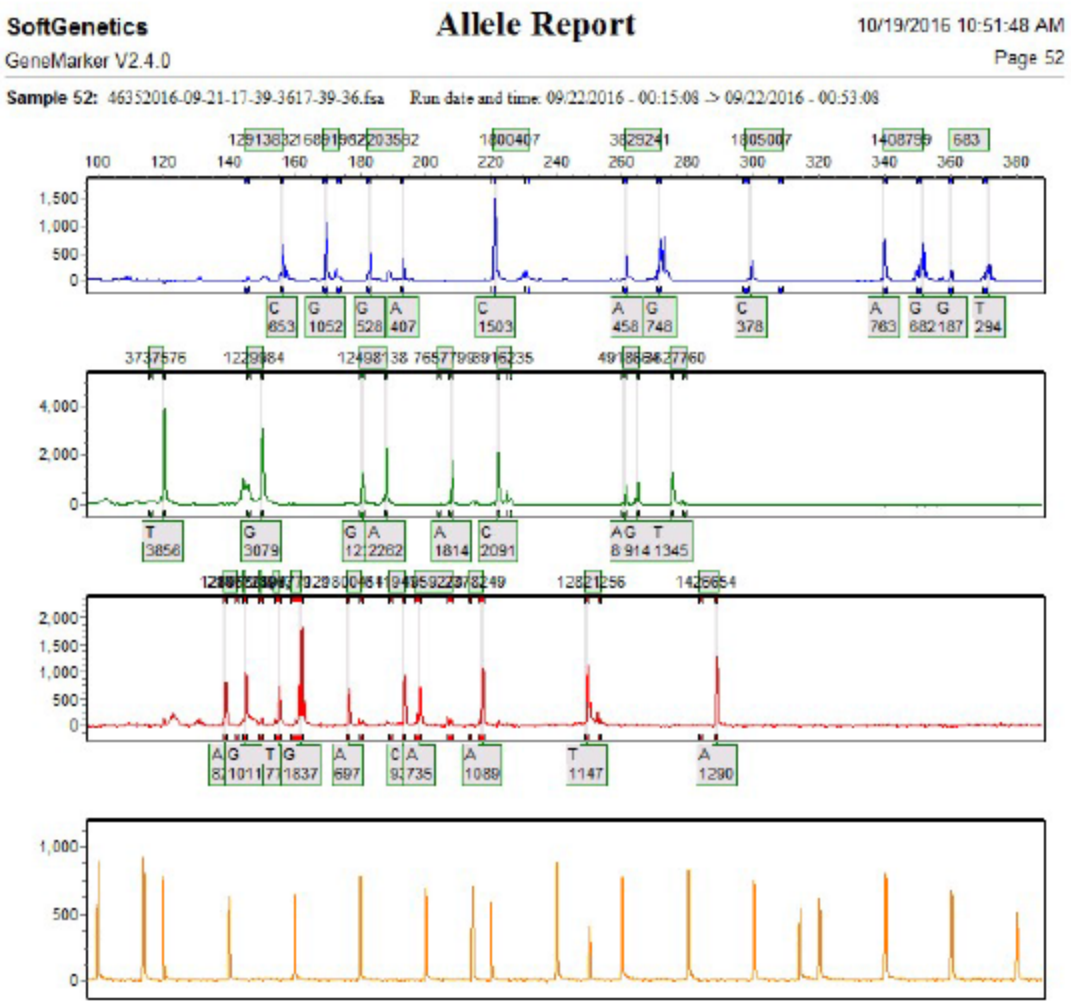

SoftGenetics
Allele Report
10/19/2016 10:51:48 AM
GeneMarker V2.4.0
Page 50
Sample 50: 43892016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 19:21:16 -> 09/20/2016 - 19:58:52

SoftGenetics **Allele Report** 10/19/2016 10:51:48 AM

GeneMarker V2.4.0 Page 51

Sample 51: 45682016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 19:21:16 -> 09/20/2016 - 19:58:52

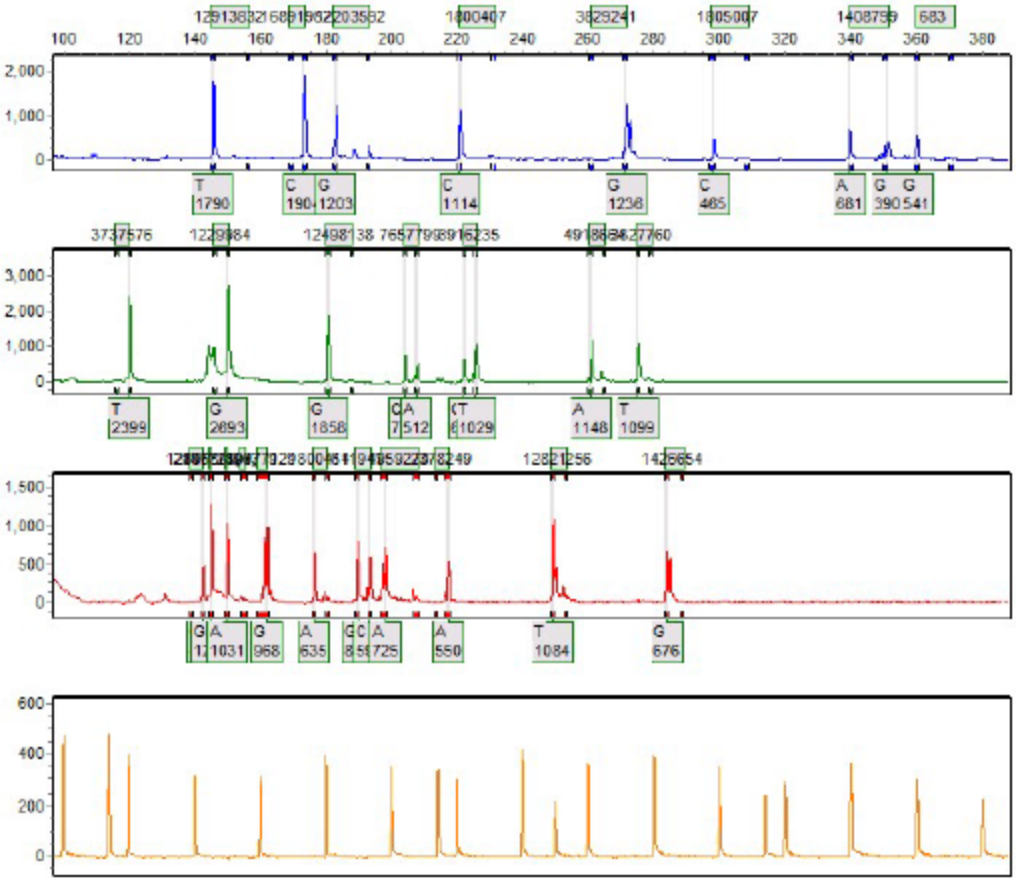



SoftGenetics
Allele Report
10/19/2016 10:51:48 AM
GeneMarker V2.4.0
Page 52
Sample 52: 46352016-09-21-17-39-3617-39-36.fsa Run date and time: 09/22/2016 - 00:15:08 -> 09/22/2016 - 00:53:08

SoftGenetics **Allele Report** 10/19/2016 10:51:48 AM

GeneMarker V2.4.0 Page 53

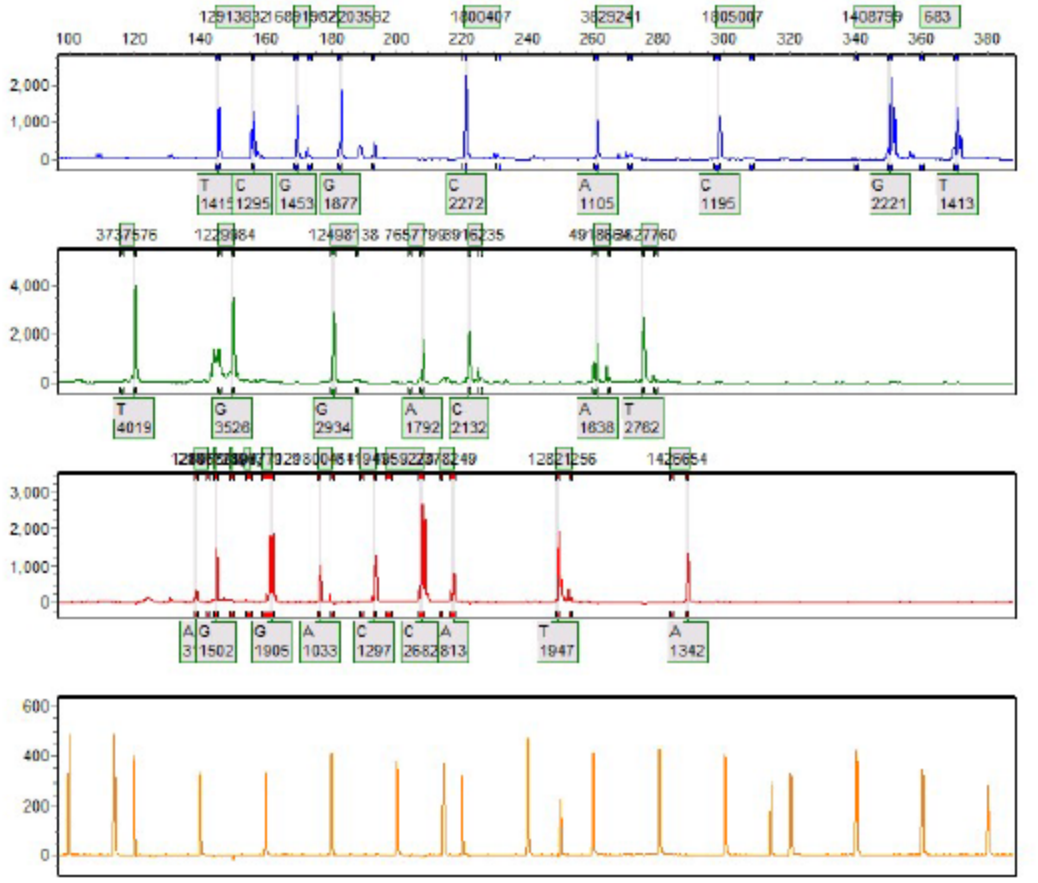
Sample 53: 47102016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 19:21:16 -> 09/20/2016 - 19:58:52

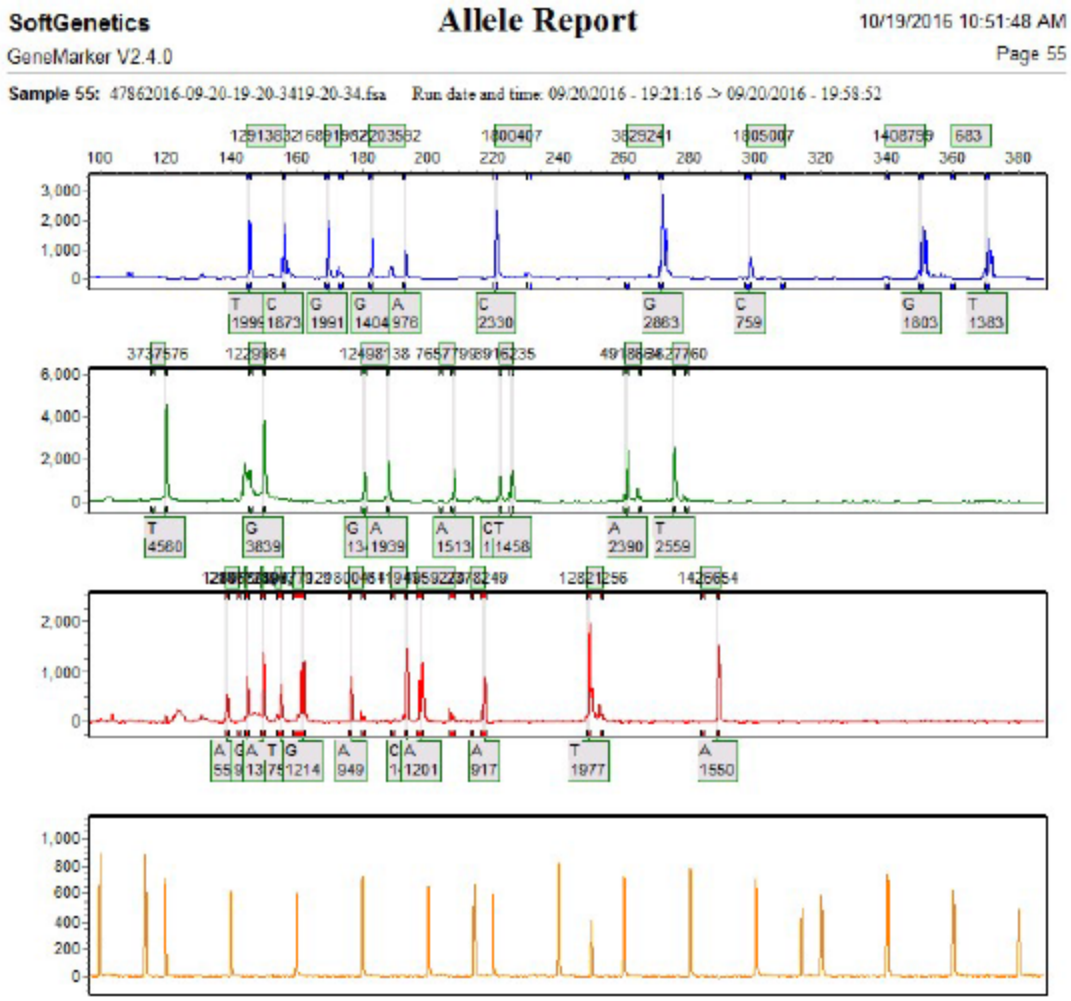


SoftGenetics **Allele Report** 10/19/2016 10:51:48 AM

GeneMarker V2.4.0 Page 54

Sample 54: 47322016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 19:21:16 -> 09/20/2016 - 19:58:52

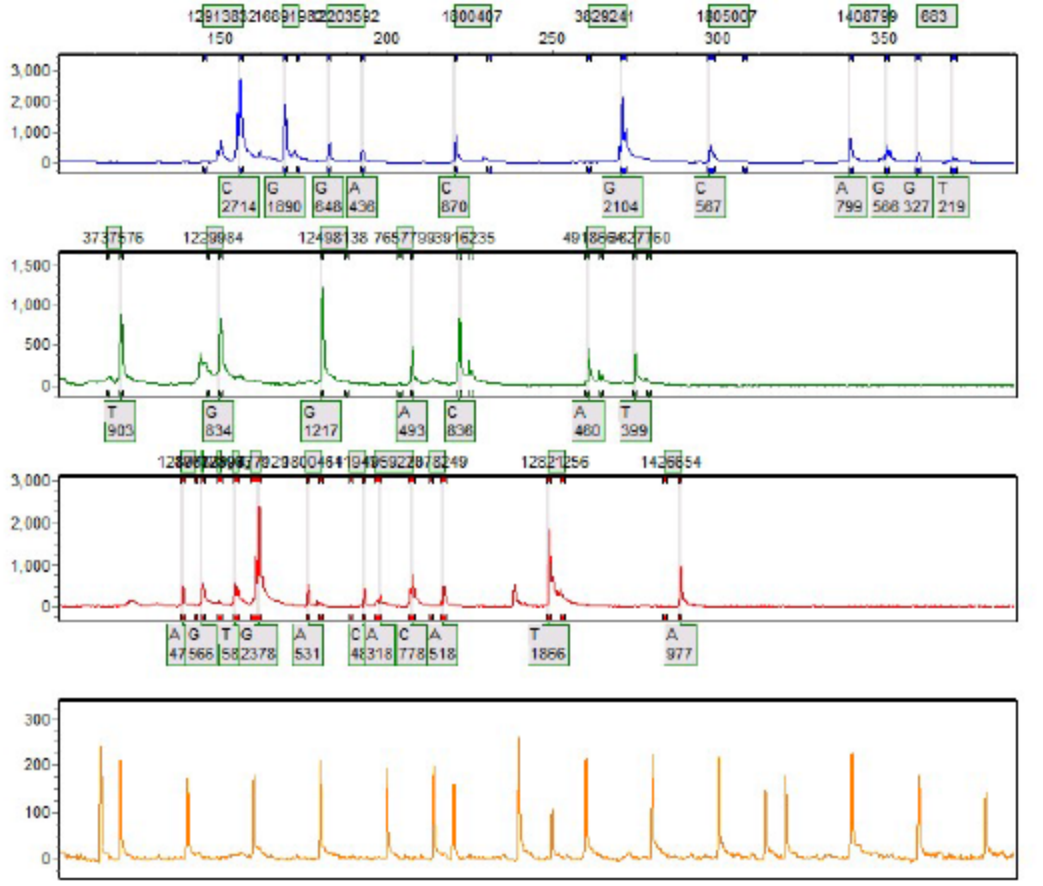


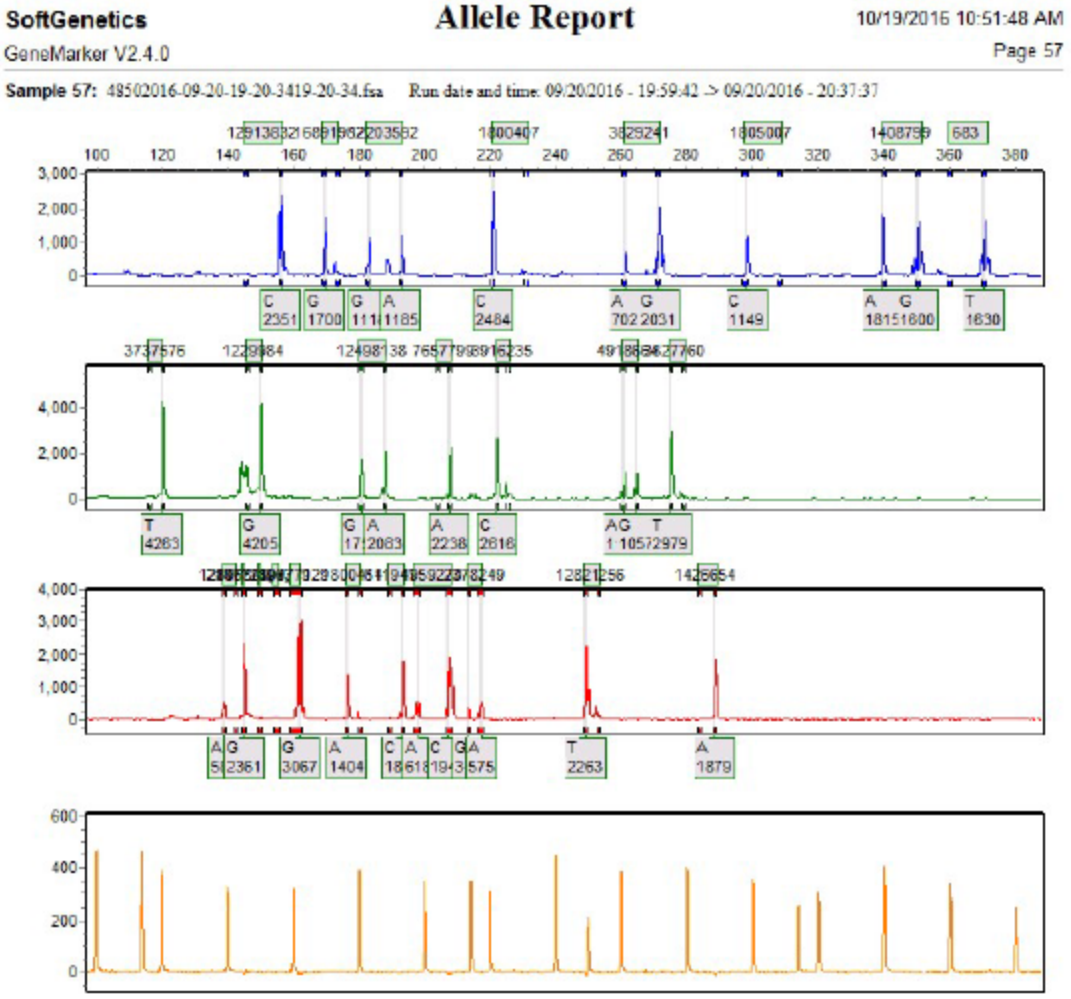

SoftGenetics
Allele Report
10/19/2016 10:51:48 AM
GeneMarker V2.4.0
Page 55
Sample 55: 47862016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 19:21:16 -> 09/20/2016 - 19:58:52

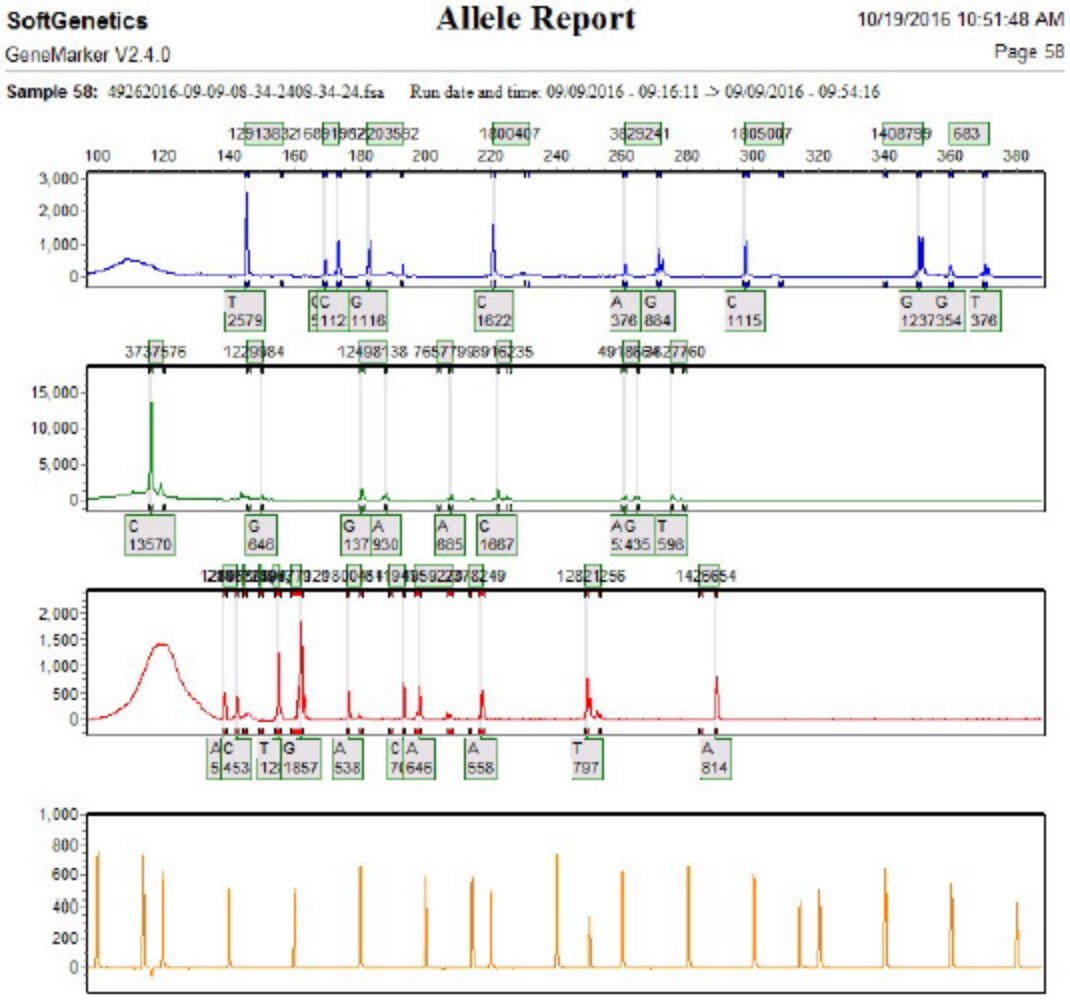
SoftGenetics **Allele Report** 10/19/2016 1:13:27 PM

GeneMarker V2.4.0 Page 1

Sample 1: 48192016-10-12-14-36-5214-36-52.fsa Run date and time: 10/12/2016 - 14:37:50 -> 10/12/2016 - 15:26:21



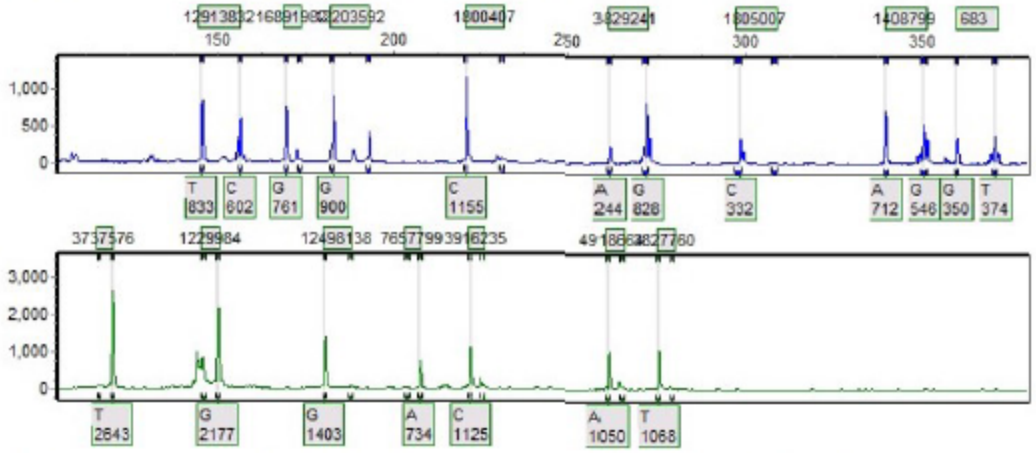

SoftGenetics
GeneMarker V2.4.0
Allele Report
10/19/2016 10:51:48 AM
Page 57
Sample 57: 485\02016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 19:59:42 -> 09/20/2016 - 20:37:37


SoftGenetics
Allele Report
10/19/2016 10:51:48 AM
GeneMarker V2.4.0
Page 58
Sample 58: 49262016-09-09-08-34-2408-34-24.fsa Run date and time: 09/09/2016 - 09:16:11 -> 09/09/2016 - 09:54:16

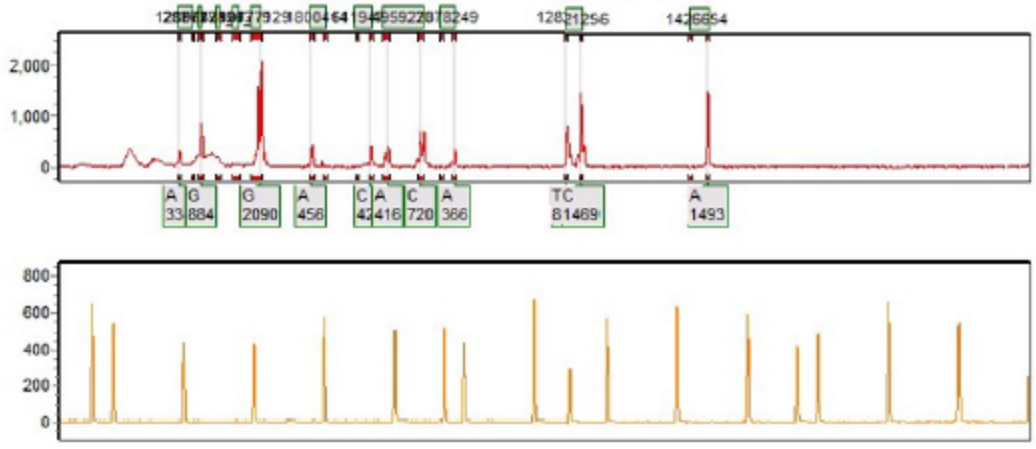
SoftGenetics **Allele Report** 10/19/2016 11:49:33 AM

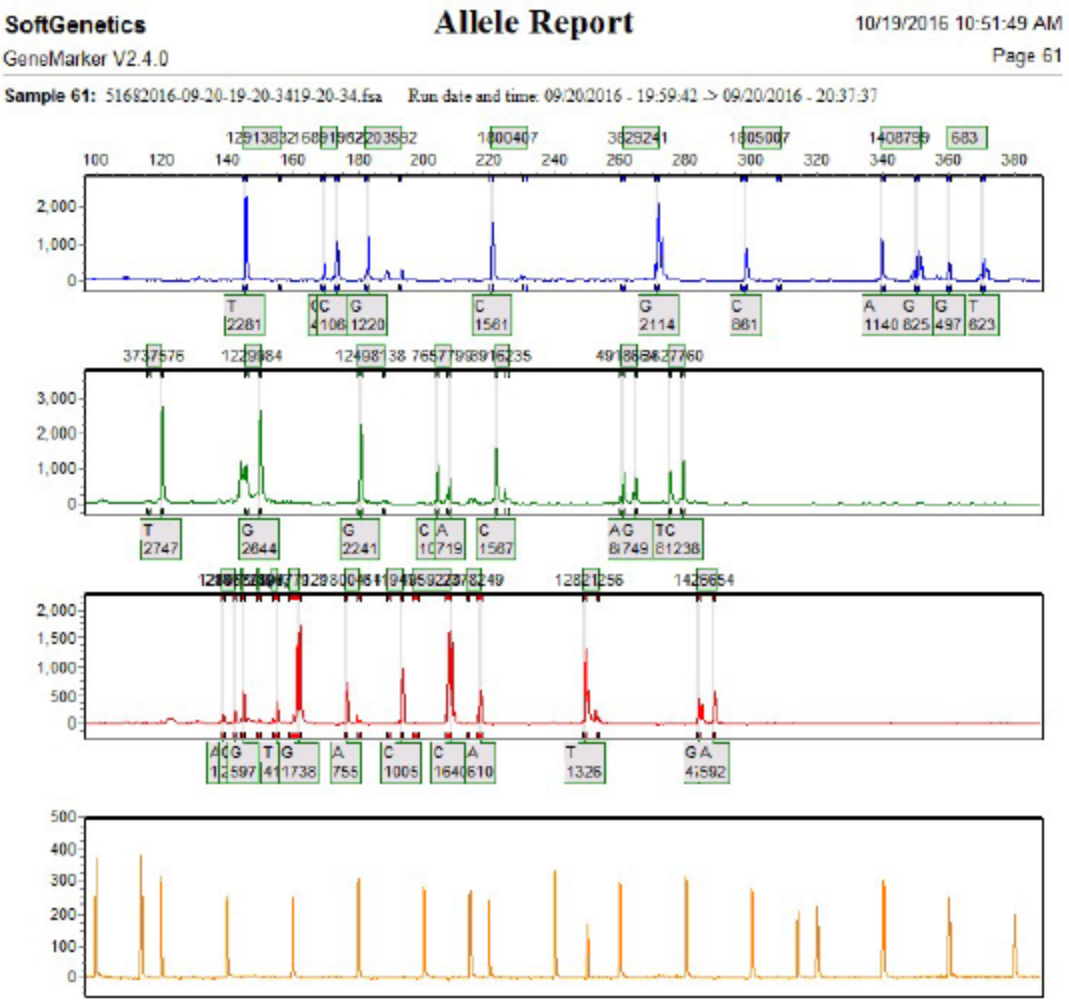
GeneMarker V2.4.0 Page 7

Sample 7: 49382016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 19:59:42 -> 09/20/2016 - 20:37:37



Sample 60: 4938P2016-10-07-07-32-3307-32-33.fsa Run date and time: 10/07/2016 - 07:33:16 -> 10/07/2016 - 08:21:37

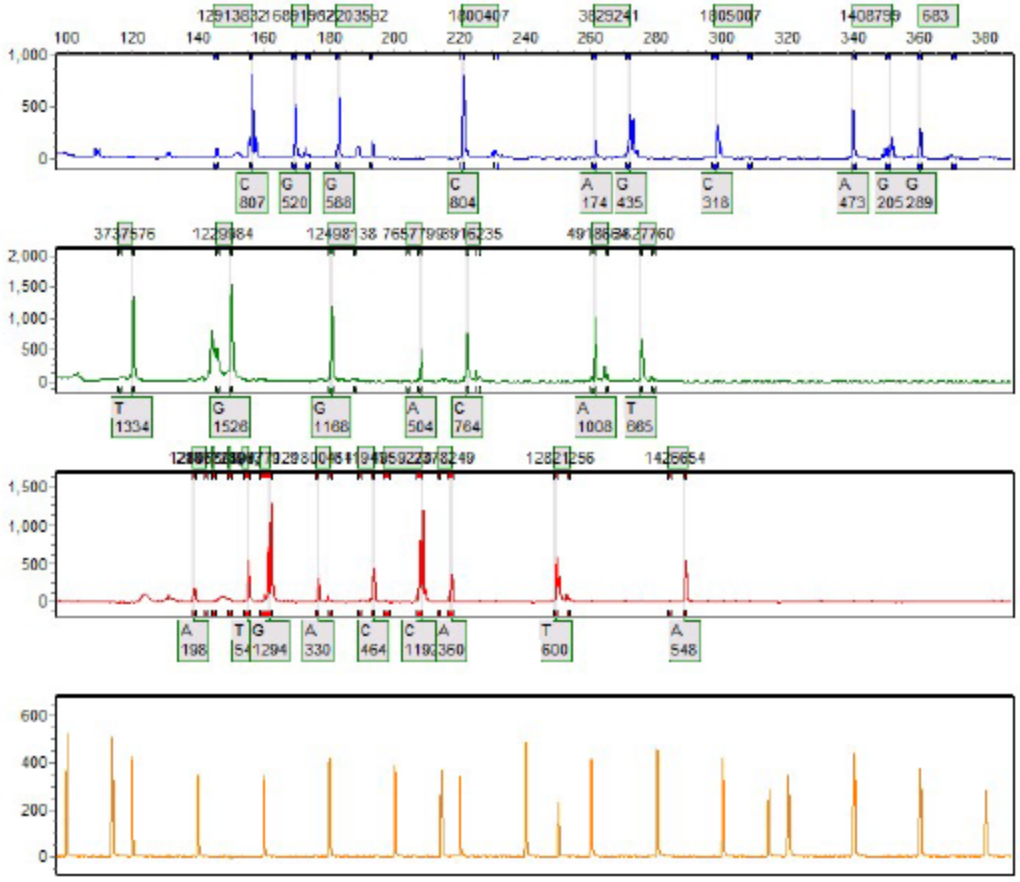


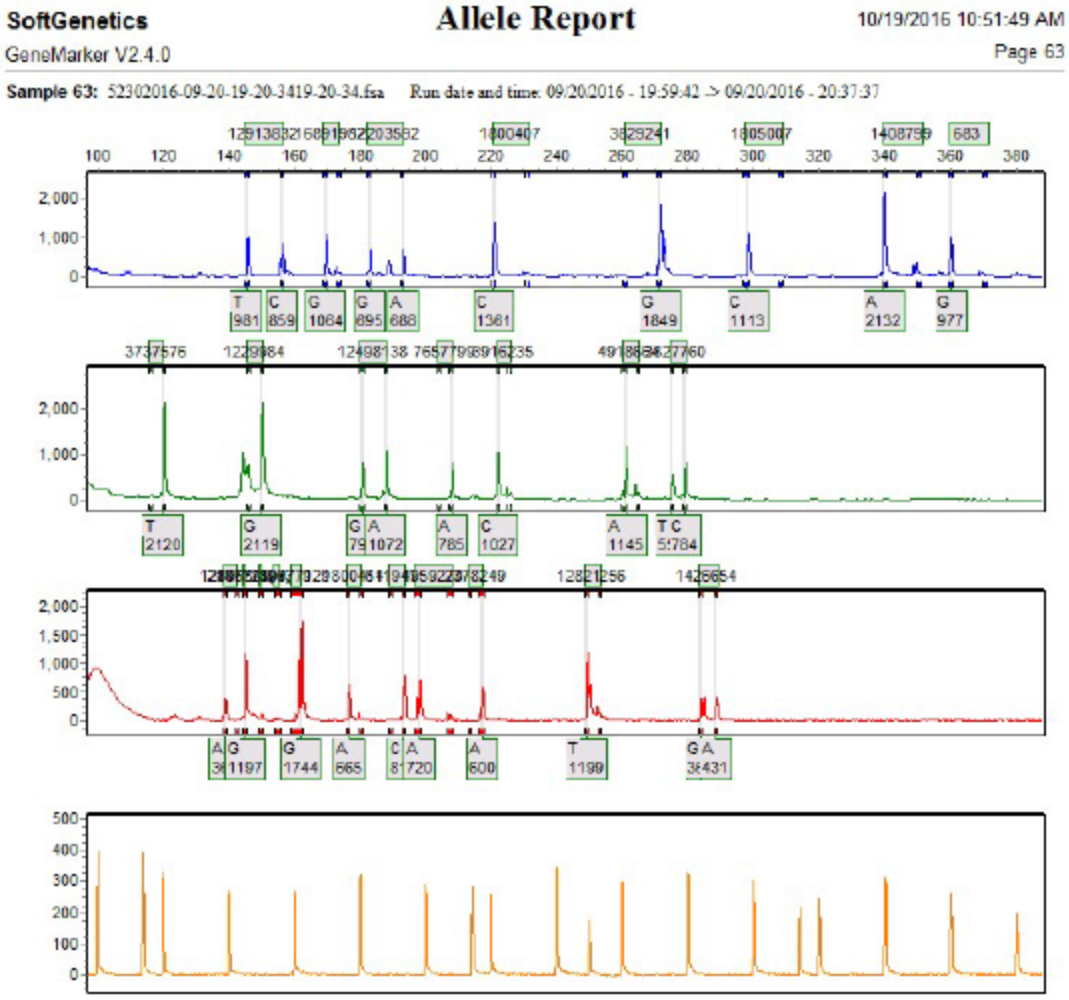

SoftGenetics
Allele Report
10/19/2016 10:51:49 AM
GeneMarker V2.4.0
Page 61
Sample 61: 51682016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 19:59:42 -> 09/20/2016 - 20:37:37

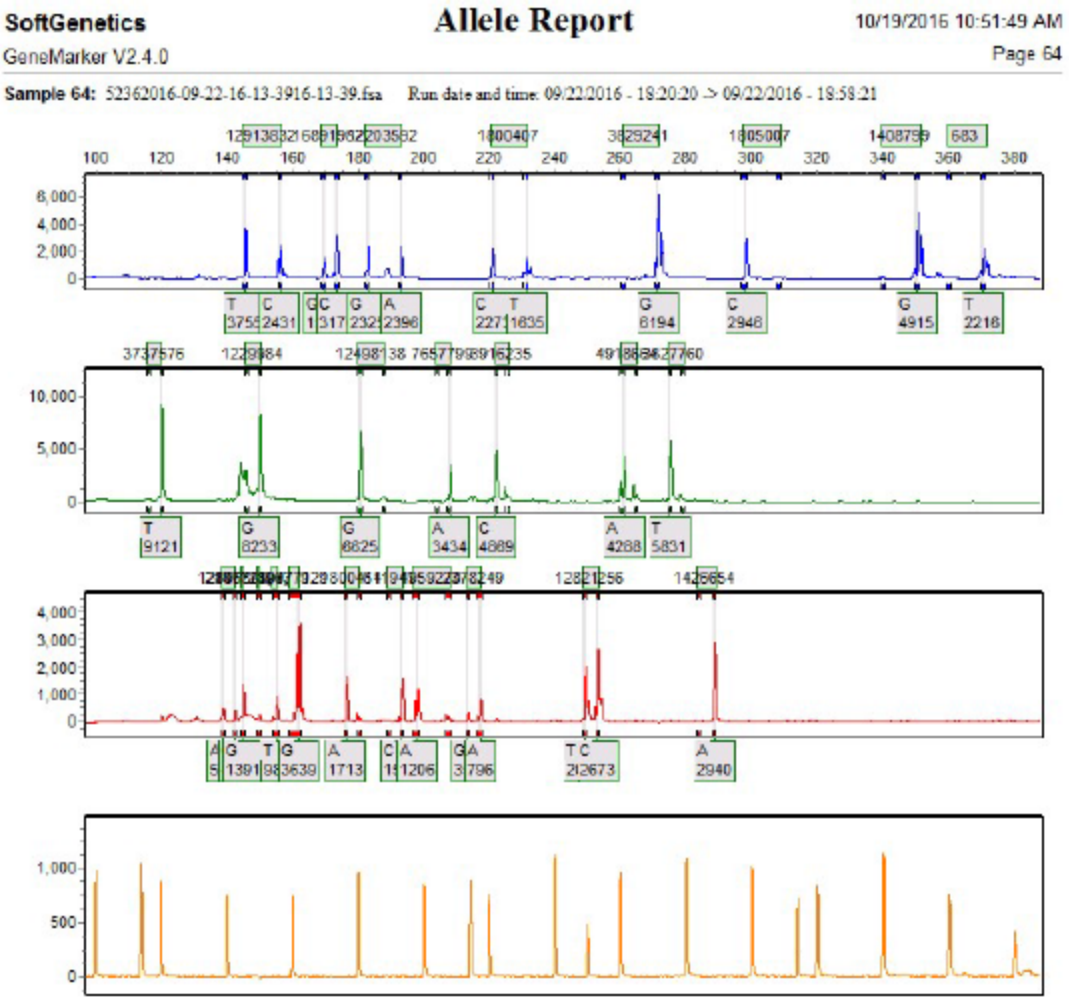
SoftGenetics **Allele Report** 10/19/2016 10:51:49 AM

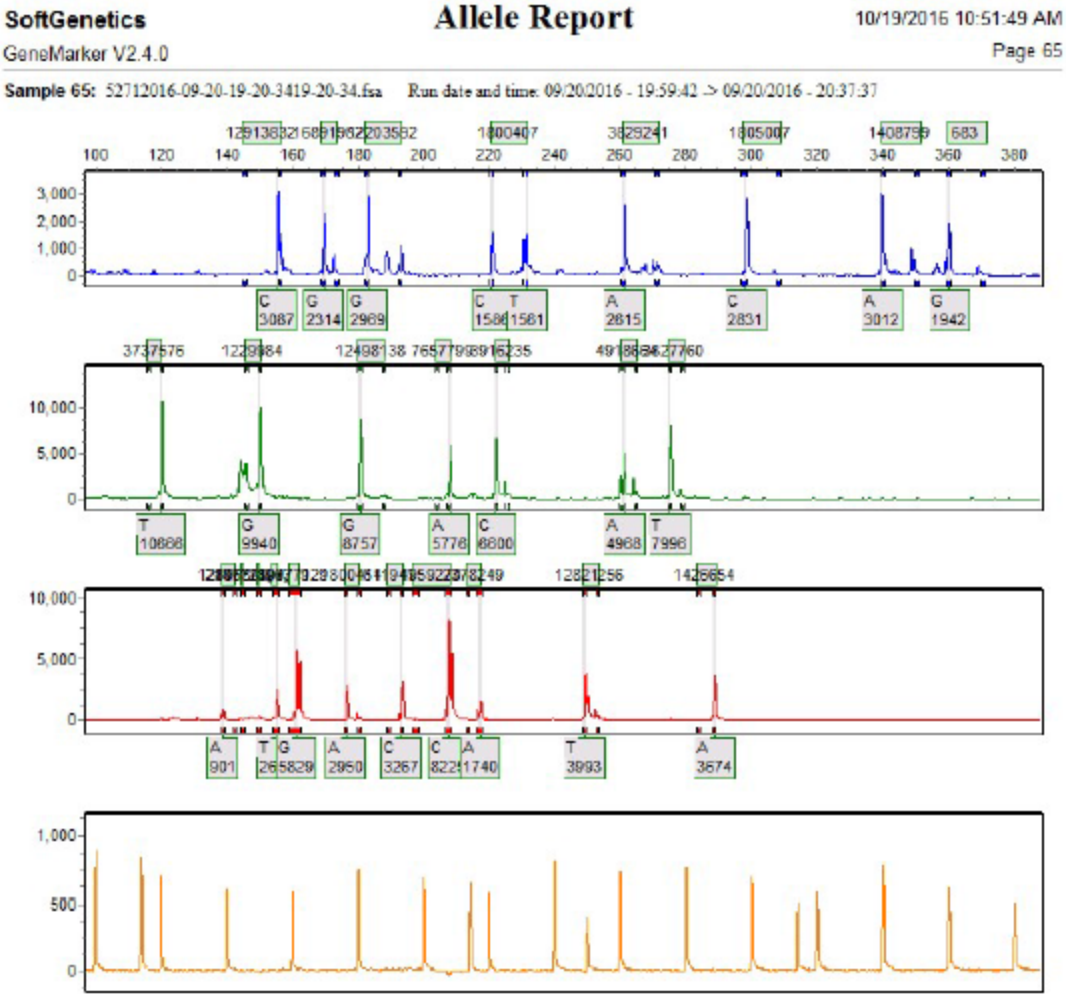
GeneMarker V2.4.0 Page 62

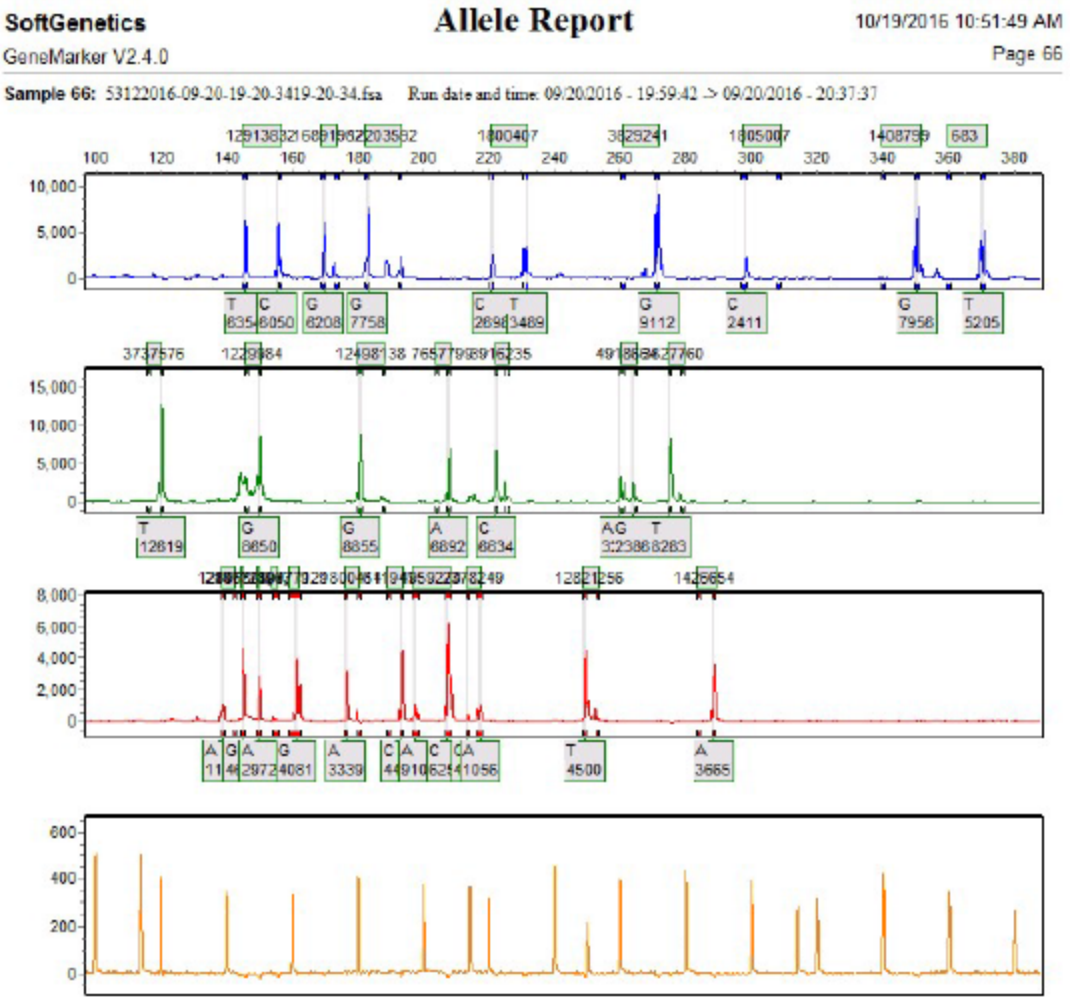
Sample 62: 51892016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 19:59:42 -> 09/20/2016 - 20:37:37




SoftGenetics
Allele Report
10/19/2016 10:51:49 AM
GeneMarker V2.4.0
Page 63
Sample 63: 52302016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 19:59:42 -> 09/20/2016 - 20:37:37


SoftGenetics
Allele Report
10/19/2016 10:51:49 AM
GeneMarker V2.4.0
Page 64
Sample 64: 52362016-09-22-16-13-3916-13-39.fsa Run date and time: 09/22/2016 - 18:20:20 -> 09/22/2016 - 18:58:21

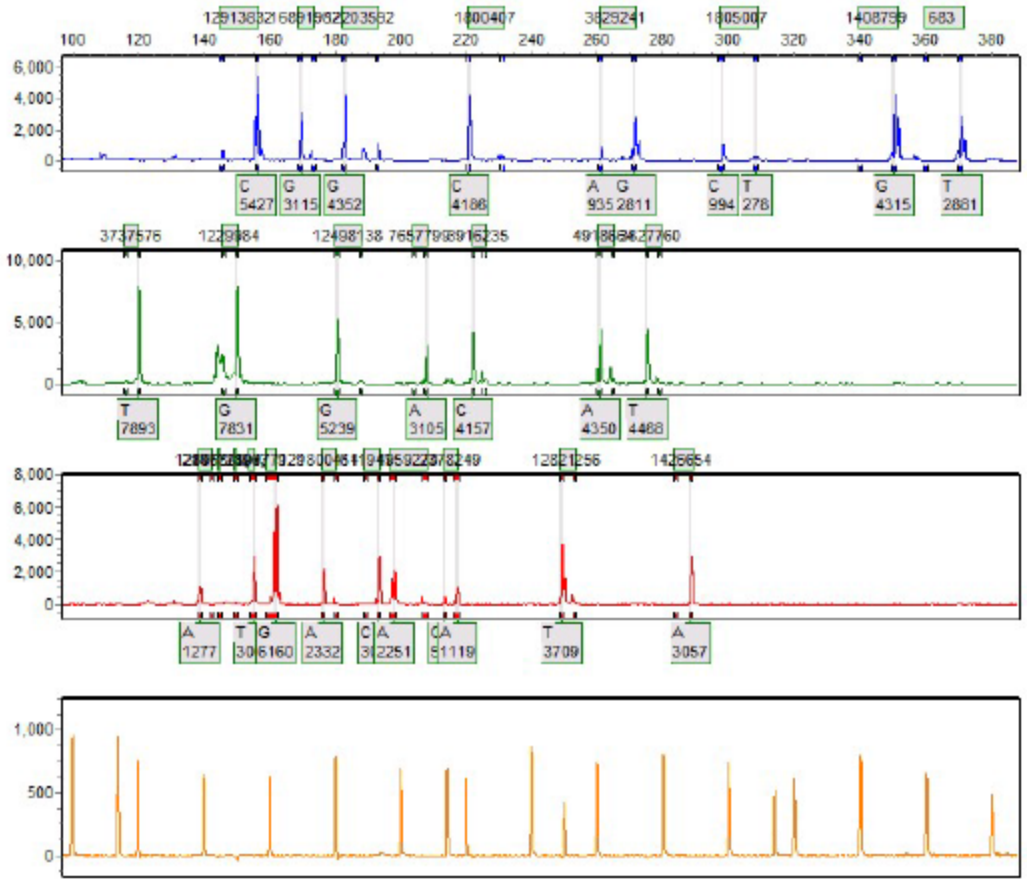

SoftGenetics
Allele Report
10/19/2016 10:51:49 AM
GeneMarker V2.4.0
Page 65
Sample 65: 52712016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 19:59:42 -> 09/20/2016 - 20:37:37

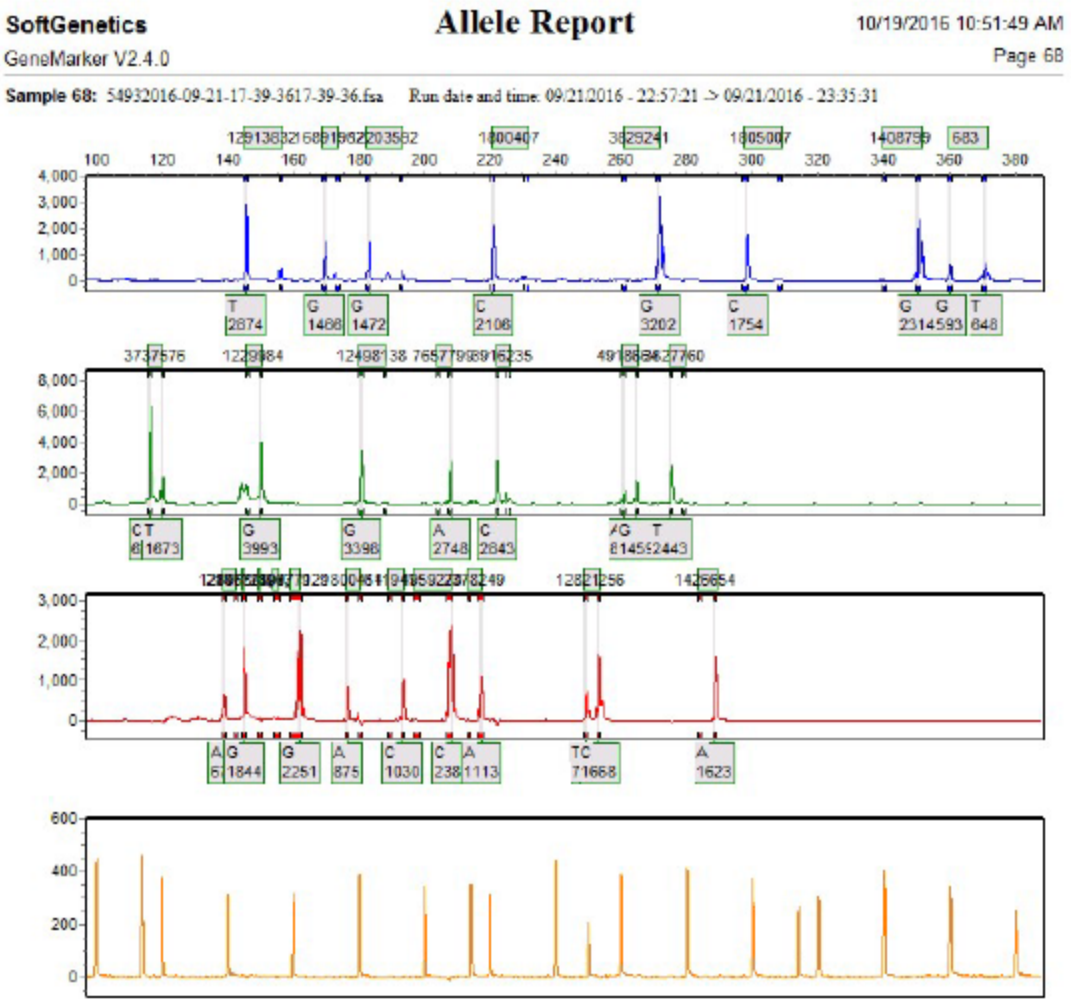

SoftGenetics
Allele Report
10/19/2016 10:51:49 AM
GeneMarker V2.4.0
Page 66
Sample 66: 53122016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 19:59:42 -> 09/20/2016 - 20:37:37

SoftGenetics **Allele Report** 10/19/2016 10:51:49 AM

GeneMarker V2.4.0 Page 67

Sample 67: 54362016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 20:38:27 -> 09/20/2016 - 21:16:32

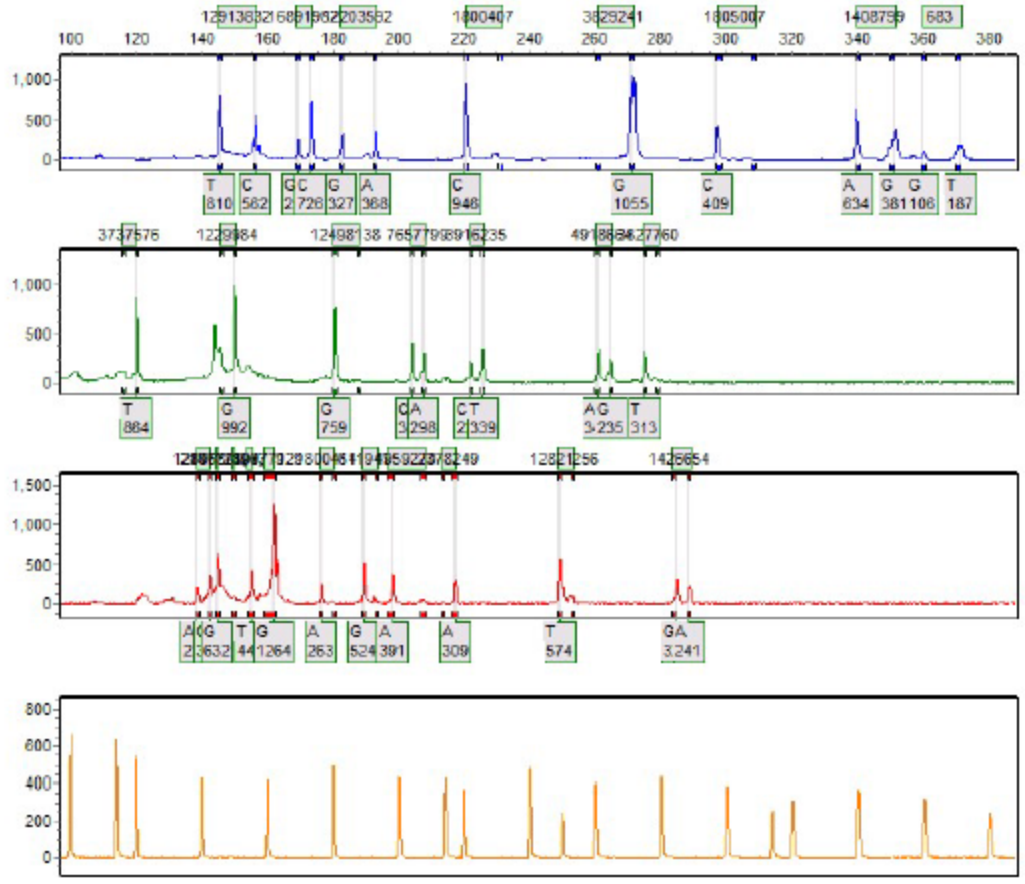



SoftGenetics
Allele Report
10/19/2016 10:51:49 AM
GeneMarker V2.4.0
Page 68
Sample 68: 54932016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 22:57:21 -> 09/21/2016 - 23:35:31

SoftGenetics **Allele Report** 10/19/2016 10:51:49 AM

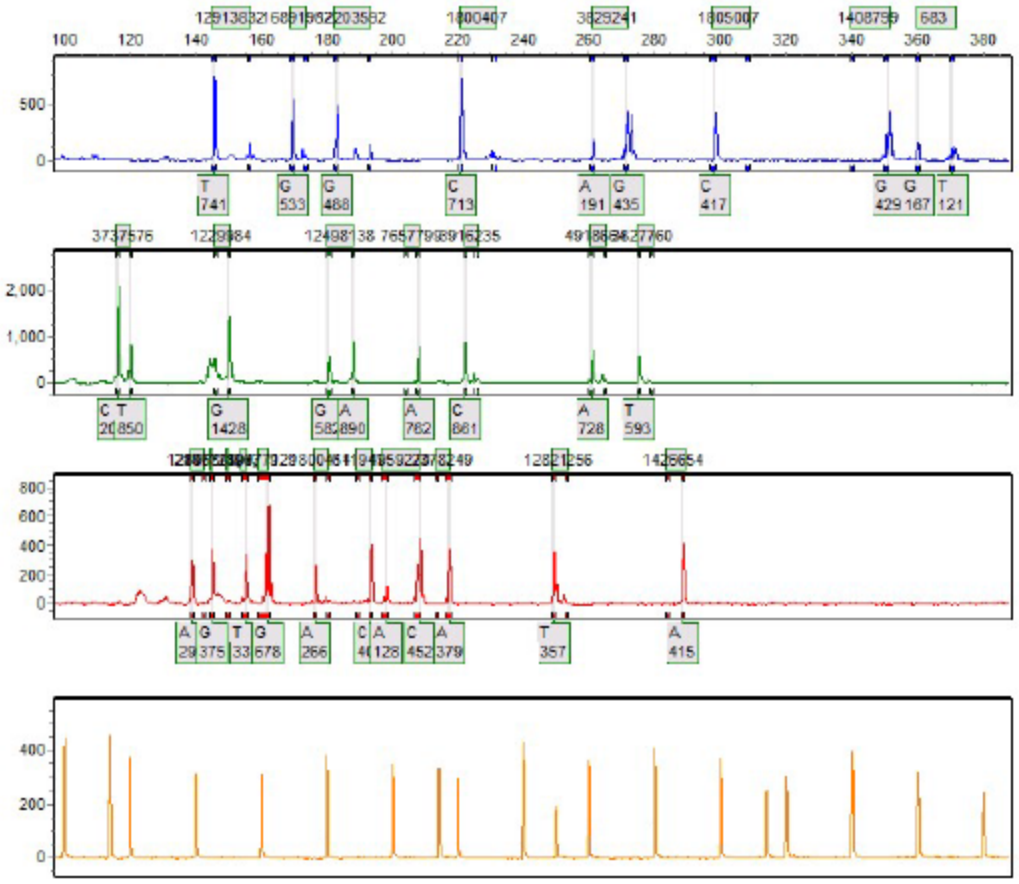
GeneMarker V2.4.0 Page 69

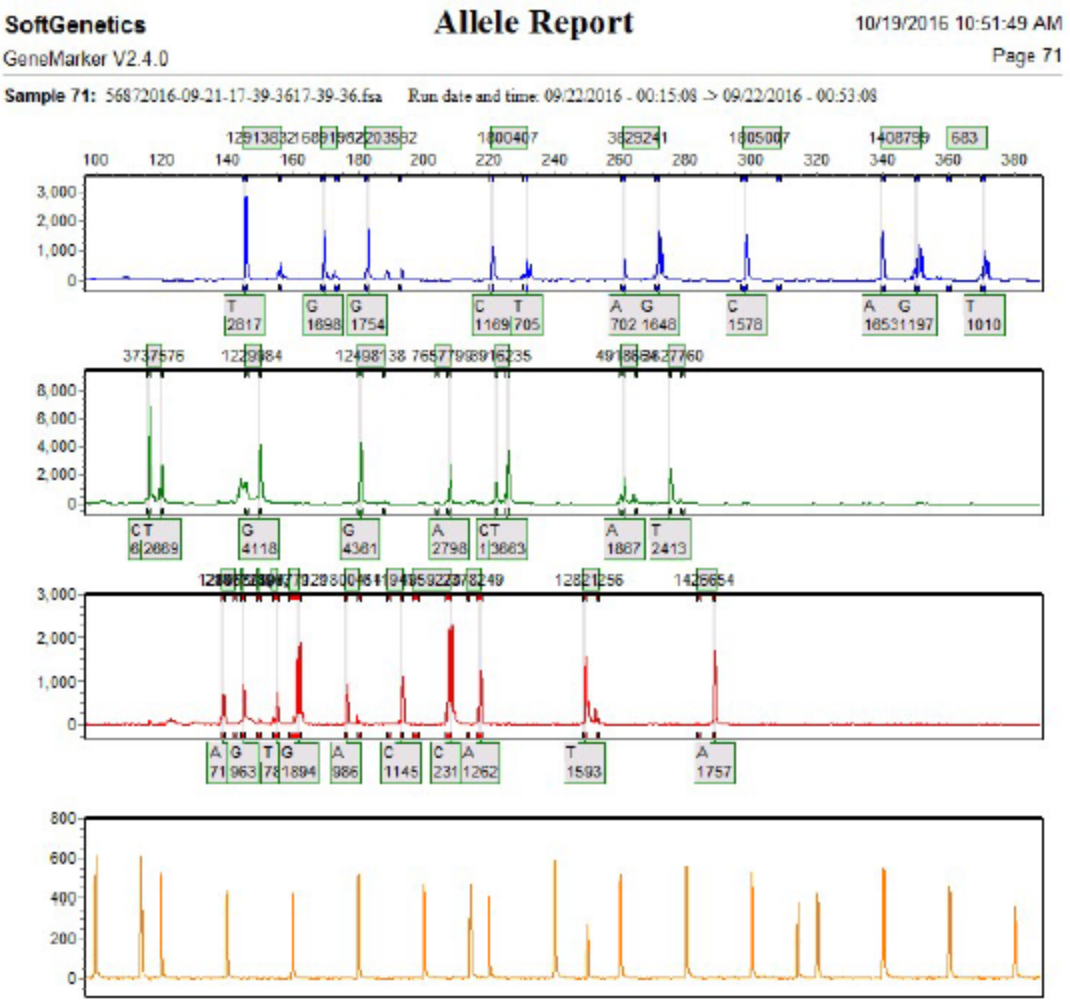
Sample 69: 5607-2016-09-09-08-34-2408-34-24.fsa Run date and time: 09/09/2016 - 08:35:03 -> 09/09/2016 - 09:15:23

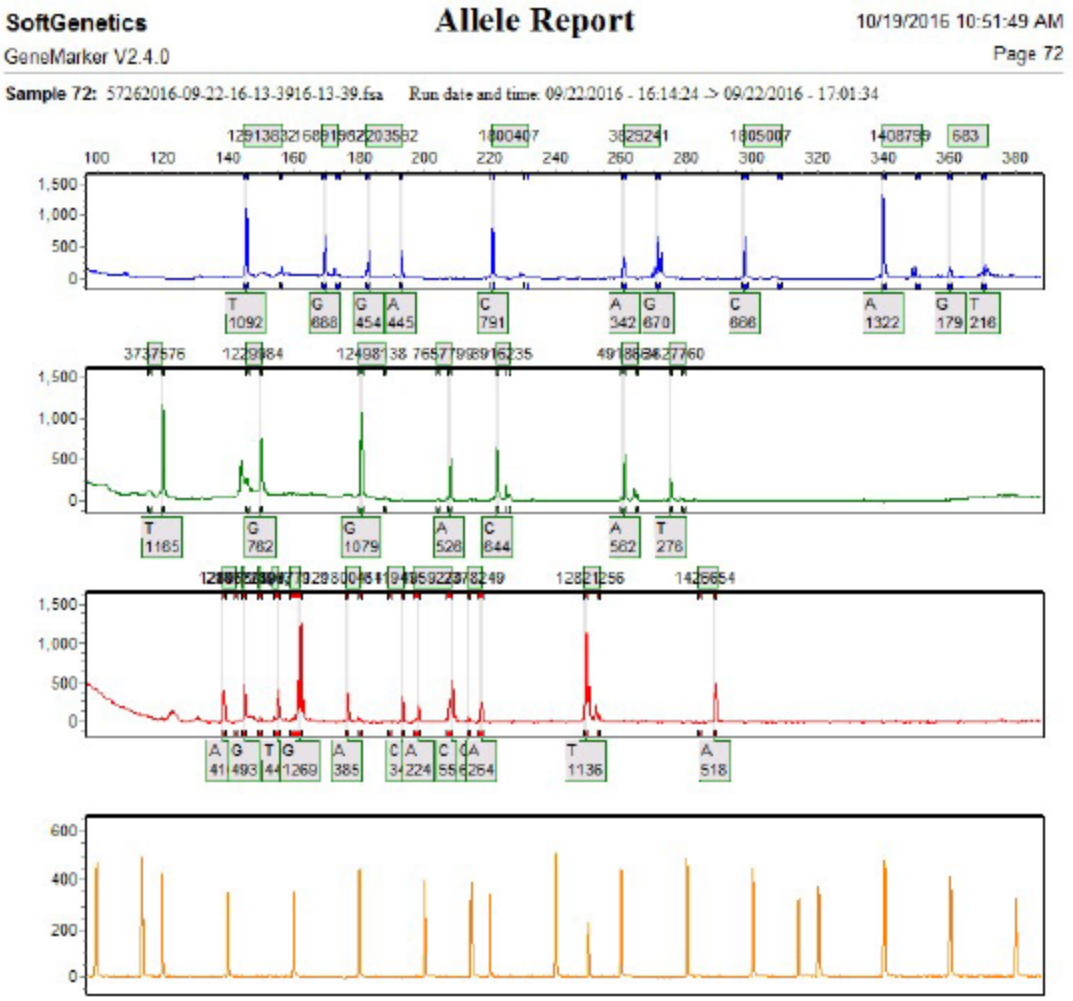


Allele Report

Sample 70: 56312016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 22:57:21 -> 09/21/2016 - 23:35:31



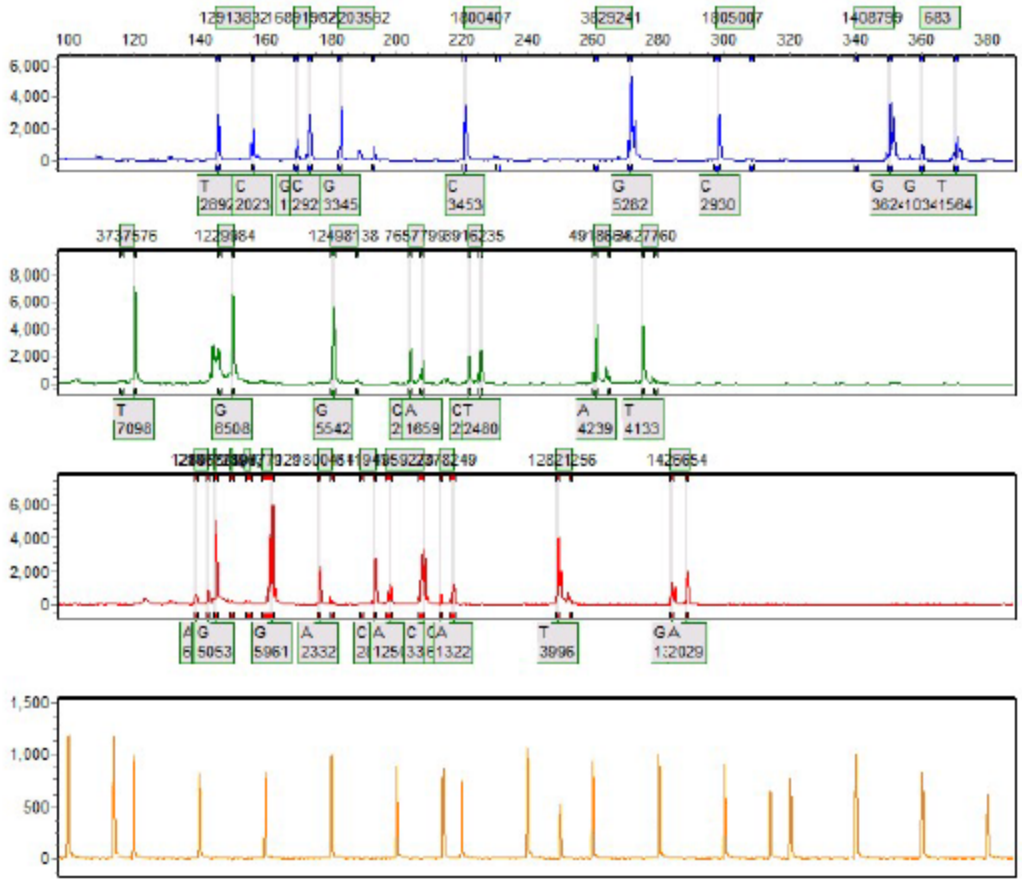

SoftGenetics
Allele Report
10/19/2016 10:51:49 AM
GeneMarker V2.4.0
Page 71
Sample 71: 56872016-09-21-17-39-3617-39-36.fsa Run date and time: 09/22/2016 - 00:15:08 -> 09/22/2016 - 00:53:08

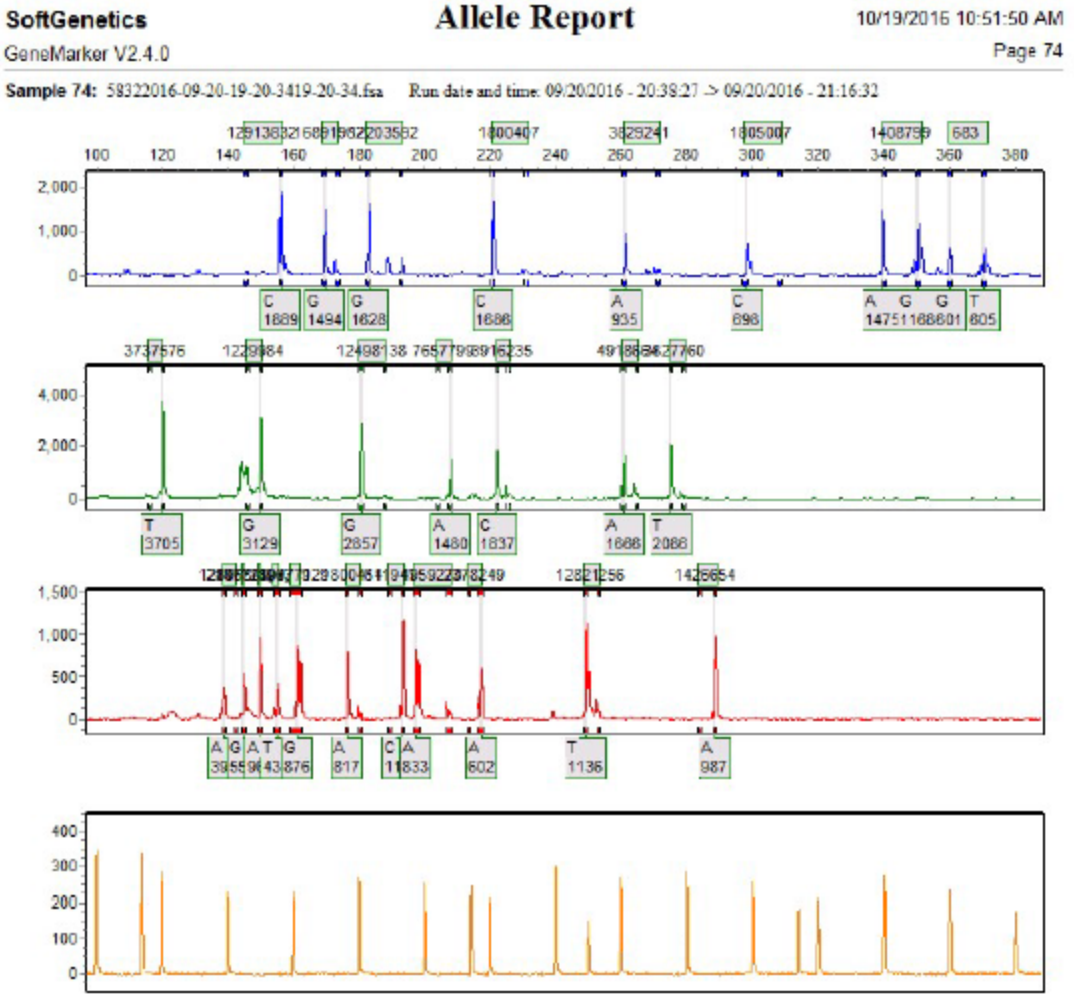

SoftGenetics
Allele Report
10/19/2016 10:51:49 AM
GeneMarker V2.4.0
Page 72
Sample 72: 57262016-09-22-16-13-3916-13-39.fsa Run date and time: 09/22/2016 - 16:14:24 -> 09/22/2016 - 17:01:34

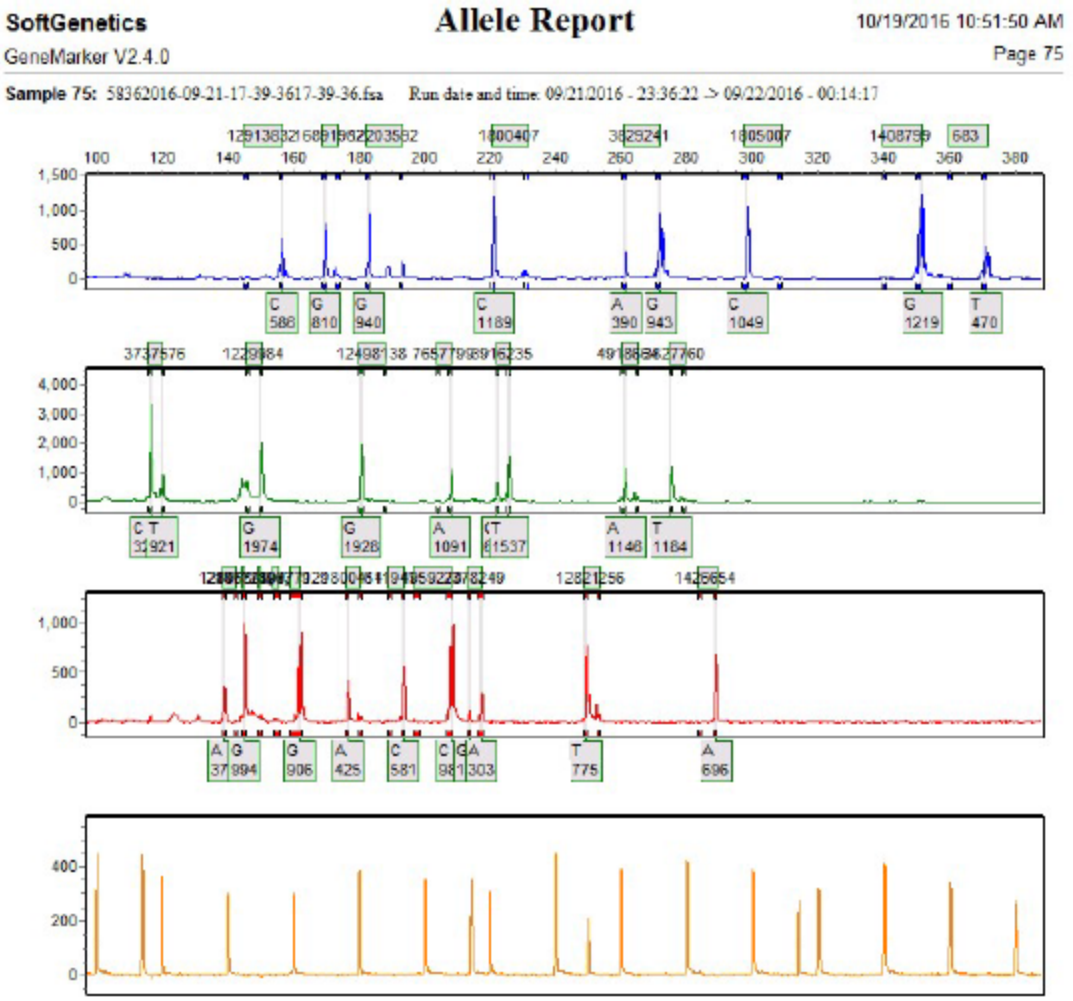
SoftGenetics **Allele Report** 10/19/2016 10:51:49 AM

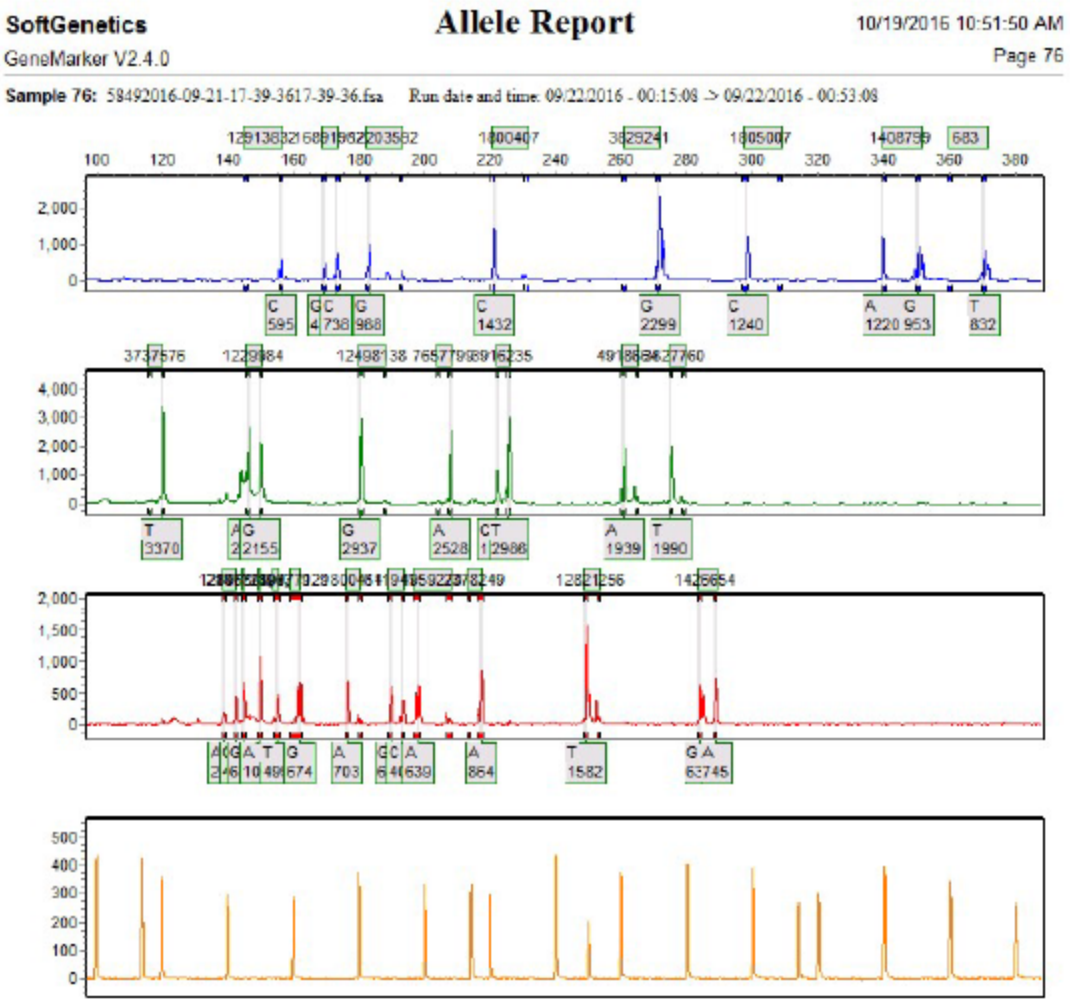
GeneMarker V2.4.0 Page 73

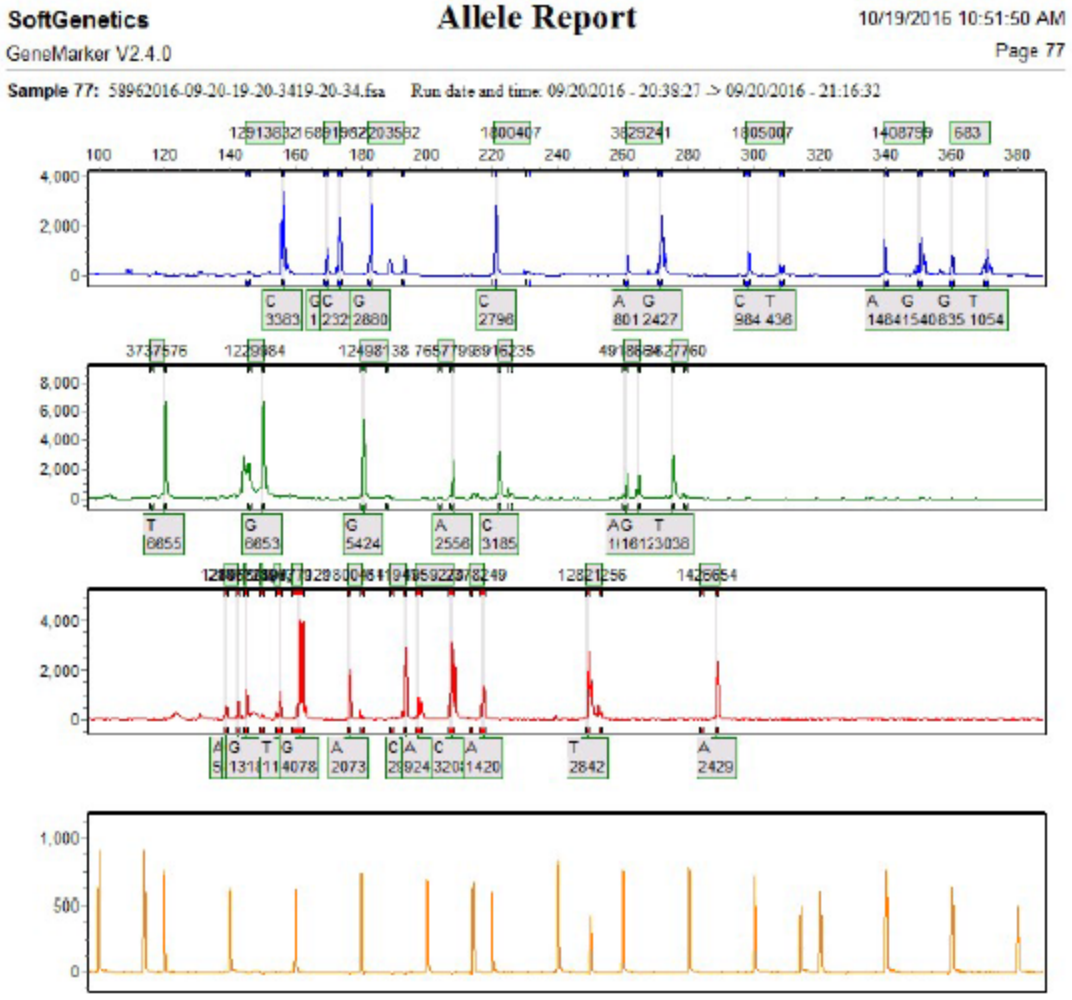
Sample 73: 58032016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 20:38:27 -> 09/20/2016 - 21:16:32

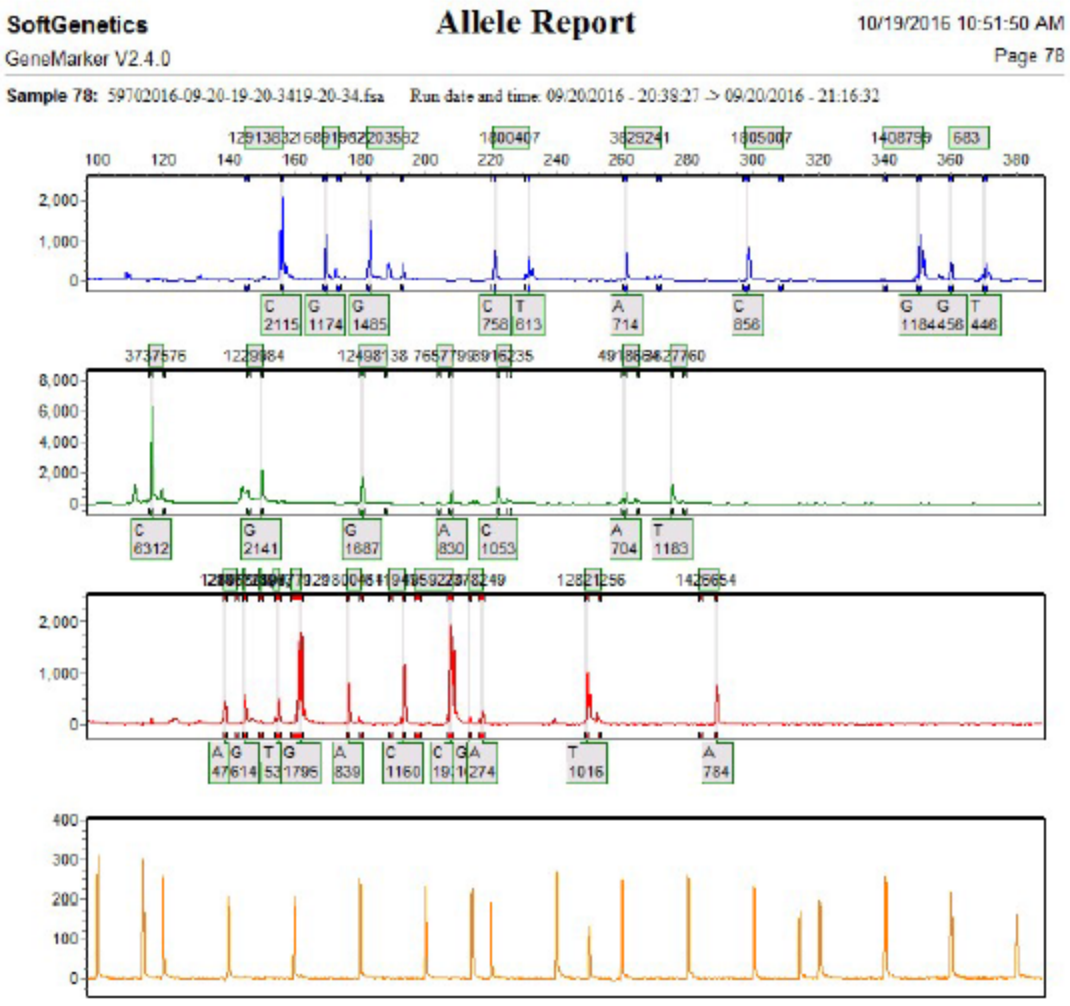



SoftGenetics
GeneMarker V2.4.0
Allele Report
10/19/2016 10:51:50 AM
Page 74
Sample 74: 58322016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 20:39:27 -> 09/20/2016 - 21:16:32


SoftGenetics
Allele Report
10/19/2016 10:51:50 AM
GeneMarker V2.4.0
Page 75
Sample 75: 58362016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 23:36:22 -> 09/22/2016 - 00:14:17


SoftGenetics
Allele Report
10/19/2016 10:51:50 AM
GeneMarker V2.4.0
Page 76
Sample 76: 58492016-09-21-17-39-3617-39-36.fsa Run date and time: 09/22/2016 - 00:15:08 -> 09/22/2016 - 00:53:08

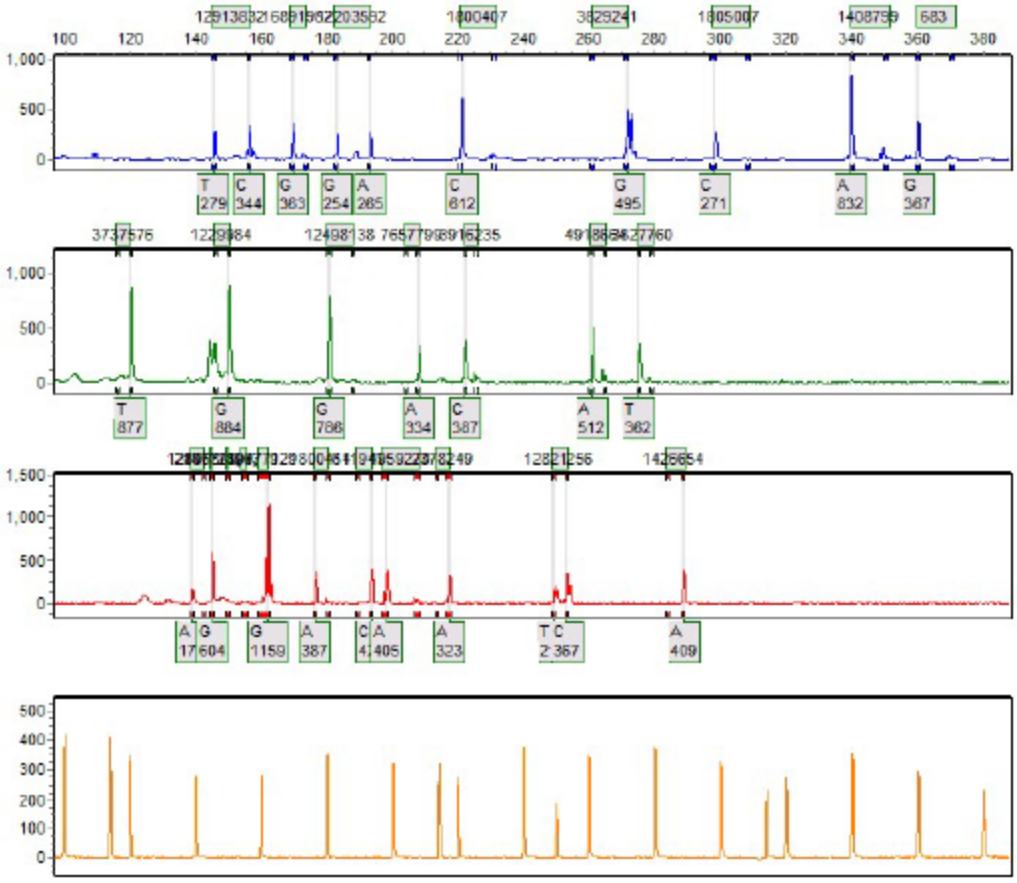

SoftGenetics
Allele Report
10/19/2016 10:51:50 AM
GeneMarker V2.4.0
Page 77
Sample 77: 58962016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 20:38:27 -> 09/20/2016 - 21:16:32


SoftGenetics
Allele Report
10/19/2016 10:51:50 AM
GeneMarker V2.4.0
Page 78
Sample 78: 59702016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 20:38:27 -> 09/20/2016 - 21:16:32

SoftGenetics **Allele Report** 10/19/2016 10:51:50 AM

GeneMarker V2.4.0 Page 79

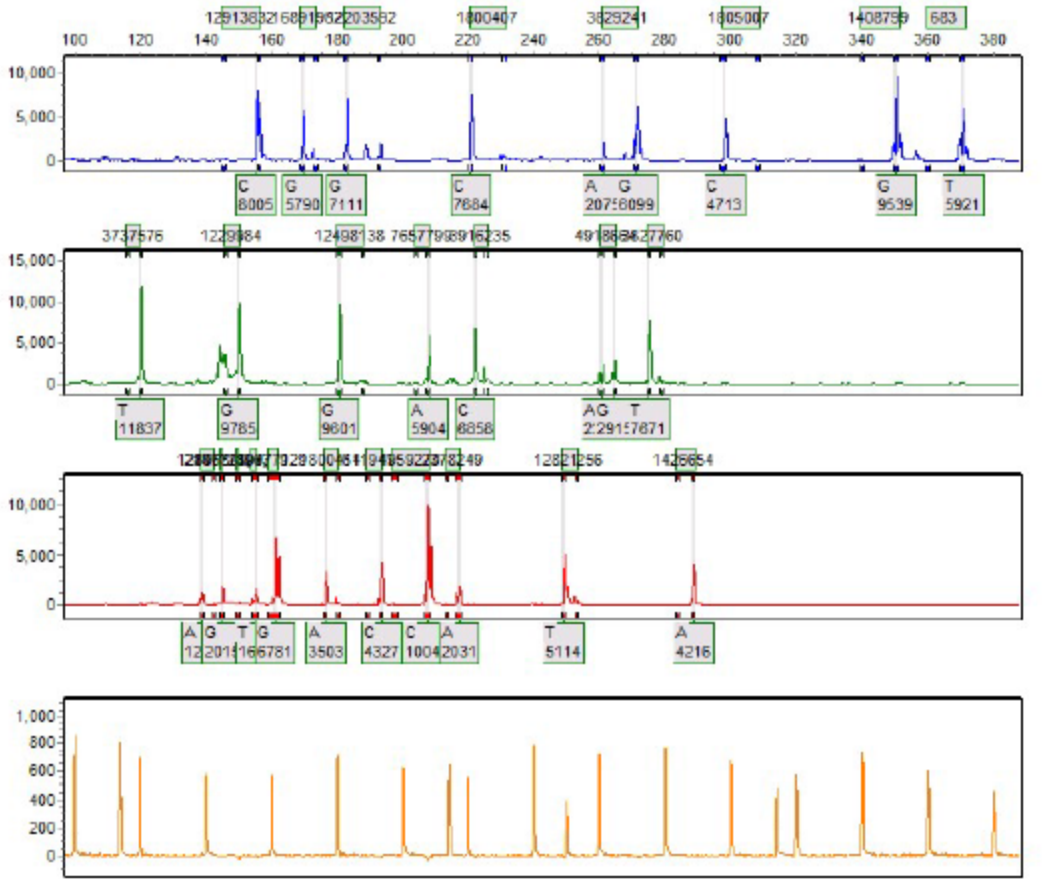
Sample 79: 60432016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 20:38:27 -> 09/20/2016 - 21:16:32

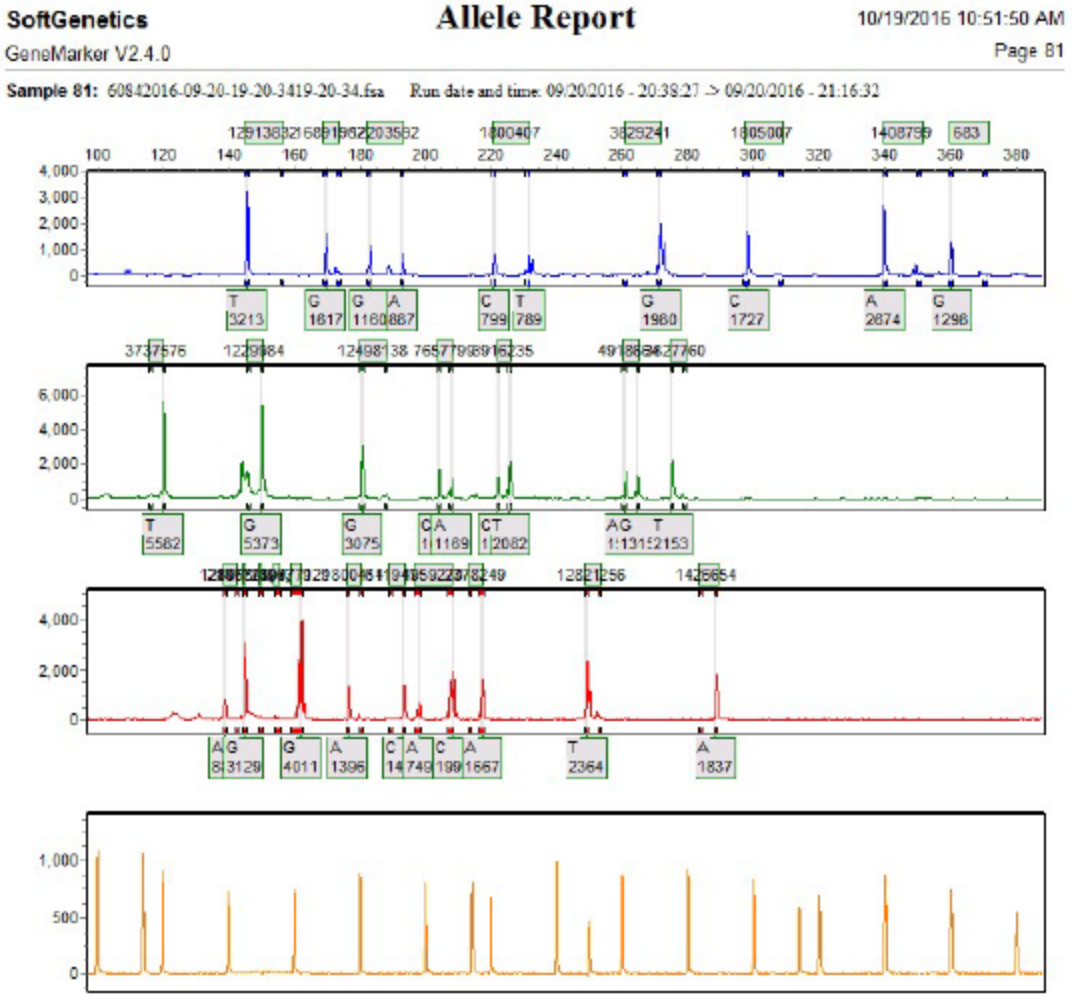


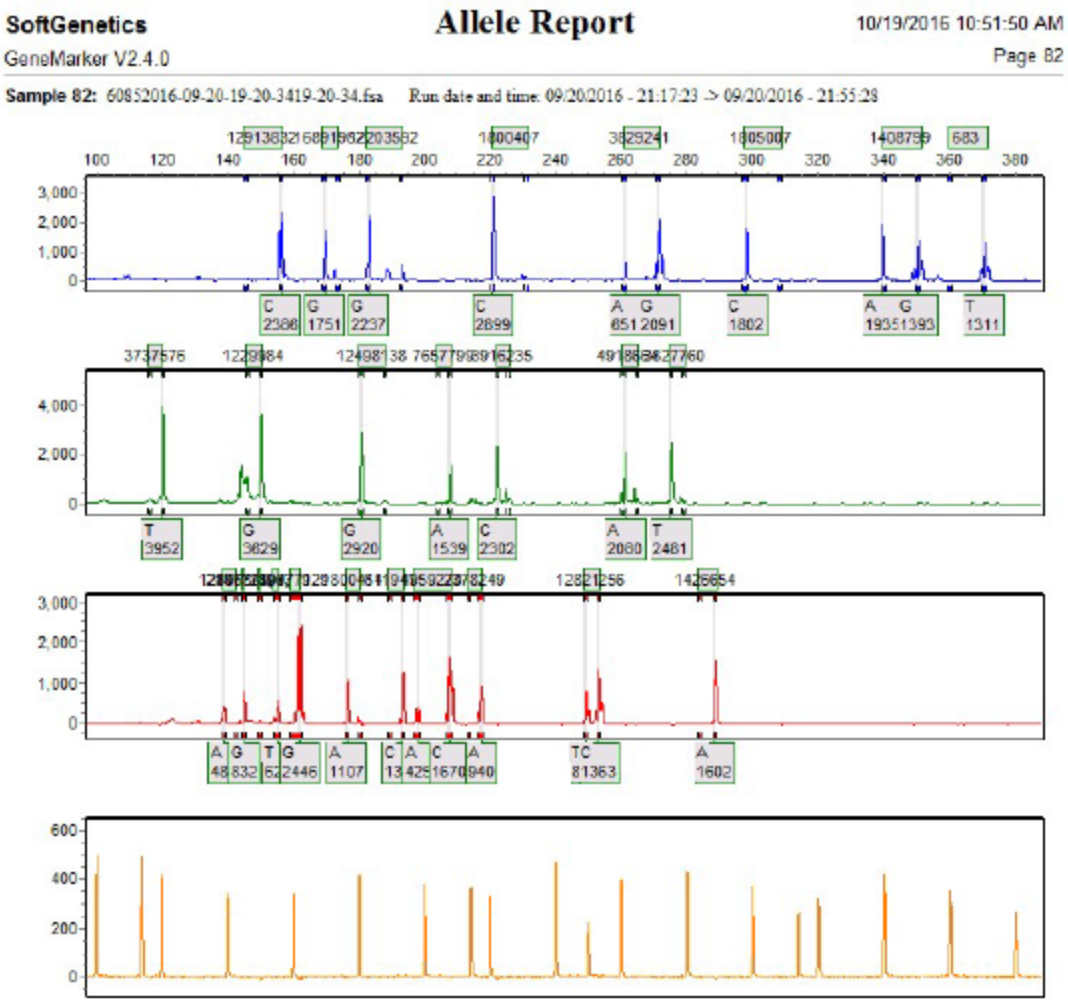
SoftGenetics **Allele Report** 10/19/2016 10:51:50 AM

GeneMarker V2.4.0 Page 80

Sample 80: 60572016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 20:38:27 -> 09/20/2016 - 21:16:32



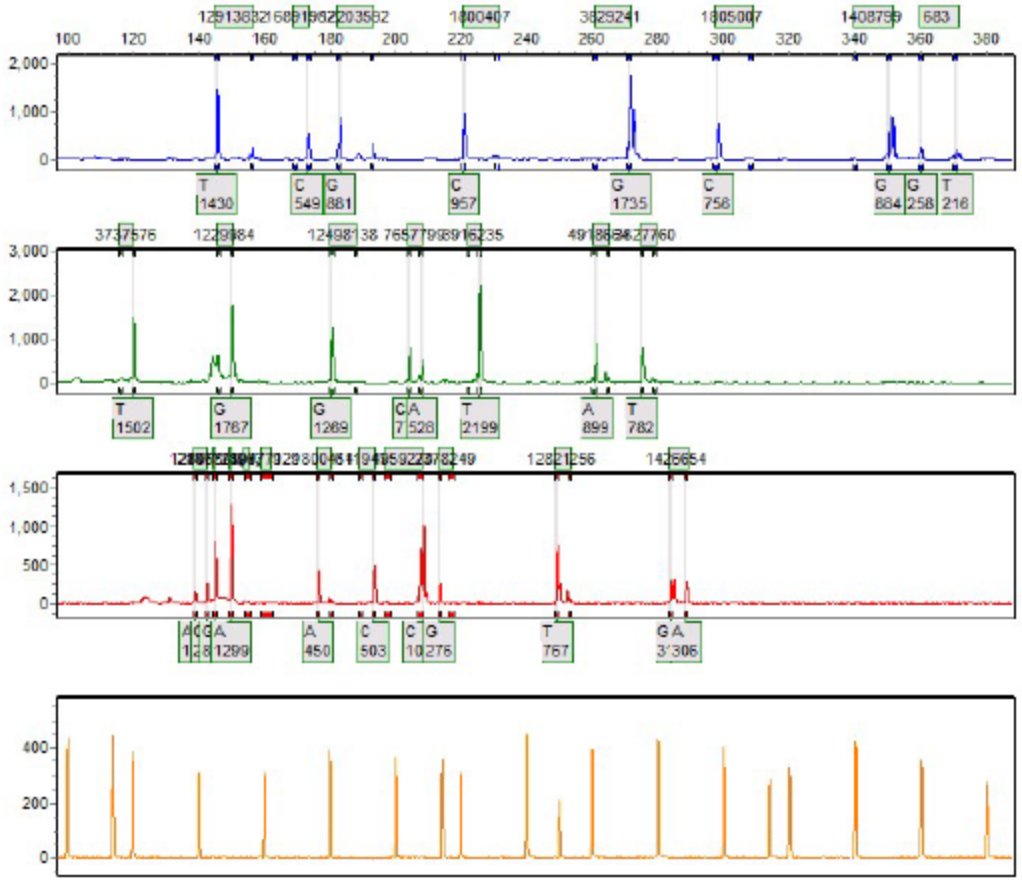

SoftGenetics
Allele Report
10/19/2016 10:51:50 AM
GeneMarker V2.4.0
Page 81
Sample 81: 6084 2016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 20:38:27 -> 09/20/2016 - 21:16:32

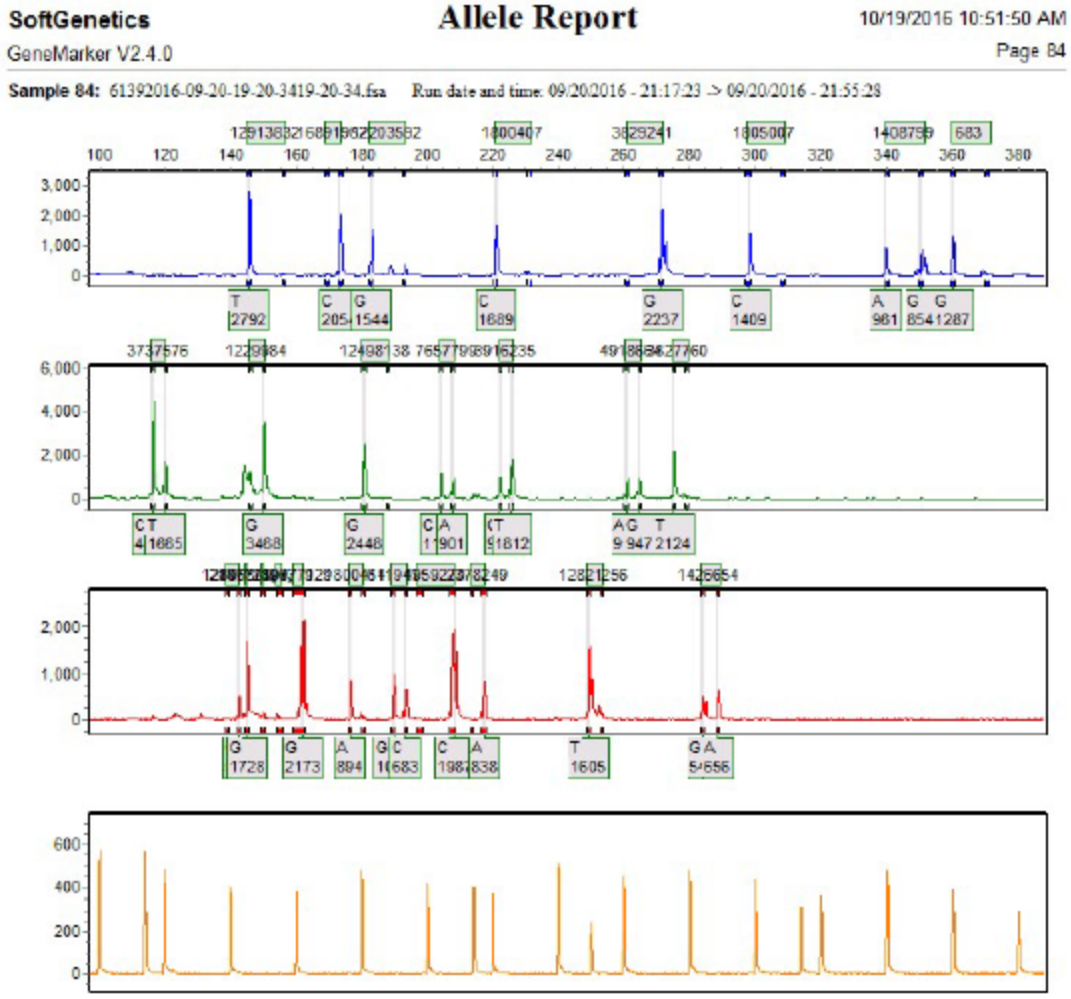

SoftGenetics
Allele Report
10/19/2016 10:51:50 AM
GeneMarker V2.4.0
Page 82
Sample 82: 60852016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 21:17:23 -> 09/20/2016 - 21:55:28

SoftGenetics **Allele Report** 10/19/2016 10:51:50 AM

GeneMarker V2.4.0 Page 83

Sample 83: 61352016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 22:57:21 -> 09/21/2016 - 23:35:31

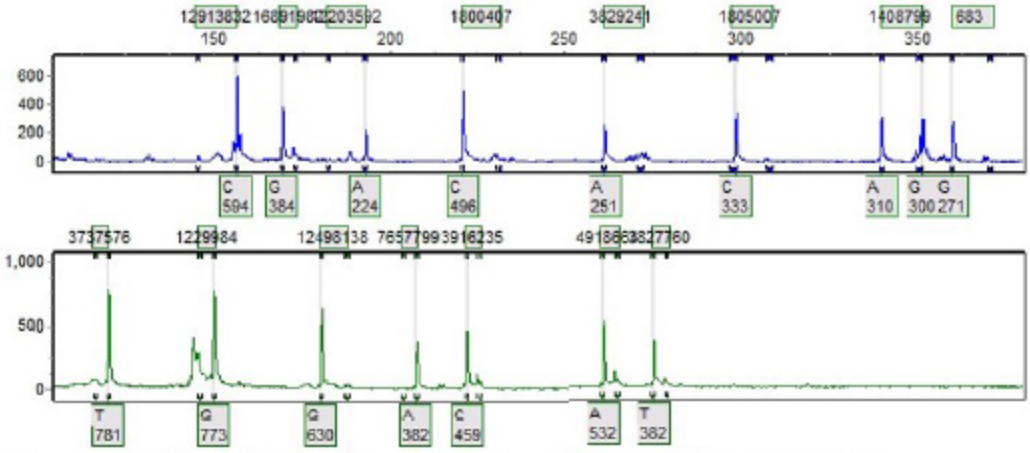



SoftGenetics
Allele Report
10/19/2016 10:51:50 AM
GeneMarker V2.4.0
Page 84
Sample 84: 61392016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 21:17:23 -> 09/20/2016 - 21:55:28

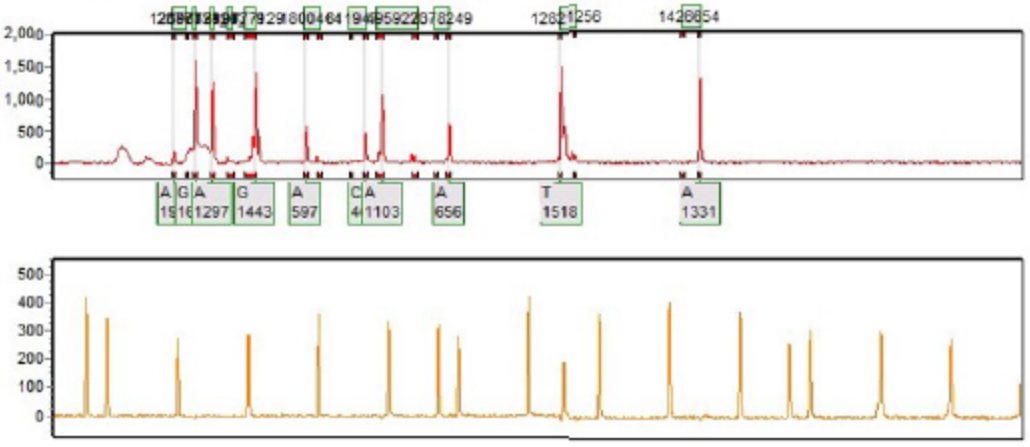
SoftGenetics **Allele Report** 10/19/2016 11:49:33 AM

GeneMarker V2.4.0 Page 8

Sample 8: 6149P2016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 21:17:23 -> 09/20/2016 - 21:55:28



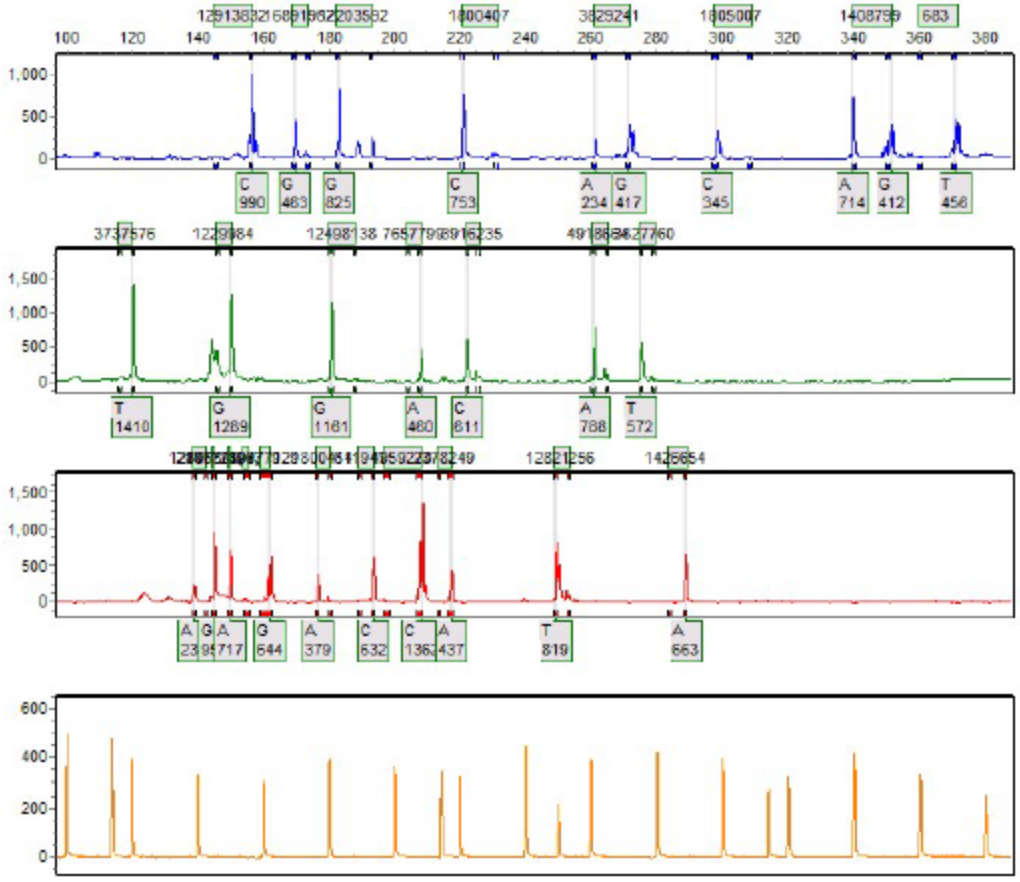
Sample 9: 6149P2016-10-07-07-32-3307-32-33.fsa Run date and time: 10/07/2016 - 07:33:16 -> 10/07/2016 - 08:21:37

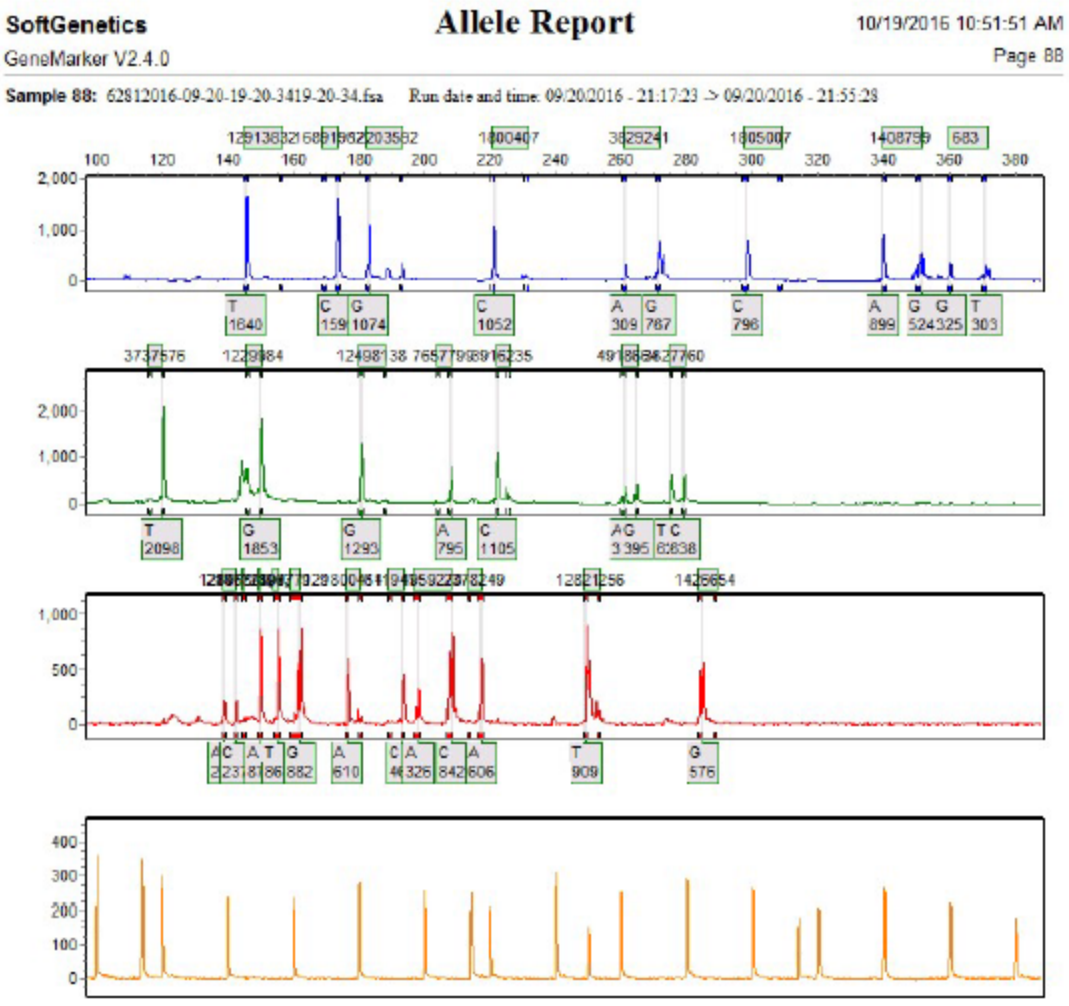


SoftGenetics **Allele Report** 10/19/2016 10:51:51 AM

GeneMarker V2.4.0 Page 87

Sample 87: 62132016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 21:17:23 -> 09/20/2016 - 21:55:28

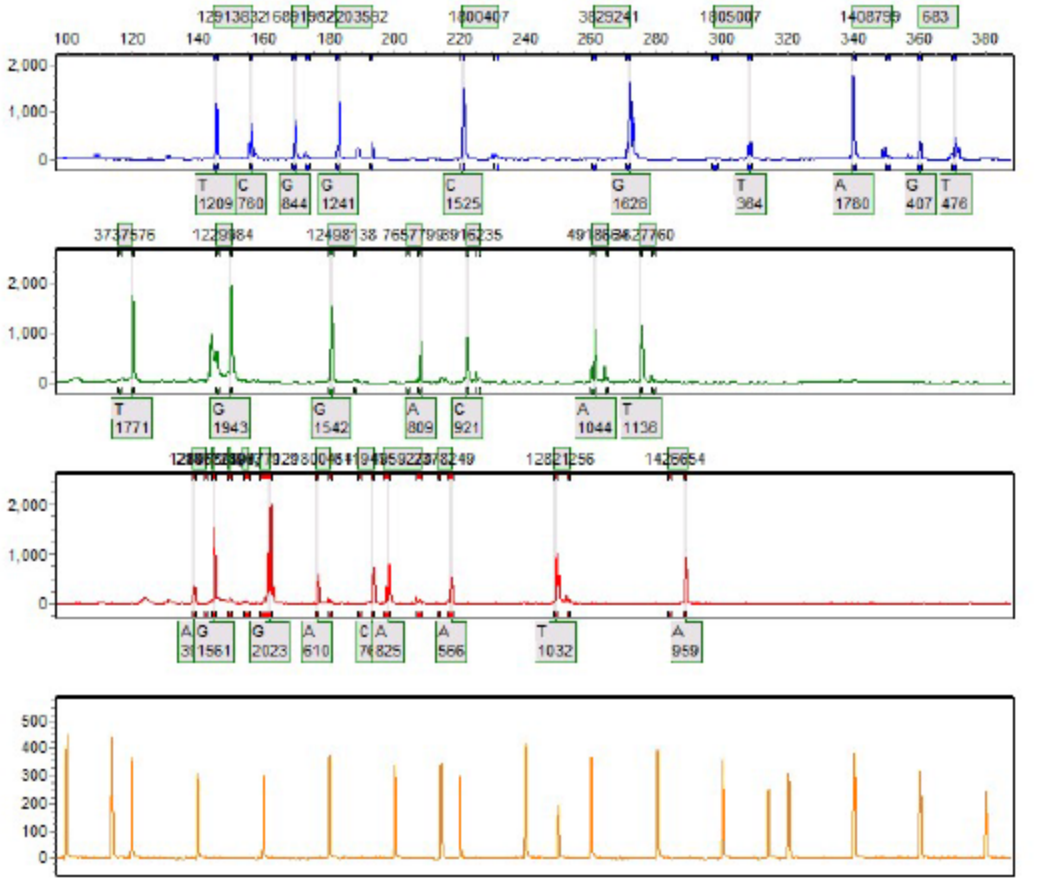


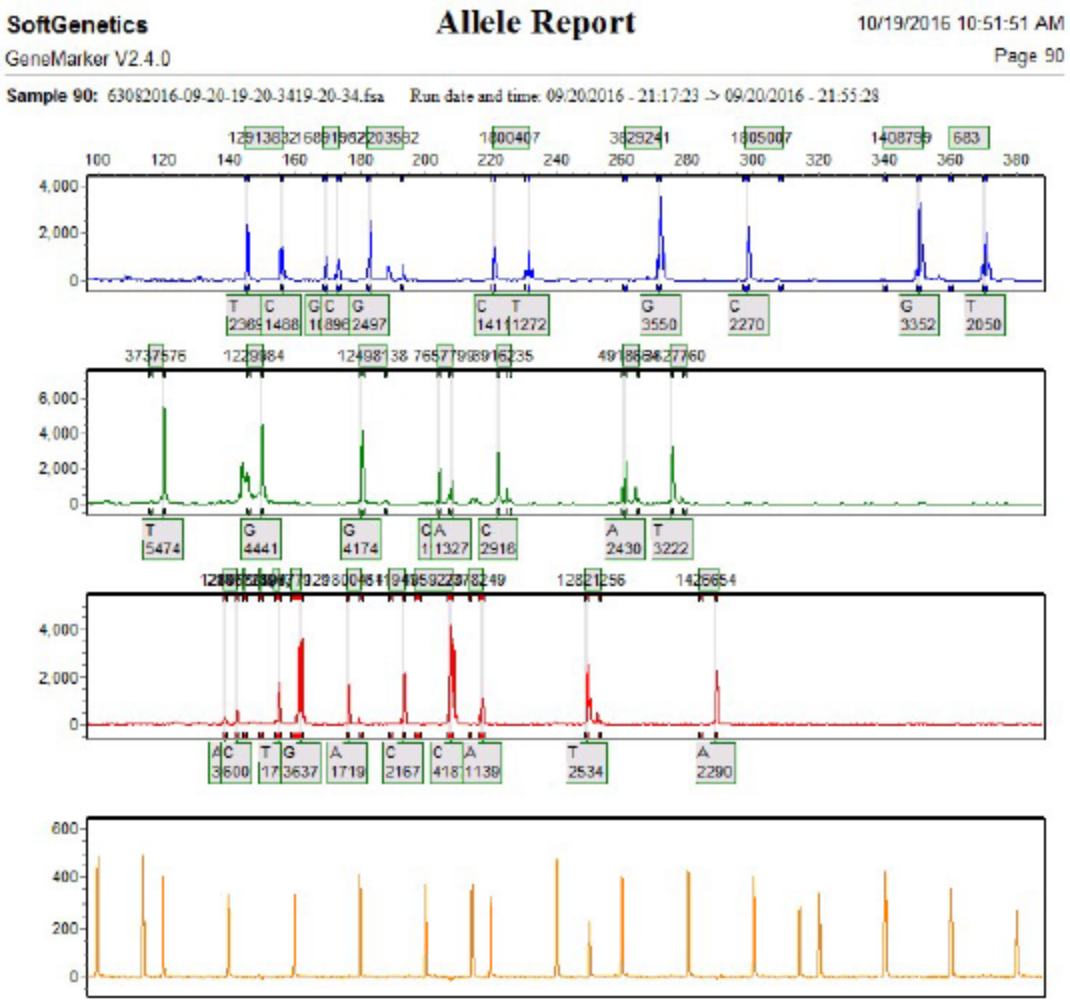

SoftGenetics
Allele Report
10/19/2016 10:51:51 AM
GeneMarker V2.4.0
Page 88
Sample 88: 62812016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 21:17:23 -> 09/20/2016 - 21:55:28

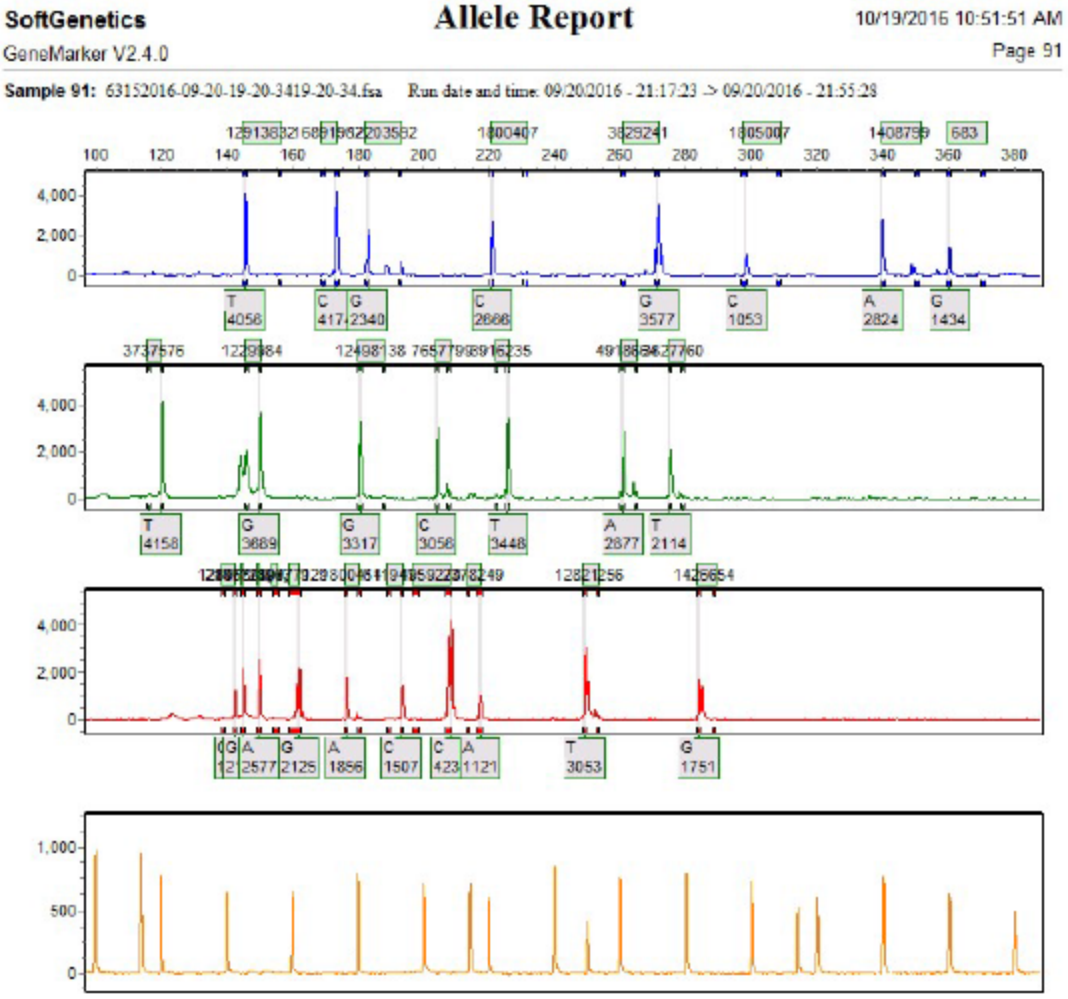
SoftGenetics **Allele Report** 10/19/2016 10:51:51 AM

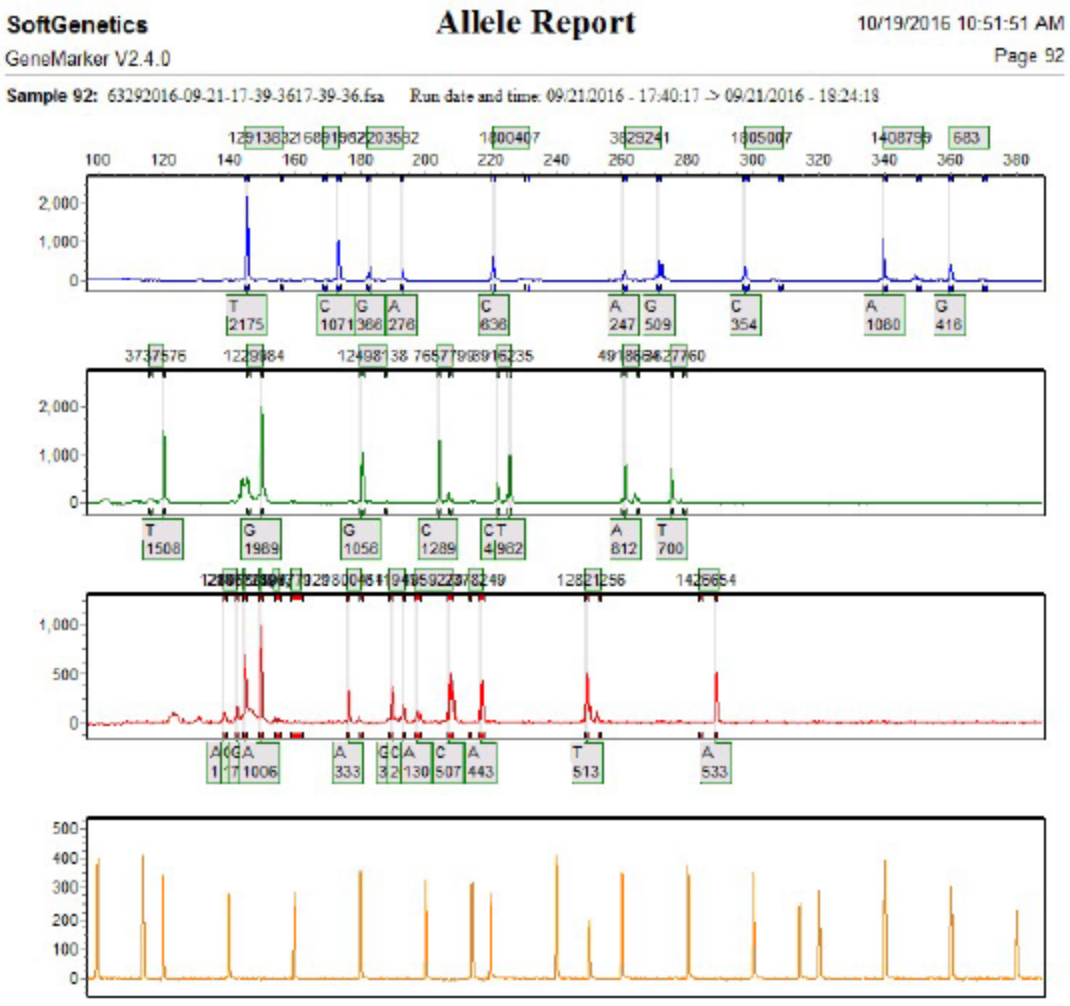
GeneMarker V2.4.0 Page 89

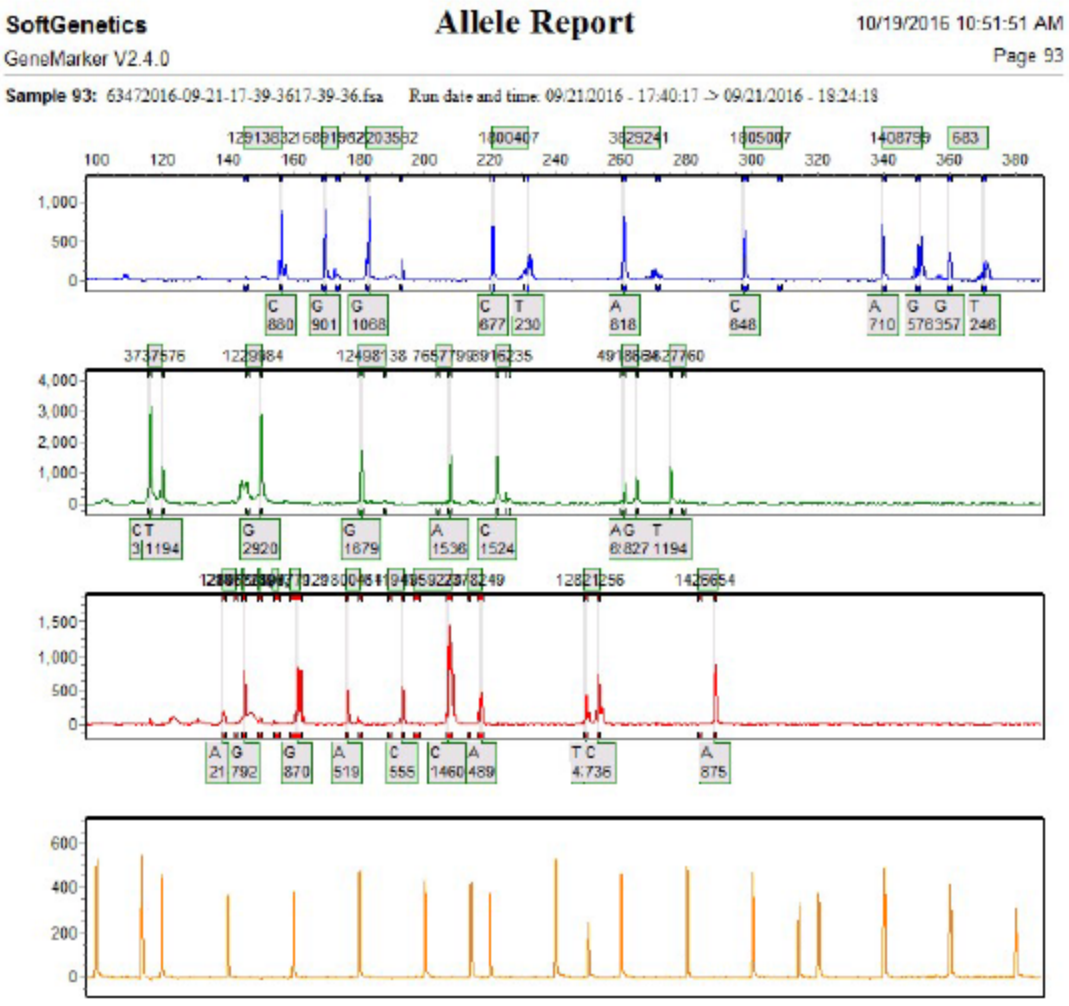
Sample 89: 63052016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 21:17:23 -> 09/20/2016 - 21:55:28




SoftGenetics
Allele Report
10/19/2016 10:51:51 AM
GeneMarker V2.4.0
Page 90
Sample 90: 63082016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 21:17:23 -> 09/20/2016 - 21:55:28


SoftGenetics
Allele Report
10/19/2016 10:51:51 AM
GeneMarker V2.4.0
Page 91
Sample 91: 63152016-09-20-19-20-3419-20-34.fsa Run date and time: 09/20/2016 - 21:17:23 -> 09/20/2016 - 21:55:28
100 120 140 160 180 200 220 240 260 280 300 320 340 360 380
4,000
2,000
0
T 4058
C 417
G 2340
C 2666
G 3577
C 1053
A 2824
G 1434
T 4158
G 3689
G 3317
C 3056
T 3448
A 2877
T 2114
G 12
A 2577
G 2125
A 1856
C 1507
C 423
A 1121
T 3053
G 1751
1,000
500
0

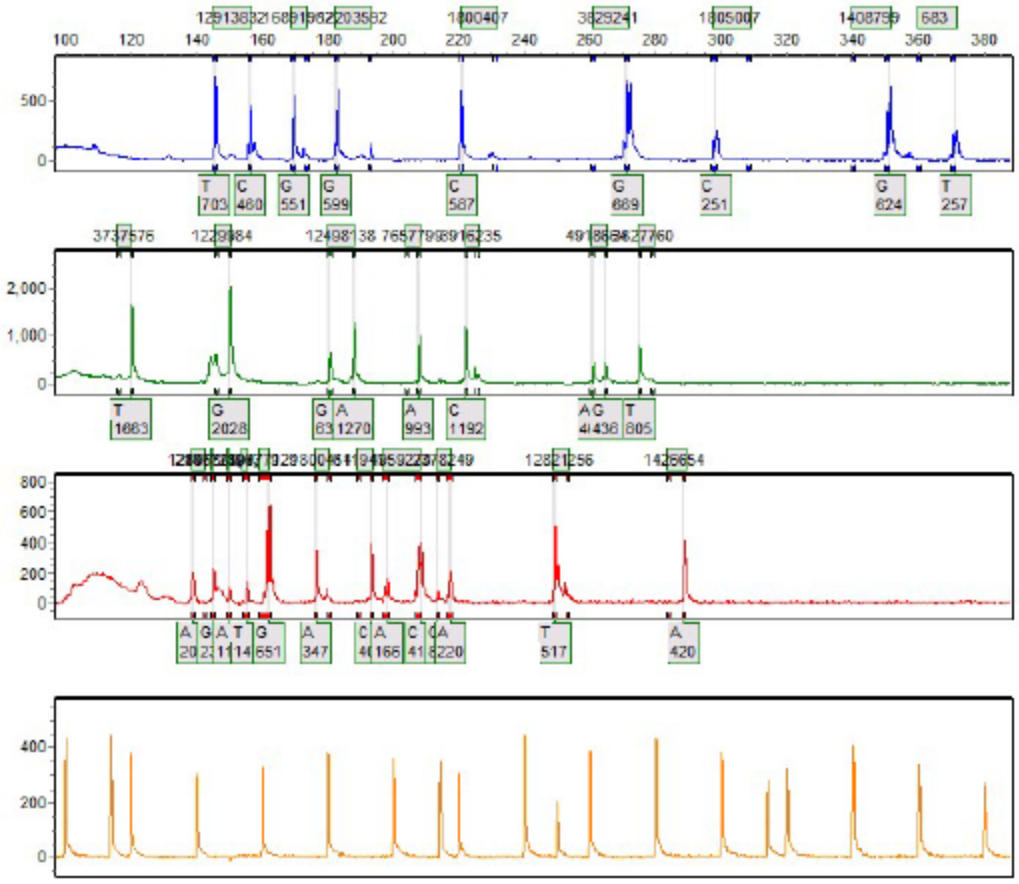

SoftGenetics
Allele Report
10/19/2016 10:51:51 AM
GeneMarker V2.4.0
Page 92
Sample 92: 63292016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 17:40:17 -> 09/21/2016 - 18:24:18


SoftGenetics
Allele Report
10/19/2016 10:51:51 AM
GeneMarker V2.4.0
Page 93
Sample 93: 63472016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 17:40:17 -> 09/21/2016 - 18:24:18

SoftGenetics **Allele Report** 10/19/2016 10:51:51 AM

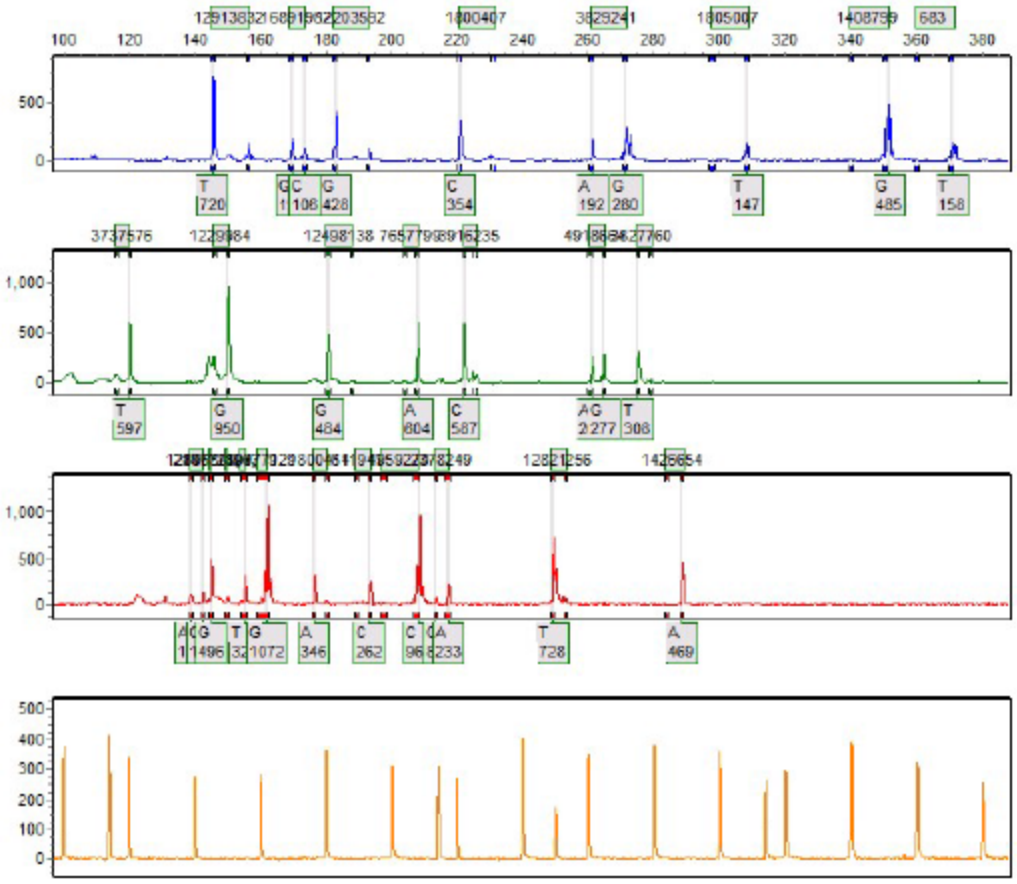
GeneMarker V2.4.0 Page 94

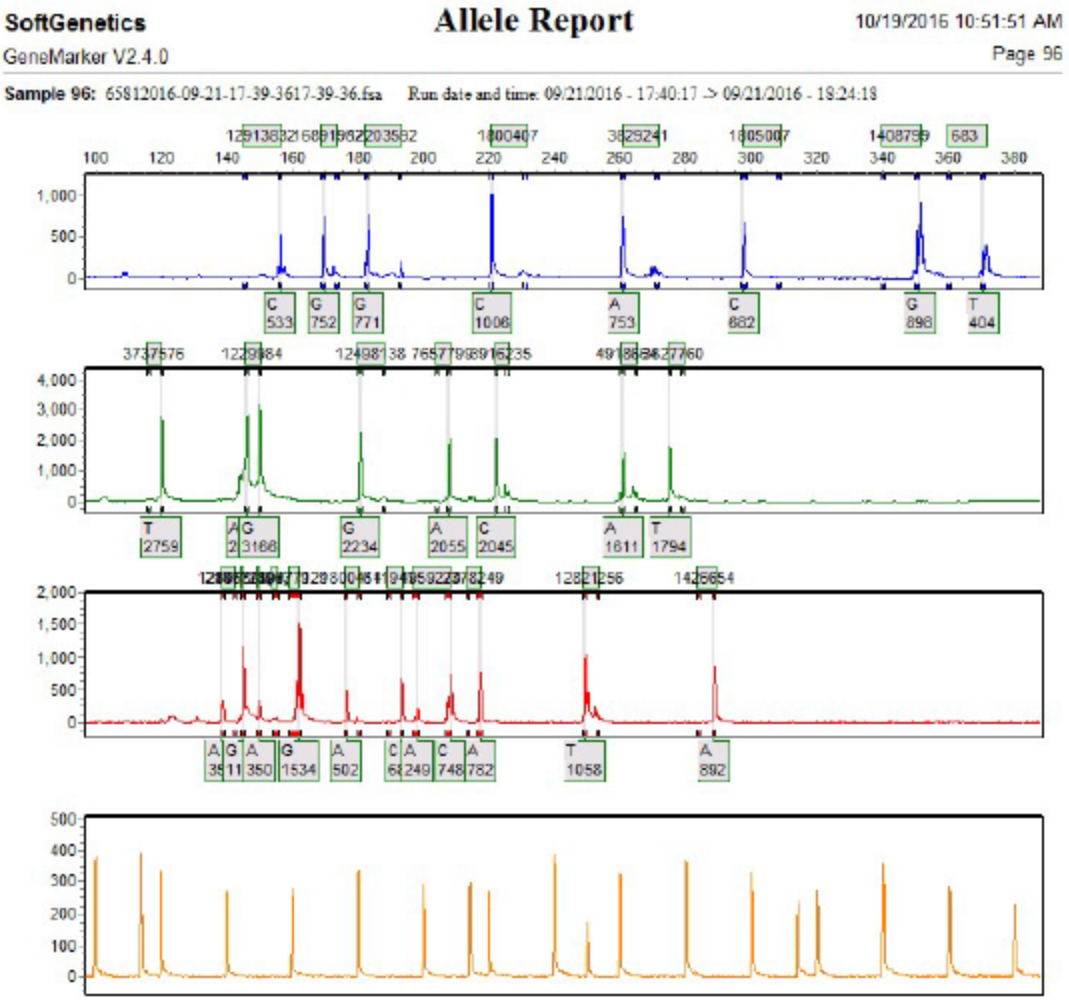
Sample 94: 63582016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 17:40:17 -> 09/21/2016 - 18:24:18



Allele Report

Sample 95: 63822016-09-22-16-13-3916-13-39.fsa Run date and time: 09/22/2016 - 18:20:20 -> 09/22/2016 - 18:58:21

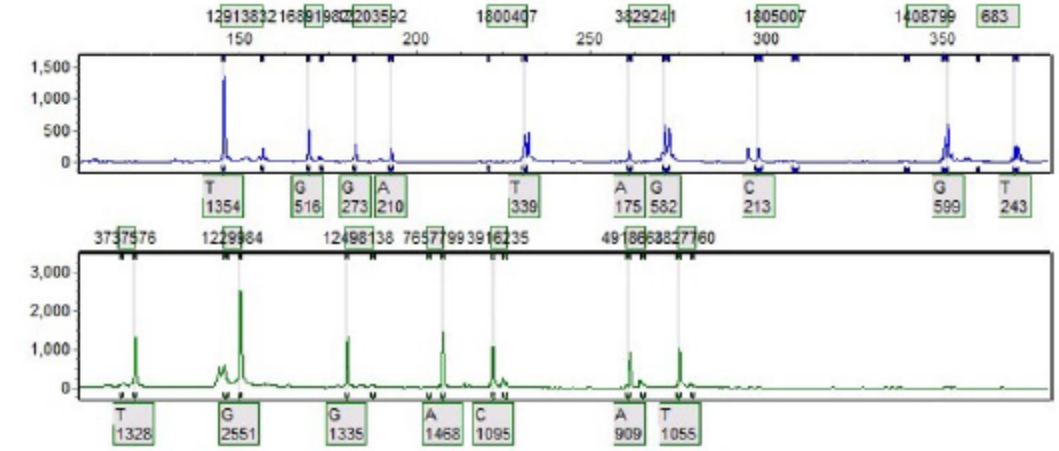



SoftGenetics
Allele Report
10/19/2016 10:51:51 AM
GeneMarker V2.4.0
Page 96
Sample 96: 65812016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 17:40:17 -> 09/21/2016 - 18:24:18
C 533
G 752
G 771
C 1008
A 753
C 882
G 898
T 404
T 2759
A G 2 3166
G 2234
A 2055
C 2045
A 1811
T 1794
A 350
G 1534
A 502
T 1058
A 892

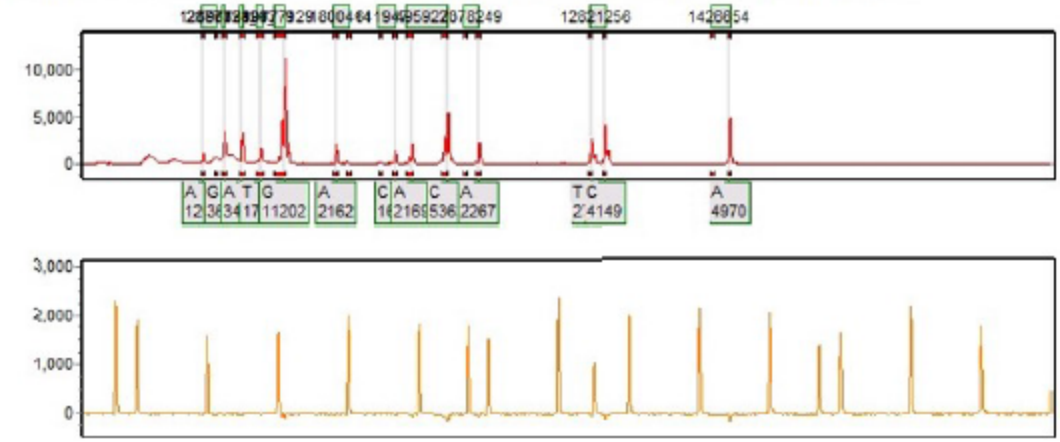
SoftGenetics **Allele Report** 10/19/2016 11:49:33 AM

GeneMarker V2.4.0 Page 10

Sample 10: 6789P2016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 17:40:17 -> 09/21/2016 - 18:24:18



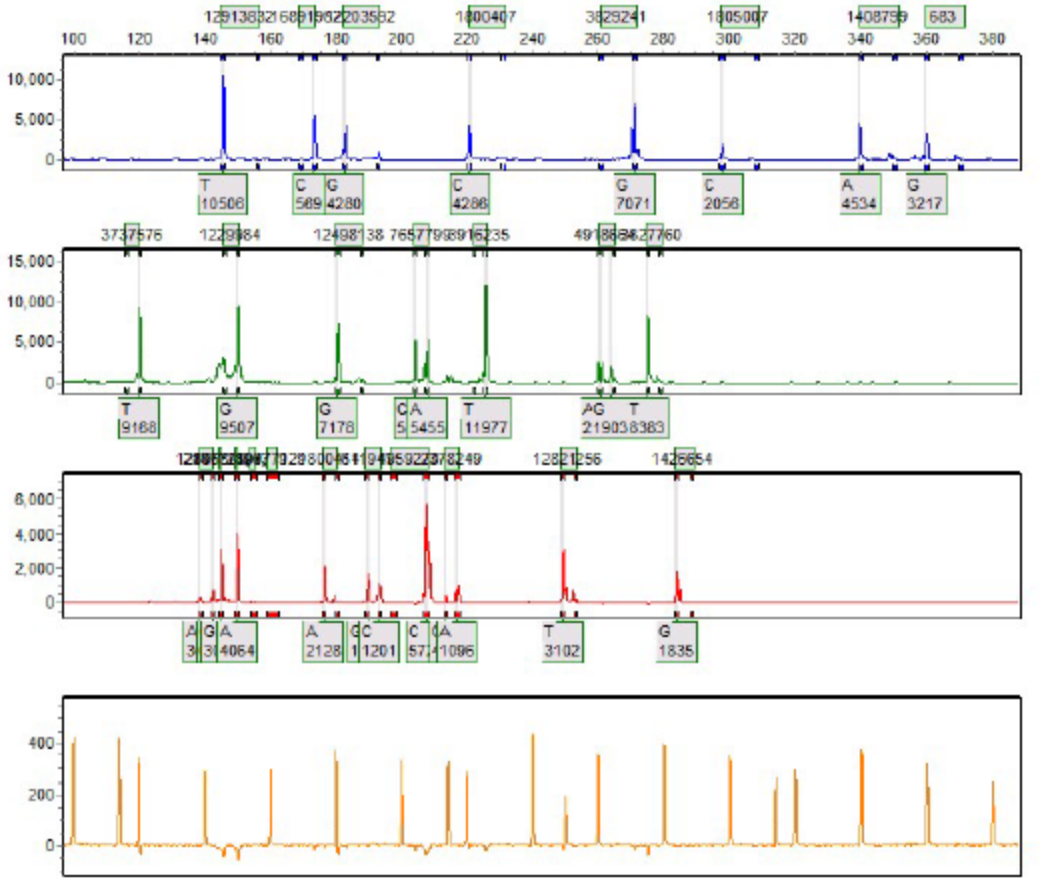
Sample 11: 6789P2016-10-07-07-32-3307-32-33.fsa Run date and time: 10/07/2016 - 08:22:28 -> 10/07/2016 - 09:00:13

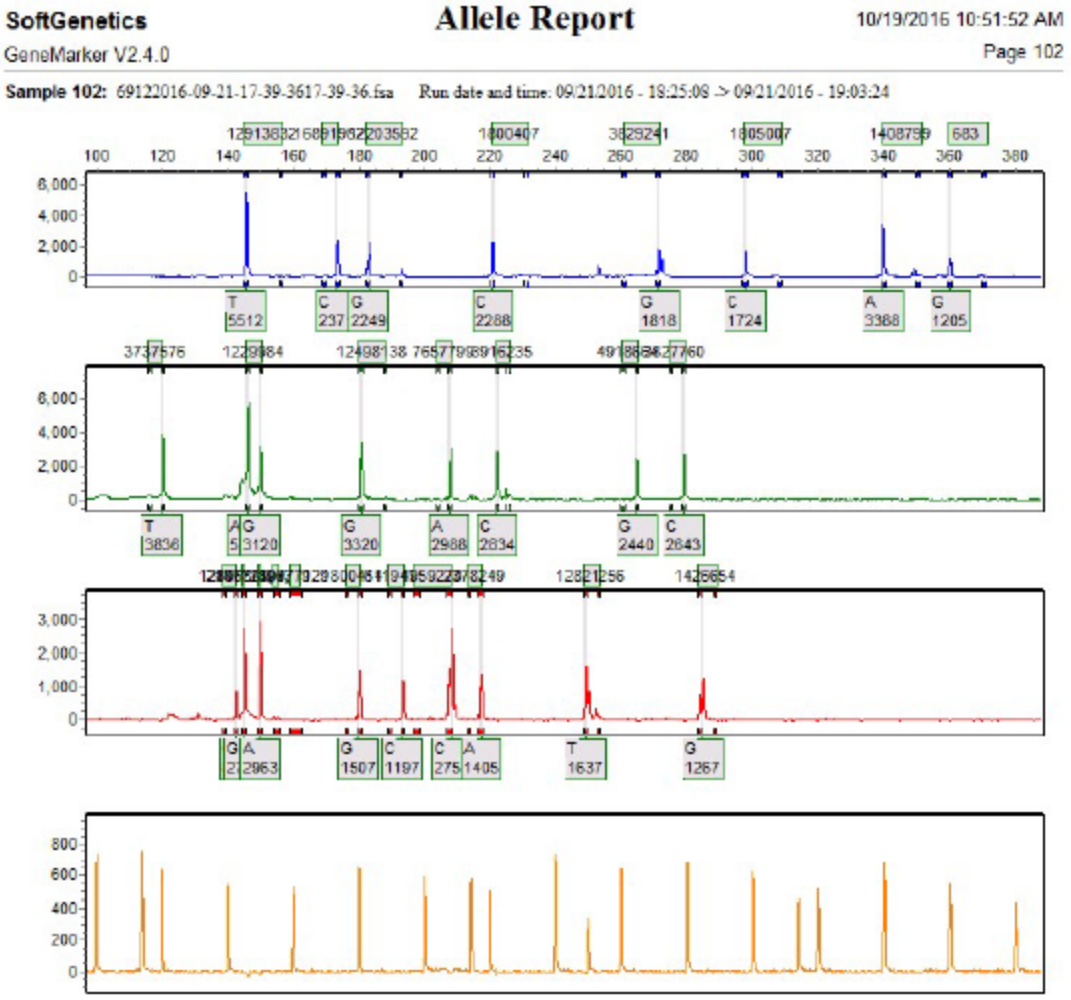


SoftGenetics **Allele Report** 10/19/2016 10:51:52 AM

GeneMarker V2.4.0 Page 101

Sample 101: 69012016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 17:40:17 -> 09/21/2016 - 18:24:18

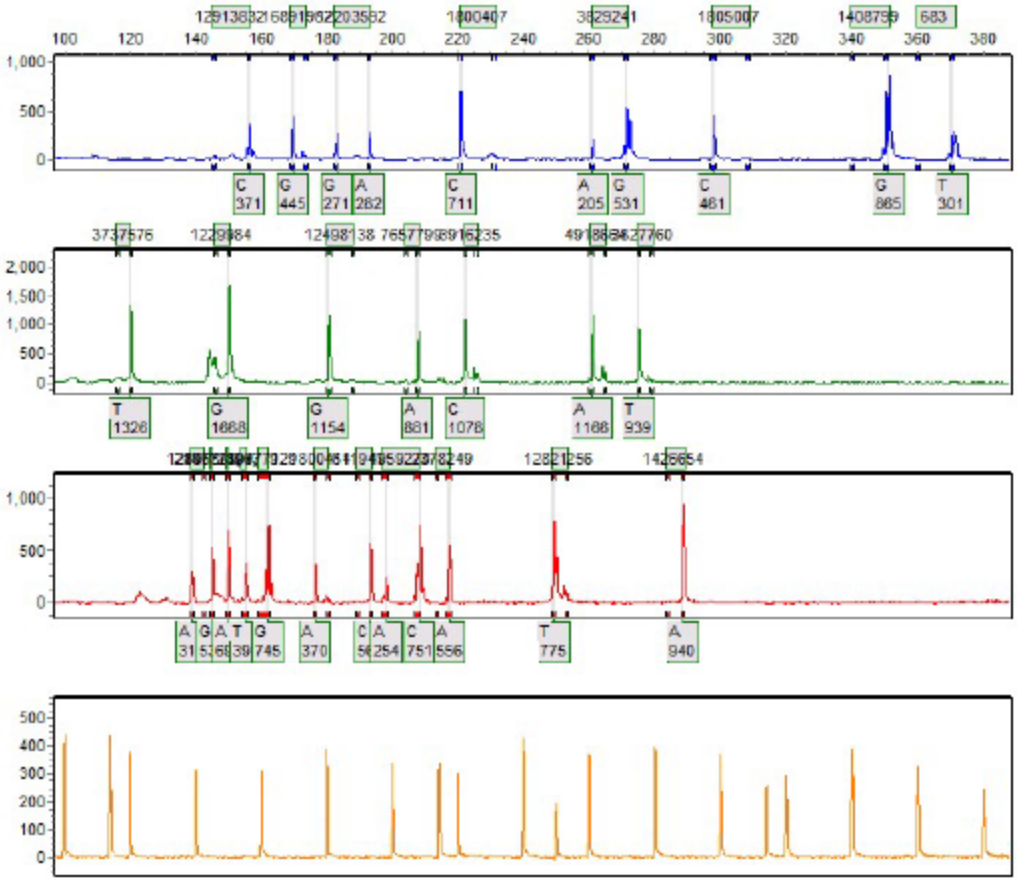



SoftGenetics
Allele Report
10/19/2016 10:51:52 AM
GeneMarker V2.4.0
Page 102
Sample 102: 69122016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 18:25:08 -> 09/21/2016 - 19:03:24

SoftGenetics **Allele Report** 10/19/2016 10:51:52 AM

GeneMarker V2.4.0 Page 103

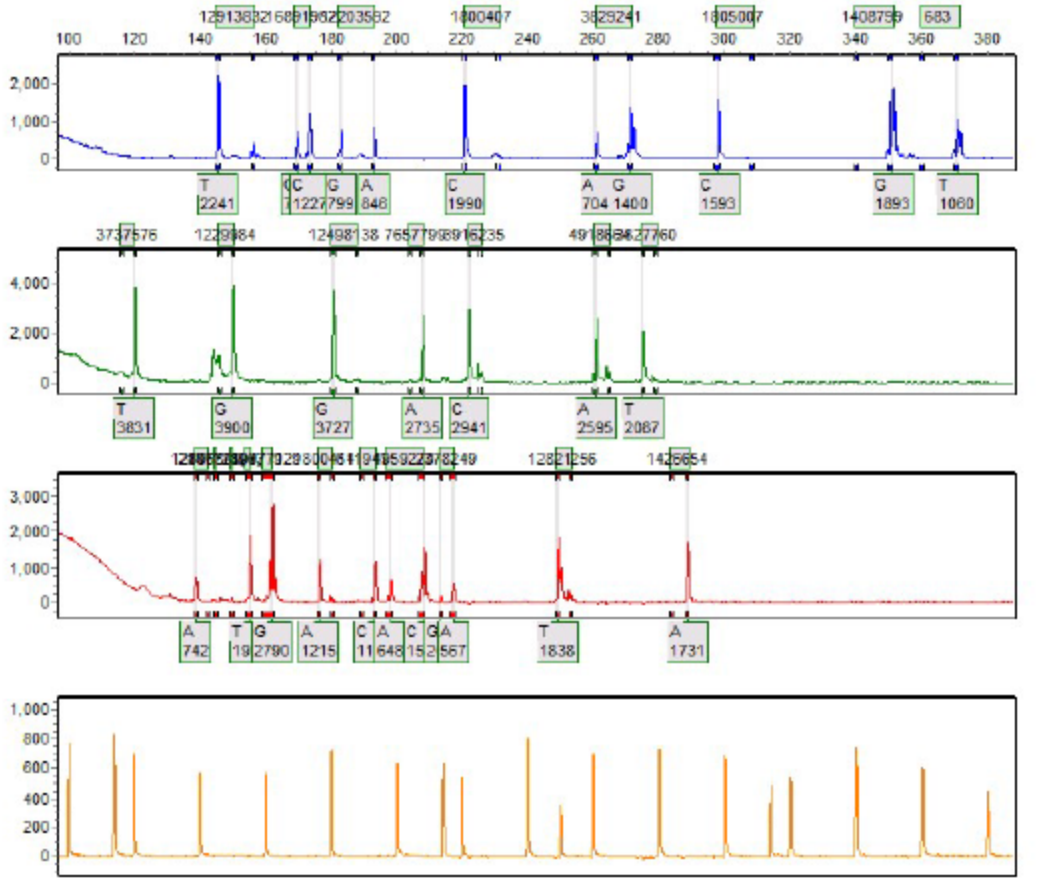
Sample 103: 69432016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 18:25:08 -> 09/21/2016 - 19:03:24



SoftGenetics **Allele Report** 10/19/2016 10:51:52 AM

GeneMarker V2.4.0 Page 104

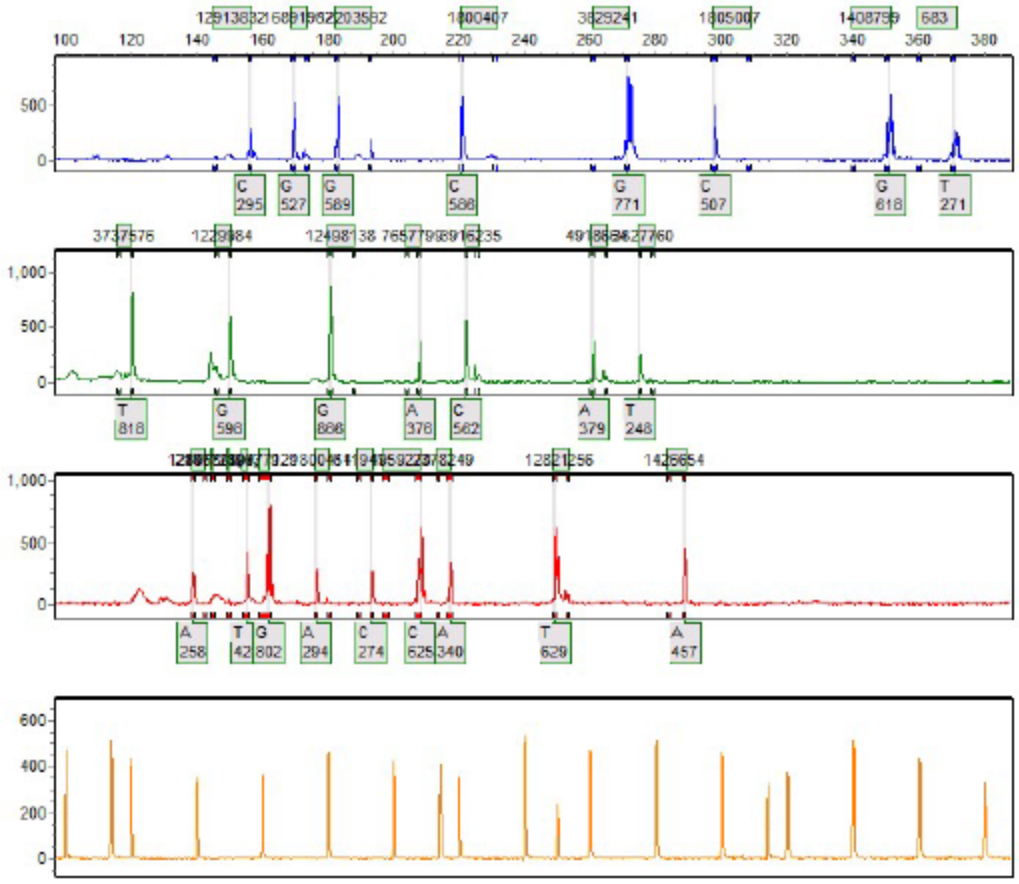
Sample 104: 69482016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 18:25:08 -> 09/21/2016 - 19:03:24

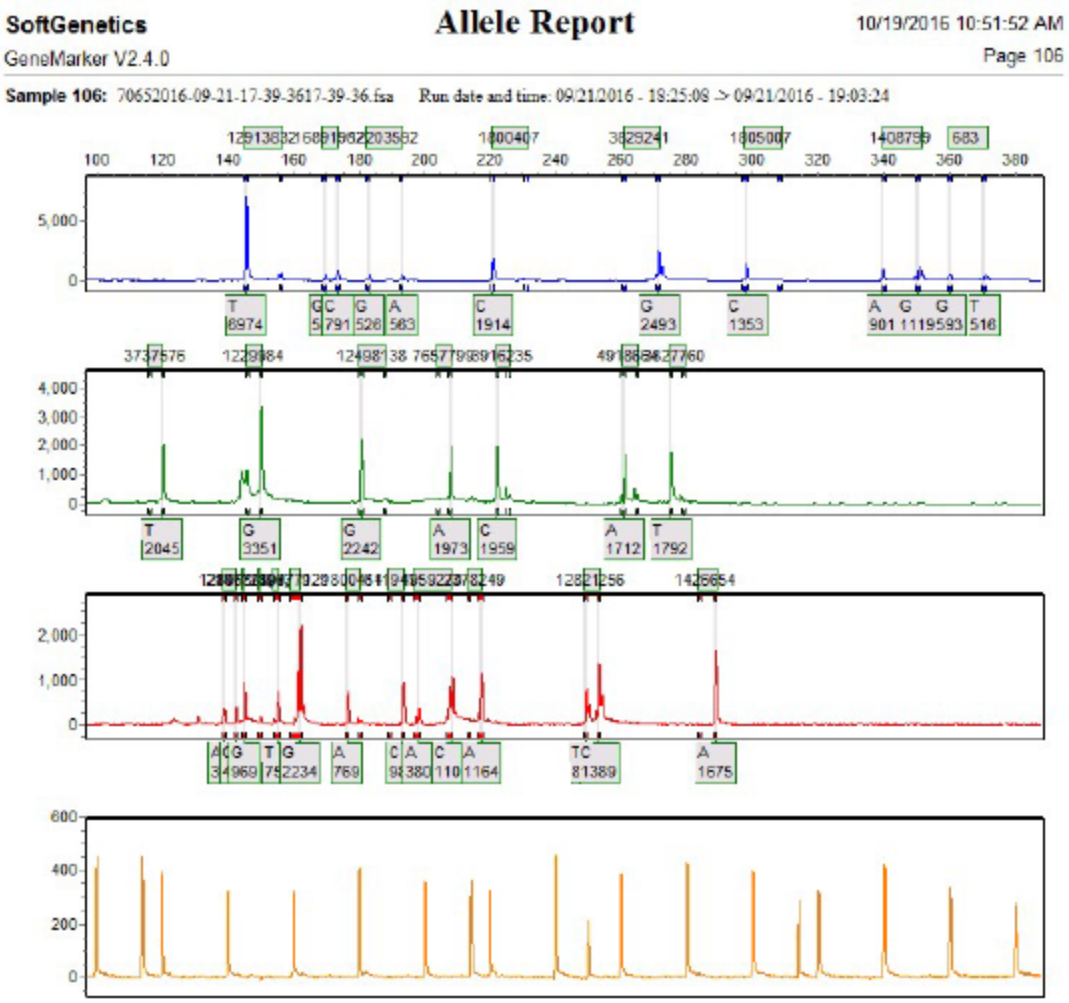


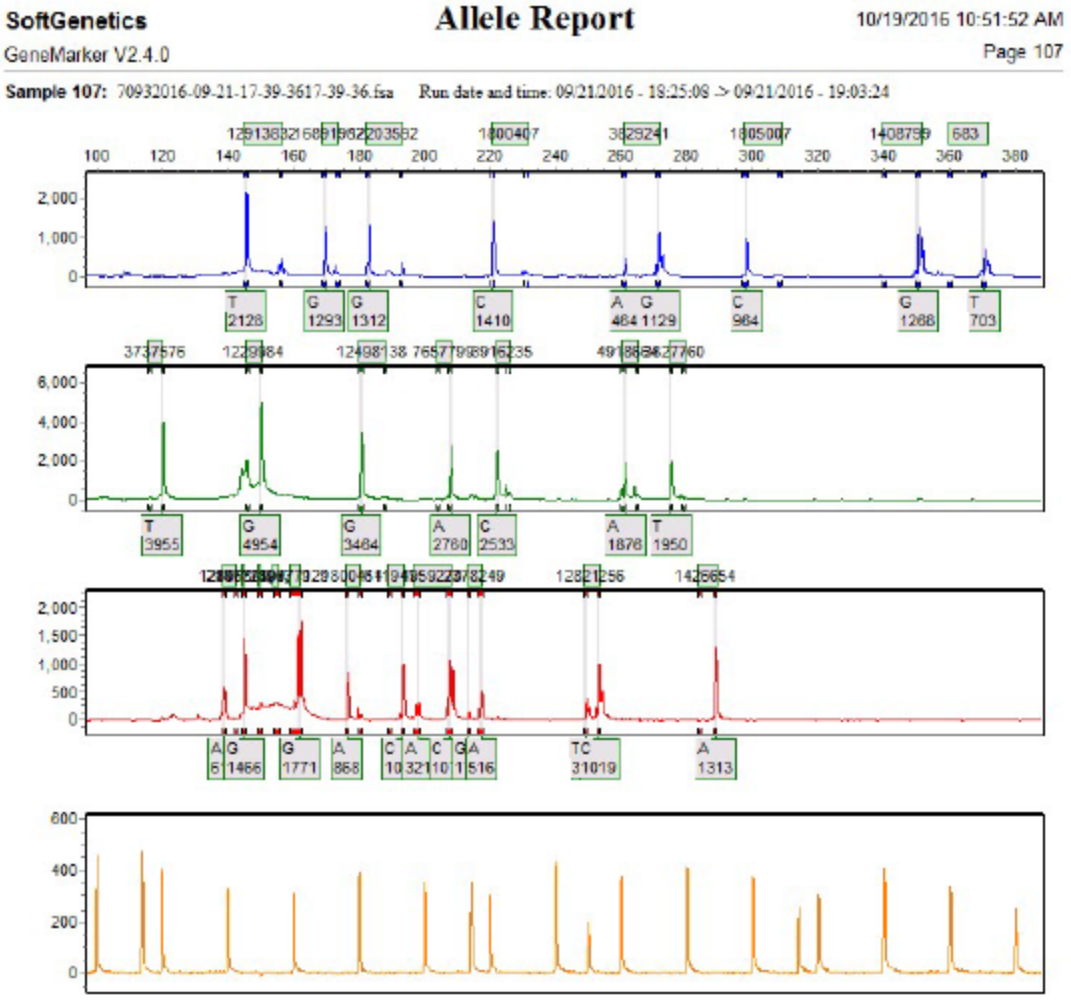
SoftGenetics **Allele Report** 10/19/2016 10:51:52 AM

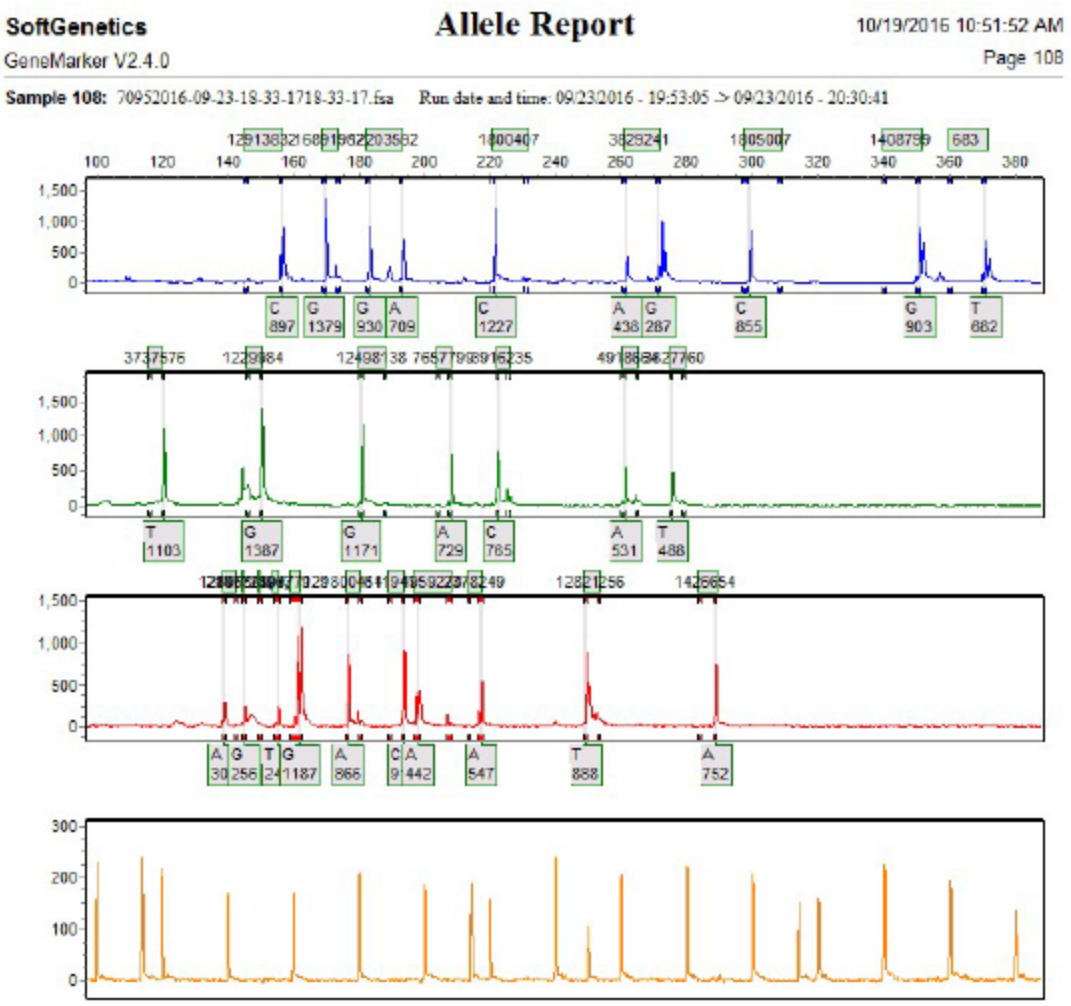
GeneMarker V2.4.0 Page 105

Sample 105: 70512016-09-22-16-13-3916-13-39.fsa Run date and time: 09/22/2016 - 17:02:24 -> 09/22/2016 - 17:40:30




SoftGenetics
Allele Report
10/19/2016 10:51:52 AM
GeneMarker V2.4.0
Page 106
Sample 106: 70652016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 18:25:08 -> 09/21/2016 - 19:03:24

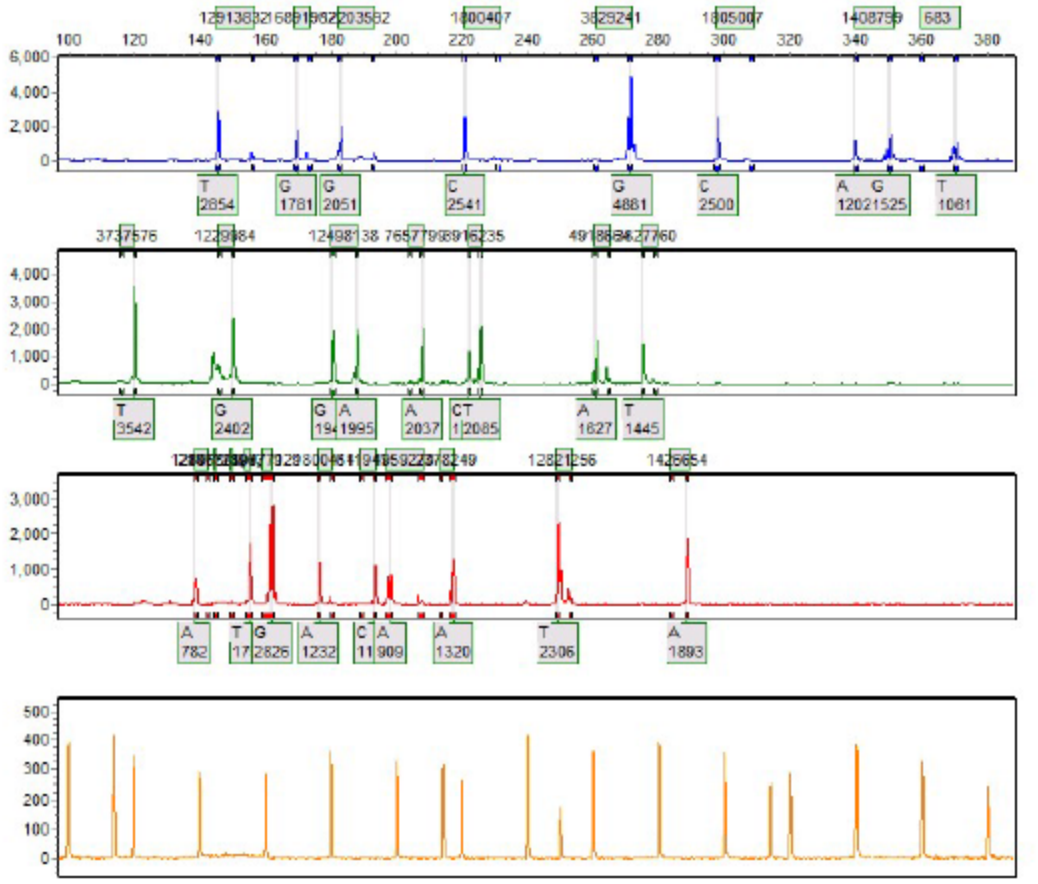

SoftGenetics
Allele Report
10/19/2016 10:51:52 AM
GeneMarker V2.4.0
Page 107
Sample 107: 70932016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 18:25:08 -> 09/21/2016 - 19:03:24

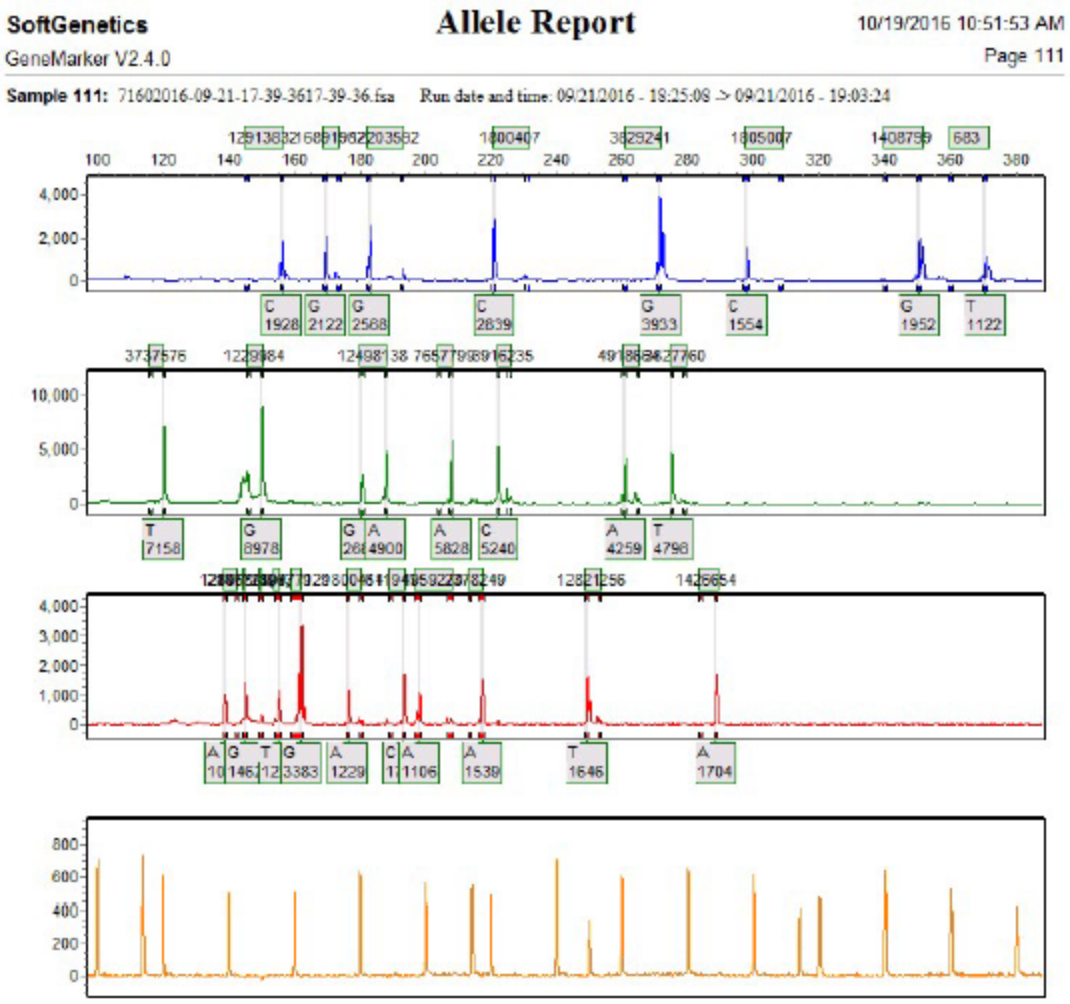

SoftGenetics
Allele Report
10/19/2016 10:51:52 AM
GeneMarker V2.4.0
Page 108
Sample 108: 70952016-09-23-18-33-1718-33-17.fsa Run date and time: 09/23/2016 - 19:53:05 -> 09/23/2016 - 20:30:41

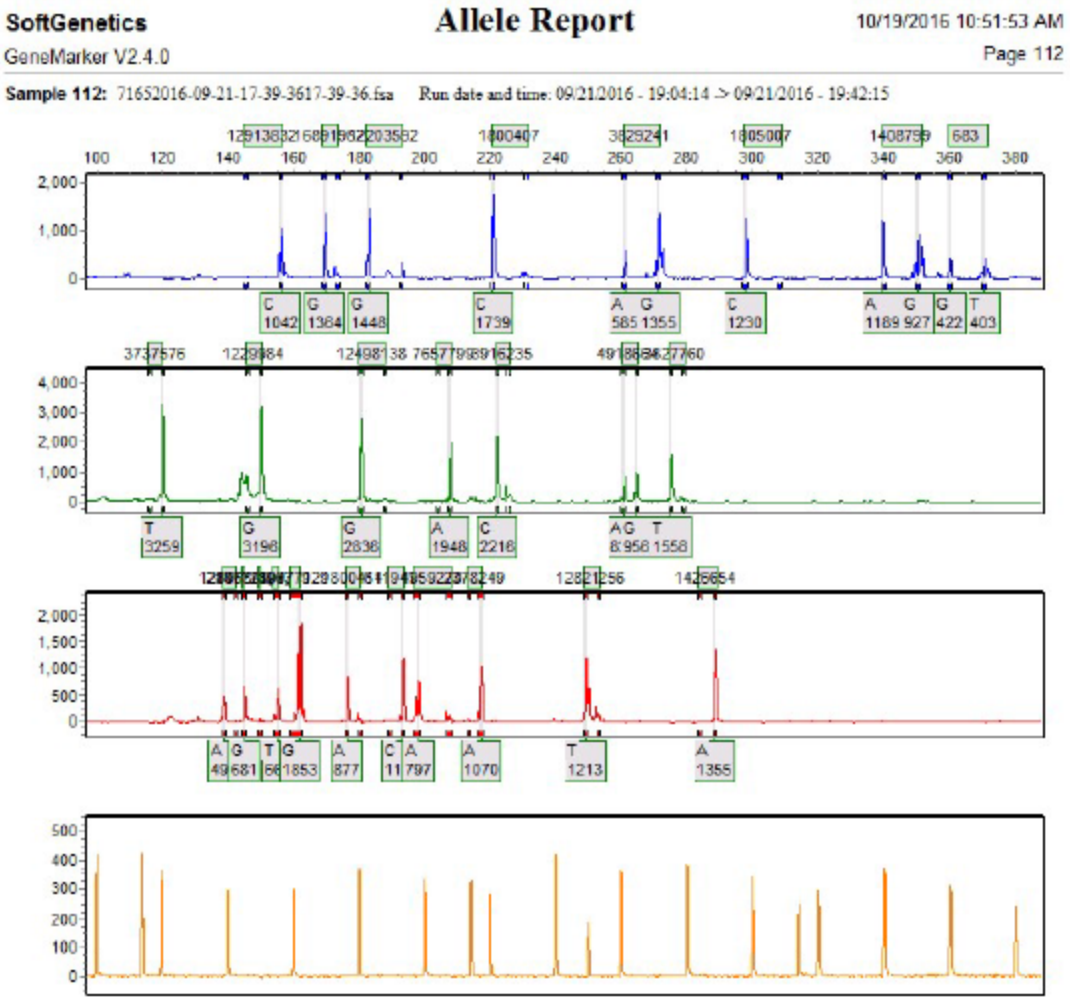
SoftGenetics **Allele Report** 10/19/2016 10:51:52 AM

GeneMarker V2.4.0 Page 109

Sample 109: 71092016-09-22-16-13-3916-13-39.fsa Run date and time: 09/22/2016 - 17:02:24 -> 09/22/2016 - 17:40:30



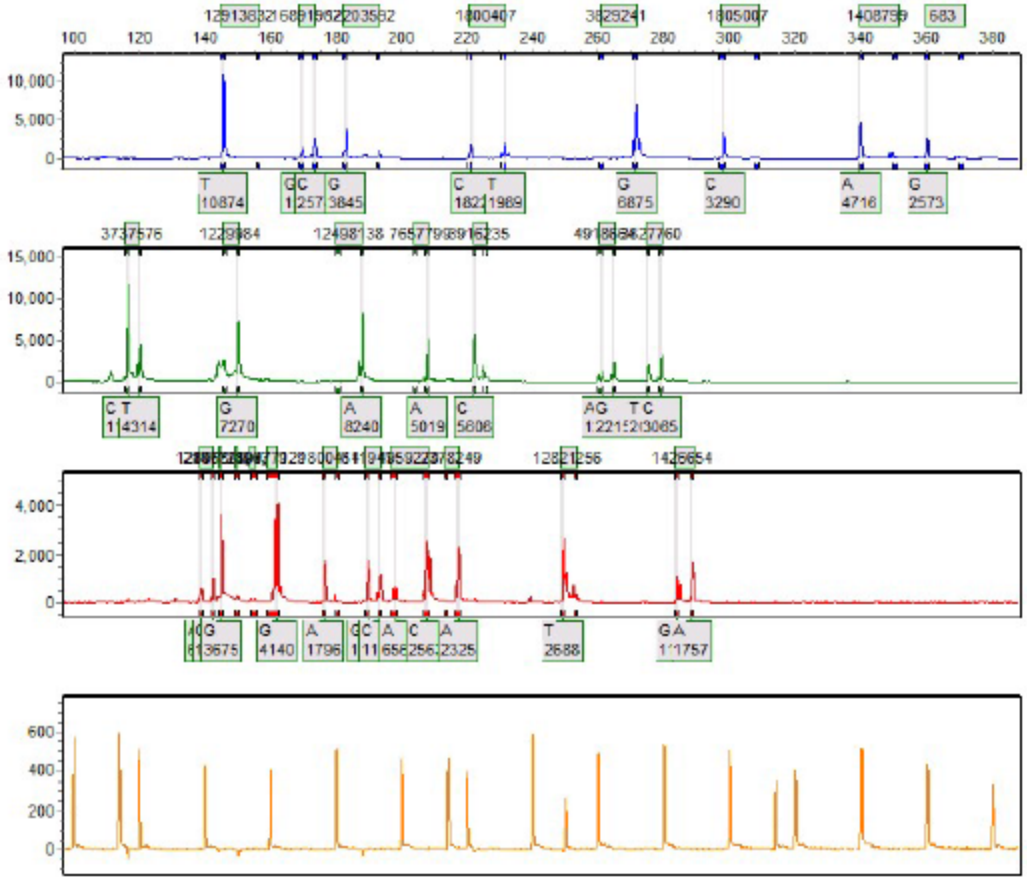

SoftGenetics
Allele Report
10/19/2016 10:51:53 AM
GeneMarker V2.4.0
Page 111
Sample 111: 71602016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 18:25:08 -> 09/21/2016 - 19:03:24


SoftGenetics
Allele Report
10/19/2016 10:51:53 AM
GeneMarker V2.4.0
Page 112
Sample 112: 71652016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 19:04:14 -> 09/21/2016 - 19:42:15

SoftGenetics **Allele Report** 10/19/2016 10:51:53 AM

GeneMarker V2.4.0 Page 113

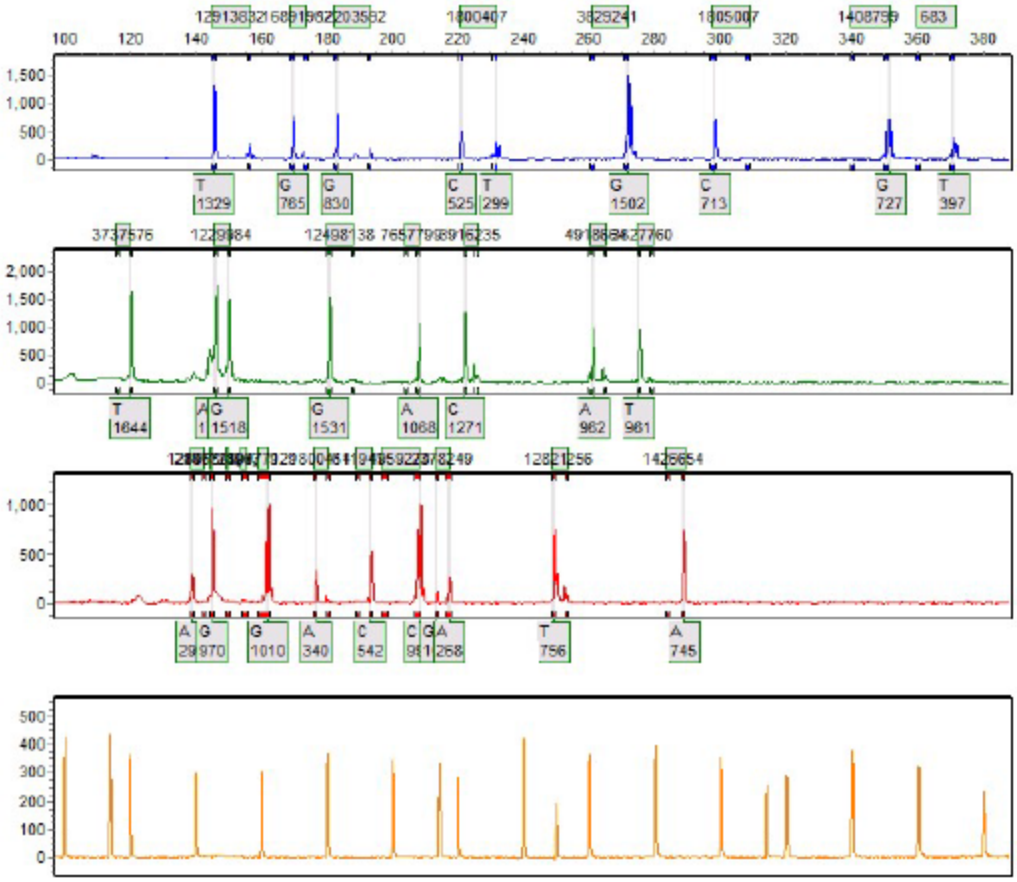
Sample 113: 71812016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 19:04:14 -> 09/21/2016 - 19:42:15



SoftGenetics **Allele Report** 10/19/2016 10:51:53 AM

GeneMarker V2.4.0 Page 114

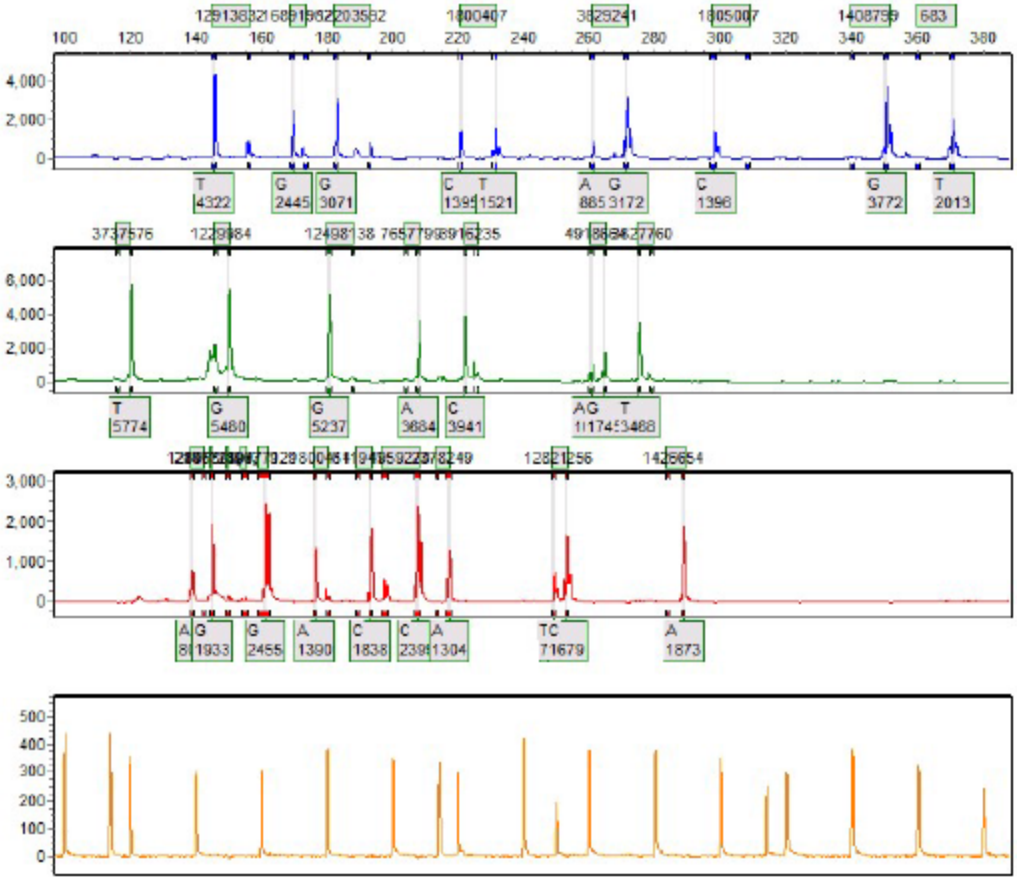
Sample 114: 72592016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 23:36:22 -> 09/22/2016 - 00:14:17



SoftGenetics **Allele Report** 10/19/2016 10:51:53 AM

GeneMarker V2.4.0 Page 115

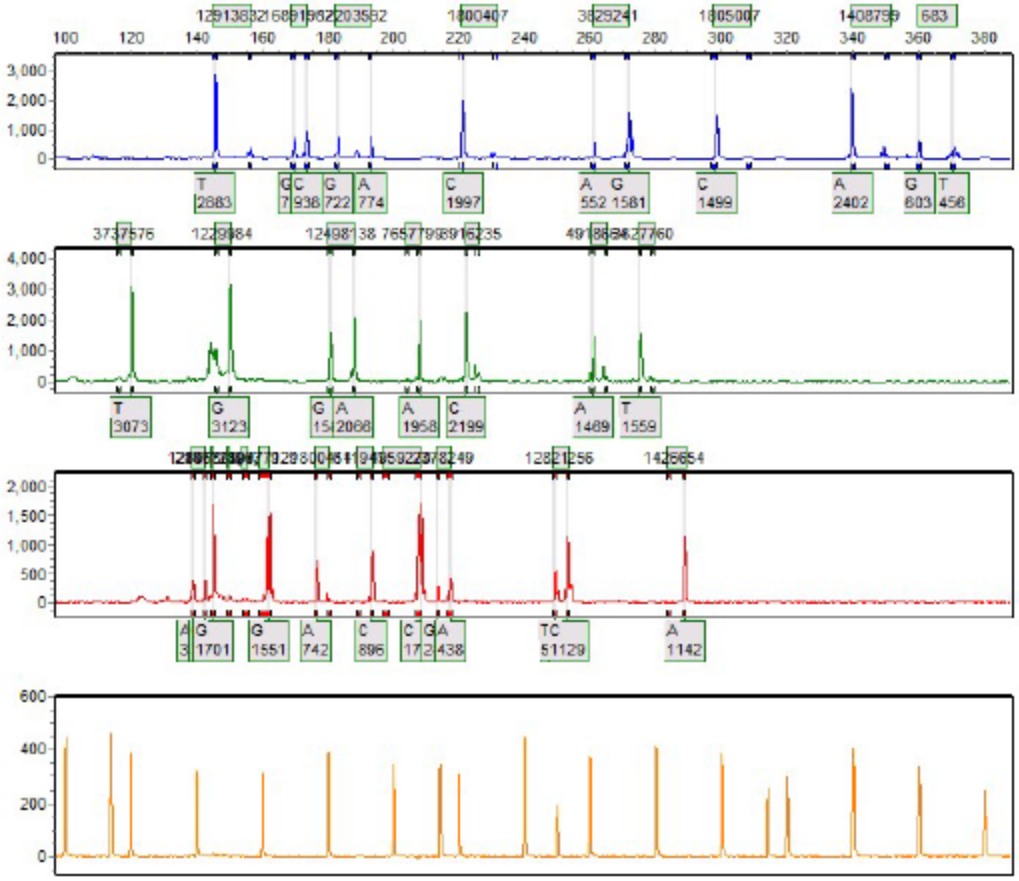
Sample 115: 72632016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 19:04:14 -> 09/21/2016 - 19:42:15



SoftGenetics **Allele Report** 10/19/2016 10:51:53 AM

GeneMarker V2.4.0 Page 116

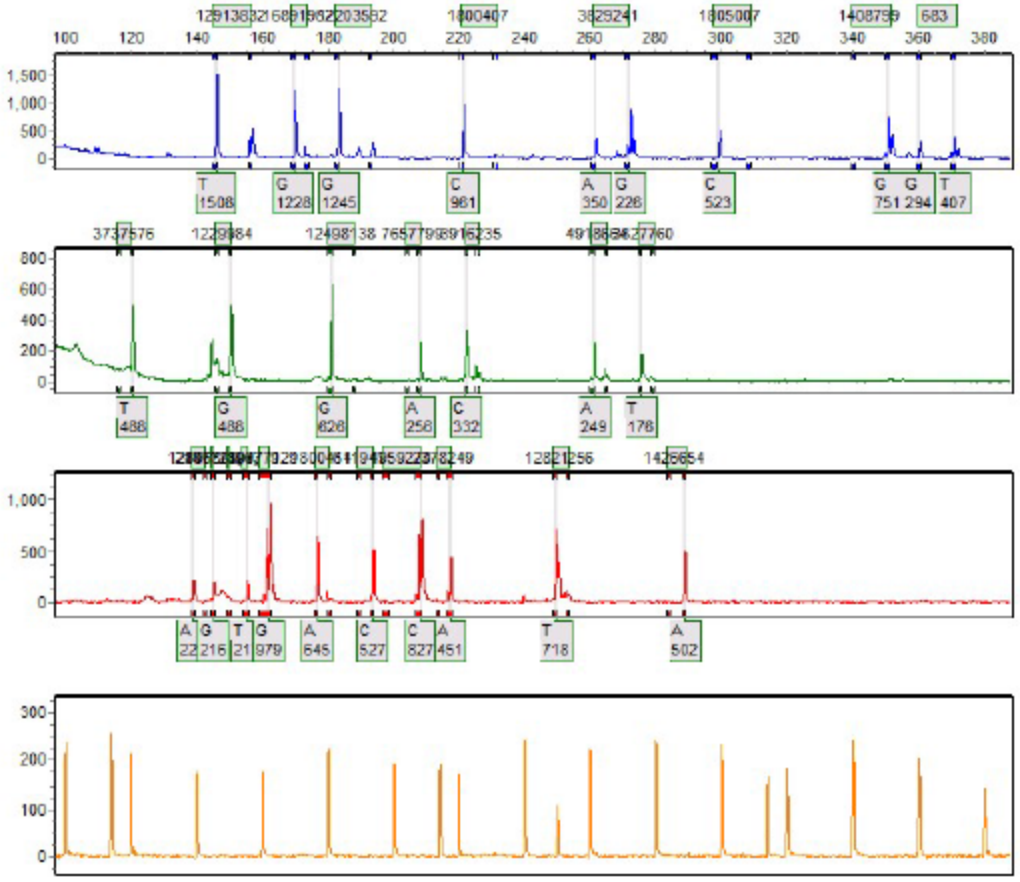
Sample 116: 72802016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 23:36:22 -> 09/22/2016 - 00:14:17



SoftGenetics **Allele Report** 10/19/2016 10:51:53 AM

GeneMarker V2.4.0 Page 117

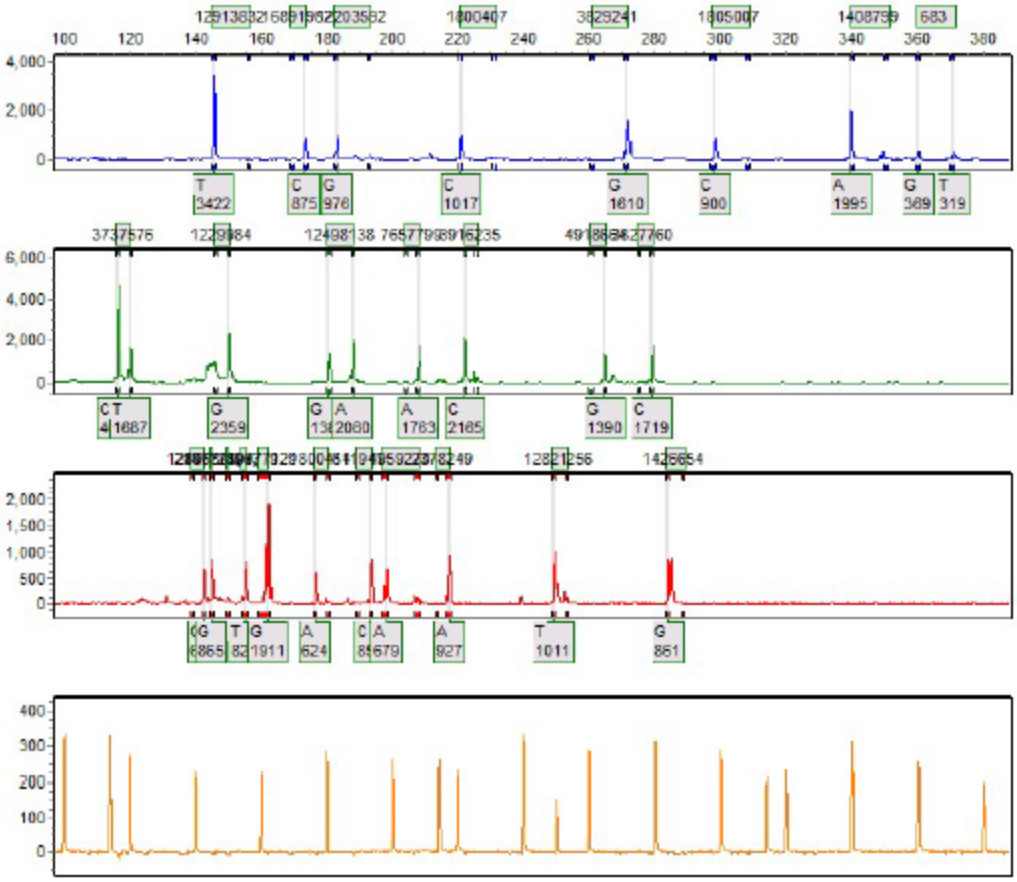
Sample 117: 72942016-09-23-18-33-1718-33-17.fsa Run date and time: 09/23/2016 - 19:53:05 -> 09/23/2016 - 20:30:41



SoftGenetics **Allele Report** 10/19/2016 10:51:53 AM

GeneMarker V2.4.0 Page 118

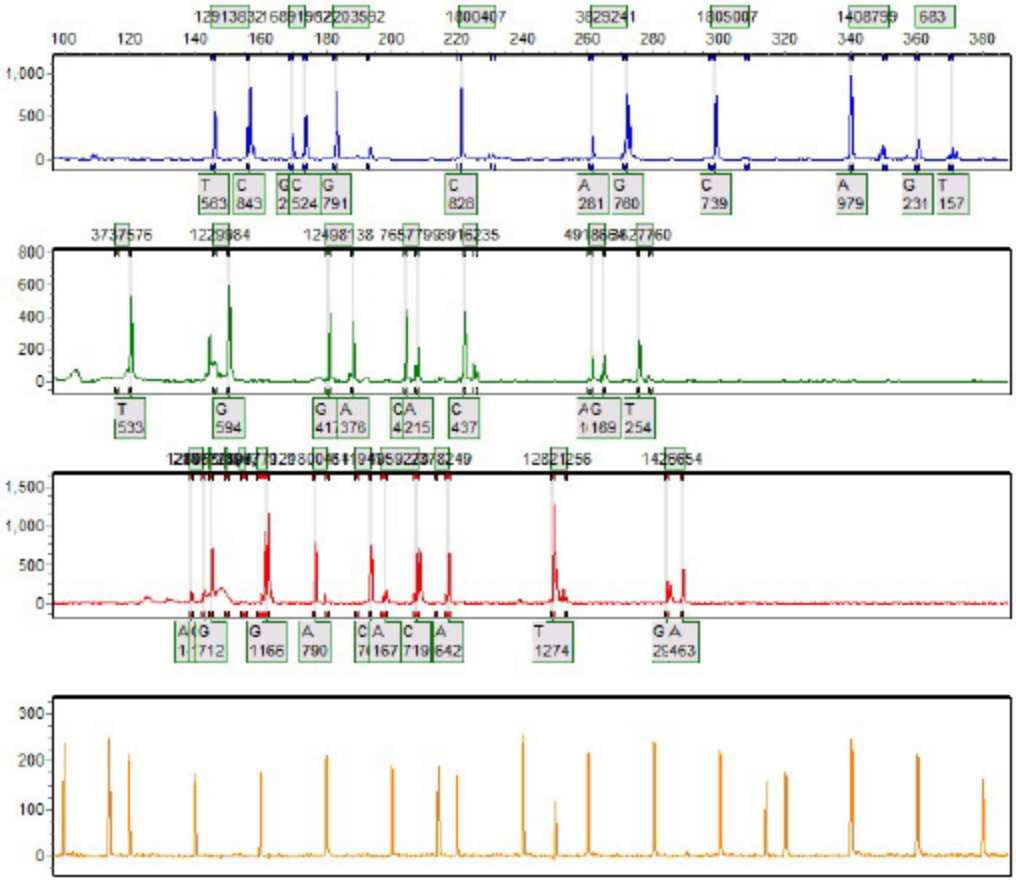
Sample 118: 73452016-09-21-17-39-3617-39-36.fsa Run date and time: 09/22/2016 - 00:53:59 -> 09/22/2016 - 01:32:04

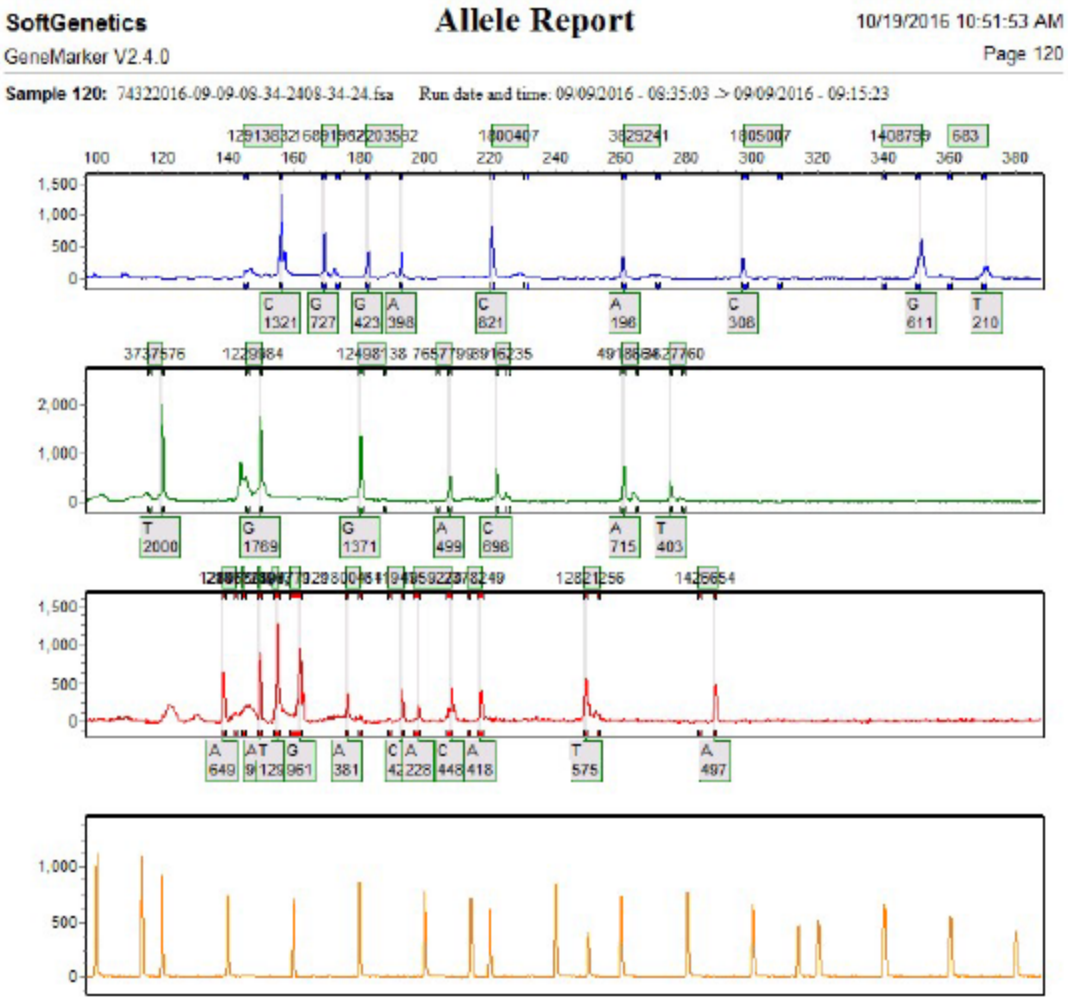


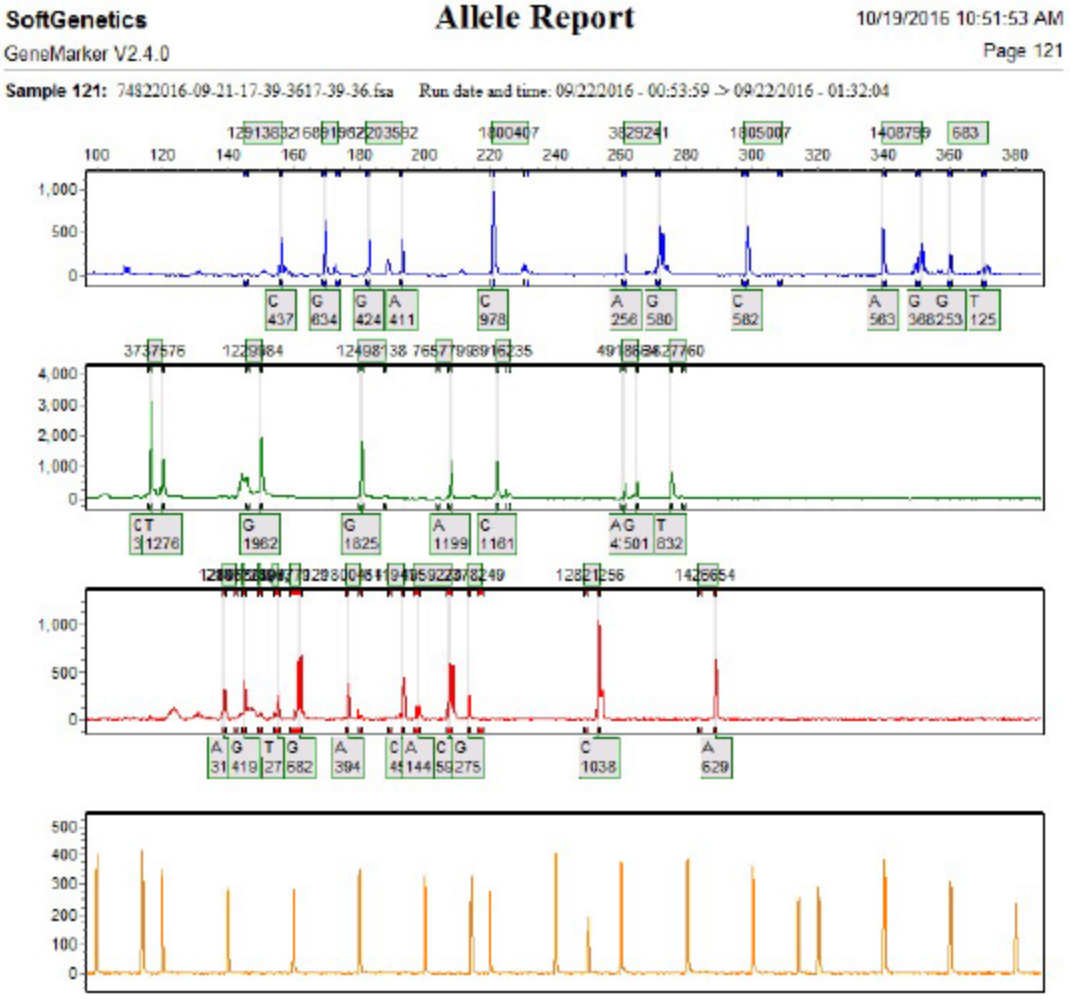
SoftGenetics **Allele Report** 10/19/2016 10:51:53 AM

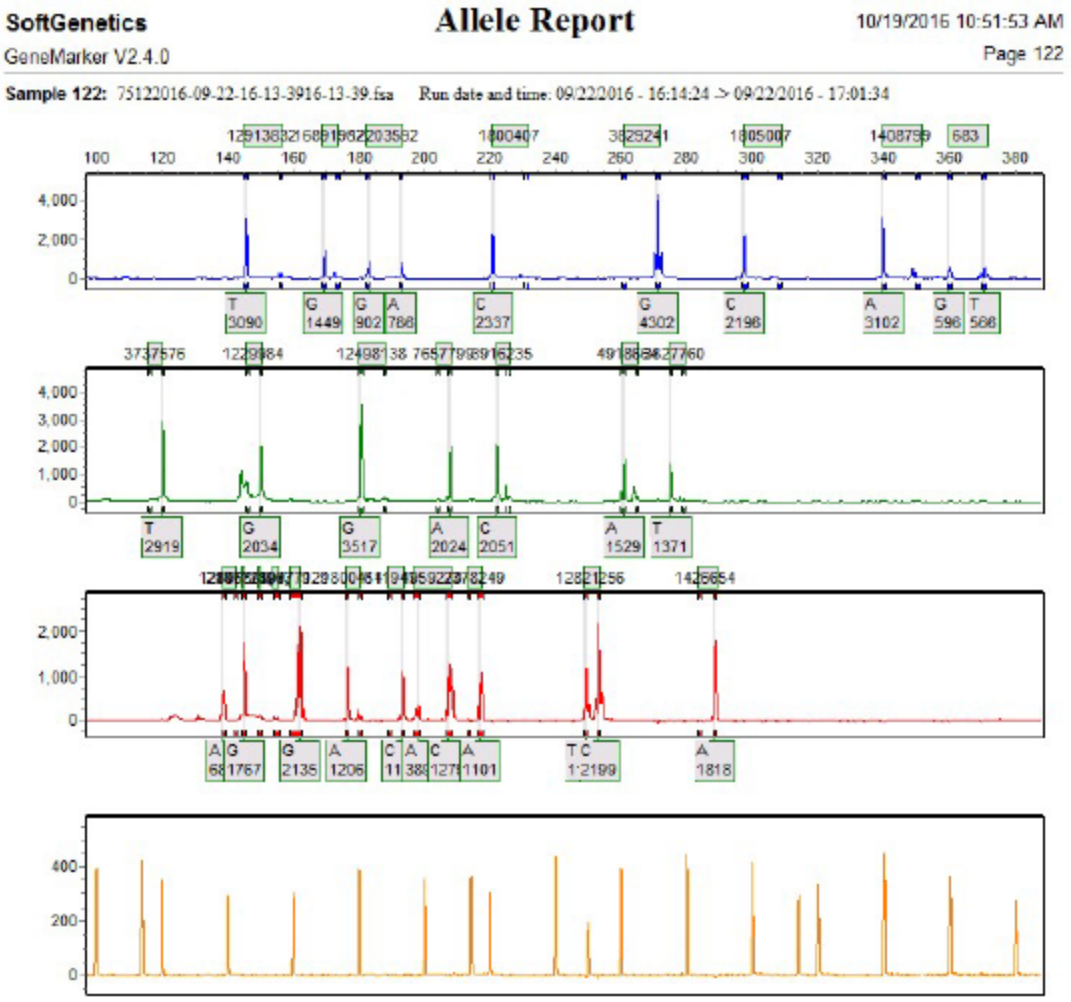
GeneMarker V2.4.0 Page 119

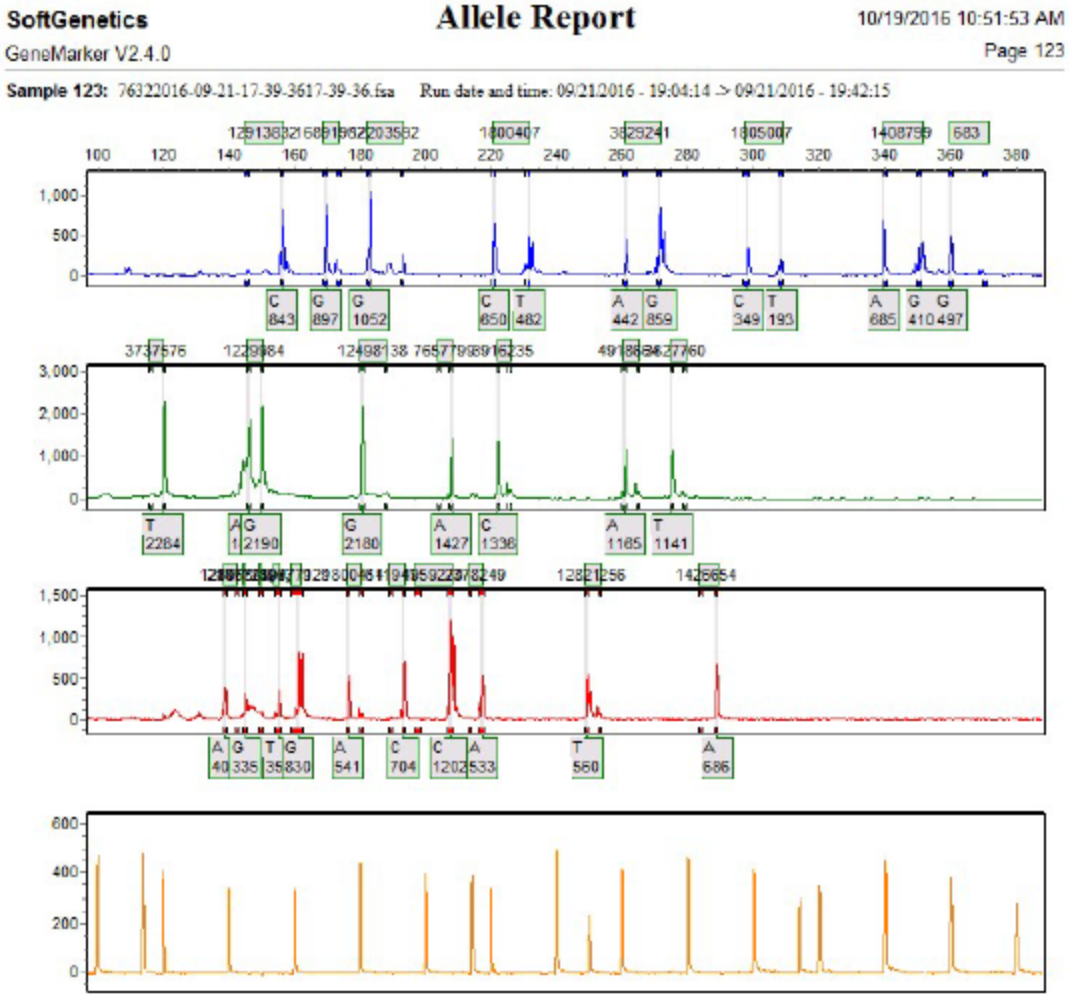
Sample 119: 74012016-09-24-12-28-1612-28-16.fsa Run date and time: 09/24/2016 - 12:28:57 -> 09/24/2016 - 13:17:27

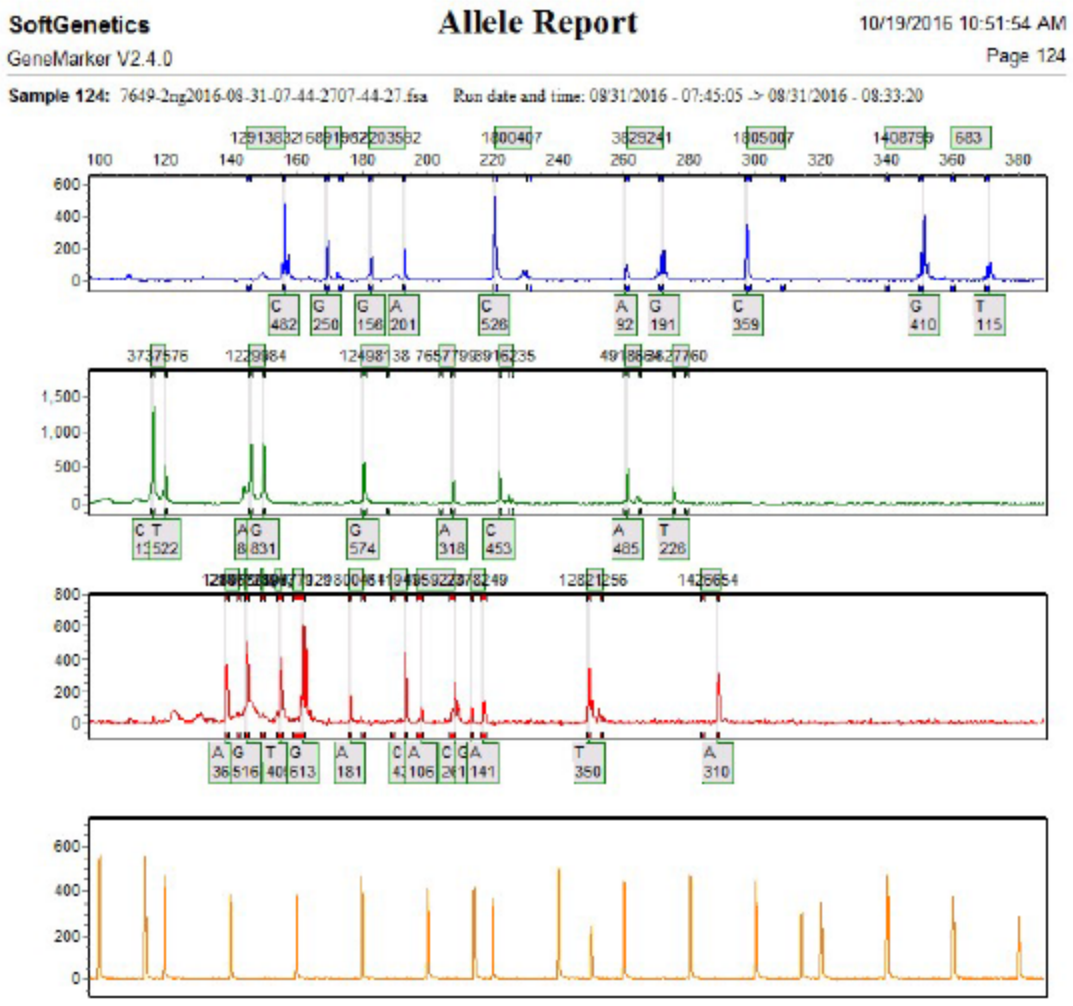



SoftGenetics
Allele Report
10/19/2016 10:51:53 AM
GeneMarker V2.4.0
Page 120
Sample 120: 74322016-09-09-08-34-2408-34-24.fsa Run date and time: 09/09/2016 - 08:35:03 -> 09/09/2016 - 09:15:23


SoftGenetics
Allele Report
10/19/2016 10:51:53 AM
GeneMarker V2.4.0
Page 121
Sample 121: 74822016-09-21-17-39-3617-39-36.fsa Run date and time: 09/22/2016 - 00:53:59 -> 09/22/2016 - 01:32:04


SoftGenetics
Allele Report
10/19/2016 10:51:53 AM
GeneMarker V2.4.0
Page 122
Sample 122: 75122016-09-22-16-13-3916-13-39.fsa Run date and time: 09/22/2016 - 16:14:24 -> 09/22/2016 - 17:01:34

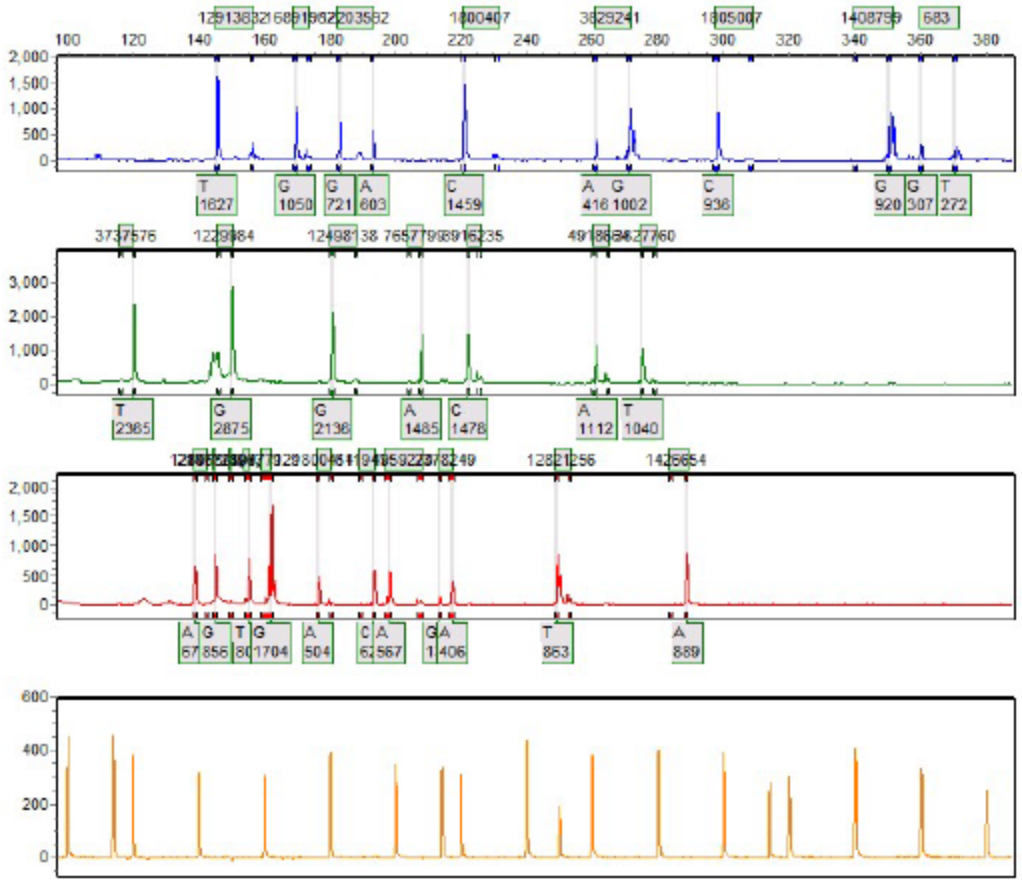

SoftGenetics
Allele Report
10/19/2016 10:51:53 AM
GeneMarker V2.4.0
Page 123
Sample 123: 76322016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 19:04:14 -> 09/21/2016 - 19:42:15

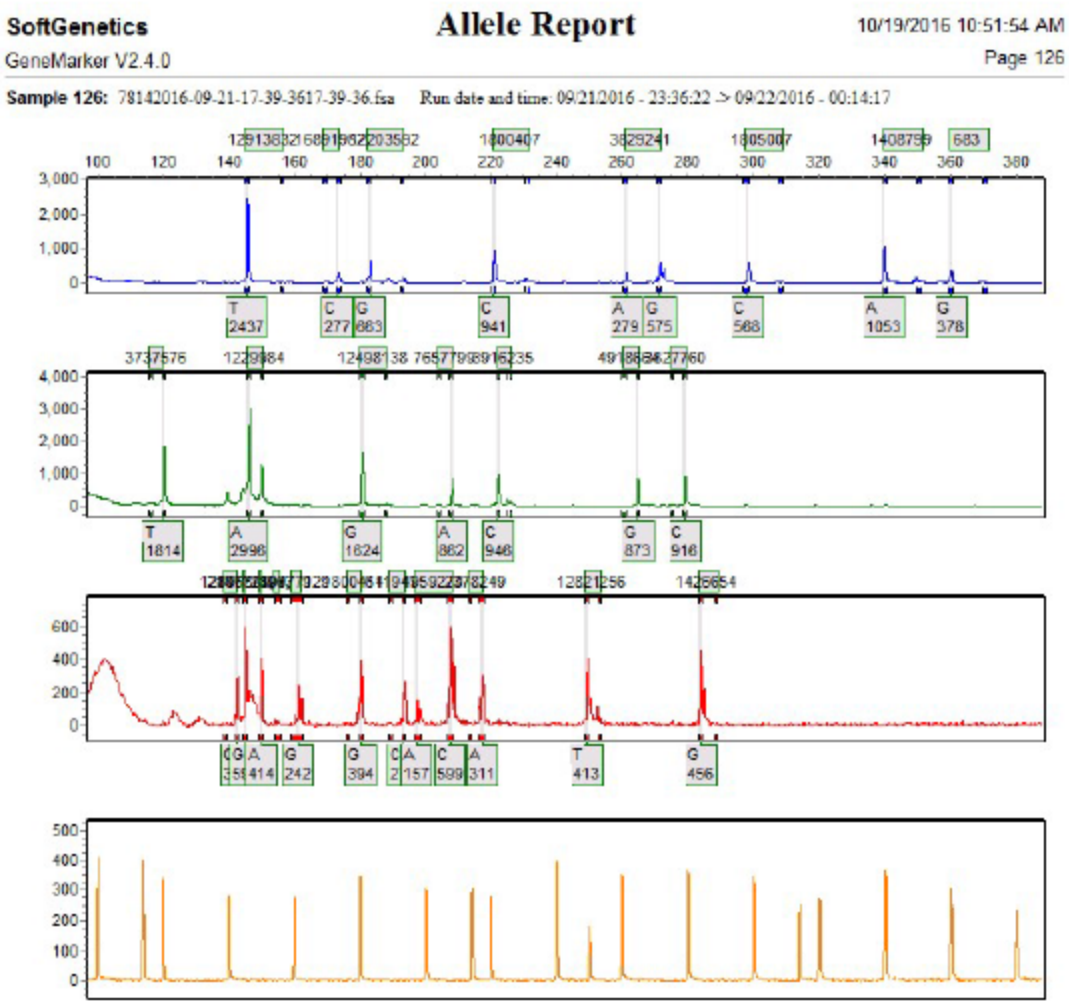

SoftGenetics
Allele Report
10/19/2016 10:51:54 AM
GeneMarker V2.4.0
Page 124
Sample 124: 7649-2ng2016-08-31-07-44-2707-44-27.fsa Run date and time: 08/31/2016 - 07:45:05 -> 08/31/2016 - 08:33:20

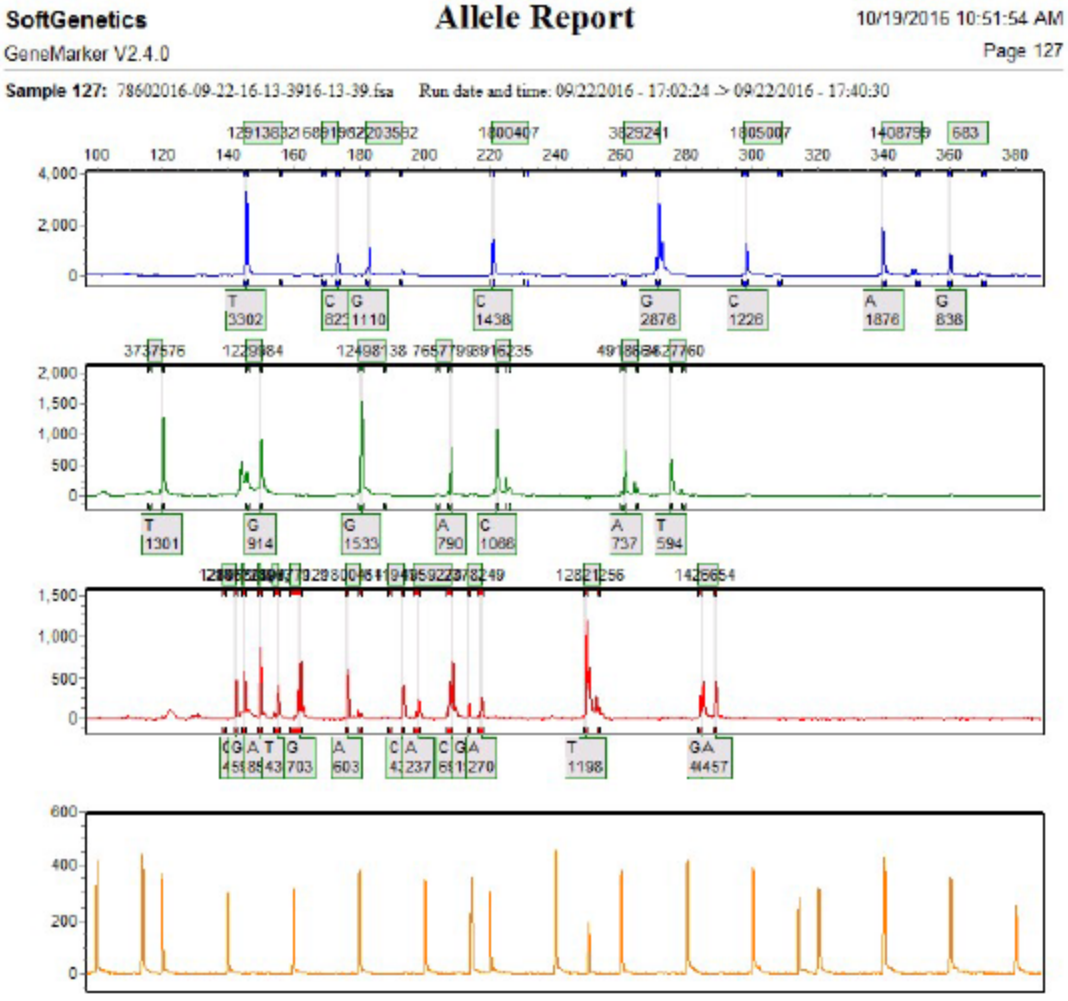
SoftGenetics **Allele Report** 10/19/2016 10:51:54 AM

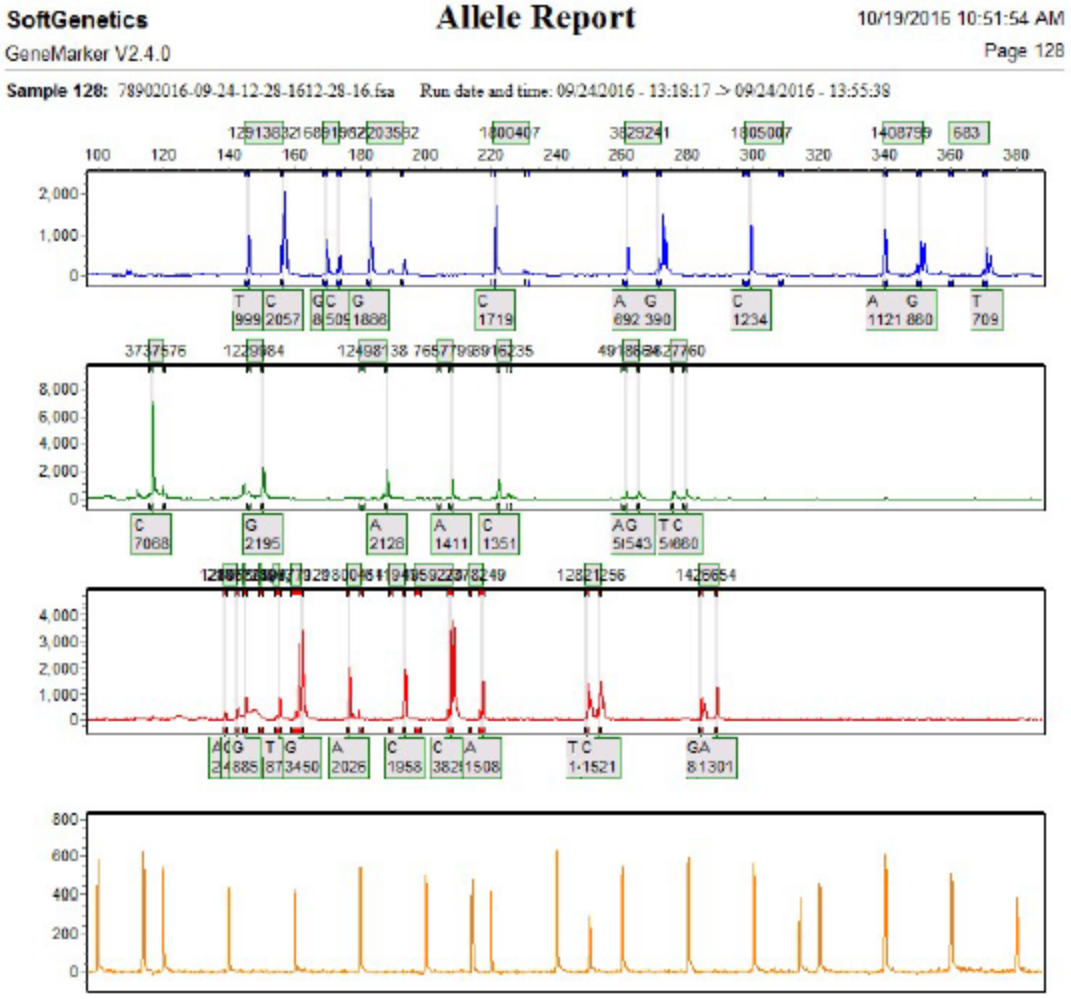
GeneMarker V2.4.0 Page 125

Sample 125: 76592016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 19:04:14 -> 09/21/2016 - 19:42:15




SoftGenetics
Allele Report
10/19/2016 10:51:54 AM
GeneMarker V2.4.0
Page 126
Sample 126: 78142016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 23:36:22 -> 09/22/2016 - 00:14:17

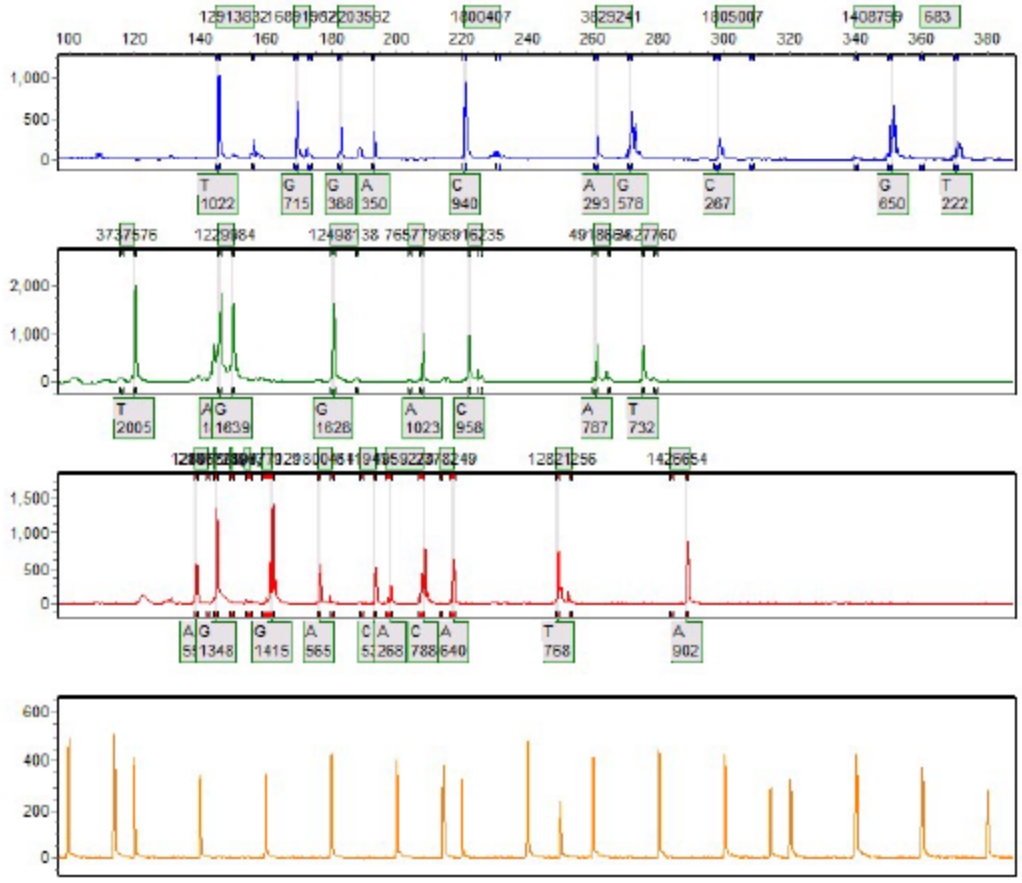

SoftGenetics
Allele Report
10/19/2016 10:51:54 AM
GeneMarker V2.4.0
Page 127
Sample 127: 78602016-09-22-16-13-3916-13-39.fsa Run date and time: 09/22/2016 - 17:02:24 -> 09/22/2016 - 17:40:30


SoftGenetics
Allele Report
10/19/2016 10:51:54 AM
GeneMarker V2.4.0
Page 128
Sample 128: 78902016-09-24-12-28-1612-28-16.fsa Run date and time: 09/24/2016 - 13:18:17 -> 09/24/2016 - 13:55:38

SoftGenetics **Allele Report** 10/19/2016 10:51:54 AM

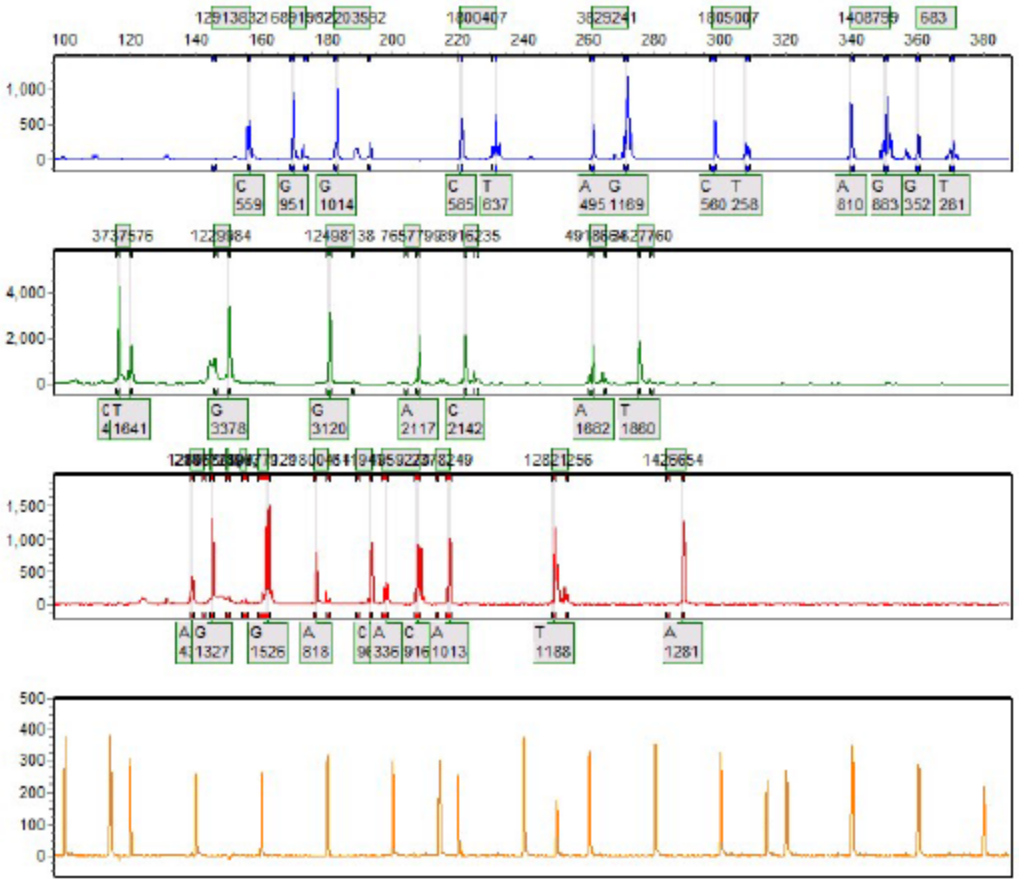
GeneMarker V2.4.0 Page 129

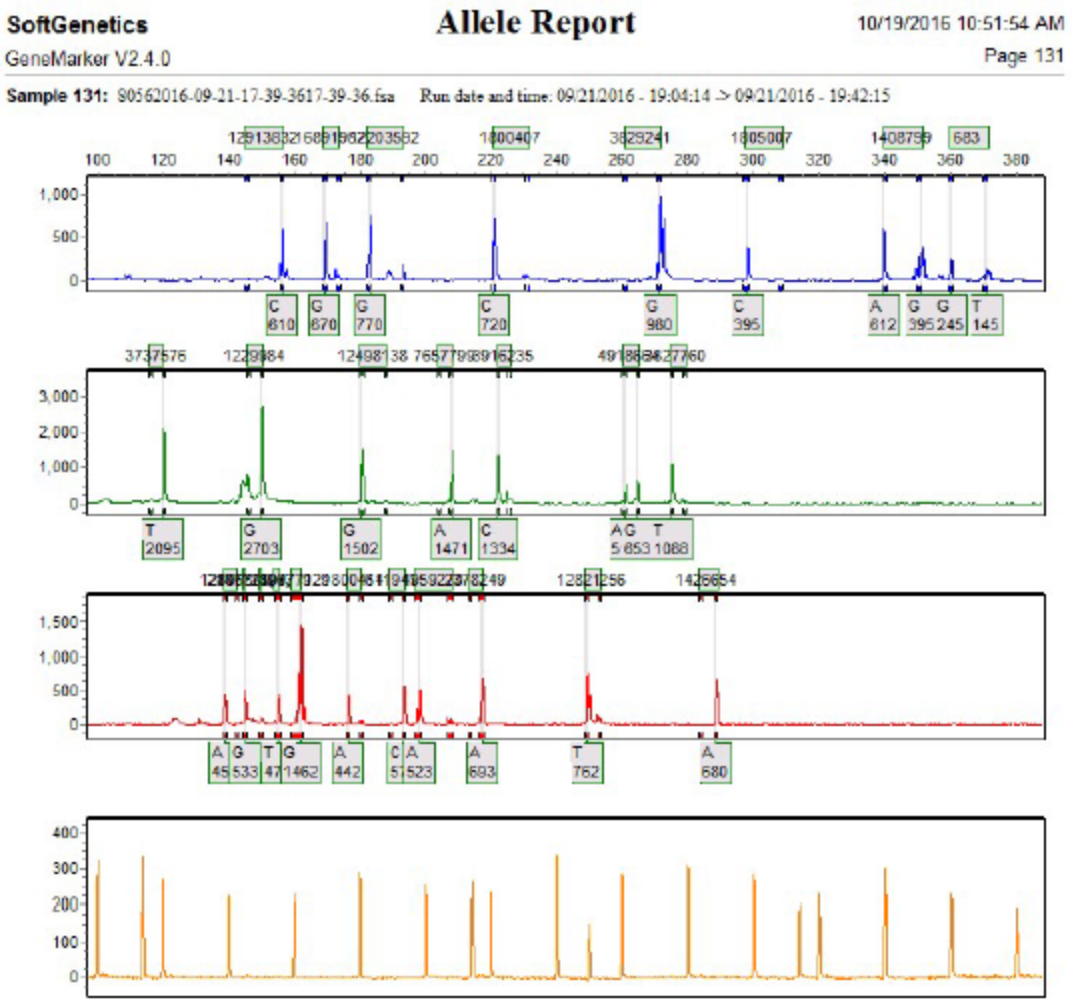
Sample 129: 79152016-09-21-17-39-3617-39-36.fsa Run date and time: 09/22/2016 - 00:53:59 -> 09/22/2016 - 01:32:04

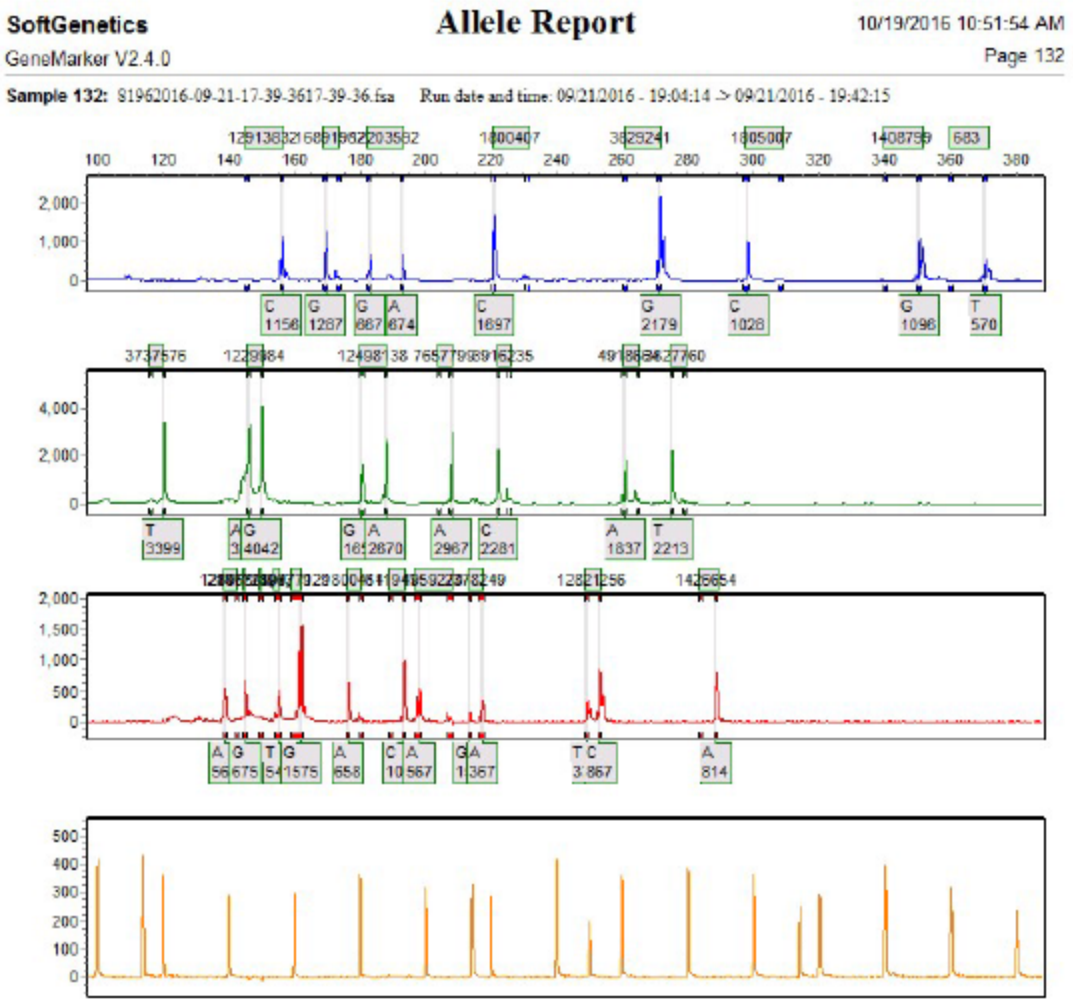


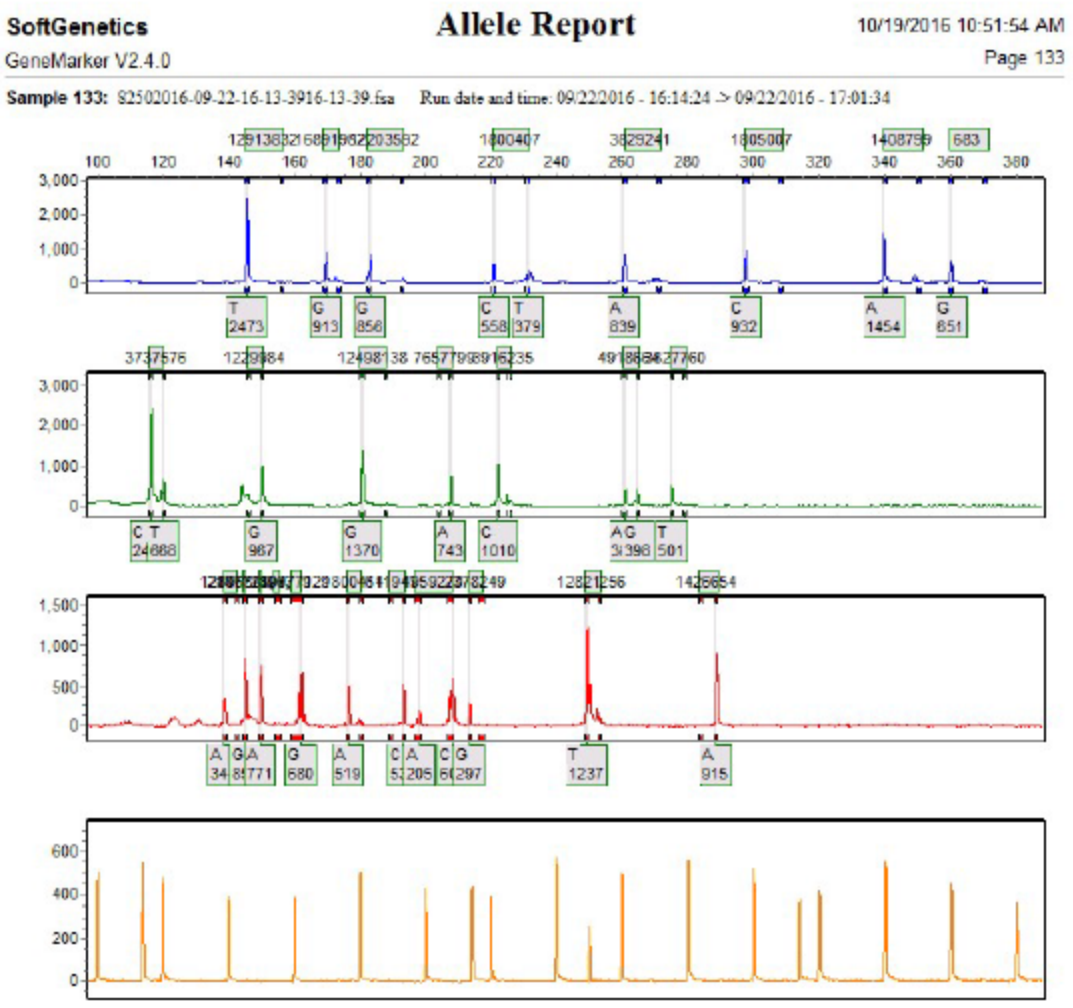
Allele Report

Sample 130: 79802016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 19:04:14 -> 09/21/2016 - 19:42:15



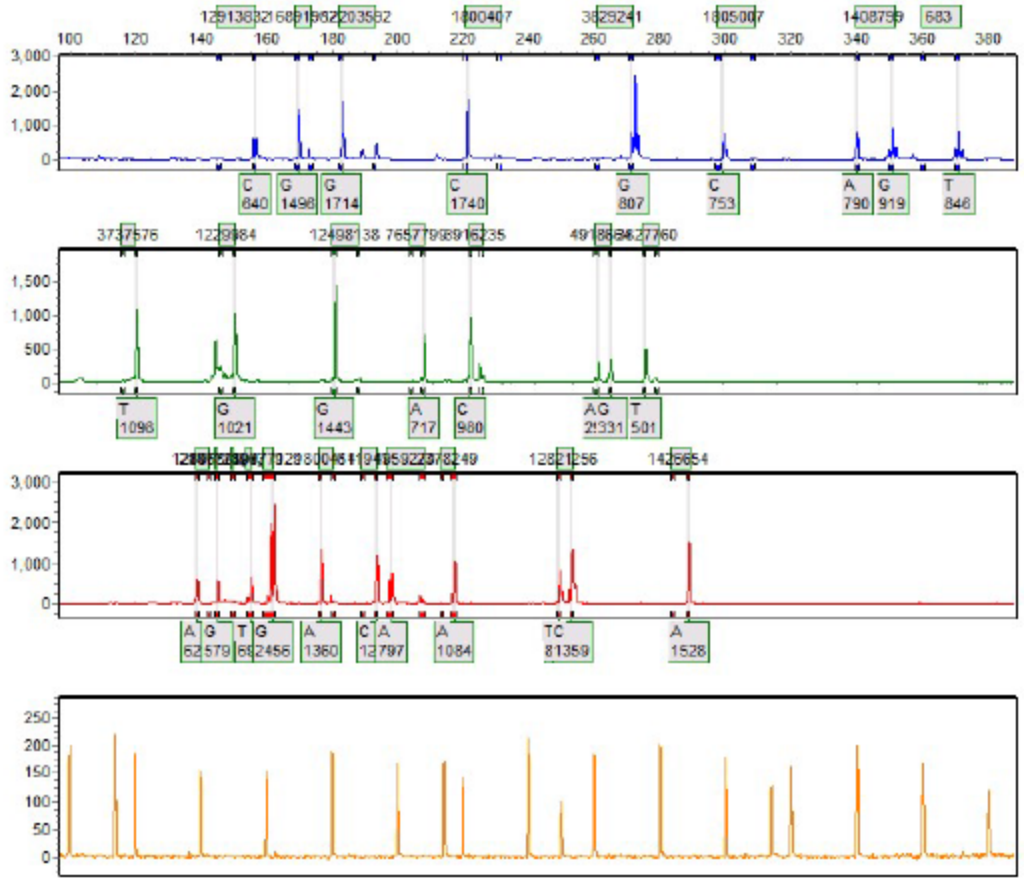

SoftGenetics
Allele Report
10/19/2016 10:51:54 AM
GeneMarker V2.4.0
Page 131
Sample 131: 80562016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 19:04:14 -> 09/21/2016 - 19:42:15

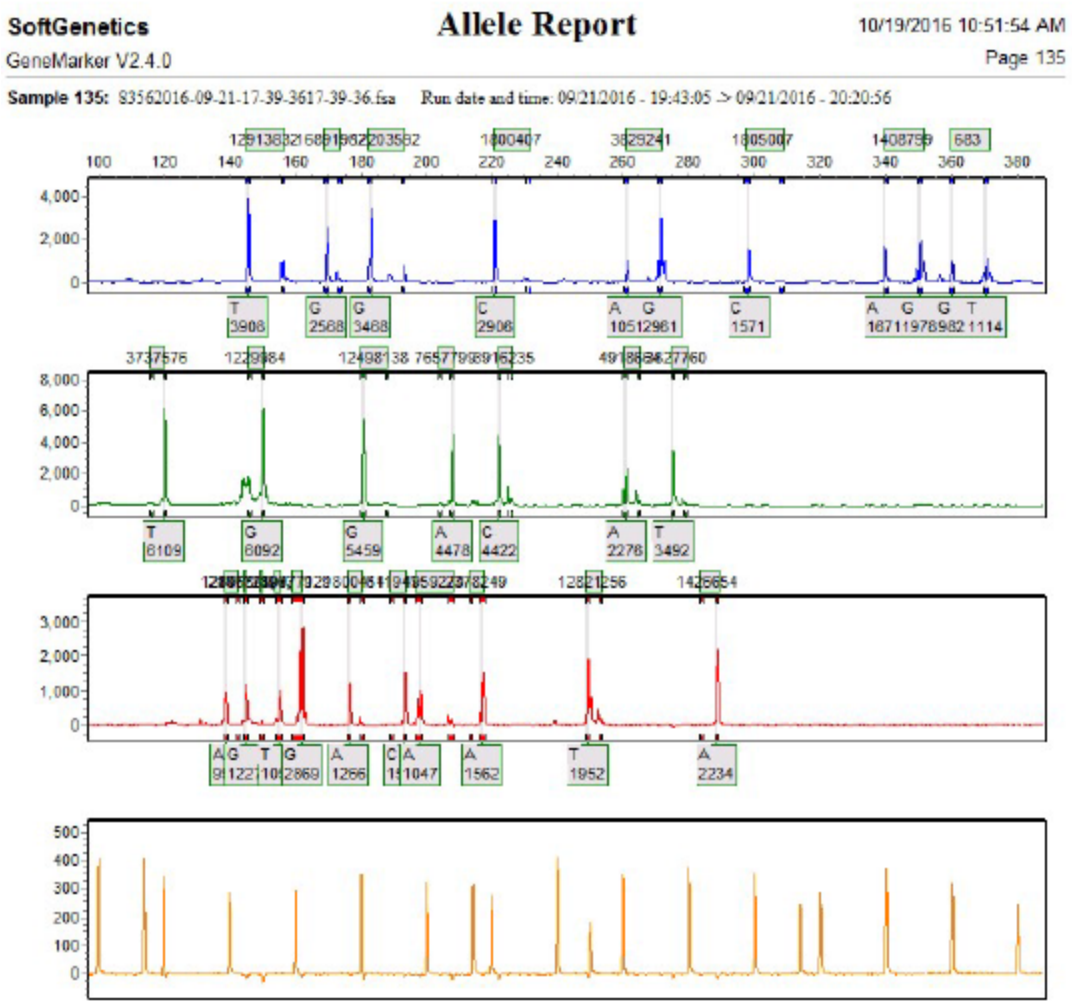

SoftGenetics
Allele Report
10/19/2016 10:51:54 AM
GeneMarker V2.4.0
Page 132
Sample 132: 81962016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 19:04:14 -> 09/21/2016 - 19:42:15


SoftGenetics
Allele Report
10/19/2016 10:51:54 AM
GeneMarker V2.4.0
Page 133
Sample 133: S2502016-09-22-16-13-3916-13-39.fsa Run date and time: 09/22/2016 - 16:14:24 -> 09/22/2016 - 17:01:34

Allele Report

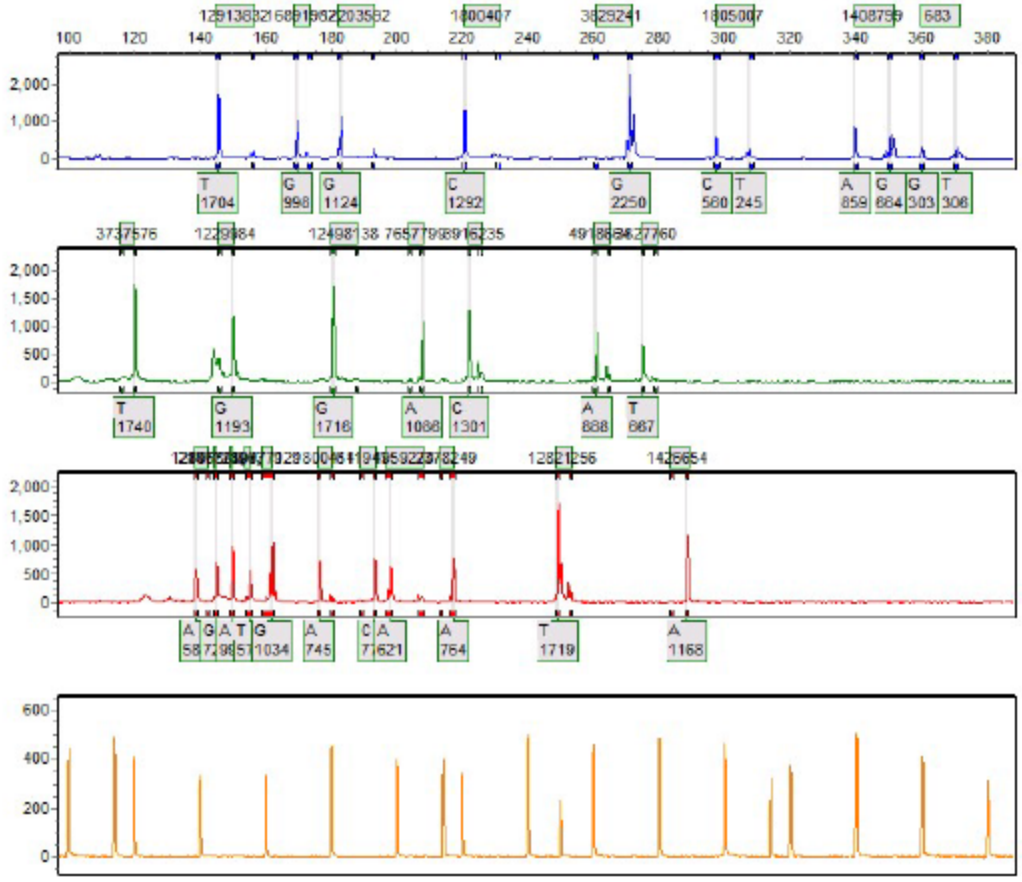
Sample 134: 82672016-09-23-18-33-1718-33-17.fsa Run date and time: 09/23/2016 - 19:53:05 -> 09/23/2016 - 20:30:41




SoftGenetics
Allele Report
10/19/2016 10:51:54 AM
GeneMarker V2.4.0
Page 135
Sample 135: 83562016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 19:43:05 -> 09/21/2016 - 20:20:56

Allele Report

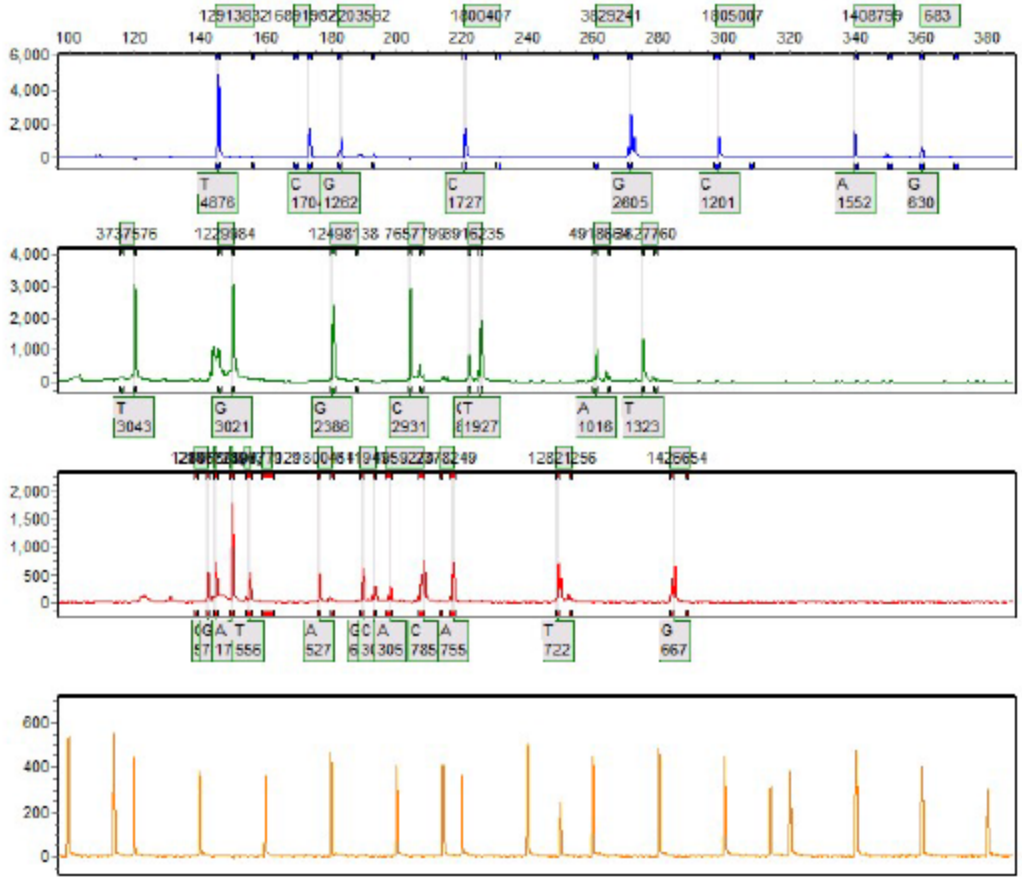
Sample 136: 83602016-09-22-16-13-3916-13-39.fsa Run date and time: 09/22/2016 - 16:14:24 -> 09/22/2016 - 17:01:34

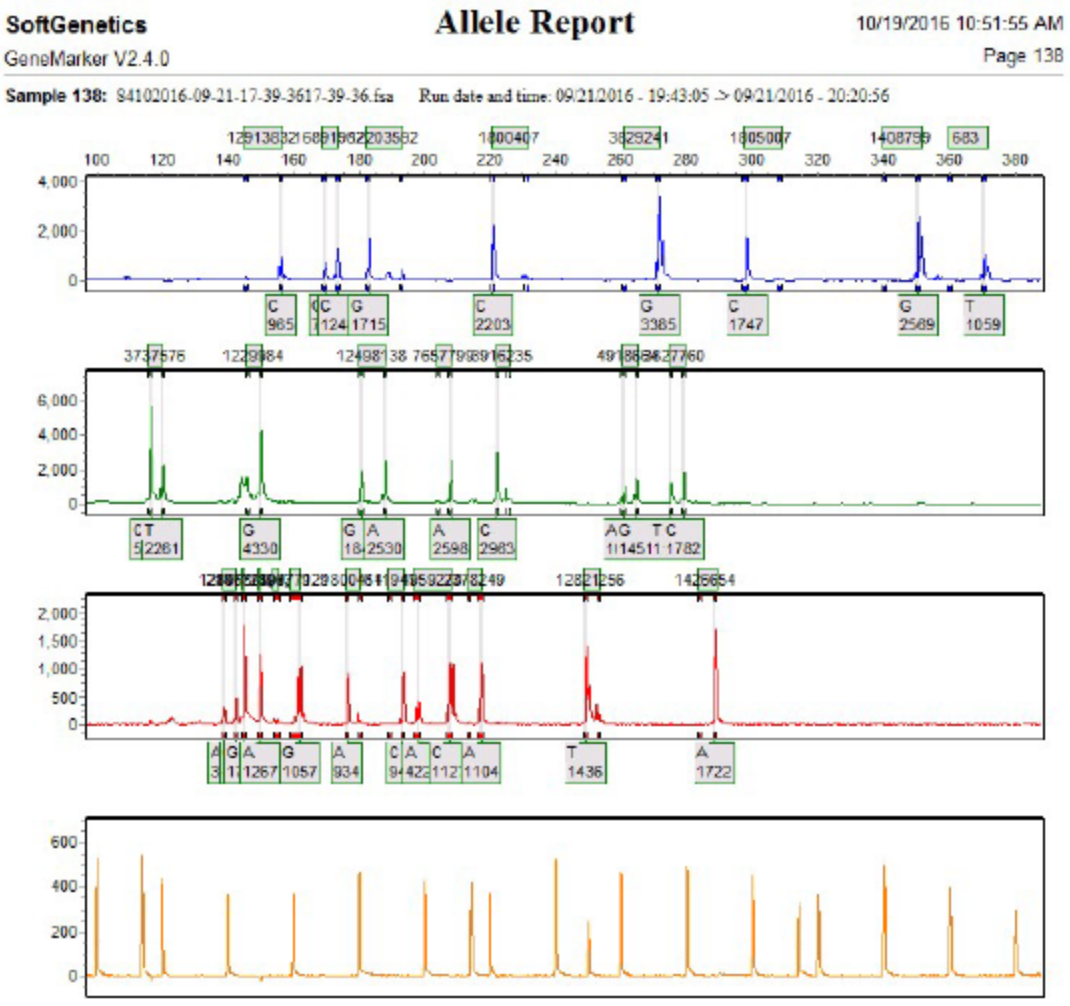


SoftGenetics **Allele Report** 10/19/2016 10:51:55 AM

GeneMarker V2.4.0 Page 137

Sample 137: 83952016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 19:43:05 -> 09/21/2016 - 20:20:56

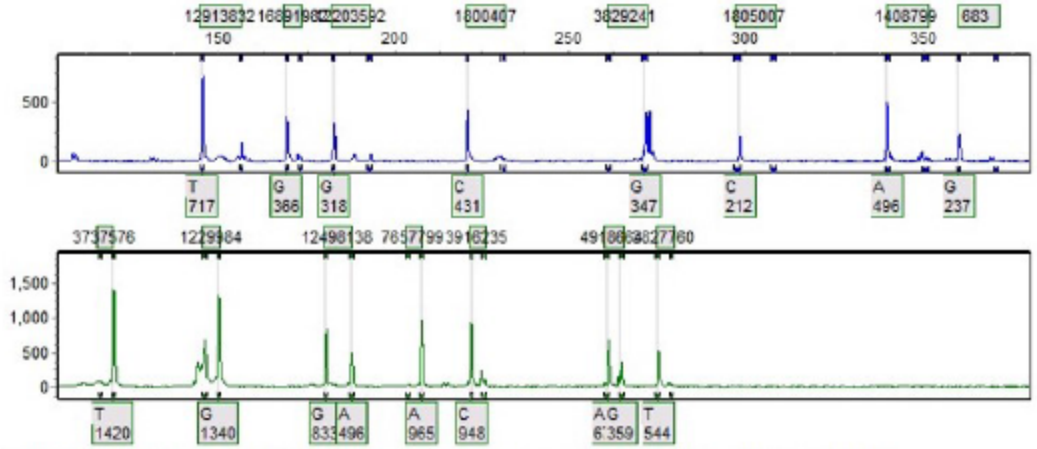



SoftGenetics
Allele Report
10/19/2016 10:51:55 AM
GeneMarker V2.4.0
Page 138
Sample 138: 84102016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 19:43:05 -> 09/21/2016 - 20:20:56

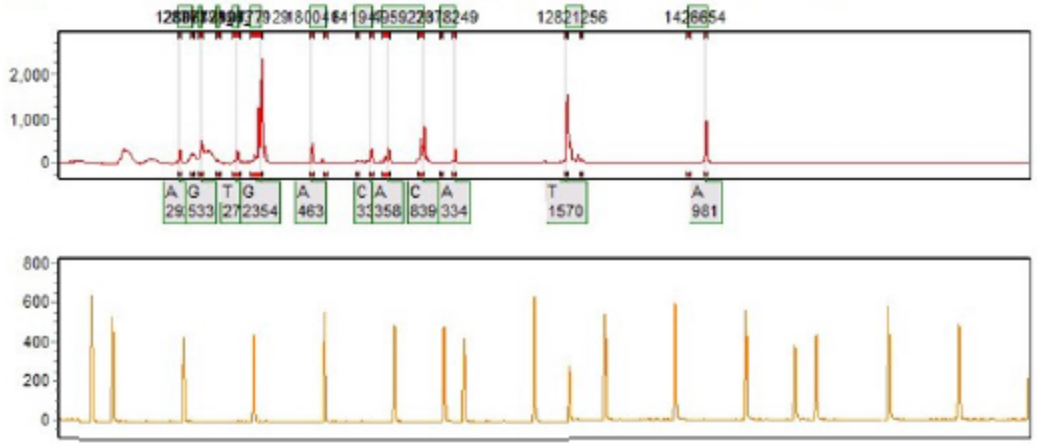
SoftGenetics **Allele Report** 10/19/2016 11:49:33 AM

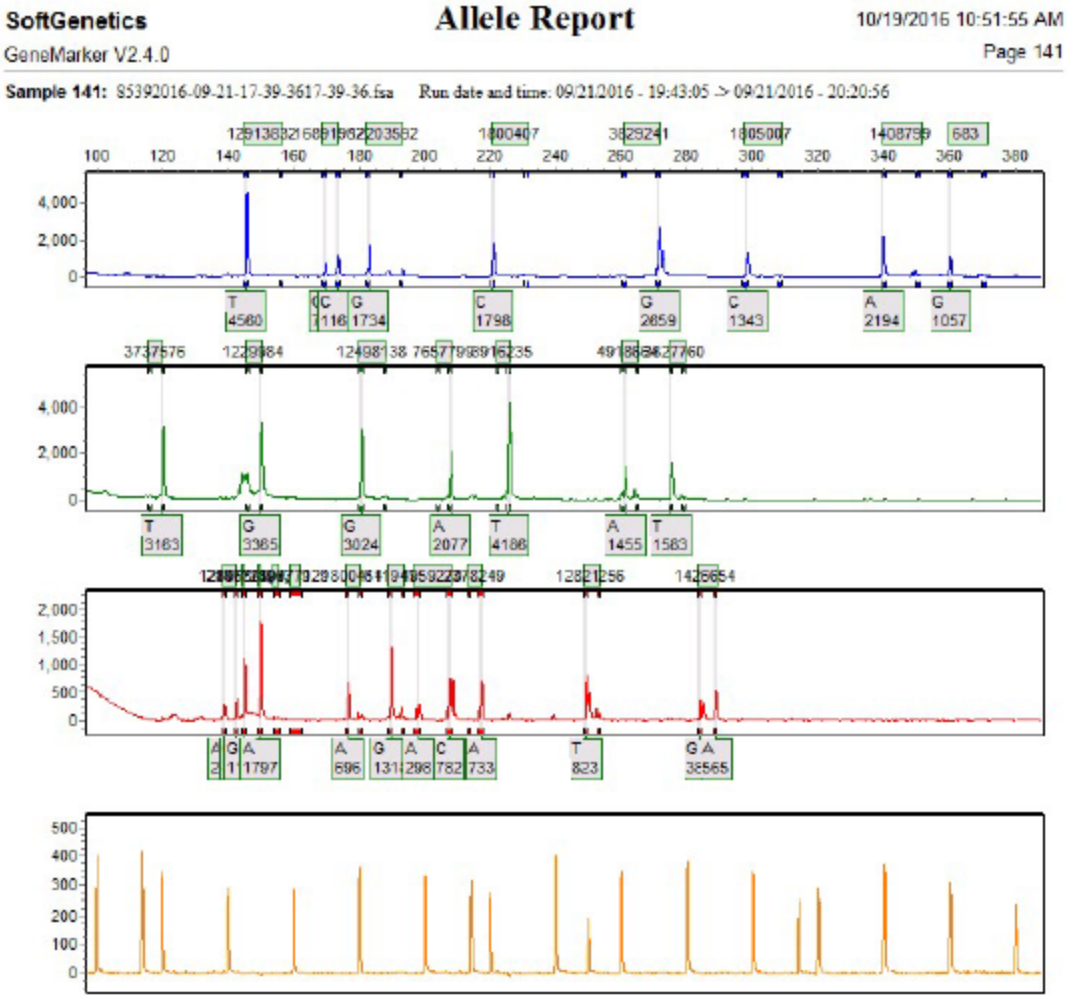
GeneMarker V2.4.0 Page 14

Sample 14: 84352016-09-21-17-39-3617-39-36.fsa Run date and time: 09/22/2016 - 00:53:59 -> 09/22/2016 - 01:32:04



Sample 15: 8435P2016-10-07-07-32-3307-32-33.fsa Run date and time: 10/07/2016 - 08:22:28 -> 10/07/2016 - 09:00:13

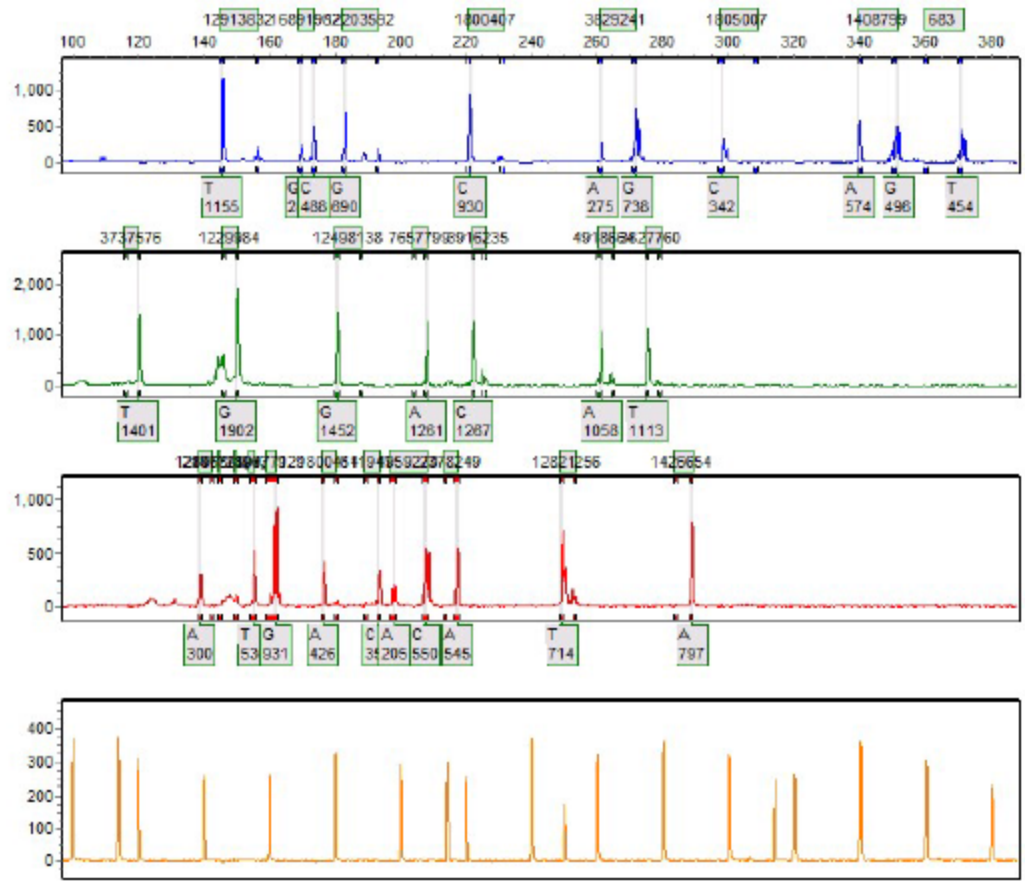


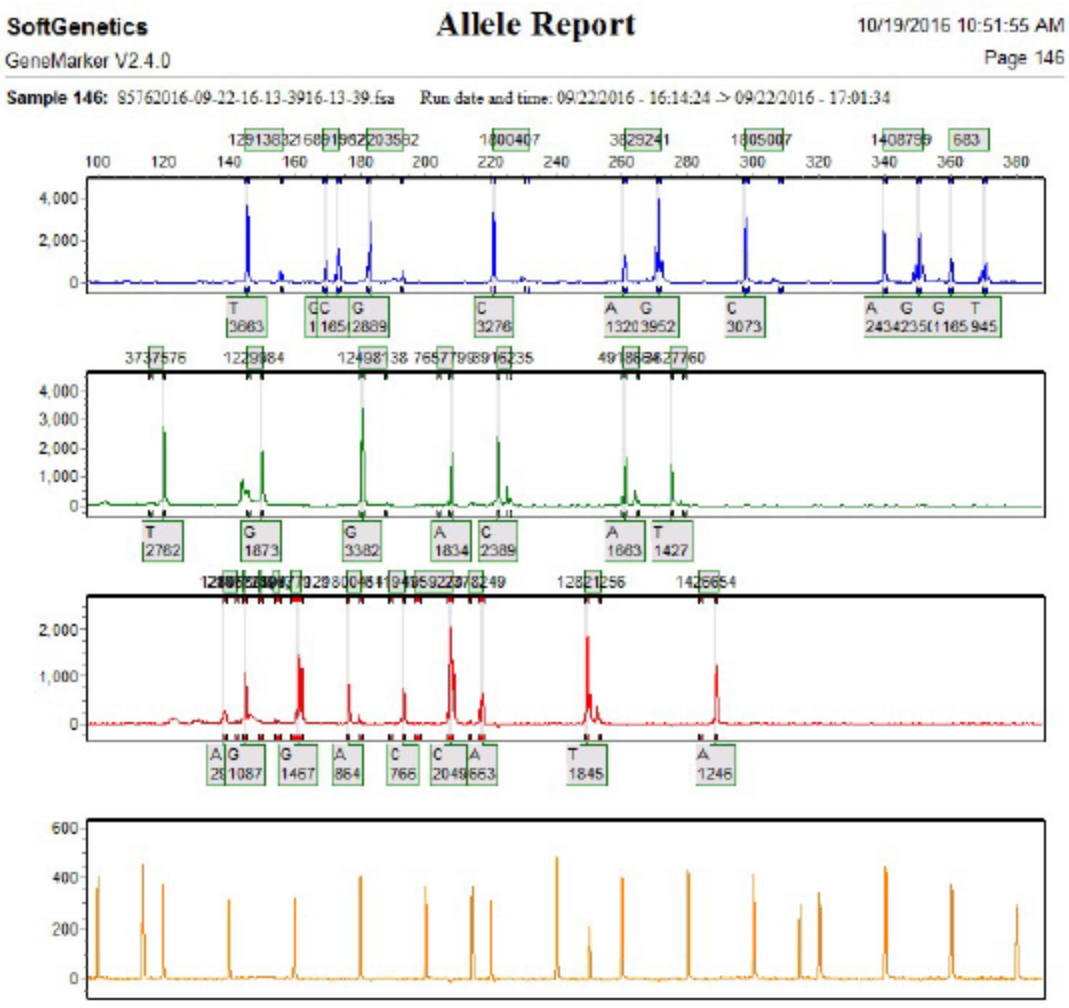

SoftGenetics
Allele Report
10/19/2016 10:51:55 AM
GeneMarker V2.4.0
Page 141
Sample 141: 85392016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 19:43:05 -> 09/21/2016 - 20:20:56

SoftGenetics **Allele Report** 10/19/2016 10:51:55 AM

GeneMarker V2.4.0 Page 145

Sample 145: 85722016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 19:43:05 -> 09/21/2016 - 20:20:56

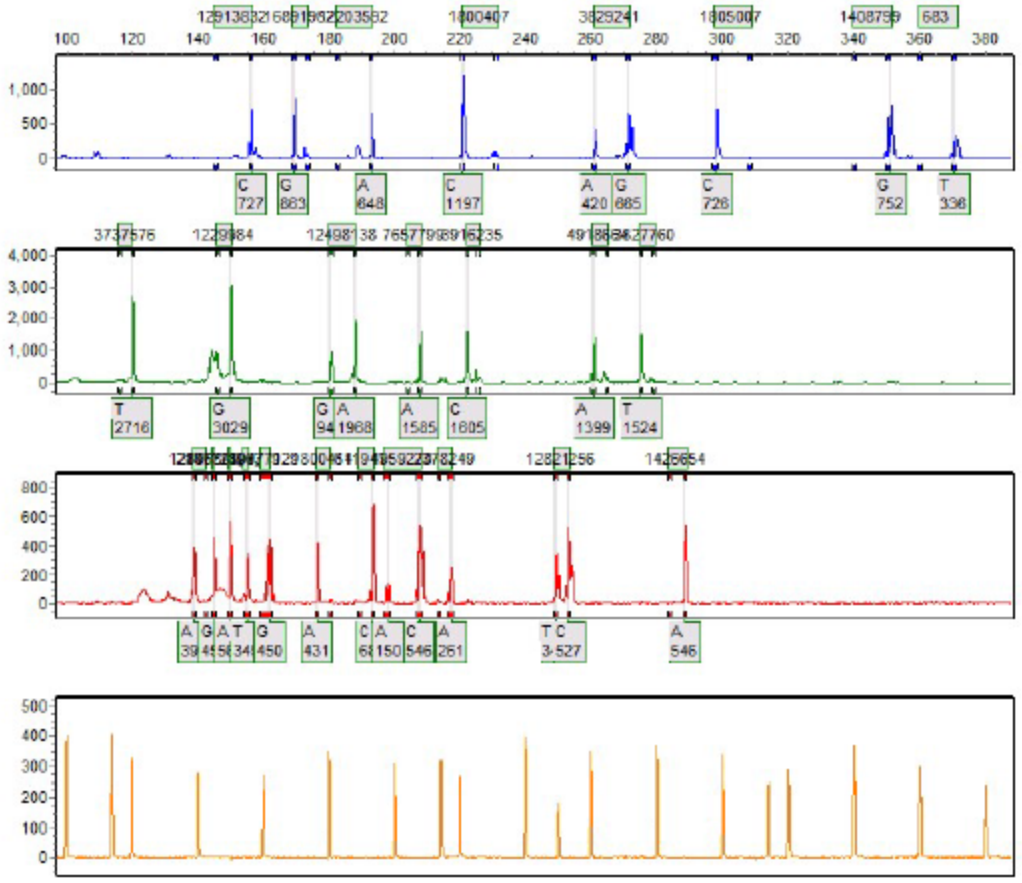


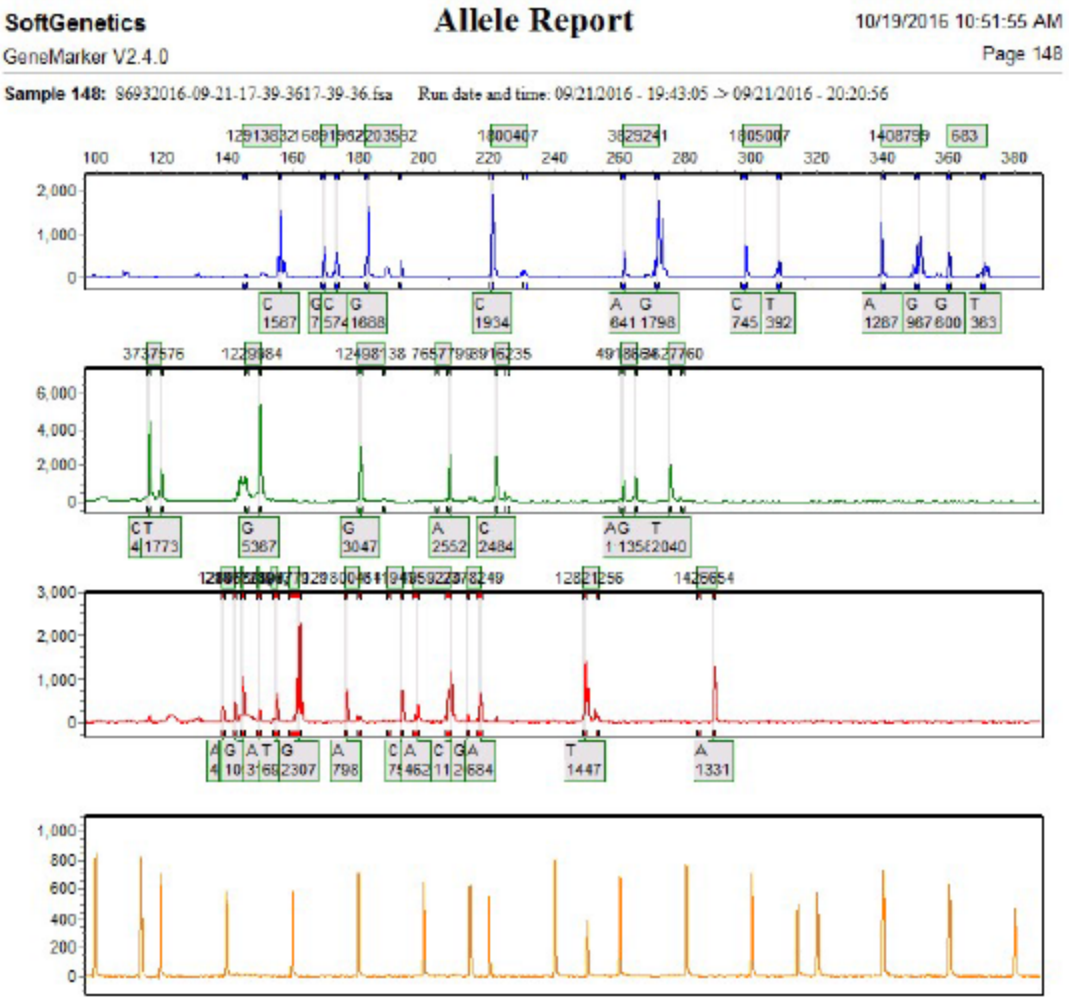

SoftGenetics
Allele Report
10/19/2016 10:51:55 AM
GeneMarker V2.4.0
Page 146
Sample 146: 85762016-09-22-16-13-3916-13-39.fsa Run date and time: 09/22/2016 - 16:14:24 -> 09/22/2016 - 17:01:34

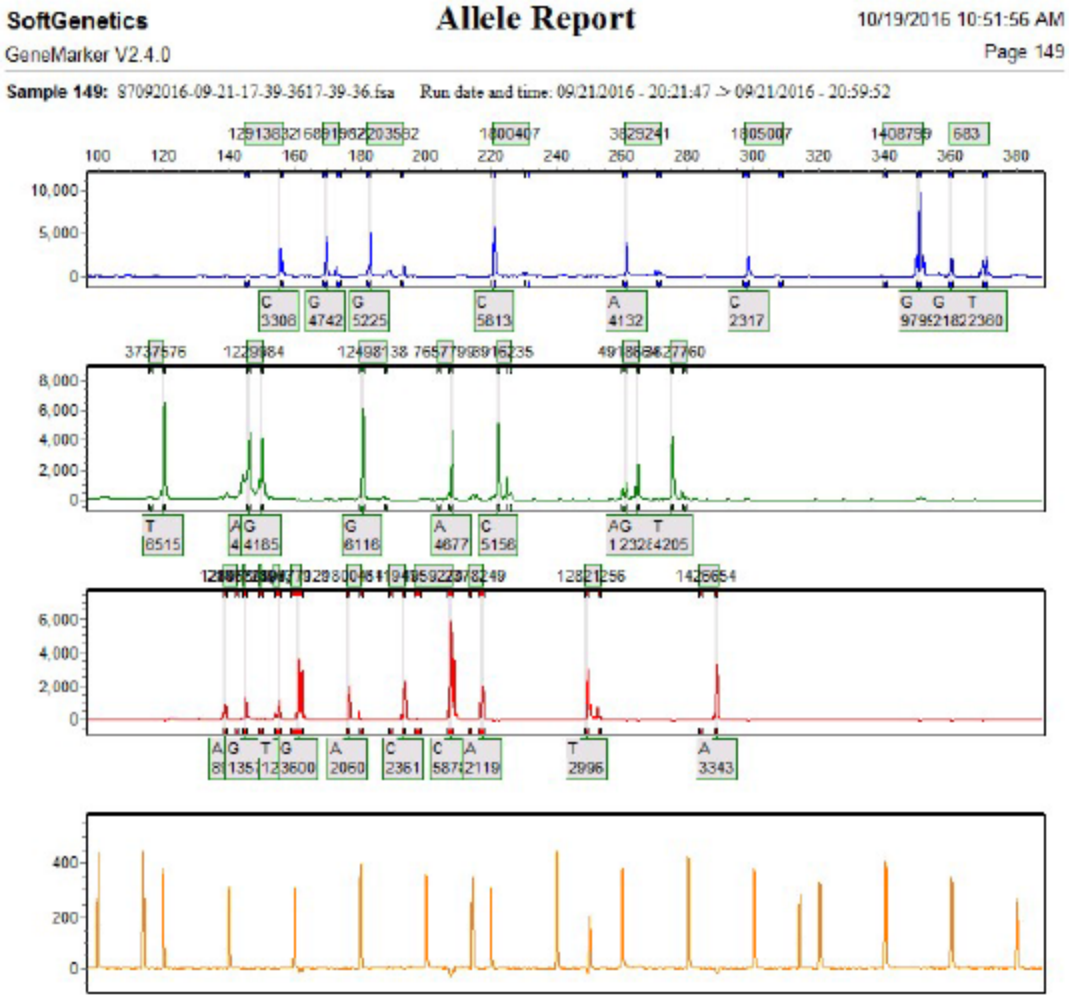
SoftGenetics **Allele Report** 10/19/2016 10:51:55 AM

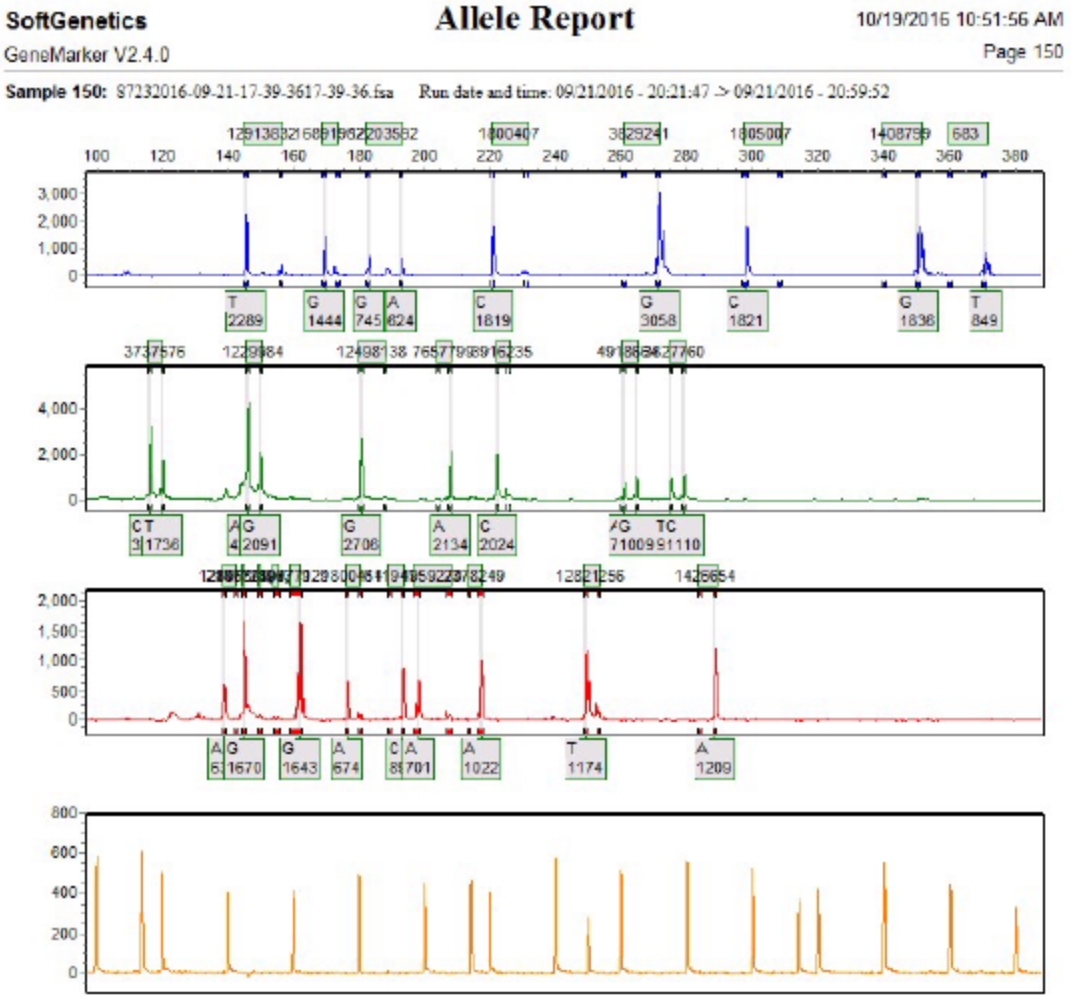
GeneMarker V2.4.0 Page 147

Sample 147: 96172016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 19:43:05 -> 09/21/2016 - 20:20:56




SoftGenetics
Allele Report
10/19/2016 10:51:55 AM
GeneMarker V2.4.0
Page 148
Sample 148: 96932016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 19:43:05 -> 09/21/2016 - 20:20:56

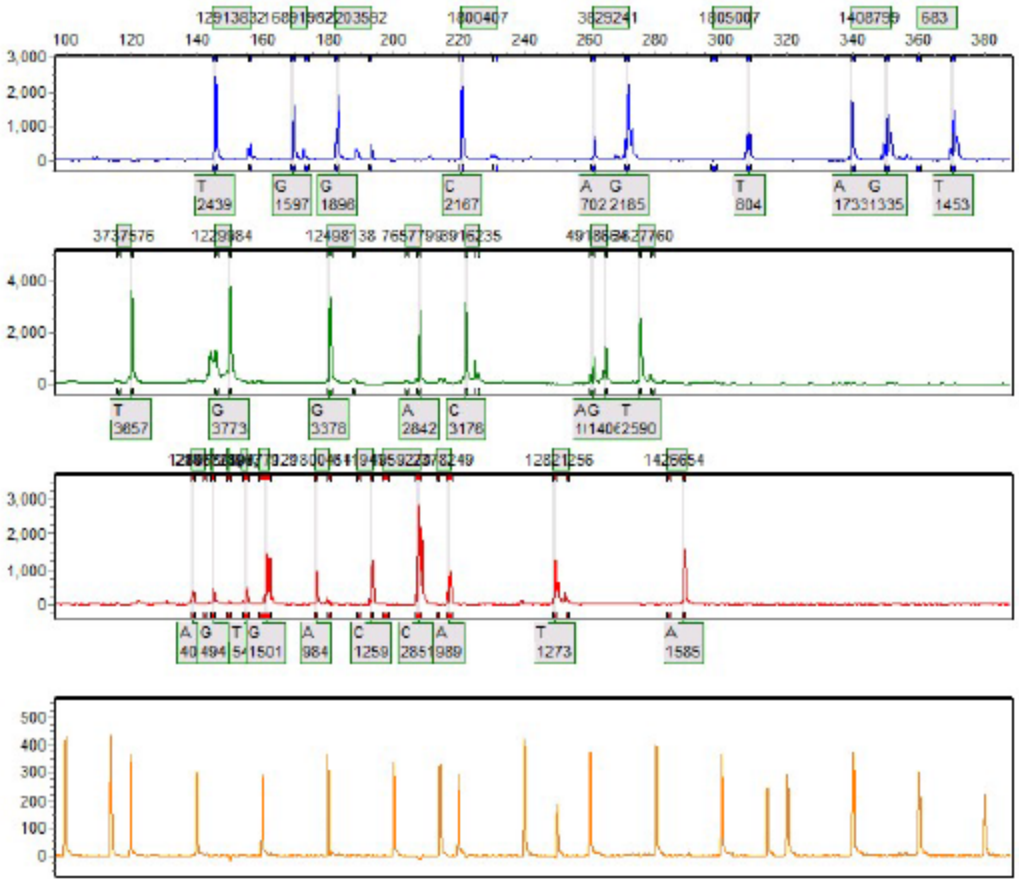

SoftGenetics
GeneMarker V2.4.0
Allele Report
10/19/2016 10:51:56 AM
Page 149
Sample 149: S7092016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 20:21:47 -> 09/21/2016 - 20:59:52


SoftGenetics
Allele Report
10/19/2016 10:51:56 AM
GeneMarker V2.4.0
Page 150
Sample 150: 87232016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 20:21:47 -> 09/21/2016 - 20:59:52

SoftGenetics **Allele Report** 10/19/2016 10:51:56 AM

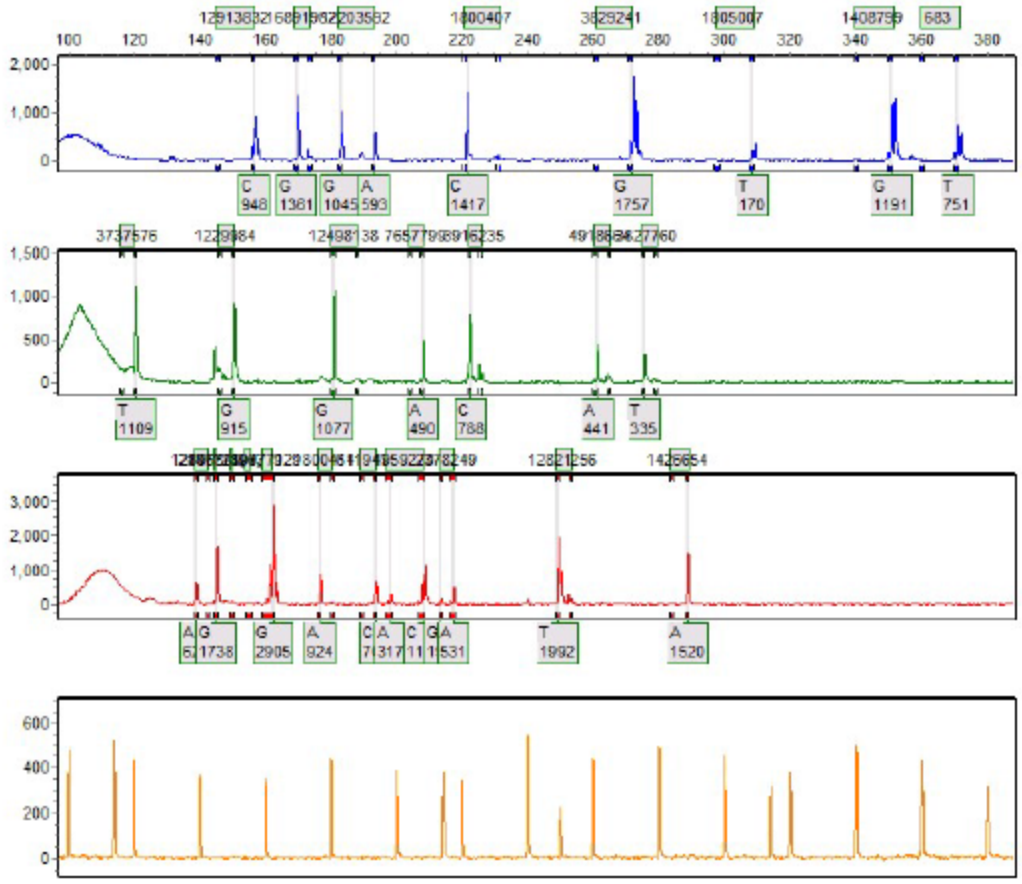
GeneMarker V2.4.0 Page 151

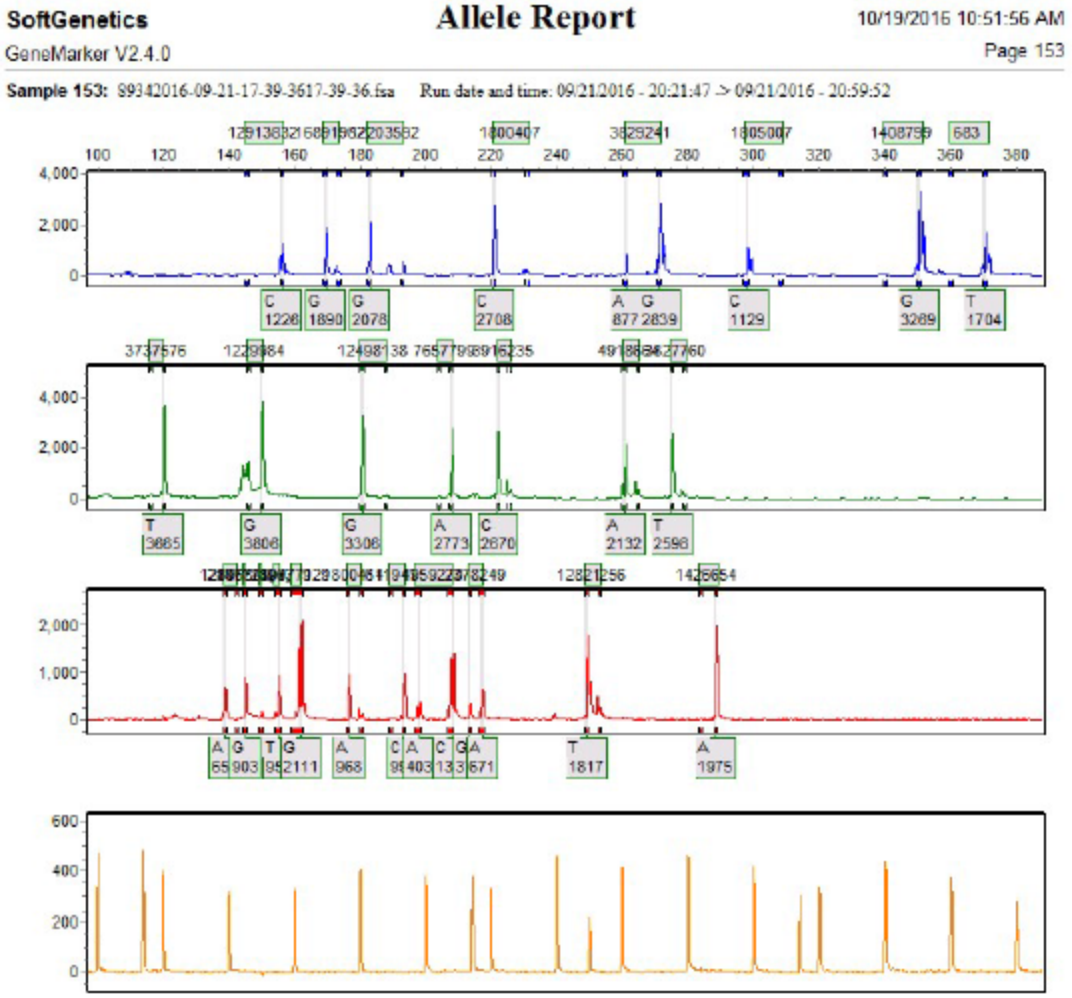
Sample 151: 87302016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 20:21:47 -> 09/21/2016 - 20:59:52

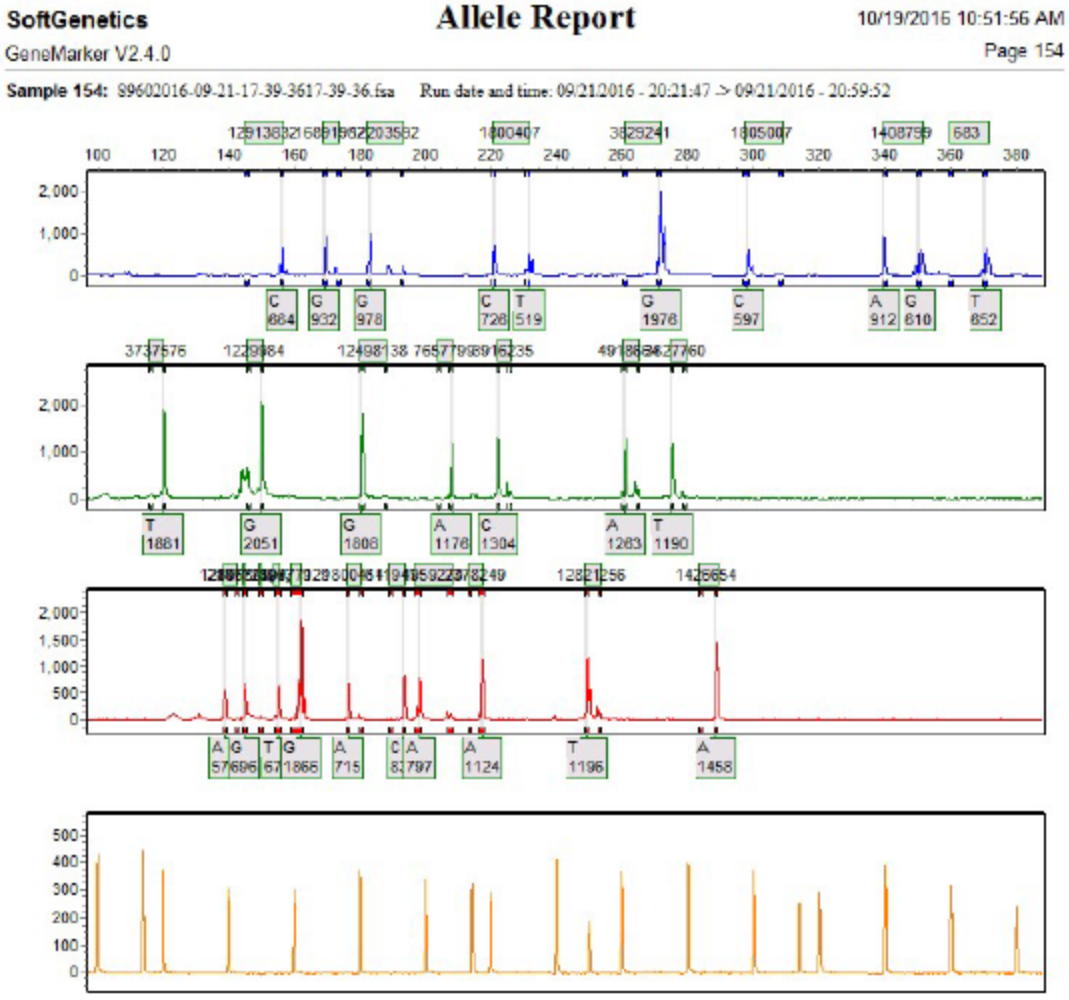


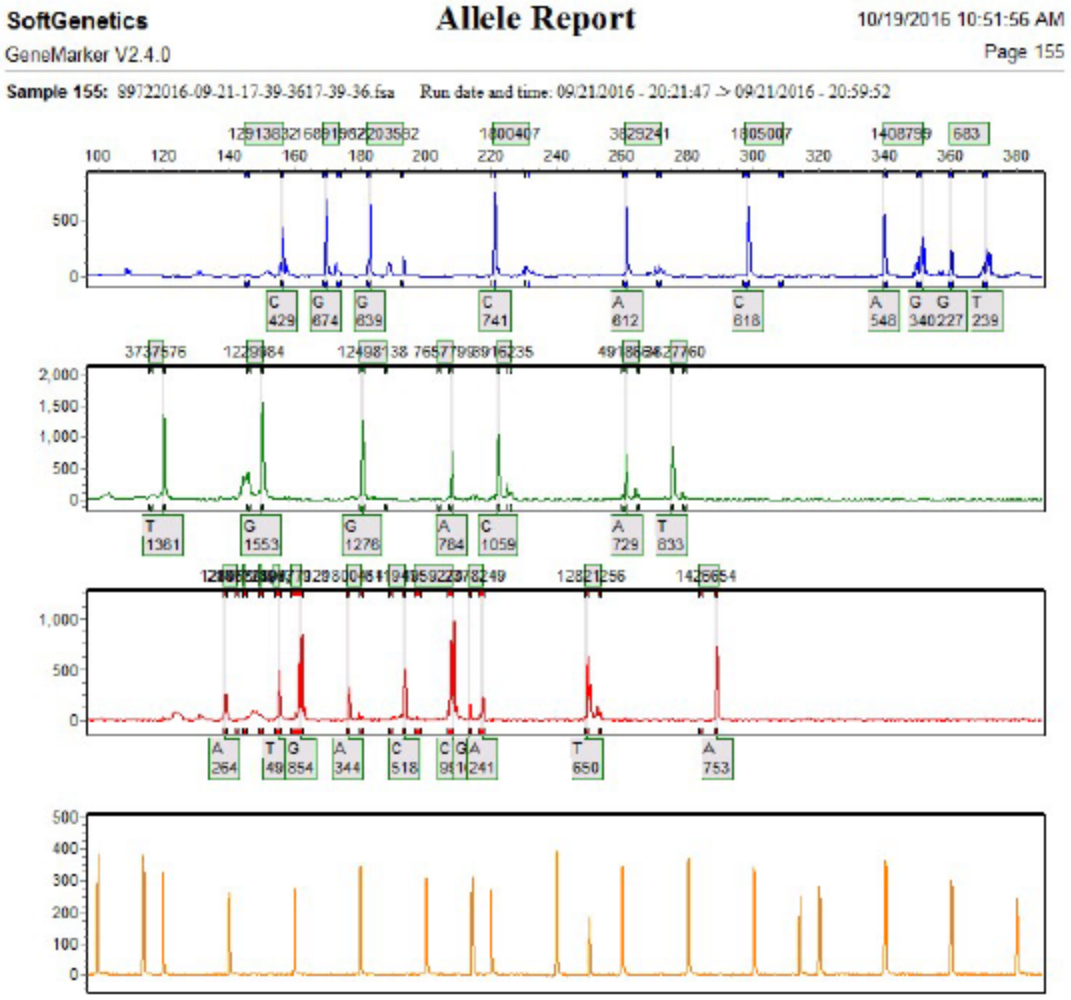
Allele Report

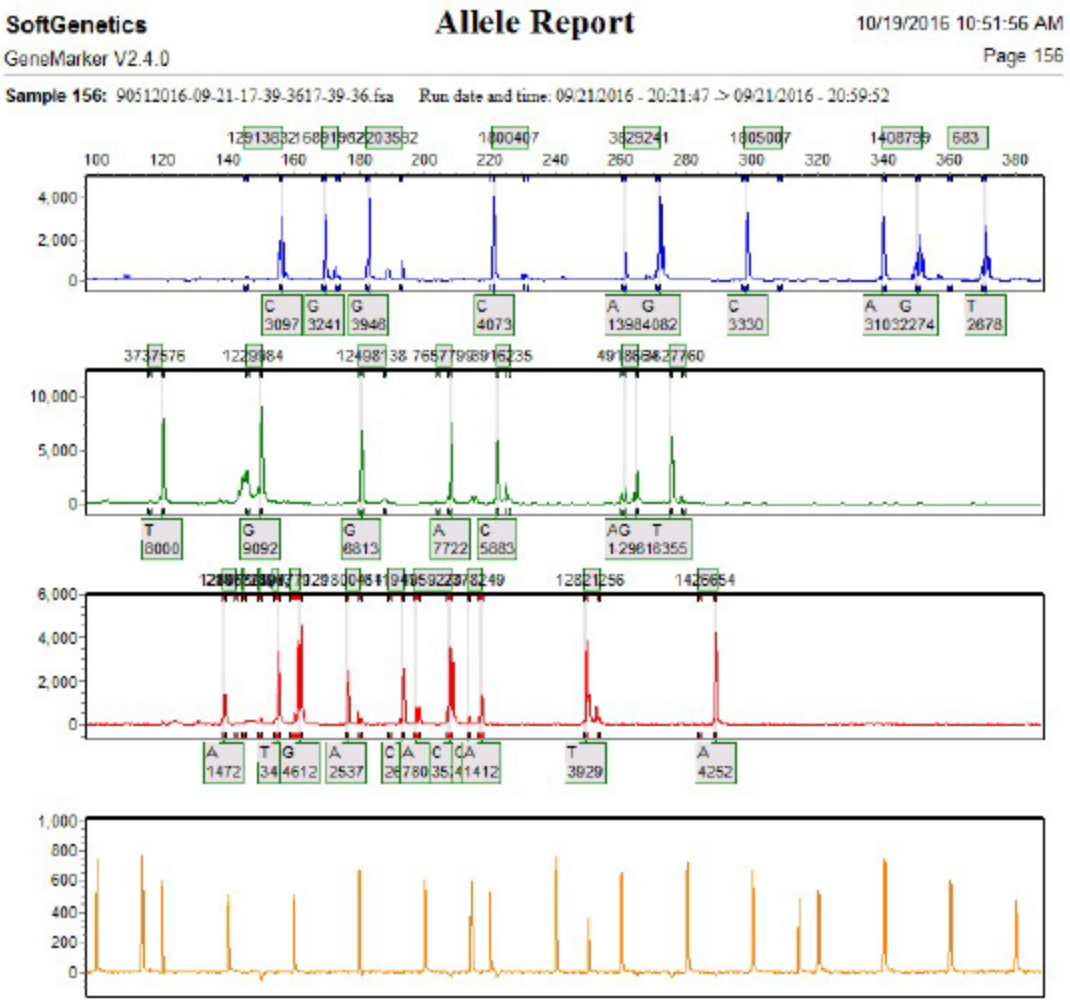
Sample 152: 89012016-09-24-12-28-1612-28-16.fsa Run date and time: 09/24/2016 - 13:18:17 -> 09/24/2016 - 13:55:38

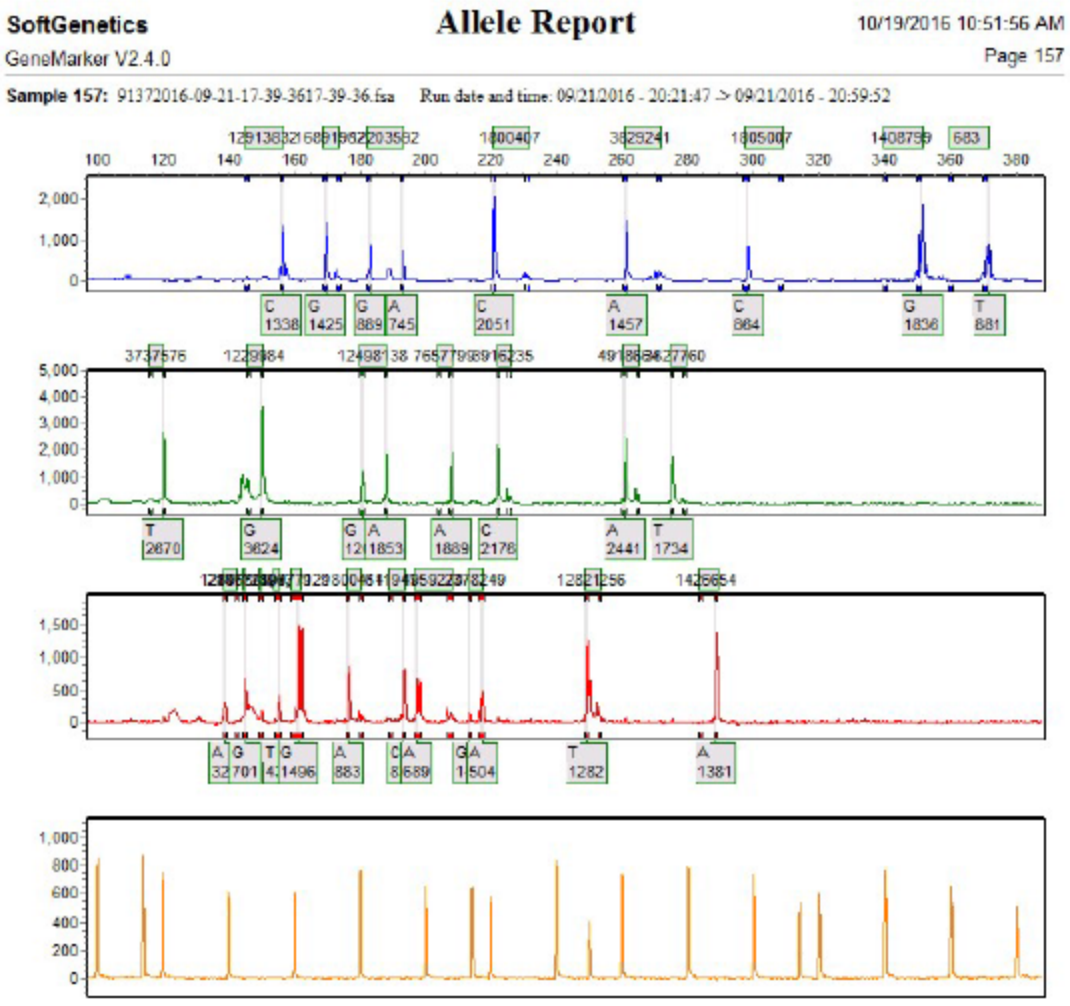


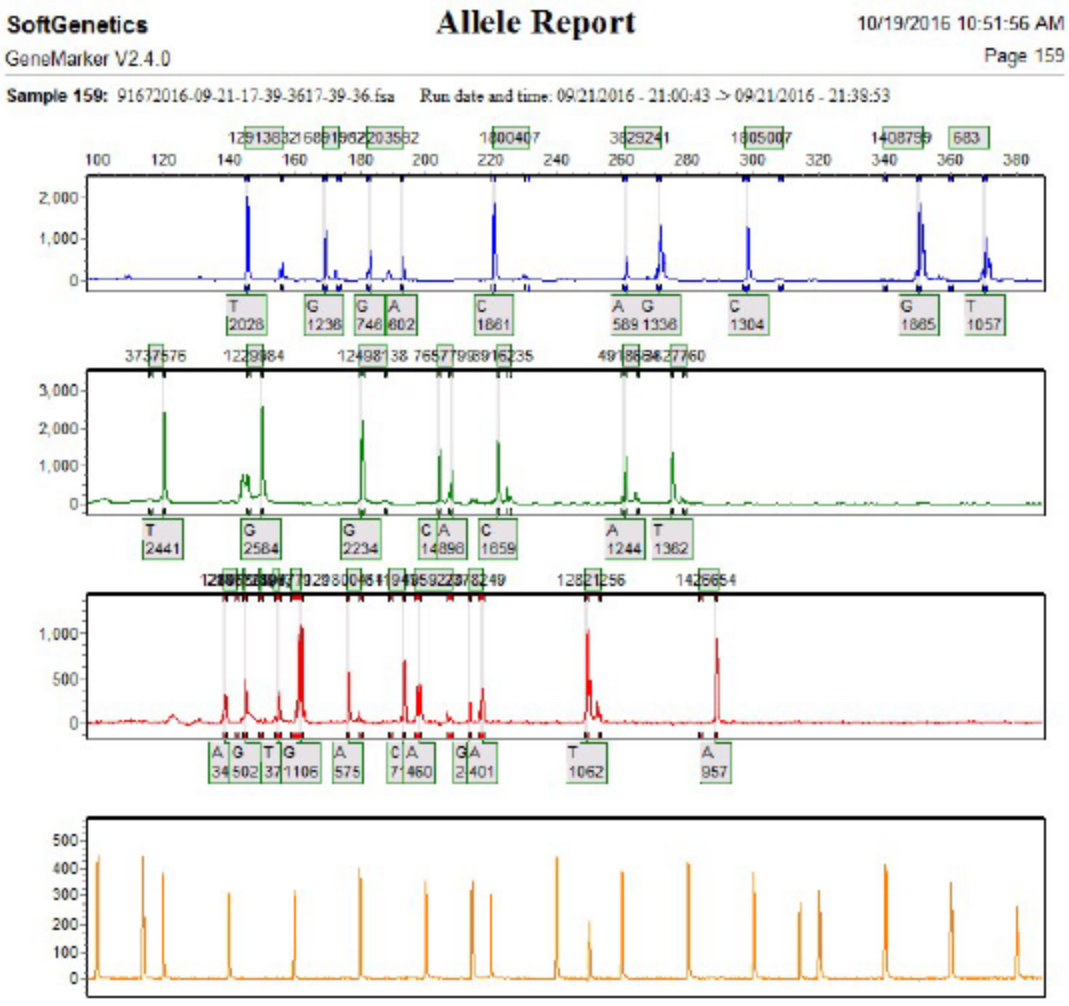

SoftGenetics
Allele Report
10/19/2016 10:51:56 AM
GeneMarker V2.4.0
Page 153
Sample 153: 89342016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 20:21:47 -> 09/21/2016 - 20:59:52

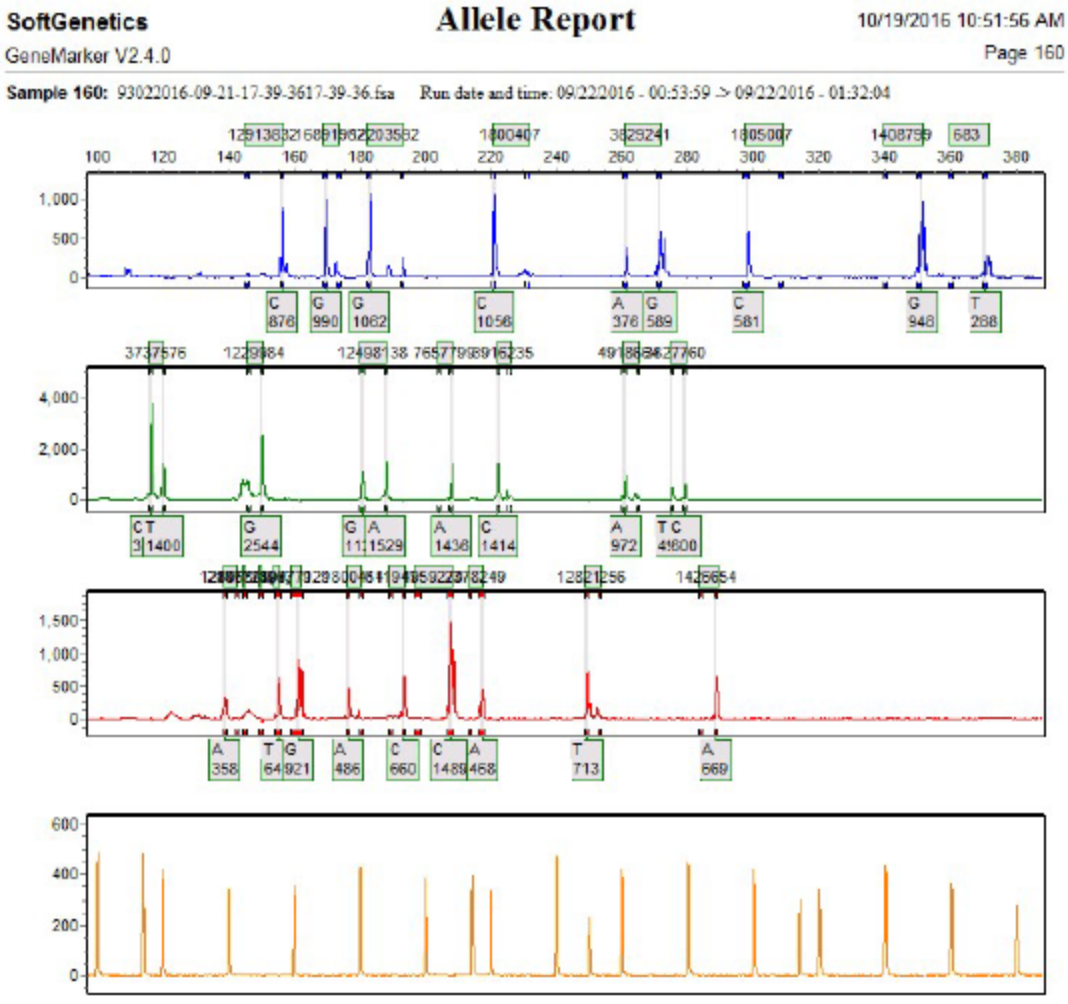

SoftGenetics
Allele Report
10/19/2016 10:51:56 AM
GeneMarker V2.4.0
Page 154
Sample 154: 89602016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 20:21:47 -> 09/21/2016 - 20:59:52

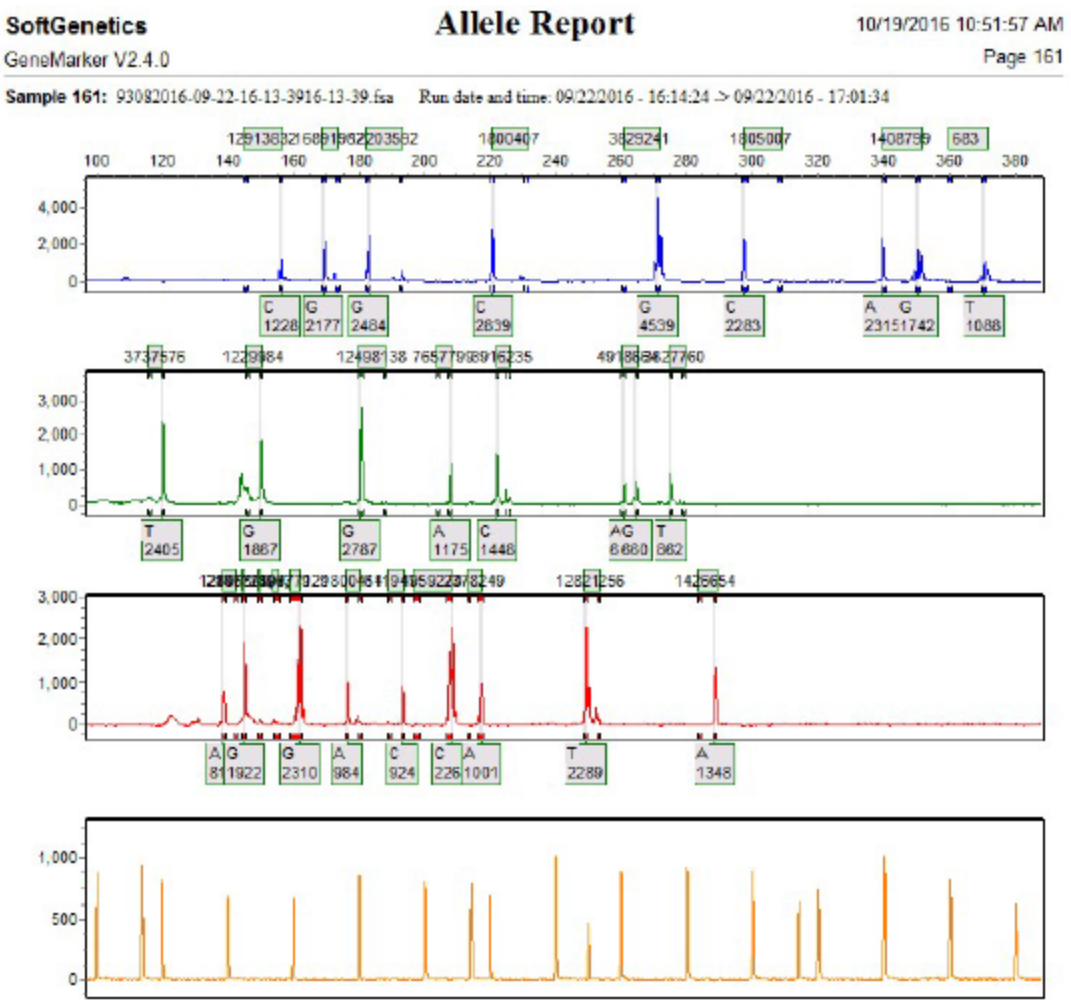

SoftGenetics
Allele Report
10/19/2016 10:51:56 AM
GeneMarker V2.4.0
Page 155
Sample 155: 89722016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 20:21:47 -> 09/21/2016 - 20:59:52


SoftGenetics
Allele Report
10/19/2016 10:51:56 AM
GeneMarker V2.4.0
Page 156
Sample 156: 90512016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 20:21:47 -> 09/21/2016 - 20:59:52

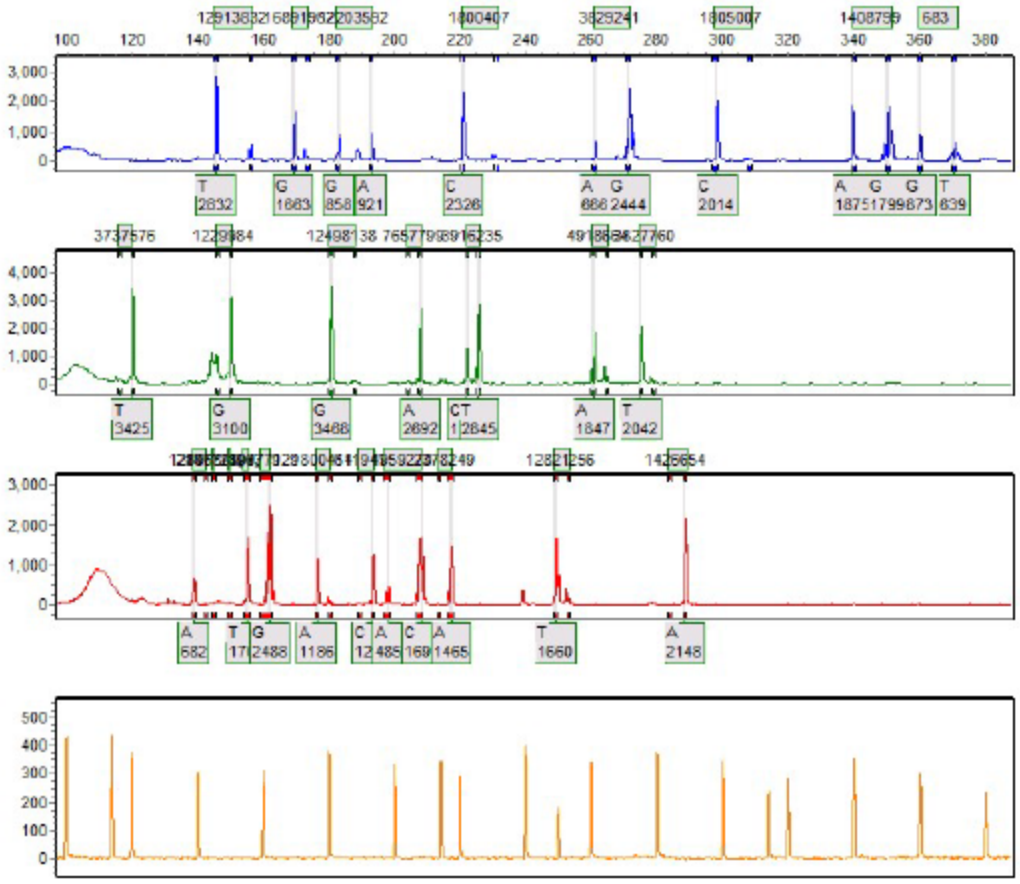

SoftGenetics
Allele Report
10/19/2016 10:51:56 AM
GeneMarker V2.4.0
Page 157
Sample 157: 91372016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 20:21:47 -> 09/21/2016 - 20:59:52


SoftGenetics
Allele Report
10/19/2016 10:51:56 AM
GeneMarker V2.4.0
Page 159
Sample 159: 91672016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 21:00:43 -> 09/21/2016 - 21:38:53


SoftGenetics
Allele Report
10/19/2016 10:51:56 AM
GeneMarker V2.4.0
Page 160
Sample 160: 93022016-09-21-17-39-3617-39-36.fsa Run date and time: 09/22/2016 - 00:53:59 -> 09/22/2016 - 01:32:04


SoftGenetics
Allele Report
10/19/2016 10:51:57 AM
GeneMarker V2.4.0
Page 161
Sample 161: 93082016-09-22-16-13-3916-13-39.fsa Run date and time: 09/22/2016 - 16:14:24 -> 09/22/2016 - 17:01:34

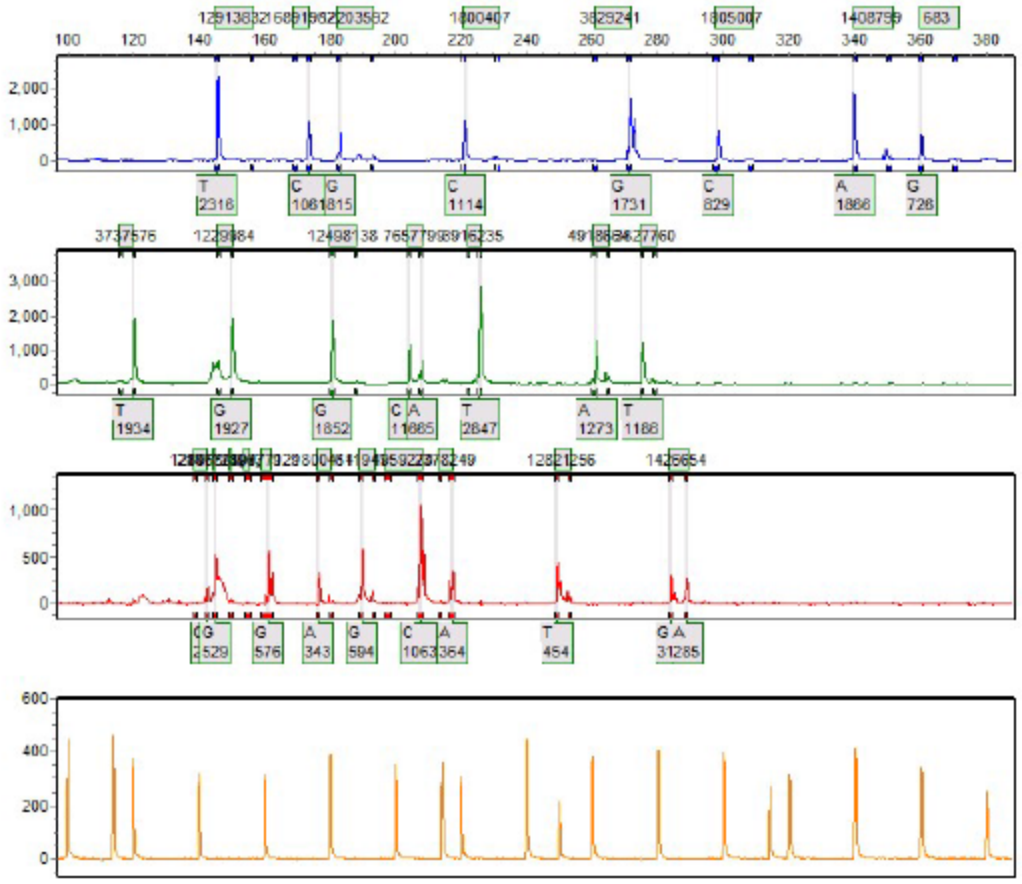
Sample 162: 93452016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 21:00:43 -> 09/21/2016 - 21:38:53



SoftGenetics **Allele Report** 10/19/2016 10:51:57 AM

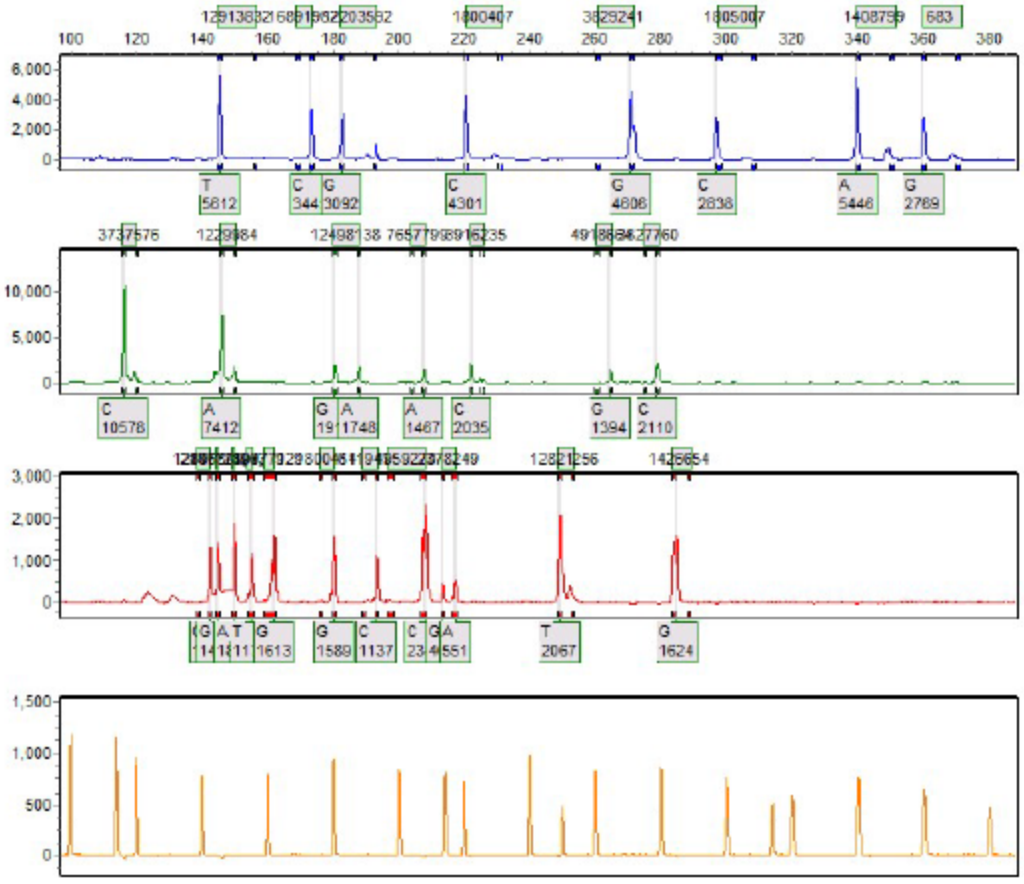
GeneMarker V2.4.0 Page 163

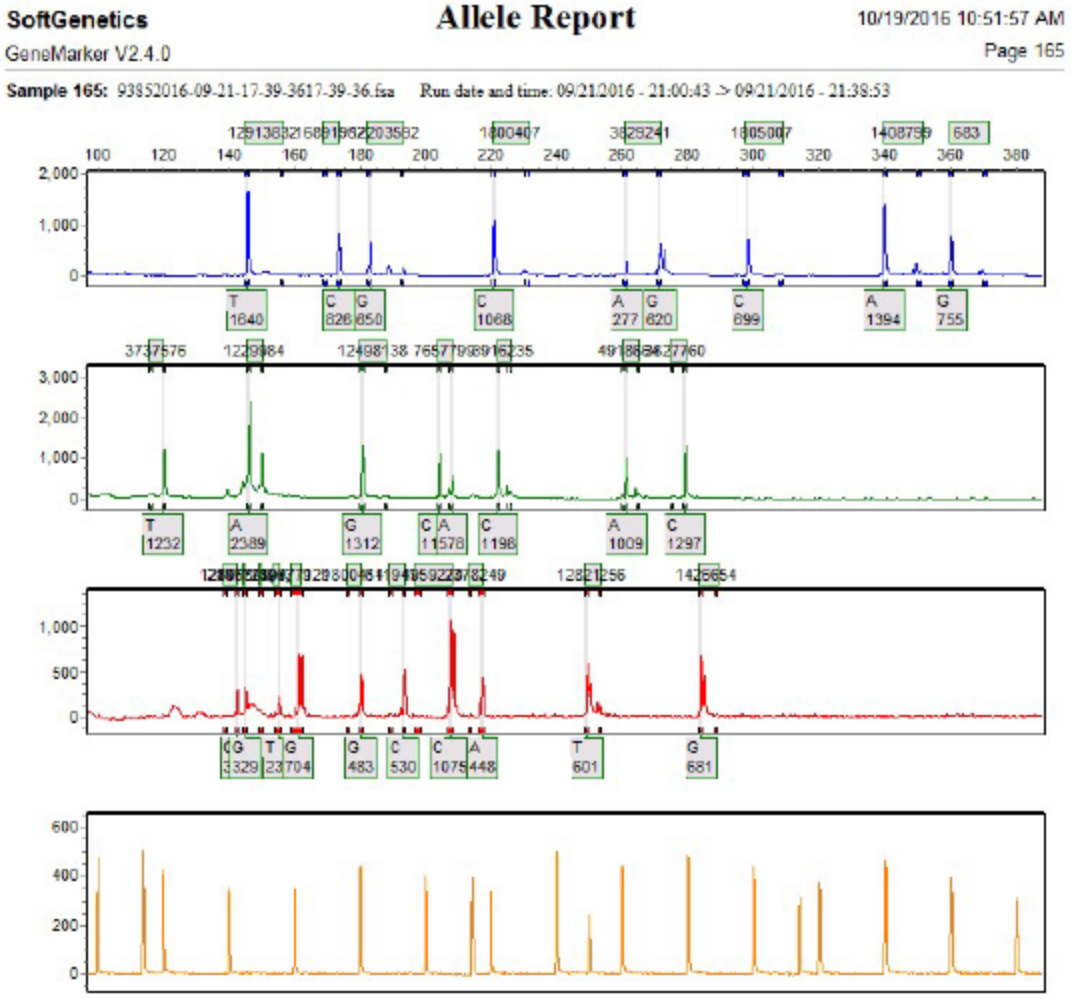
Sample 163: 93562016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 21:00:43 -> 09/21/2016 - 21:38:53

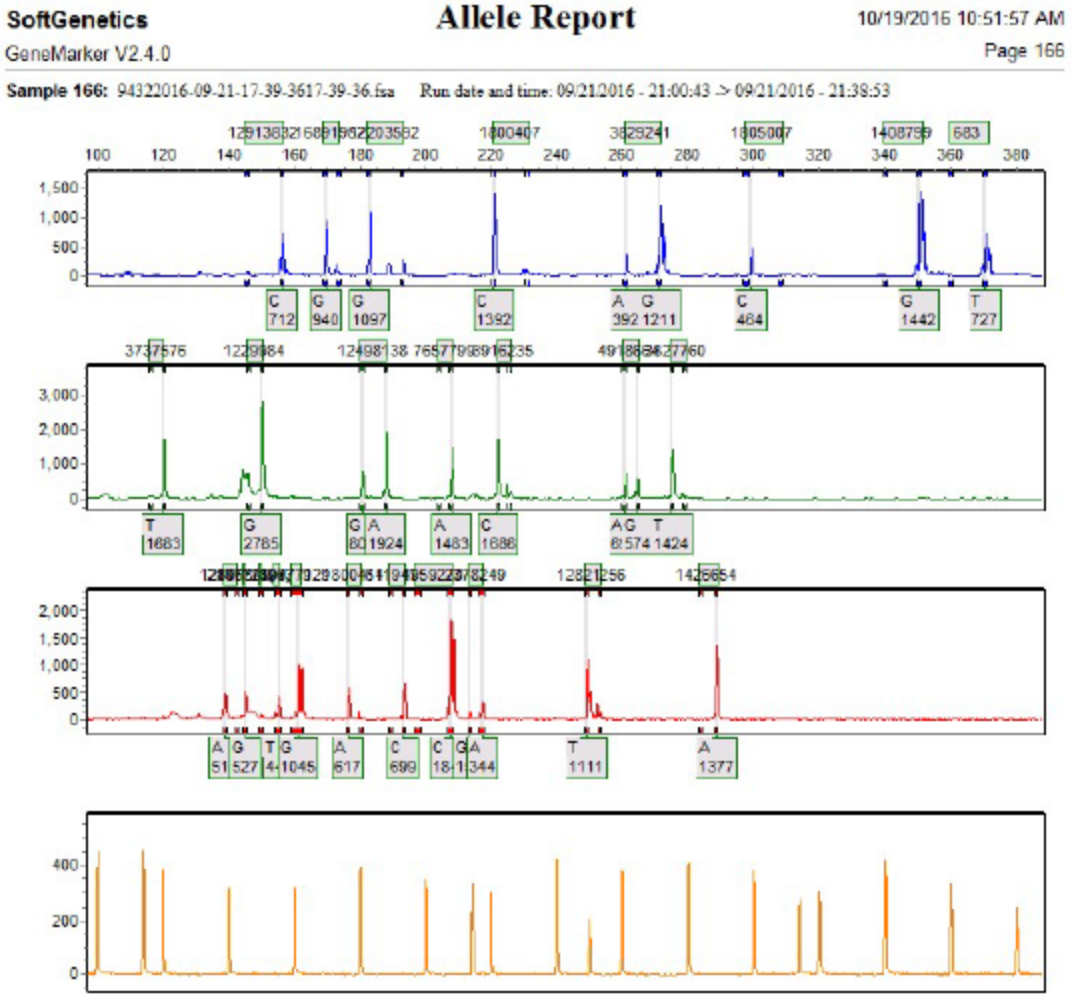


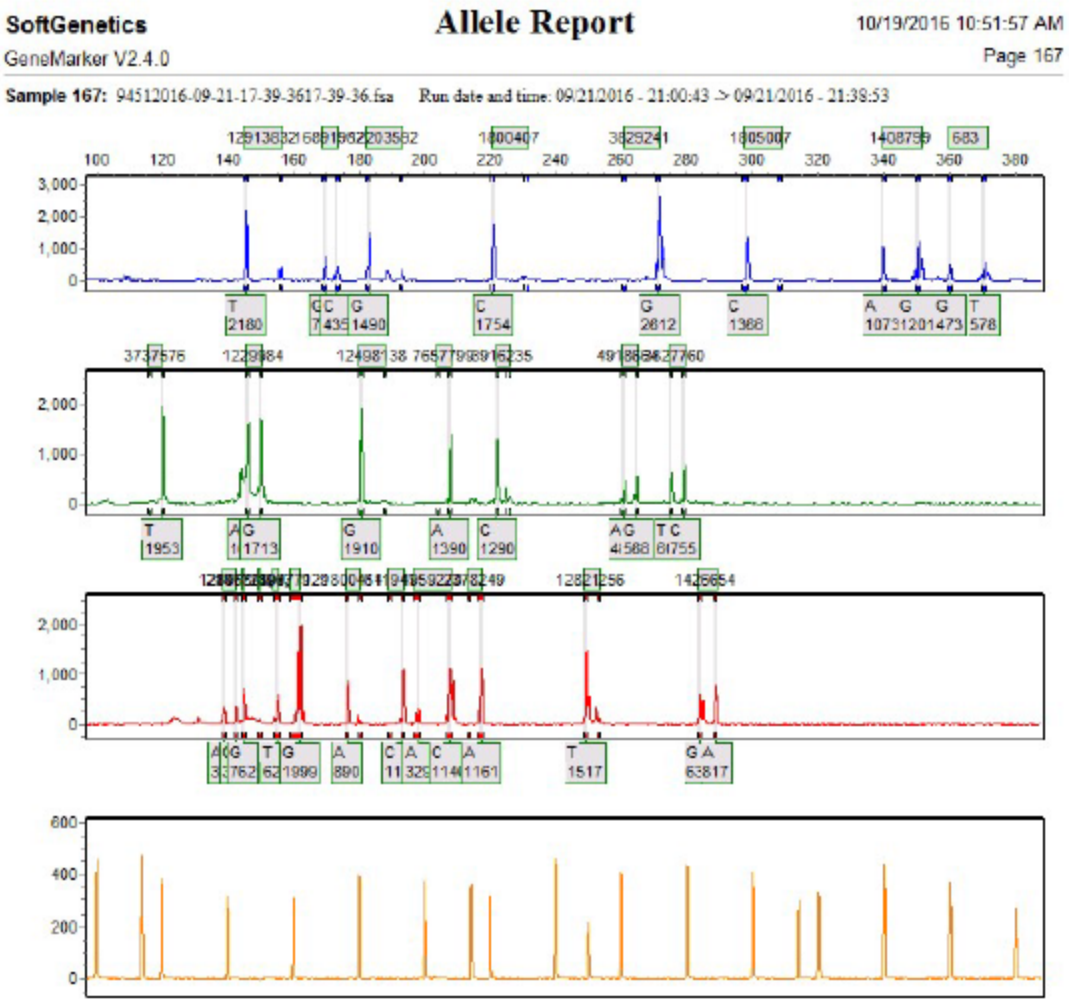
Allele Report

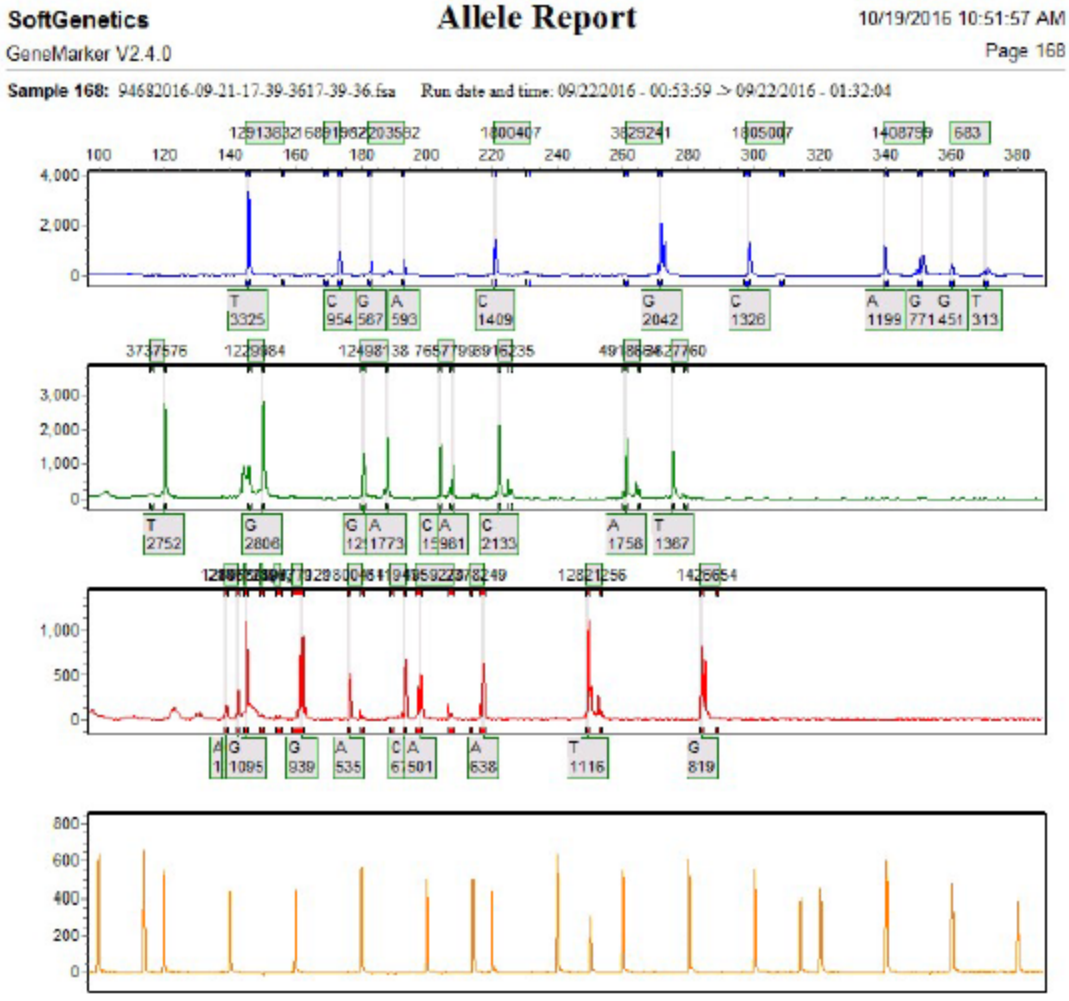
Sample 164: 93762016-09-09-08-34-2408-34-24.fsa Run date and time: 09/09/2016 - 08:35:03 -> 09/09/2016 - 09:15:23

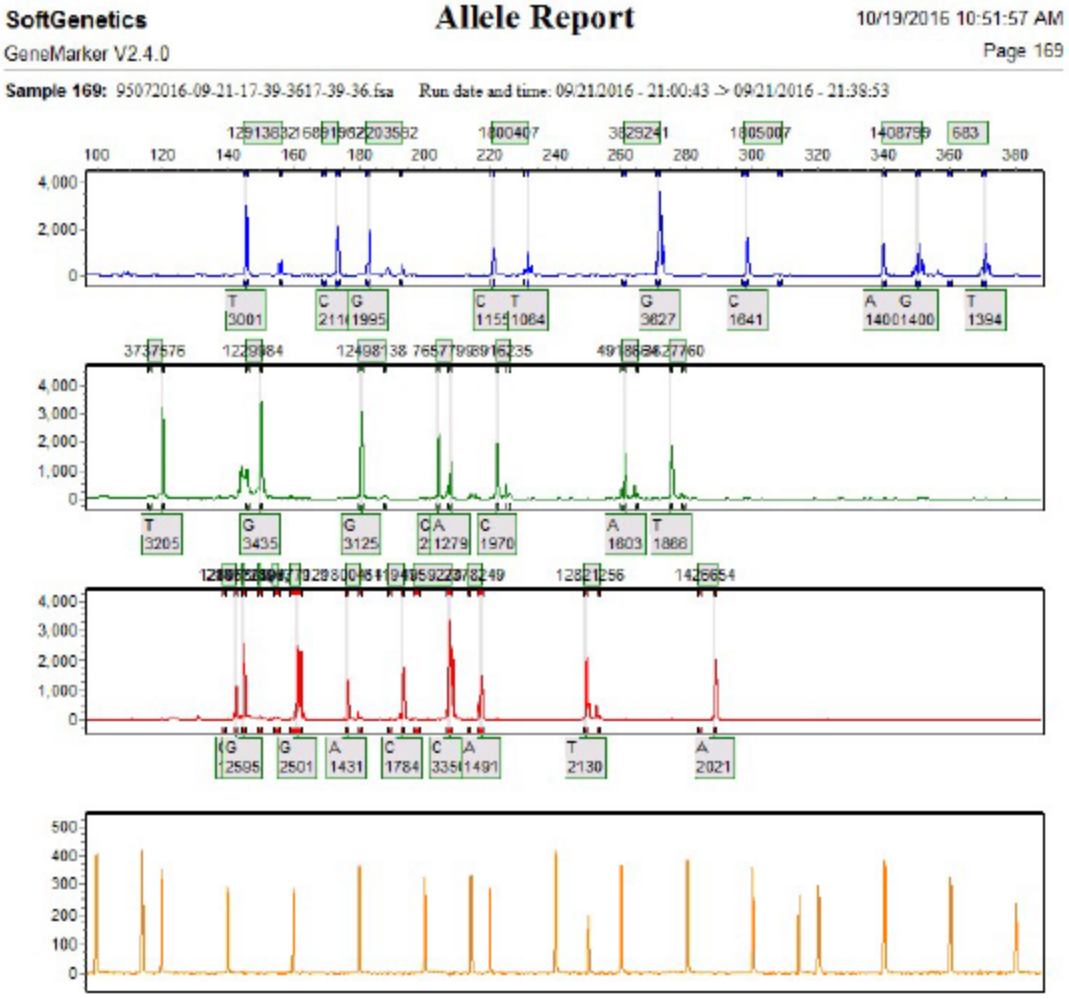



SoftGenetics
Allele Report
10/19/2016 10:51:57 AM
GeneMarker V2.4.0
Page 165
Sample 165: 93852016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 21:00:43 -> 09/21/2016 - 21:38:53


SoftGenetics
Allele Report
10/19/2016 10:51:57 AM
GeneMarker V2.4.0
Page 166
Sample 166: 94322016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 21:00:43 -> 09/21/2016 - 21:38:53


SoftGenetics
Allele Report
10/19/2016 10:51:57 AM
GeneMarker V2.4.0
Page 167
Sample 167: 94512016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 21:00:43 -> 09/21/2016 - 21:38:53

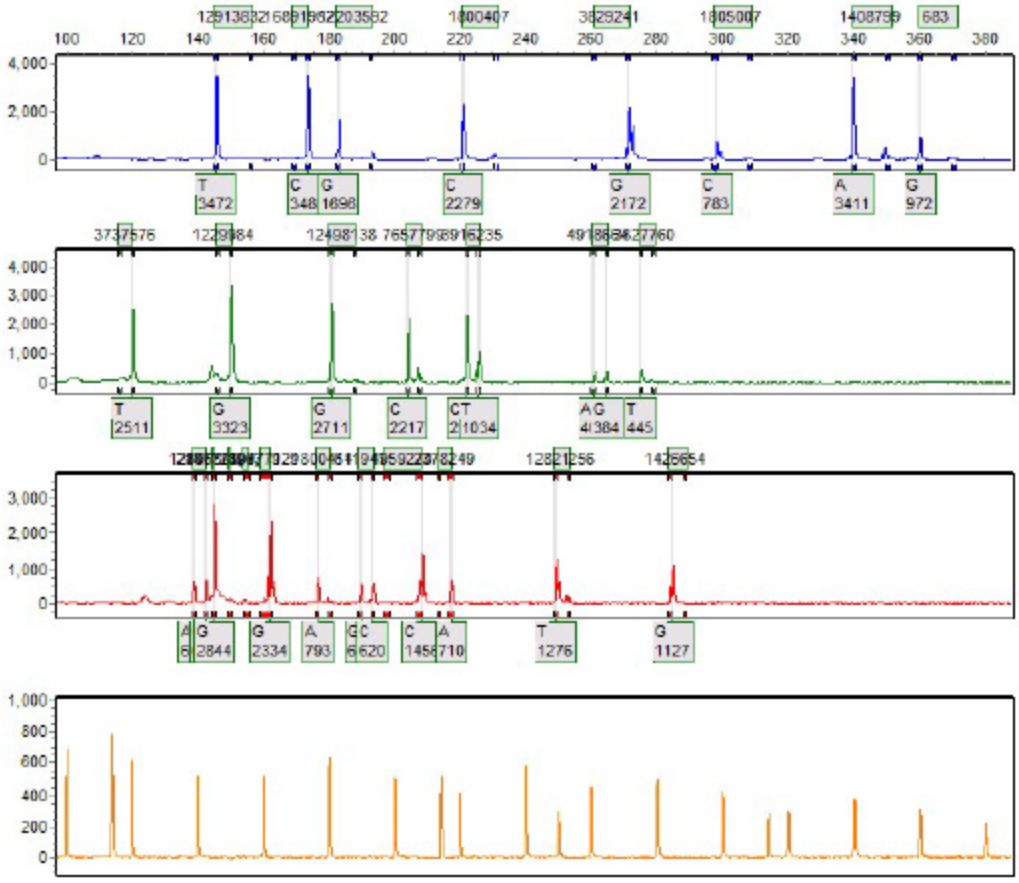

SoftGenetics
Allele Report
10/19/2016 10:51:57 AM
GeneMarker V2.4.0
Page 168
Sample 168: 94682016-09-21-17-39-3617-39-36.fsa Run date and time: 09/22/2016 - 00:53:59 -> 09/22/2016 - 01:32:04

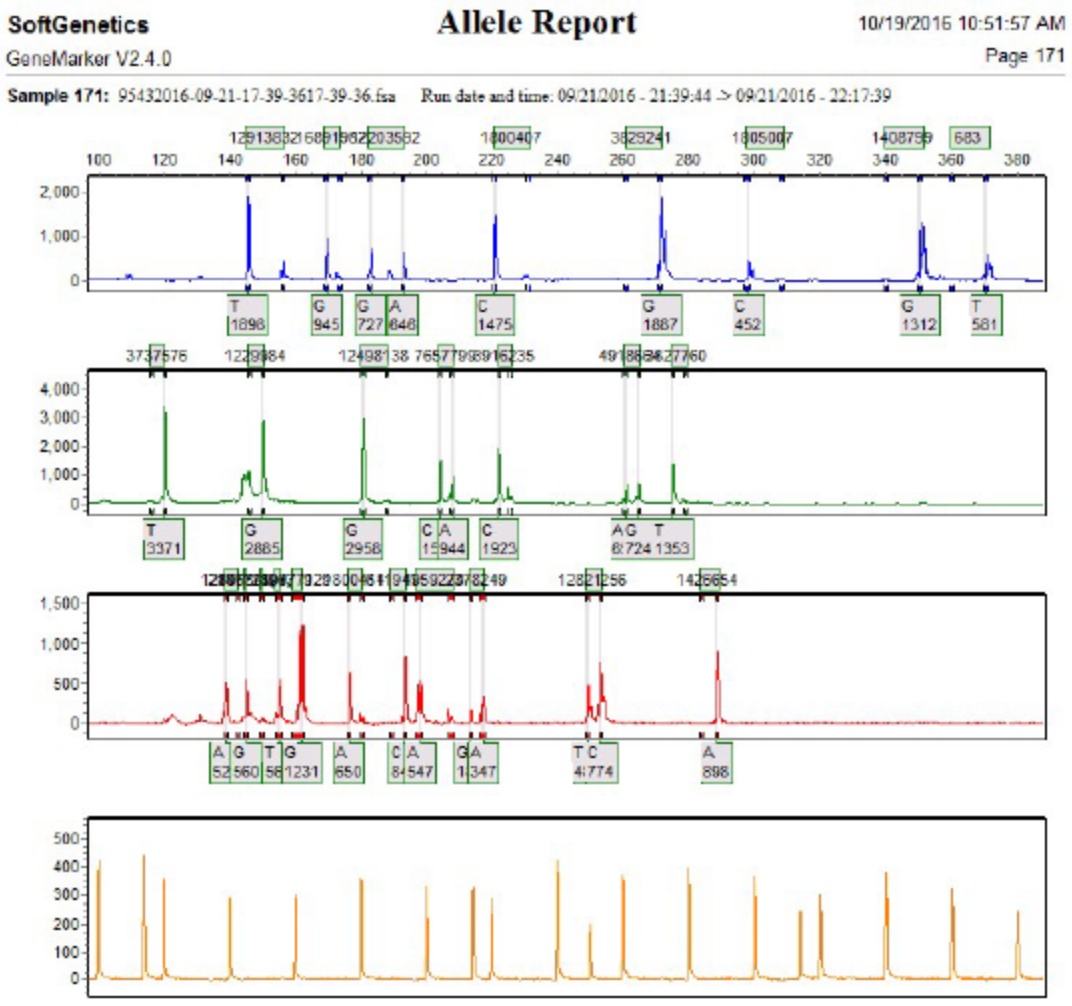

SoftGenetics
Allele Report
10/19/2016 10:51:57 AM
GeneMarker V2.4.0
Page 169
Sample 169: 95072016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 21:00:43 -> 09/21/2016 - 21:38:53

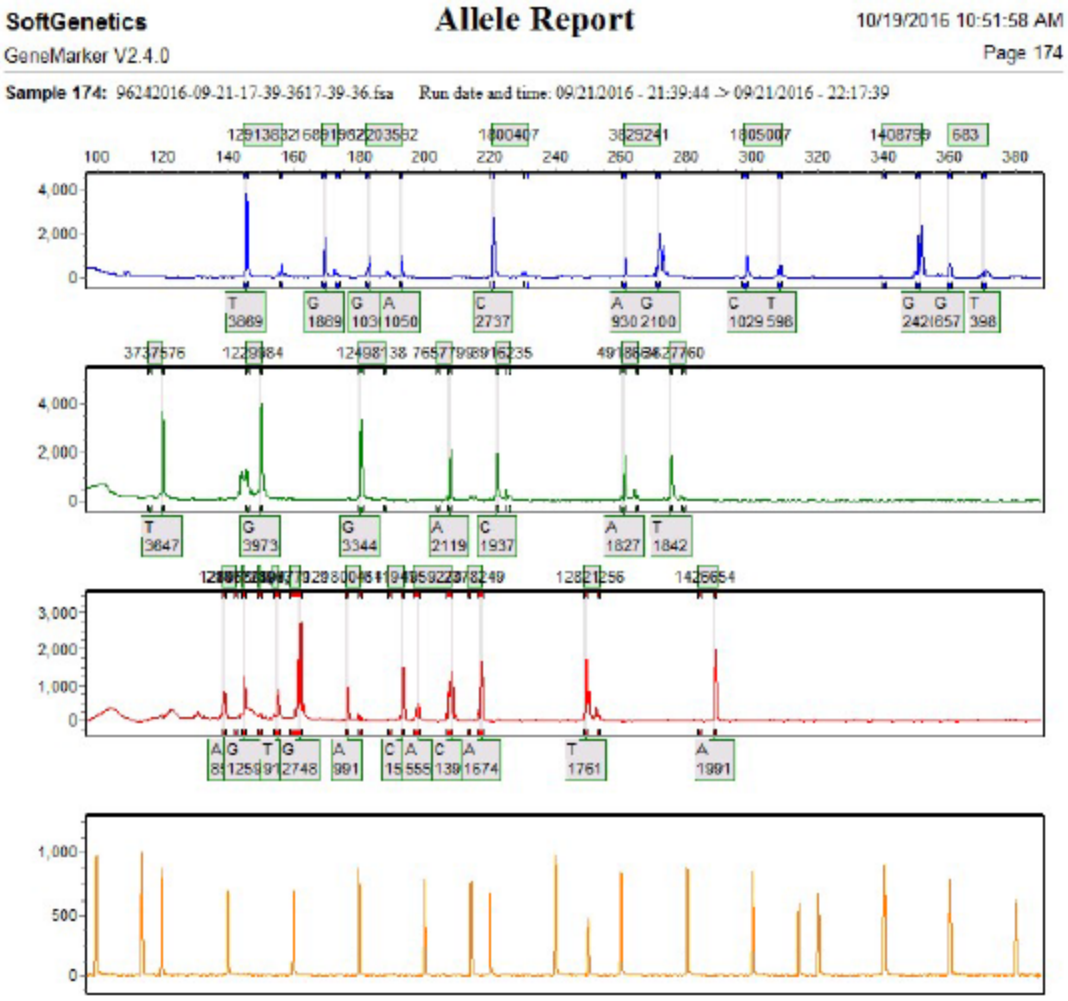
SoftGenetics **Allele Report** 10/19/2016 10:51:57 AM

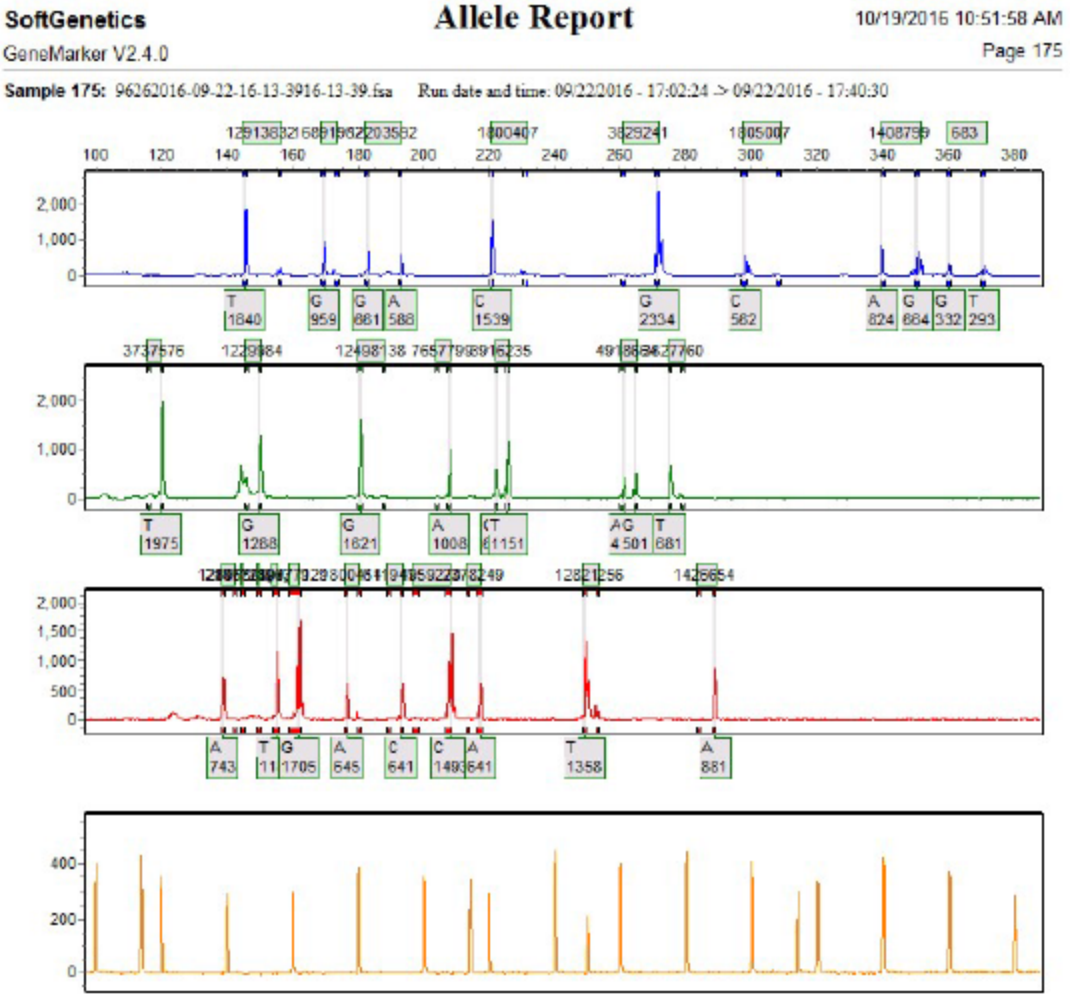
GeneMarker V2.4.0 Page 170

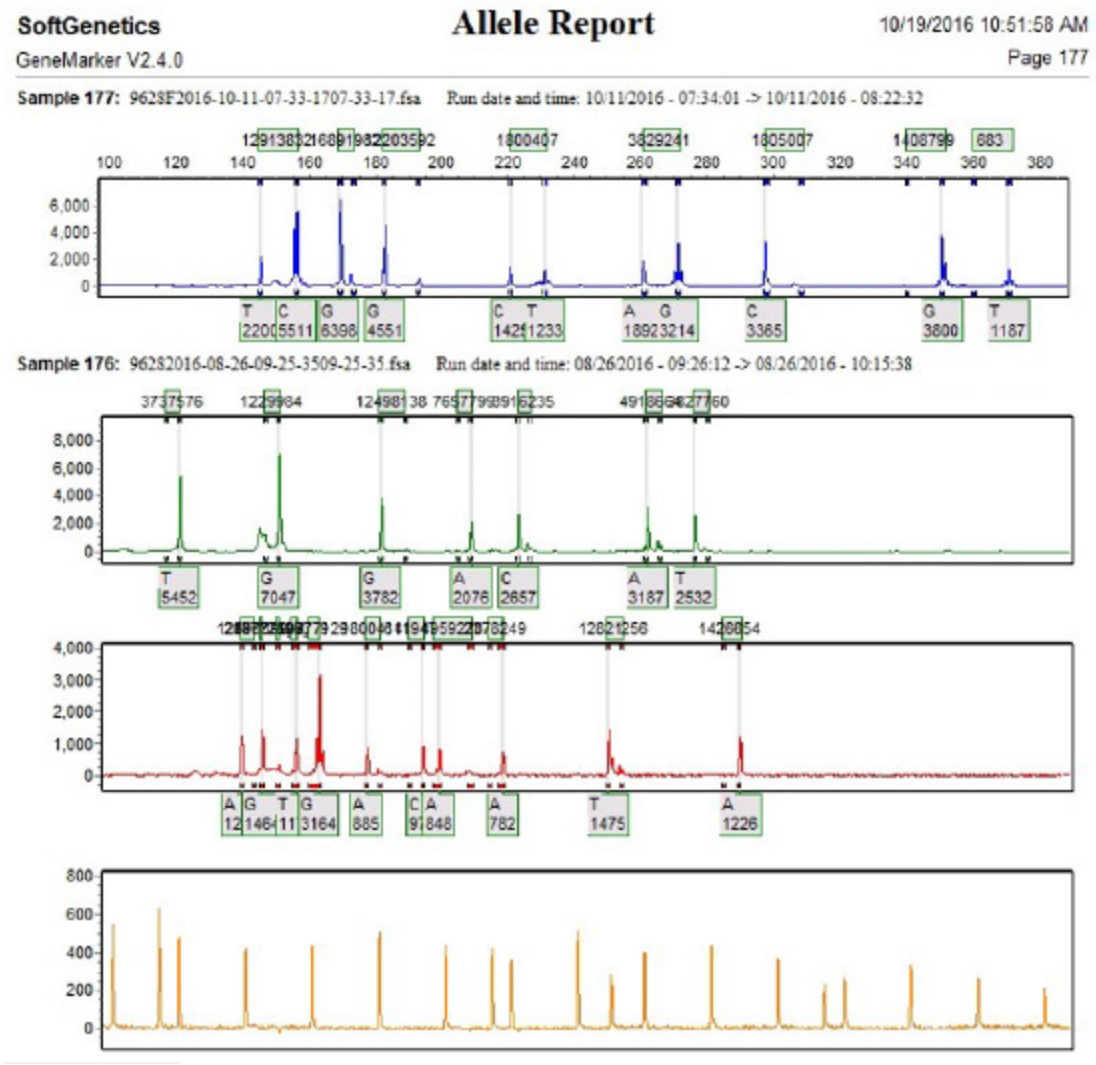
Sample 170: 95212016-08-26-09-25-3509-25-35.fsa Run date and time: 08/26/2016 - 10:16:25 -> 08/26/2016 - 10:54:30

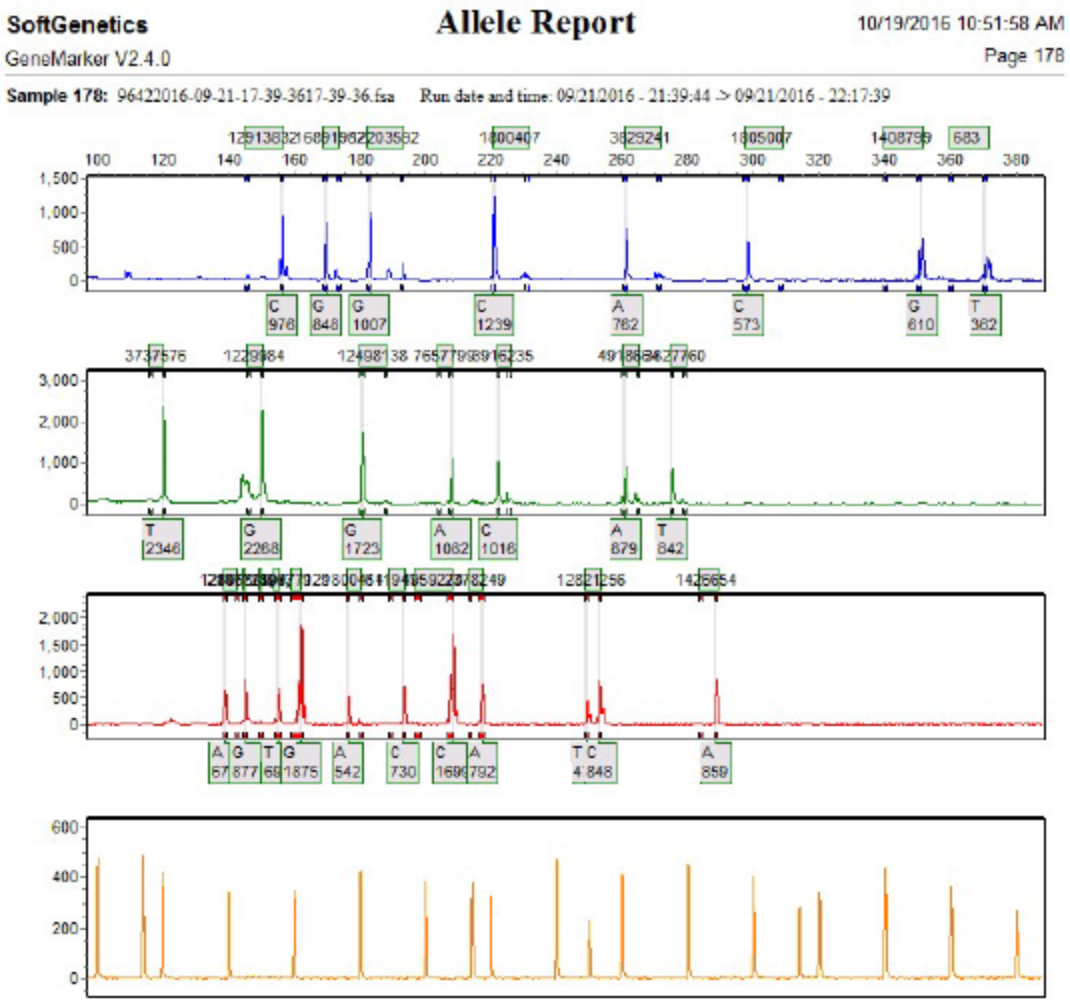


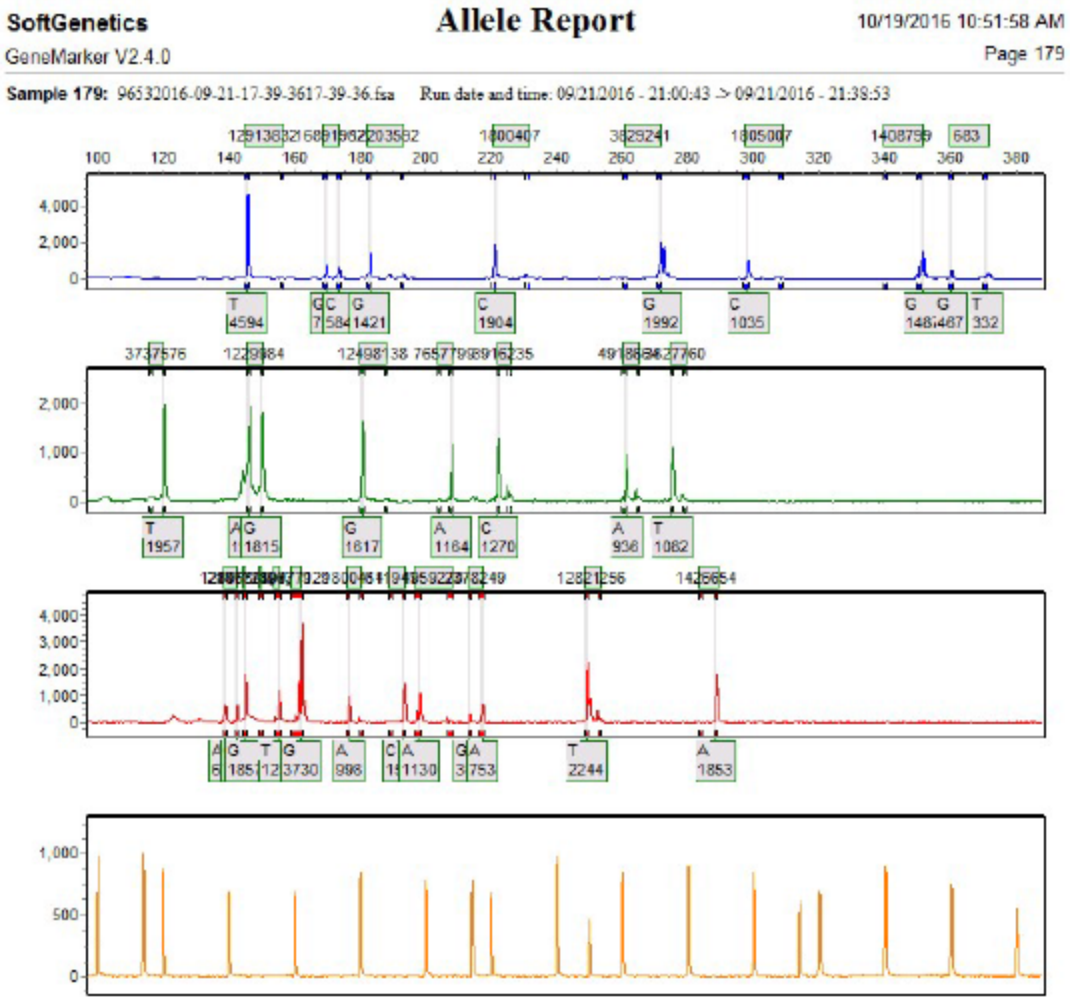

SoftGenetics
Allele Report
10/19/2016 10:51:57 AM
GeneMarker V2.4.0
Page 171
Sample 171: 95432016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 21:39:44 -> 09/21/2016 - 22:17:39


SoftGenetics
Allele Report
10/19/2016 10:51:58 AM
GeneMarker V2.4.0
Page 174
Sample 174: 96242016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 21:39:44 -> 09/21/2016 - 22:17:39


SoftGenetics
Allele Report
10/19/2016 10:51:58 AM
GeneMarker V2.4.0
Page 175
Sample 175: 96262016-09-22-16-13-3916-13-39.fsa Run date and time: 09/22/2016 - 17:02:24 -> 09/22/2016 - 17:40:30


SoftGenetics
Allele Report
10/19/2016 10:51:58 AM
GeneMarker V2.4.0
Page 177
Sample 177: 9628F2016-10-11-07-33-1707-33-17.fsa Run date and time: 10/11/2016 - 07:34:01 -> 10/11/2016 - 08:22:32
Sample 176: 9628 2016-08-26-09-25-3509-25-35.fsa Run date and time: 08/26/2016 - 09:26:12 -> 08/26/2016 - 10:15:38

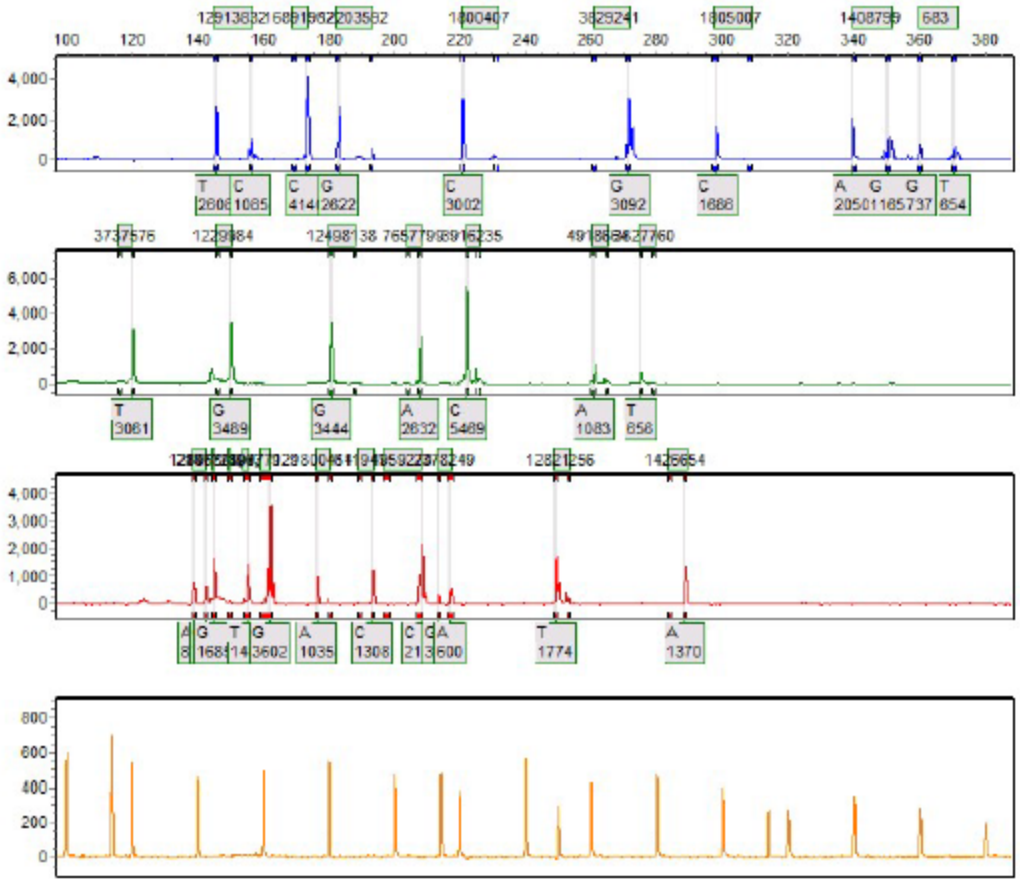

SoftGenetics
Allele Report
10/19/2016 10:51:58 AM
GeneMarker V2.4.0
Page 178
Sample 178: 96422016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 21:39:44 -> 09/21/2016 - 22:17:39

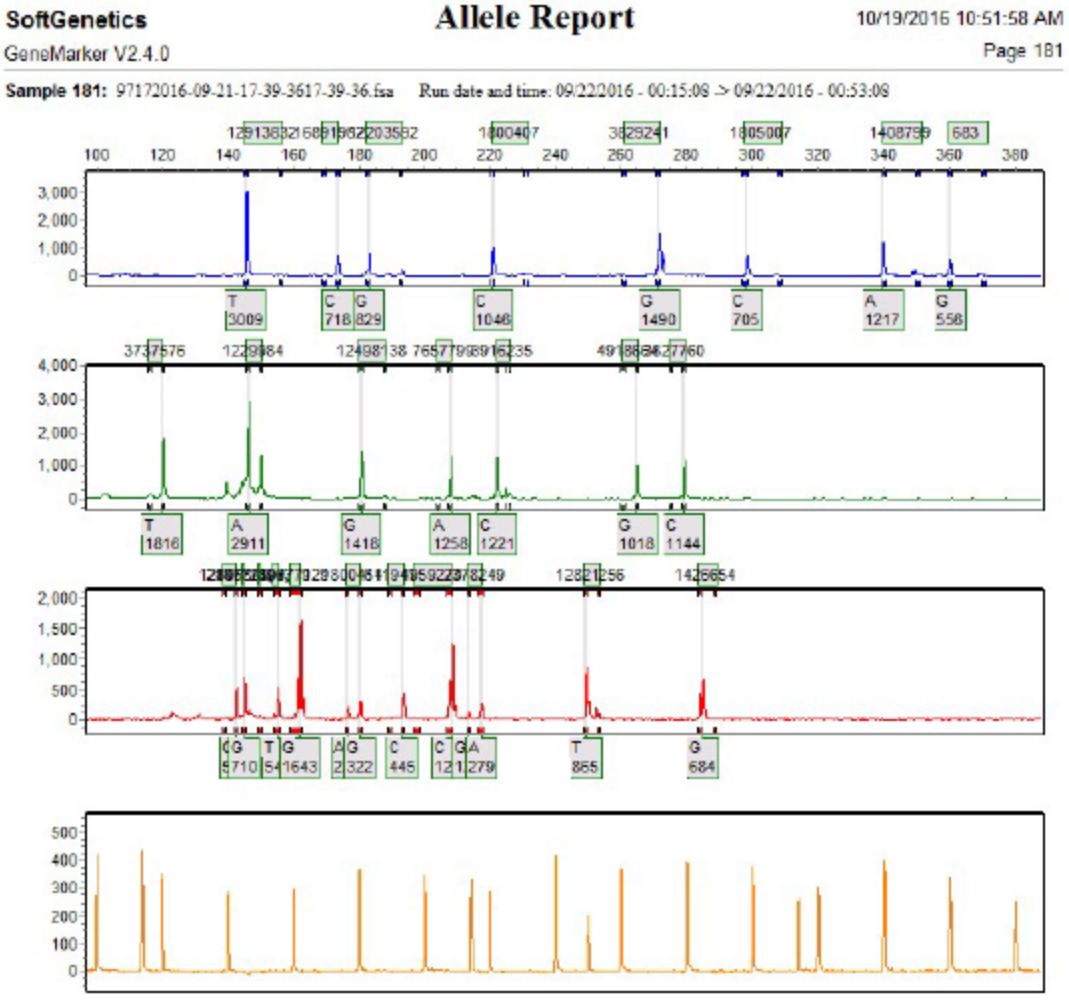

SoftGenetics
Allele Report
10/19/2016 10:51:58 AM
GeneMarker V2.4.0
Page 179
Sample 179: 96532016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 21:00:43 -> 09/21/2016 - 21:38:53

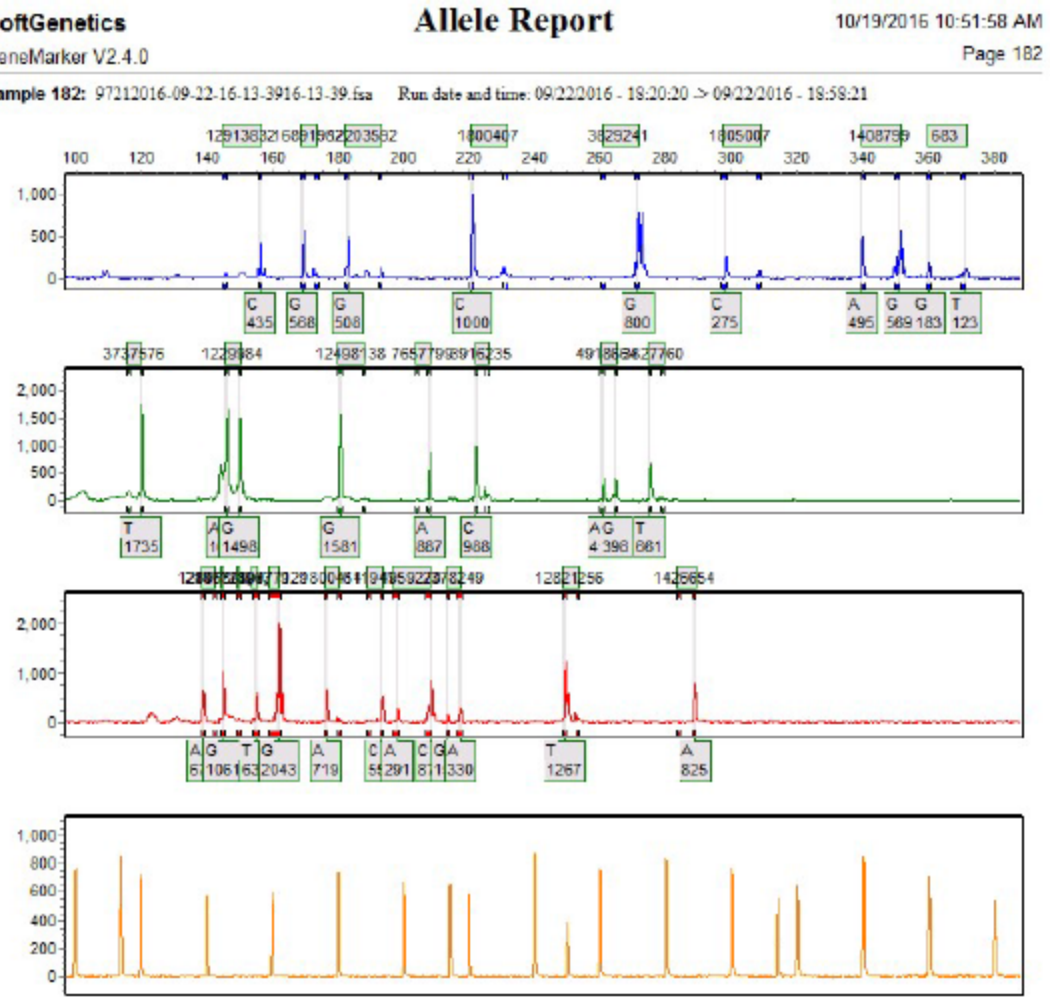
SoftGenetics **Allele Report** 10/19/2016 10:51:58 AM

GeneMarker V2.4.0 Page 180

Sample 180: 97152016-08-26-09-25-3509-25-35.fsa Run date and time: 08/26/2016 - 10:16:25 -> 08/26/2016 - 10:54:30



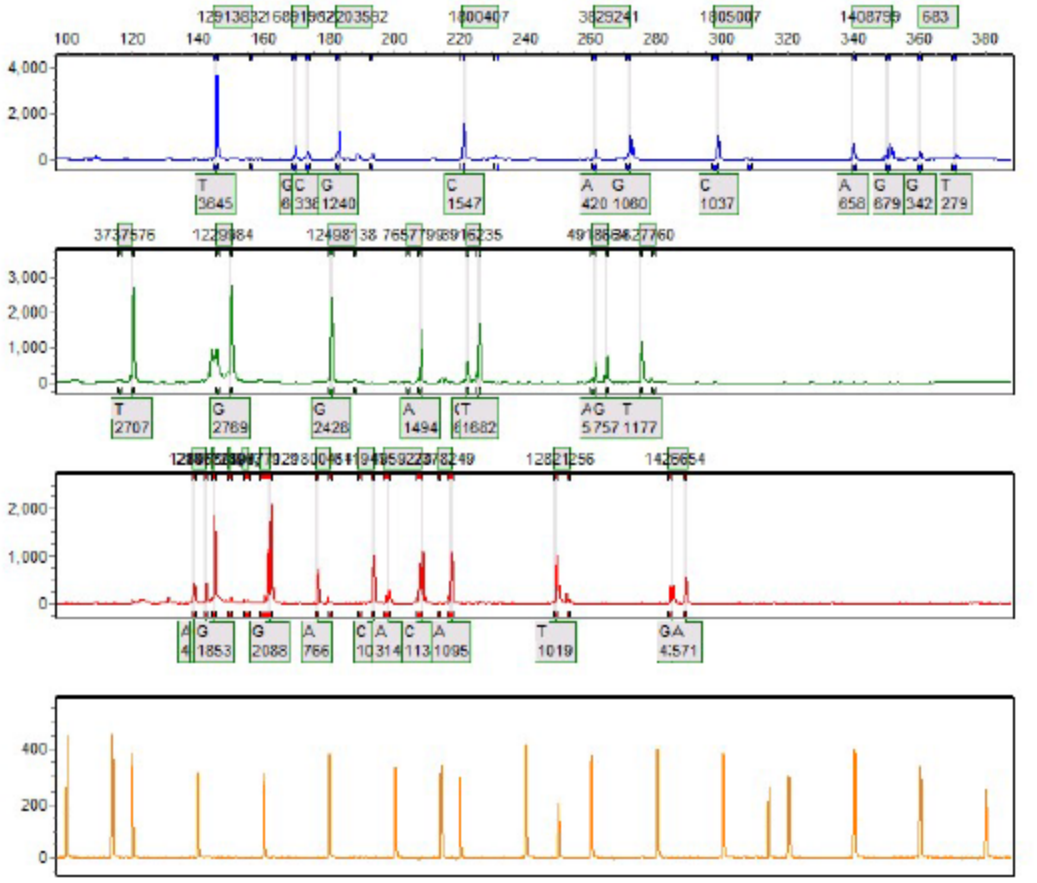

SoftGenetics
Allele Report
10/19/2016 10:51:58 AM
GeneMarker V2.4.0
Page 181
Sample 181: 97172016-09-21-17-39-3617-39-36.fsa Run date and time: 09/22/2016 - 00:15:08 -> 09/22/2016 - 00:53:08


SoftGenetics
Allele Report
10/19/2016 10:51:58 AM
GeneMarker V2.4.0
Page 182
Sample 182: 97212016-09-22-16-13-3916-13-39.fsa Run date and time: 09/22/2016 - 18:20:20 -> 09/22/2016 - 18:58:21

SoftGenetics **Allele Report** 10/19/2016 10:51:58 AM

GeneMarker V2.4.0 Page 183

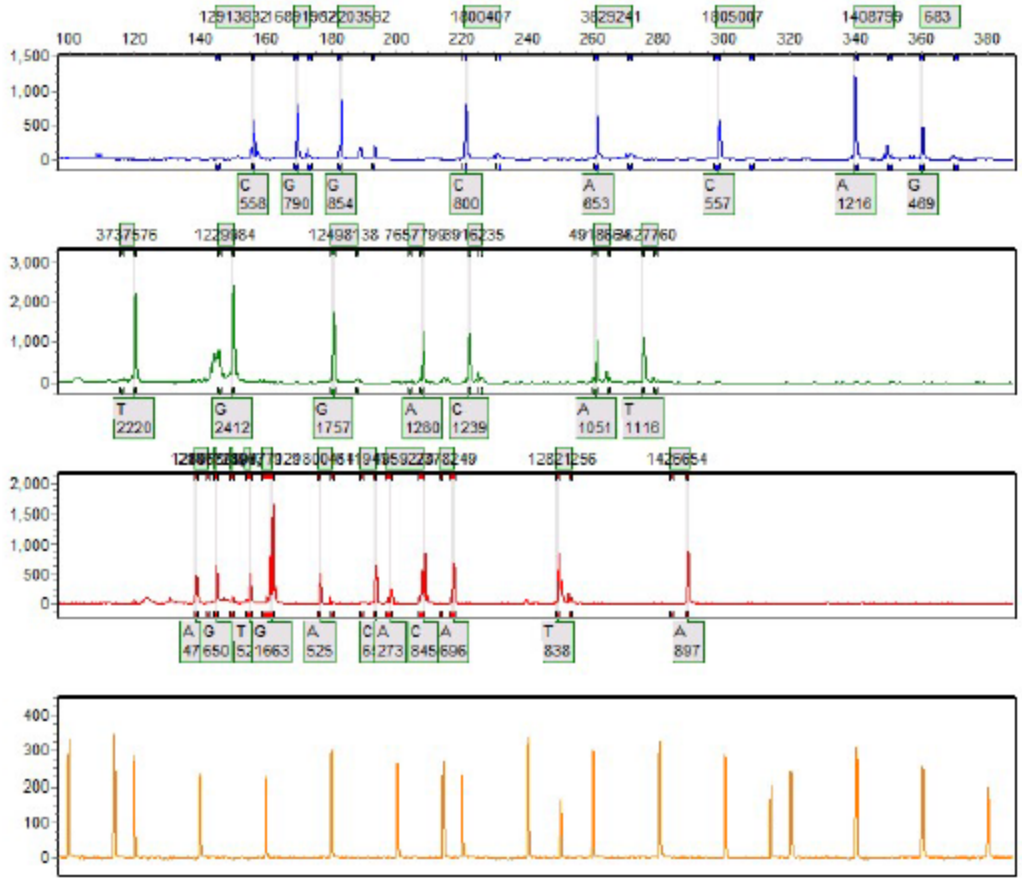
Sample 183: 97482016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 21:39:44 -> 09/21/2016 - 22:17:39



SoftGenetics **Allele Report** 10/19/2016 10:51:58 AM

GeneMarker V2.4.0 Page 184

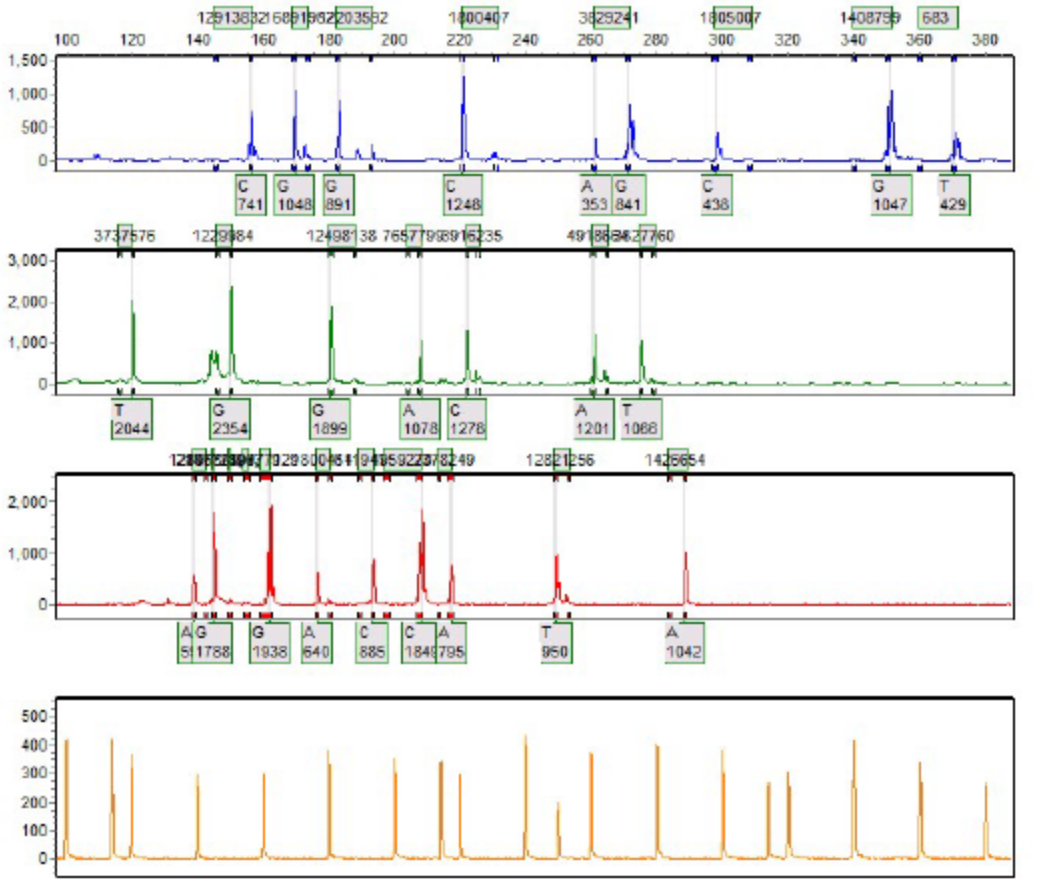
Sample 184: 97562016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 21:39:44 -> 09/21/2016 - 22:17:39

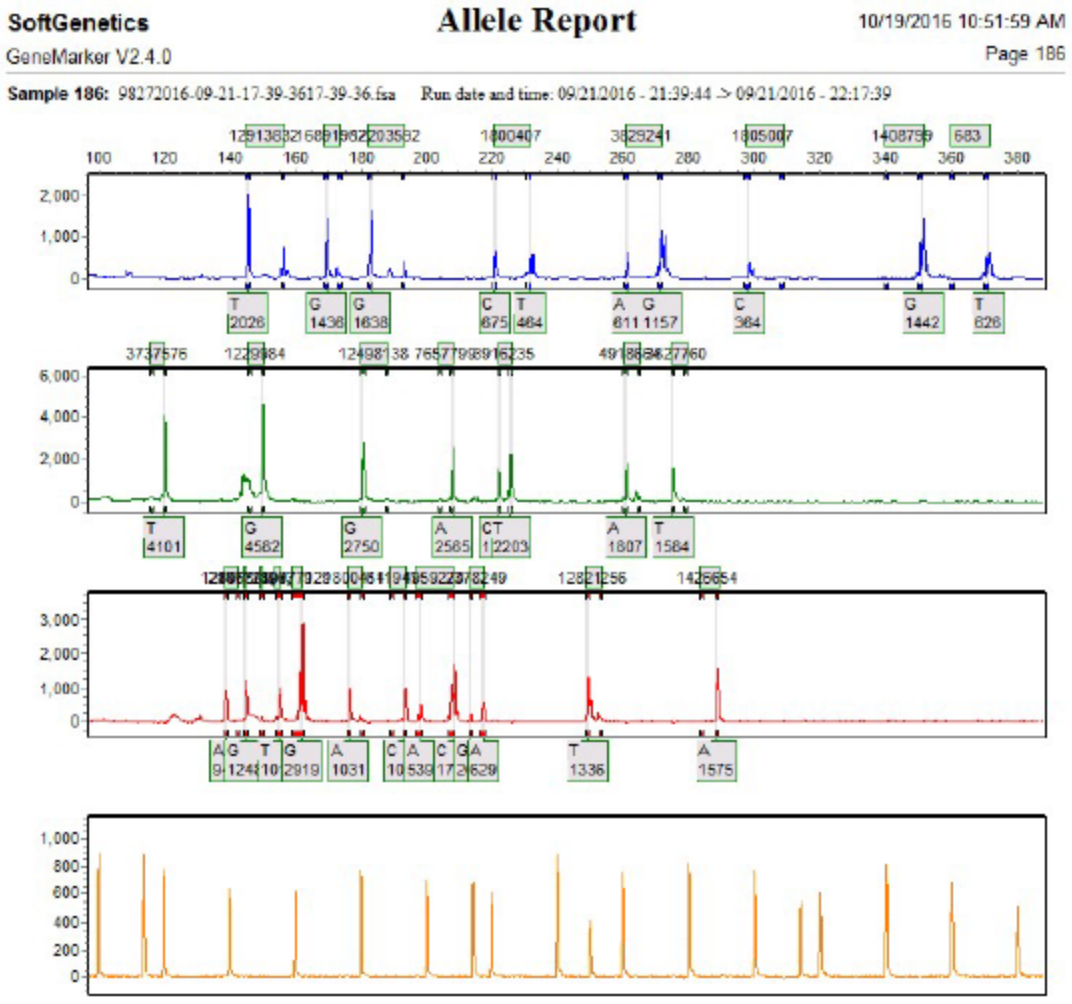


SoftGenetics **Allele Report** 10/19/2016 10:51:58 AM

GeneMarker V2.4.0 Page 185

Sample 185: 97852016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 21:39:44 -> 09/21/2016 - 22:17:39

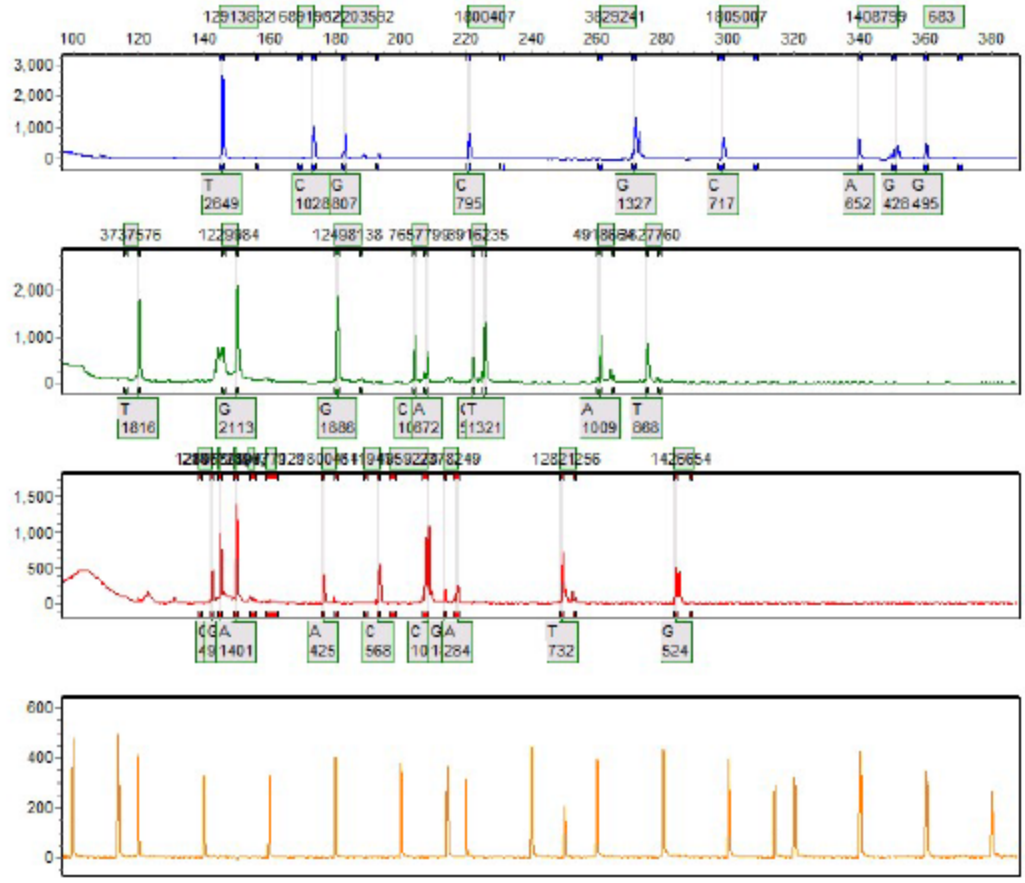



SoftGenetics
Allele Report
10/19/2016 10:51:59 AM
GeneMarker V2.4.0
Page 186
Sample 186: 98272016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 21:39:44 -> 09/21/2016 - 22:17:39

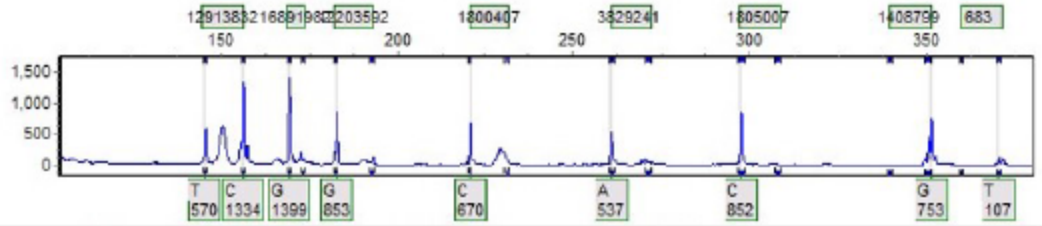
SoftGenetics **Allele Report** 10/19/2016 10:51:59 AM

GeneMarker V2.4.0 Page 187

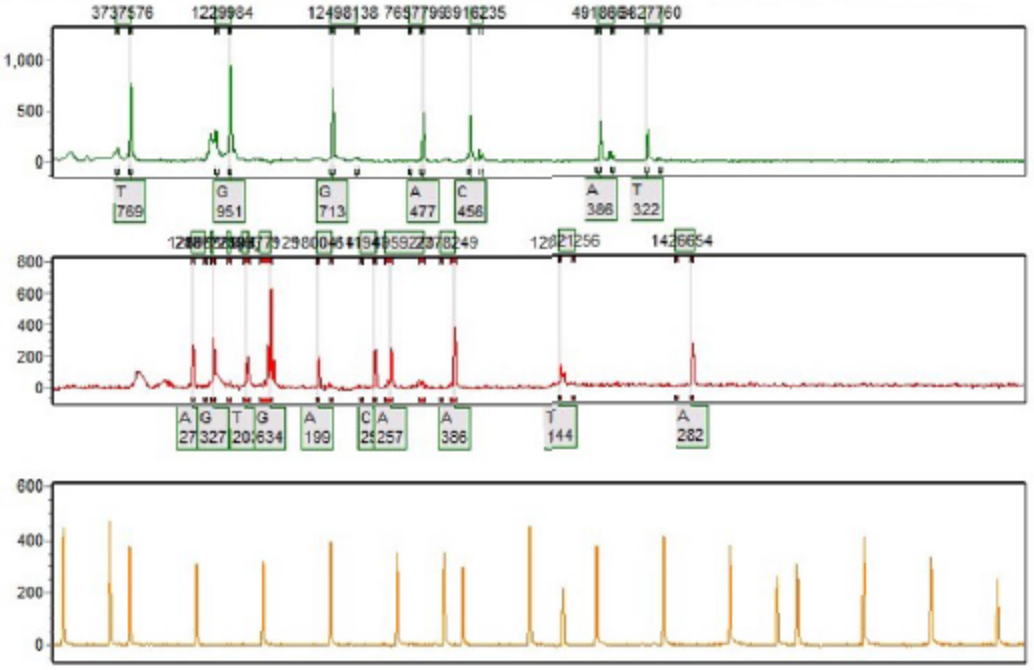
Sample 187: 98302016-09-21-17-39-3617-39-36.fsa Run date and time: 09/22/2016 - 00:15:08 -> 09/22/2016 - 00:53:08

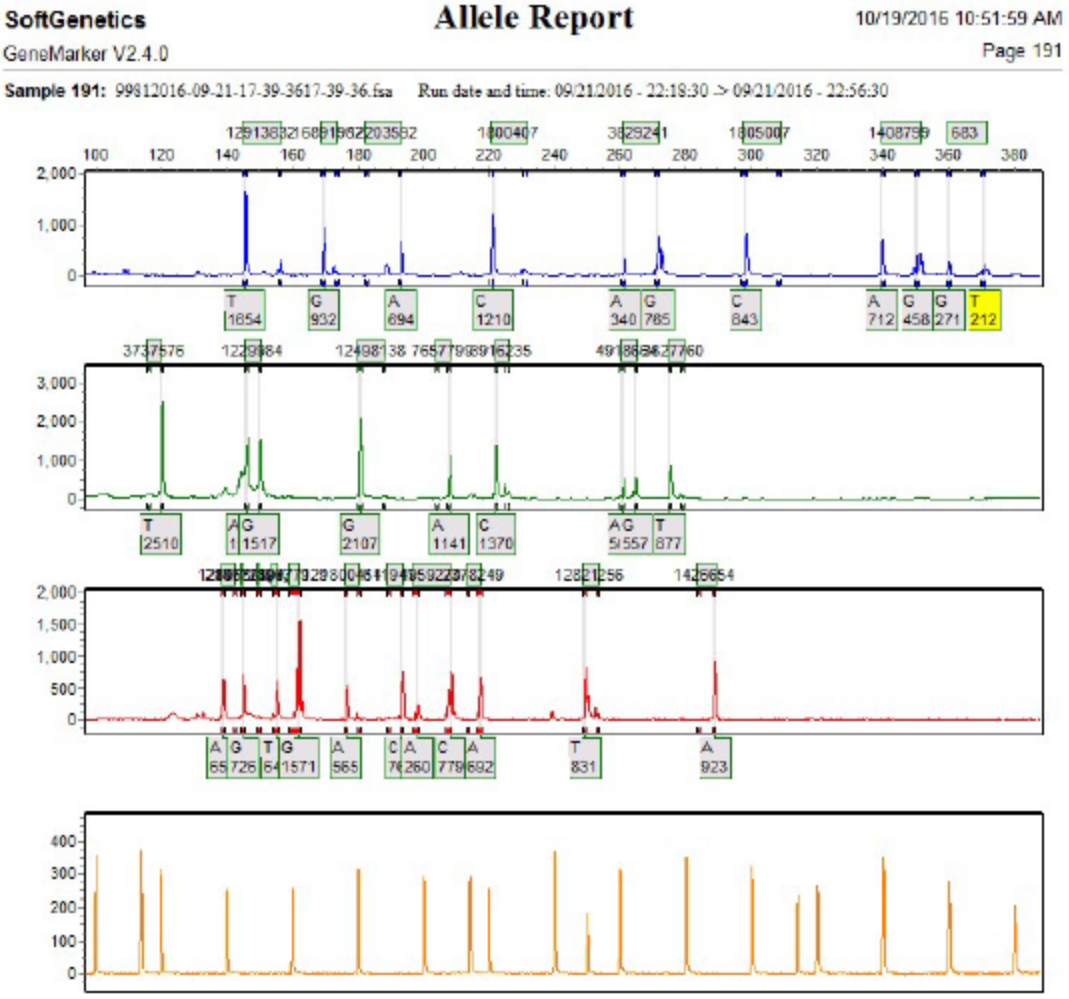


Sample 22: 9975F2016-10-11-07-33-1707-33-17.fsa Run date and time: 10/11/2016 - 07:34:01 -> 10/11/2016 - 08:22:32



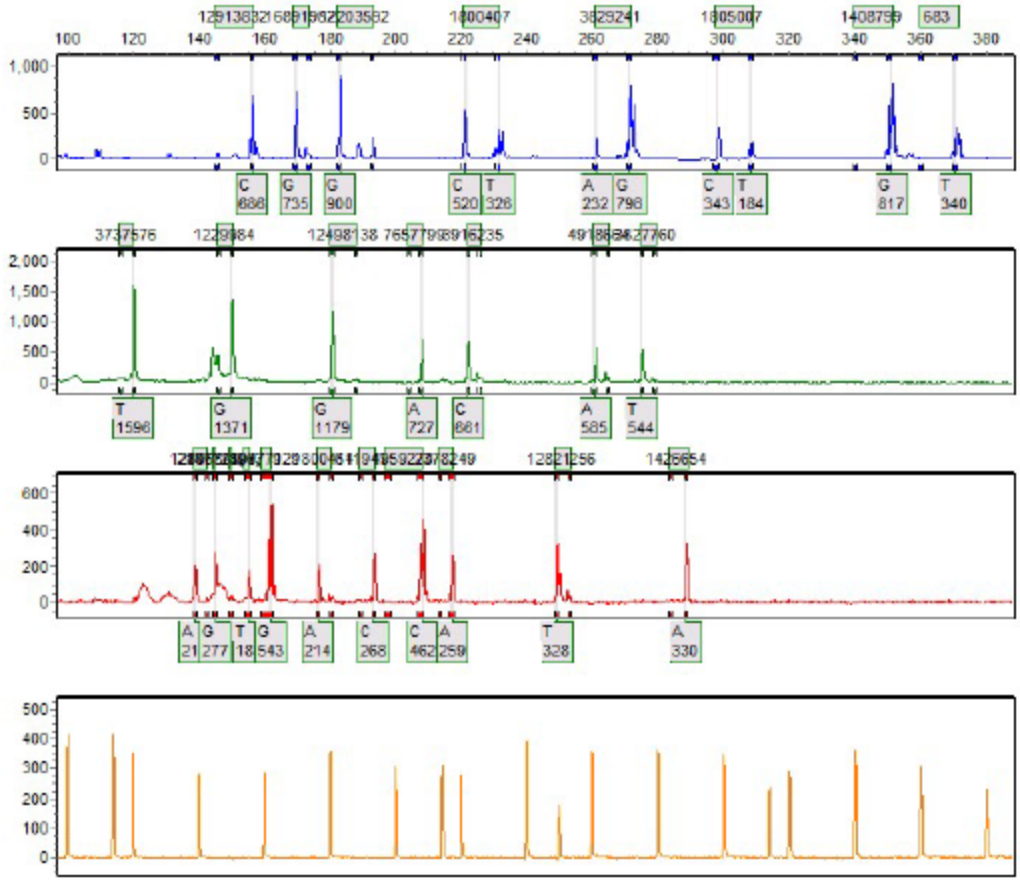
Sample 189: 99752016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 22:18:30 -> 09/21/2016 - 22:56:30




SoftGenetics
Allele Report
10/19/2016 10:51:59 AM
GeneMarker V2.4.0
Page 191
Sample 191: 99812016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 22:18:30 -> 09/21/2016 - 22:56:30

Allele Report

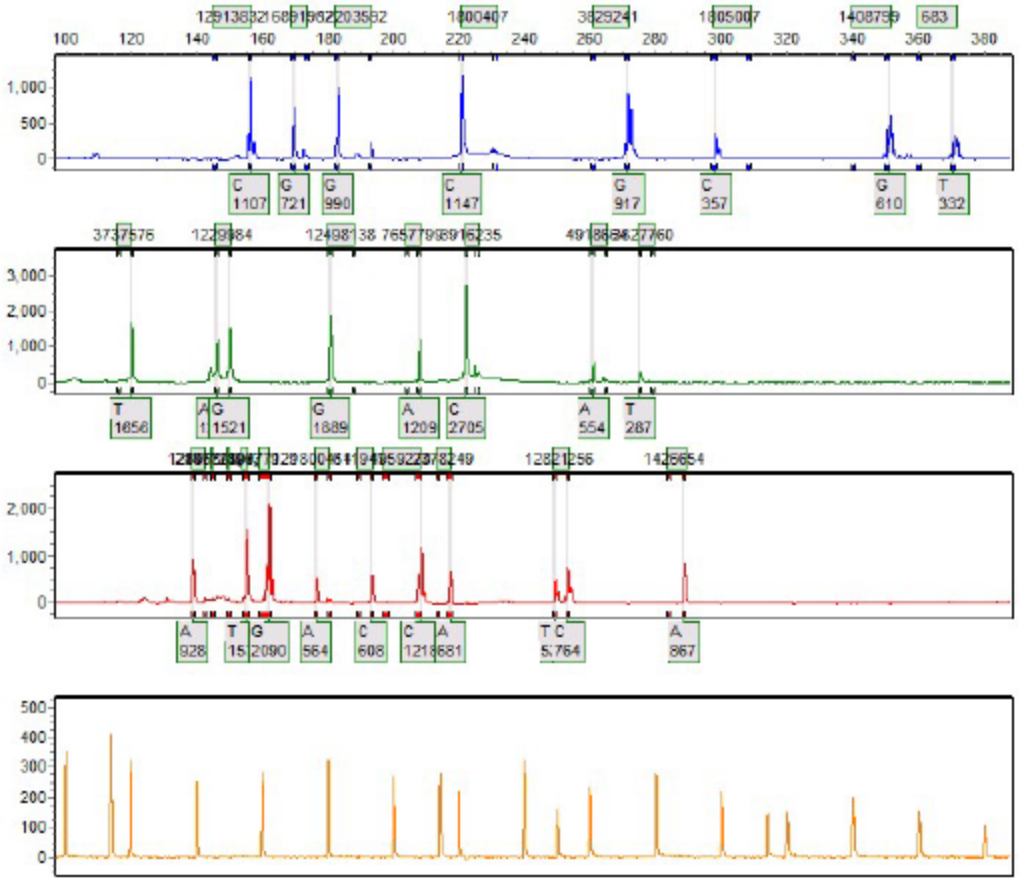
Sample 192: 99852016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 22:18:30 -> 09/21/2016 - 22:56:30



SoftGenetics **Allele Report** 10/19/2016 10:51:59 AM

GeneMarker V2.4.0 Page 193

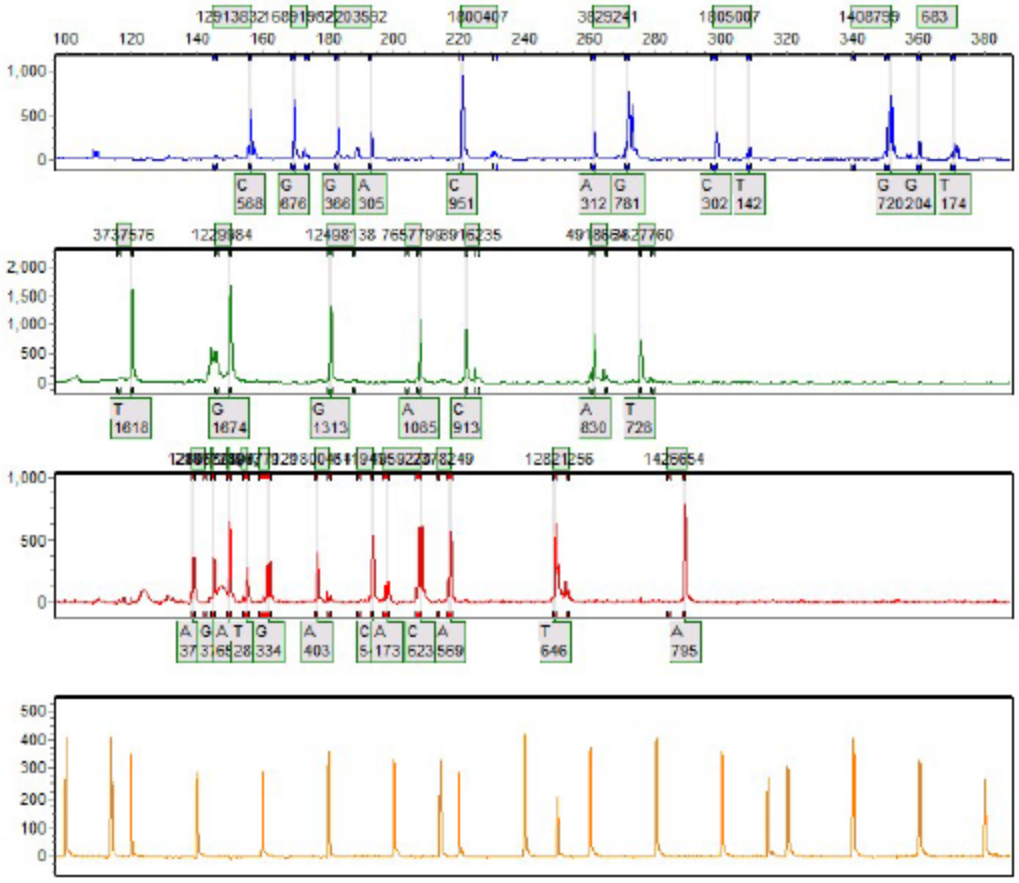
Sample 193: 9986 2016-08-26-09-25-3509-25-35.fsa Run date and time: 08/26/2016 - 10:16:25 -> 08/26/2016 - 10:54:30

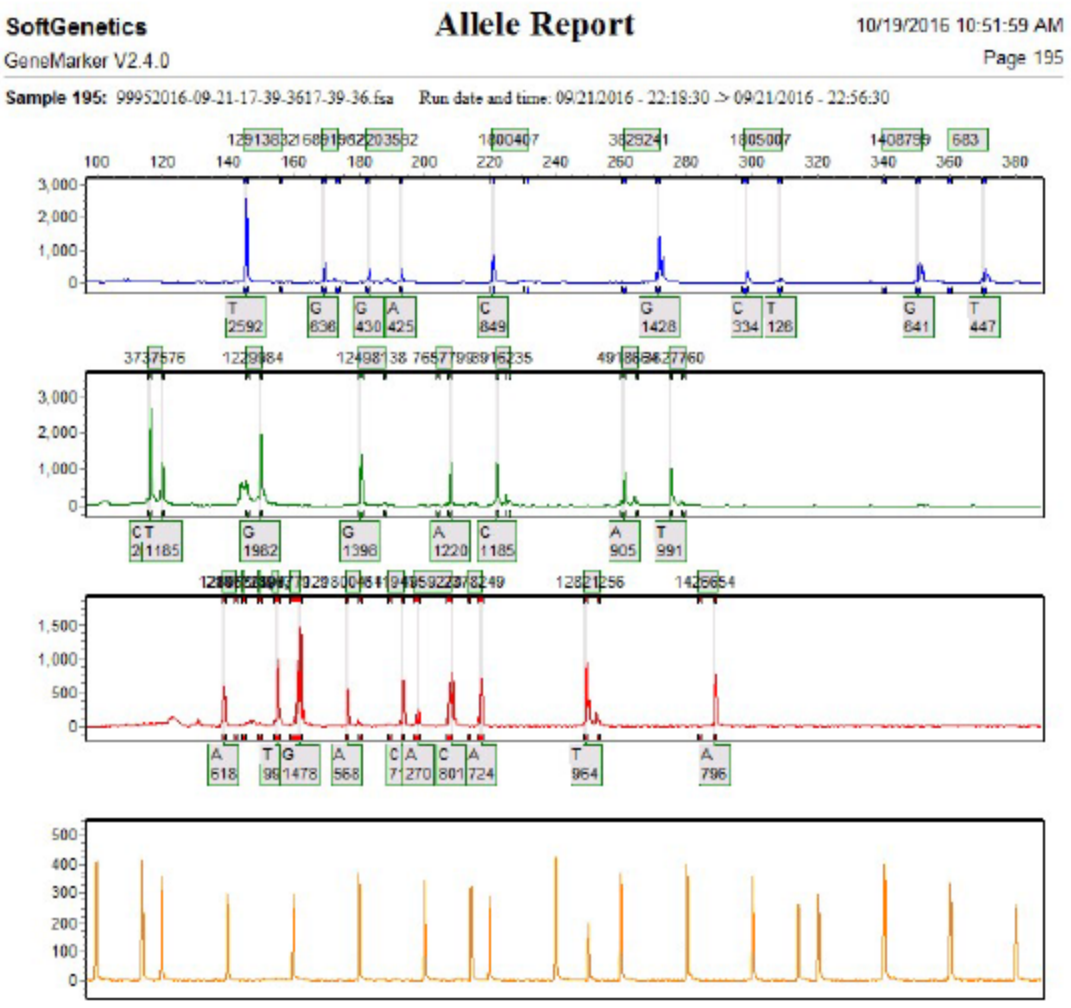


SoftGenetics **Allele Report** 10/19/2016 10:51:59 AM

GeneMarker V2.4.0 Page 194

Sample 194: 99912016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 22:18:30 -> 09/21/2016 - 22:56:30

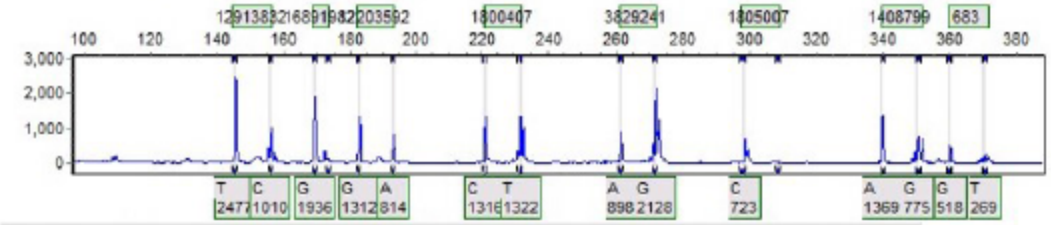



SoftGenetics
Allele Report
10/19/2016 10:51:59 AM
GeneMarker V2.4.0
Page 195
Sample 195: 99952016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 22:18:30 -> 09/21/2016 - 22:56:30

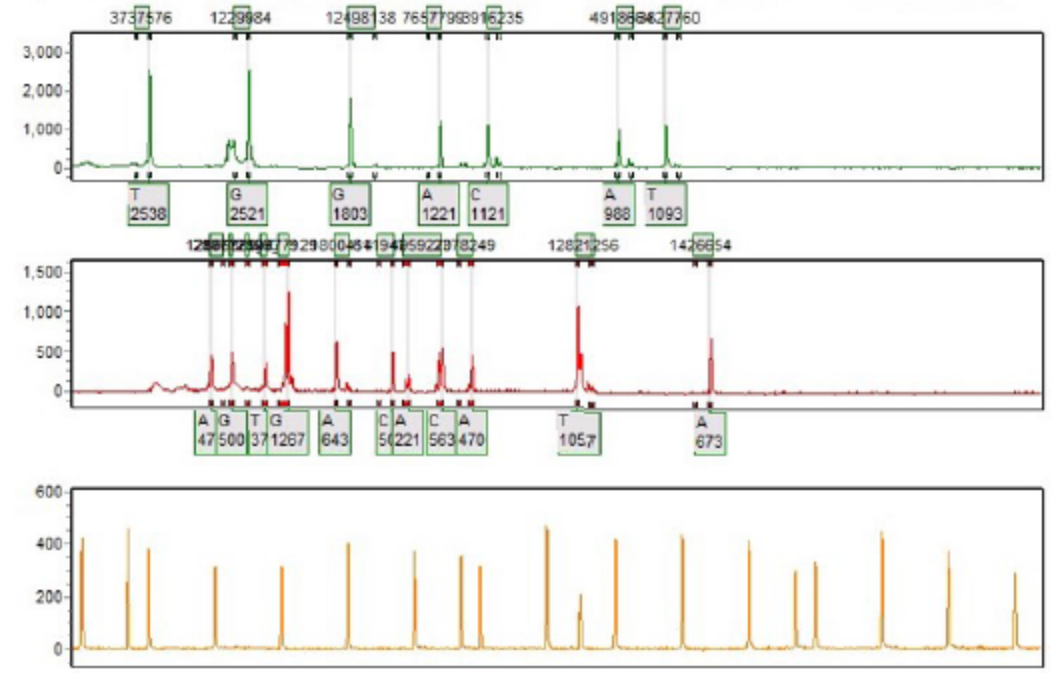
SoftGenetics **Allele Report** 10/19/2016 10:51:59 AM

GeneMarker V2.4.0 Page 197

Sample 197: 9999F2016-10-07-07-32-3307-32-33.fsa Run date and time: 10/07/2016 - 08:22:28 -> 10/07/2016 - 09:00:13



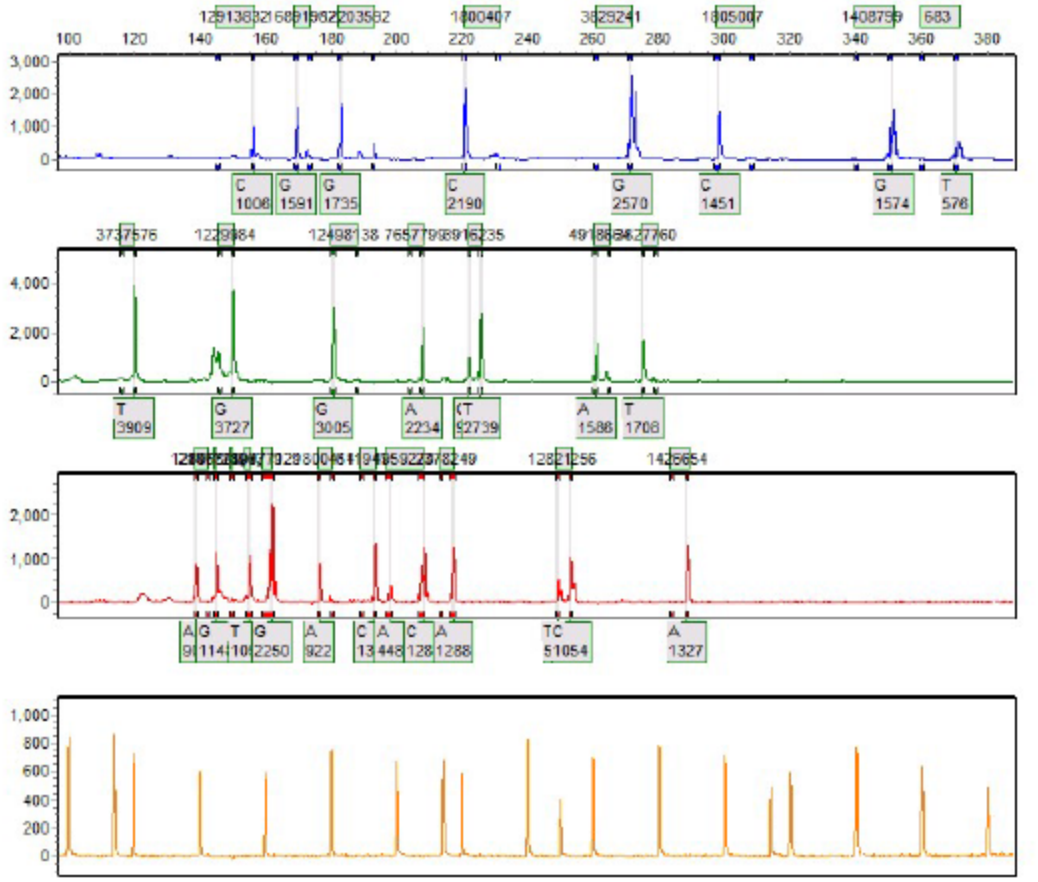
Sample 196: 99992016-09-22-16-13-3916-13-39.fsa Run date and time: 09/22/2016 - 18:20:20 -> 09/22/2016 - 18:58:21



SoftGenetics **Allele Report** 10/19/2016 10:51:59 AM

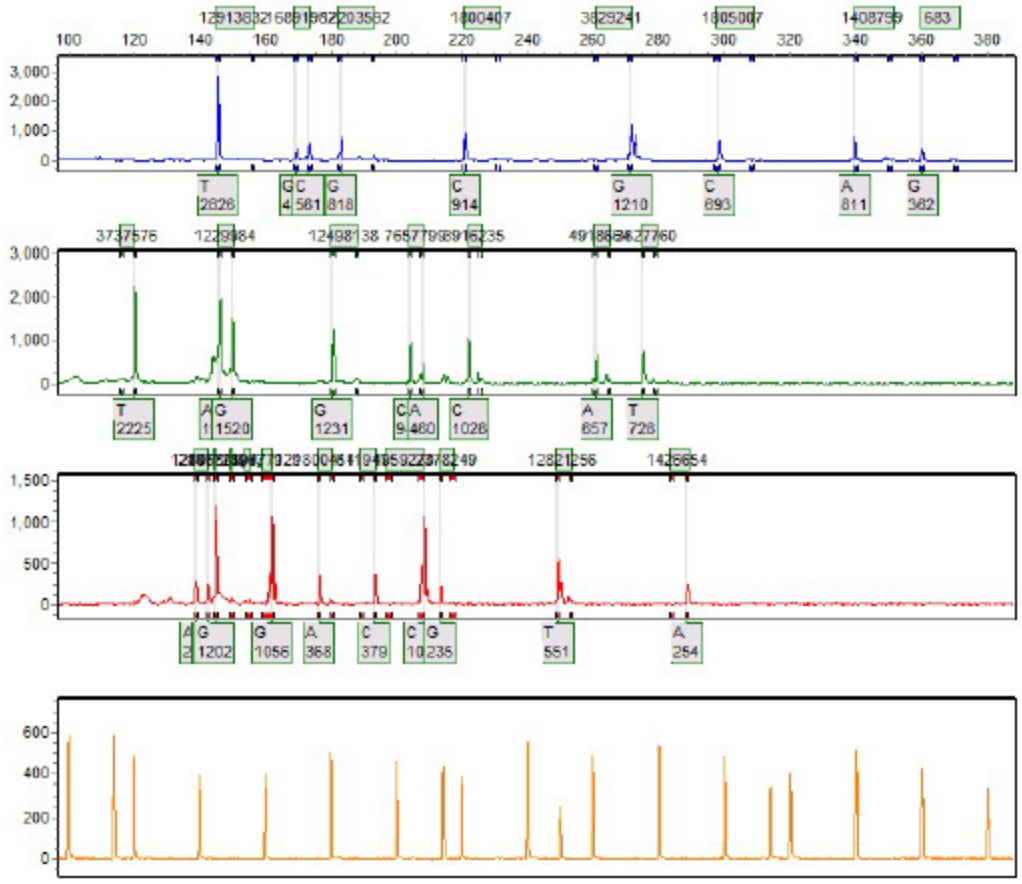
GeneMarker V2.4.0 Page 198

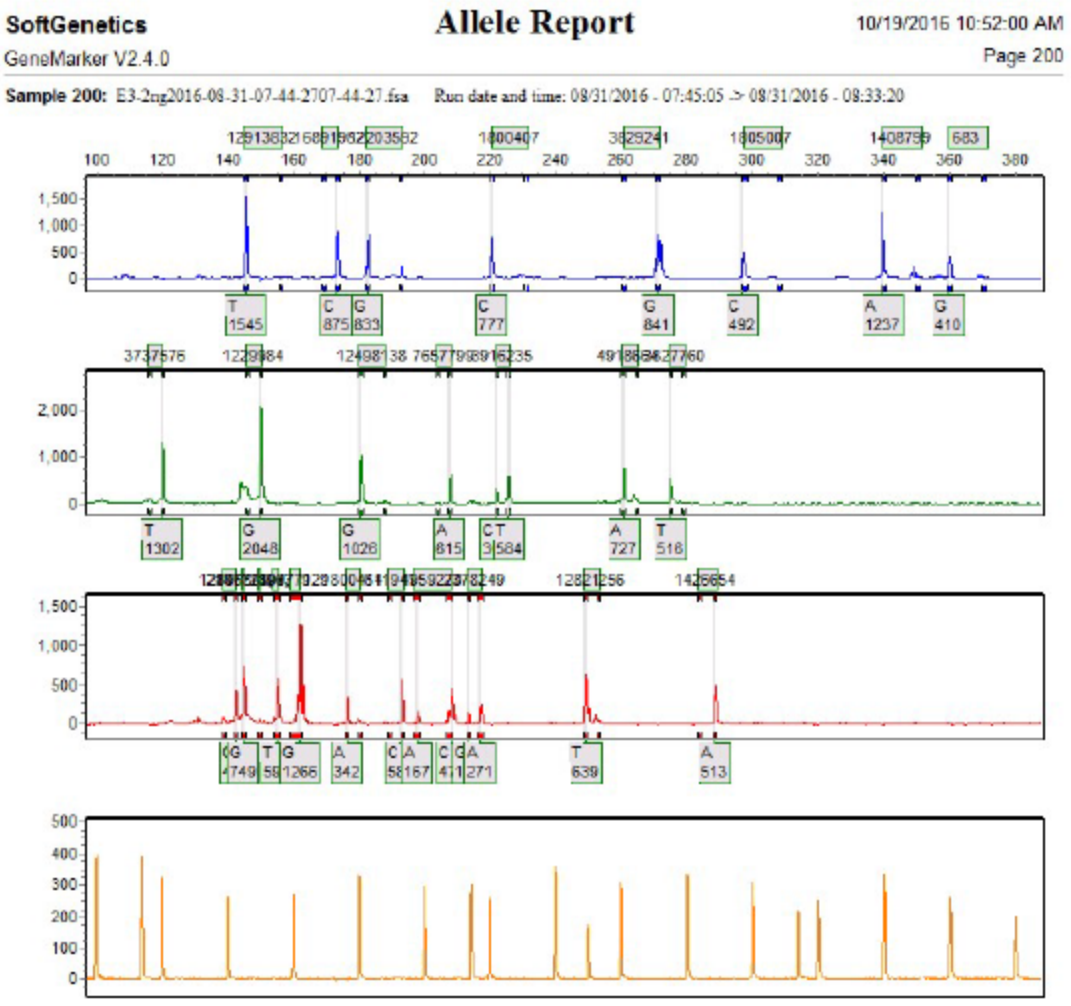
Sample 198: E12016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 22:57:21 -> 09/21/2016 - 23:35:31



Allele Report

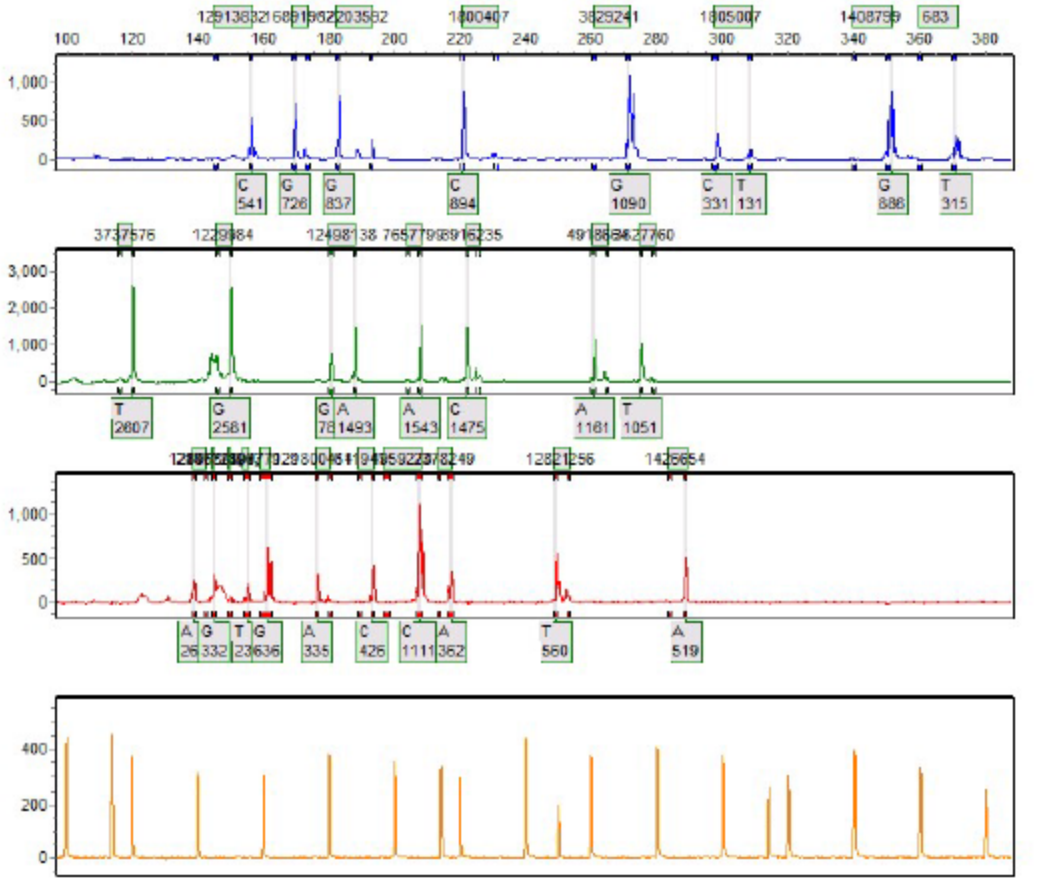
Sample 199: E22016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 22:57:21 -> 09/21/2016 - 23:35:31




SoftGenetics
Allele Report
10/19/2016 10:52:00 AM
GeneMarker V2.4.0
Page 200
Sample 200: E3-2ng2016-08-31-07-44-2707-44-27.fsa Run date and time: 08/31/2016 - 07:45:05 -> 08/31/2016 - 08:33:20

Allele Report

Sample 201: E42016-09-21-17-39-3617-39-36.fsa Run date and time: 09/21/2016 - 23:36:22 -> 09/22/2016 - 00:14:17



APPENDIX C. SEQUENCING PRIMERS

Table A.1 Sequencing primers designed for genotype validation.

SNP	Primer Sequence	Size (bp)
SLC45A2_F_2	AGTTTTTCCTGACGTCCATAGATT	
SLC45A2_R_2	GTGCACACAACTCCACAGAG	235
HERC2_F_2	AGTCTTGTAATCAACATCAGGGT	
HERC2_R_2	TCAAAGAAACGACAAGTAGACCA	259
P.OCA2F	TGAAAGGCTGCCTCTGTTCT	
rs1800407Rseq	ATGGTCACAGGCGTGAAGAG	191
SLC24A4-F	CTGGCGATCCAATTCTTTGT	
SLC24A4-R	CTTAGCCCTGGGTCTTGATG	104
IRF4-F	ACAGGGCAGCTGATCTCTTC	
rs12203592Rseq	AGACTGACAGCCGAAGCATT	250
rs1229984F	CACGTGTTCCCTGAGTGTGA	
rs1229984R	GAAGGGAAGGTAGAGAAGGGC	362
rs12498138F	ACATAGGATTTGCGAGAAACAGA	
rs12498138R	TCTGAGGTACATTGTGGGCTC	190
rs1426654F2	TCAGCCCTTGGATTGTCTCAG	
rs1426554R2	AAATCACACTGAGTAAGCAAGAAGT	130
rs3737576LF	AGTGTAGGGAACAAGAGATCGG	
rs3737576LR	AAGCTGGGAGAGATAGGAGGA	155
rs3827760F	TGCTGATGCGGTCAAAGAGT	
rs3827760R	ACTAGCCGAATGCTCAGCTC	142
rs3916235F	CACTCCACTTCACCCATCCC	
rs3916235R	TGGGCAAAGACTCTTAGTTCAGT	356
rs4918664F	AGGCAGGAATGGGAGAAAGC	
rs4918664R	TGGCAAGAGTTCTGACCAAA	172
rs7657799F	AGTTCTTGACACAAGGCCCA	
rs7657799R	GTACATTGAGAAATGCTGTAGGAA	462
rs1408799LF	AGATATTTGTAAGGTATTCTGGCCT	
rs1408799LR	AGTGCTATGAGGACAGGACC	179
rs3829241F	CCTTTAGAGGCCCCTGTGTG	
rs3829241R	TGGCTCAGCCTCTCTGTGA	92
rs12821256F	ATGCCCAAAGGATAAGGAAT	
rs12821256R	GGAGCCAAGGGCATGTTACT	118
rs4959270F	TGAGAAATCTACCCCCACGA	
rs4959270R	GTGTTCTTACCCCCTGTGGA	140
rs2378249F	CGCATAACCCATCCCTCTAA	
rs2378249R	CATTGCTTTTCAGCCCACAC	136
rs683F	CACAAAACCACCTGGTTGAA	138

Table A.1 continued

rs683R	TGAAAGGGTCTTCCCAGCTT	
rs885479F	CTGGTGAGCTTGGTGGAGA	
rs1805007R	CACCTCCTTGAGCGTCCTG	789
rs6119471F	GAAGTGTGATTCTCTTGGCTTGT	
rs6119471R	GAAGGCACTTGAGAGGAGGC	404
rs10777129F	CTGGACAACCAAGCCCTTAAA	
rs10777129R	CAGAGGCCTAGTGTTGTTGT	525
rs1800414F	TGCCAGGGACAAACGAATTG	
rs1800414R	TGTCGTGATTCCAGTTGCGTA	182
rs28777F	TACTCGTGTGGGAGTTCCAT	
rs28777R	TCTTTGATGTCCCCTTCGAT	150

APPENDIX D. PERMISSIONS

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS

This Agreement between Gina Dembinski ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	4104440192113
License date	May 08, 2017
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	BioEssays
Licensed Content Title	Human pigmentation genetics: the difference is only skin deep
Licensed Content Author	Richard A. Sturm, Neil F. Box, Michele Ramsay
Licensed Content Date	Dec 7, 1998
Licensed Content Pages	10
Type of use	Dissertation/Thesis
Requestor type	University/Academic
Format	Print and electronic
Portion	Figure/table
Number of figures/tables	1
Original Wiley figure/table number(s)	Figure 1
Will you be translating?	No
Title of your thesis / dissertation	Advancements in Forensic DNA-based Identification
Expected completion date	Aug 2017
Expected size (number of pages)	350
Requestor Location	Gina Dembinski 723 W. Michigan St., SL306 INDIANAPOLIS, IN 46202 United States Attn: Gina Dembinski
Publisher Tax ID	EU826007151
Billing Type	Invoice

Billing Address	Gina Dembinski 723 W. Michigan St., SL306 INDIANAPOLIS, IN 46202 United States Attn: Gina Dembinski
Total	0.00 USD

Terms and Conditions

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright
Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.

- You are hereby granted a personal, nonexclusive, nonsub licensable (on a stand-alone basis), nontransferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, **and any CONTENT (PDF or image file) purchased as part of your order**, is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the grant of this license (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.

- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner. **For STM**

Signatory Publishers clearing permission under the terms of the STM Permissions Guidelines **only, the terms of the license are extended to include subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and does not involve the separate exploitation of the permitted figures or extracts**, You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a standalone basis, or any of the rights granted to you hereunder to any other person.

- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto

- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NONINFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING

OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

- Any fee required for this permission shall be nonrefundable after thirty (30) days from receipt by the CCC.

- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.
- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail. WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS

Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses. The license type is clearly identified on the article.

The Creative Commons Attribution License

The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CCBY license permits commercial and noncreative

Commons Attribution NonCommercial License

The Creative Commons Attribution NonCommercial (CC-BY-NC) License permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes. (see below)

Creative Commons AttributionNonCommercialNoDerivs License

The Creative Commons Attribution NonCommercialNoDerivs License (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)

Use by commercial "for-profit" organizations

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee.

Further details can be found on Wiley Online Library http://olabout.wiley.com/WileyCDA/Section/id410895.html

APPENDIX E. ALL RGB QUANTITATIVE PREDICTIONS

Table A.2 RGB quantitative skin color predictions for all 4 BN models.

Sample	R	G	B	Actual	Pigmentation SNPs only, Error %	All SNPs, Error %	Trait SNPs only, Error %	Ancestry + Pigmentation, Error %
1370	200	156	125		10.47%	9.60%	8.47%	10.56%
1654	212	174	133		3.76%	7.96%	2.37%	3.85%
1736	187	134	83		12.56%	14.51%	17.81%	11.54%
1784	200	161	123		4.84%	5.07%	6.69%	5.39%
1892	228	183	128		4.07%	4.02%	4.95%	3.95%
1902	172	121	68		10.06%	8.86%	14.78%	10.04%
1905	231	201	170		10.65%	8.57%	7.45%	10.62%
2079	224	175	120		5.01%	5.22%	4.98%	5.23%
2093	208	168	127		3.44%	3.56%	4.75%	3.55%
2435	207	178	159		5.45%	4.59%	3.77%	5.37%
3187	213	155	131		4.71%	5.69%	6.08%	4.77%
3471	217	178	146		3.40%	4.21%	0.82%	3.23%
3542	216	164	113		8.19%	8.06%	6.89%	8.48%
4063	194	154	110		14.24%	18.06%	23.30%	14.56%
4069	200	175	140		4.36%	4.41%	4.62%	4.39%
4258	225	192	156		8.24%	8.96%	4.66%	7.52%
4389	207	165	118		7.43%	4.15%	6.66%	7.60%
4635	196	158	124		8.62%	11.23%	7.98%	8.67%
4710	148	87	53		16.09%	14.31%	24.58%	16.41%
4819	212	174	164		4.95%	5.07%	5.05%	4.93%

Table A.2 continued

5168	203	155	115		1.79%	2.00%	2.44%	1.74%
5230	208	174	148		7.03%	3.34%	10.23%	8.22%
6084	239	206	166		10.29%	12.39%	9.88%	11.84%
6149	208	165	133		3.48%	3.73%	3.55%	3.37%
6305	238	206	173		5.33%	7.01%	10.47%	5.32%
6329	212	158	119		8.60%	8.60%	5.33%	8.70%
6347	228	194	167		9.08%	6.09%	5.80%	9.25%
6789	217	179	147		3.91%	5.33%	1.62%	3.44%
7181	178	140	99		2.48%	10.43%	3.19%	3.30%
7263	234	198	156		3.52%	1.40%	5.89%	3.59%
7280	206	165	131		4.14%	6.30%	5.47%	4.48%
7294	213	182	148		0.88%	0.82%	2.42%	0.77%
7482	193	157	119		7.98%	11.08%	11.19%	7.98%
7632	207	165	133		4.32%	2.87%	4.44%	4.13%
7659	223	191	171		5.30%	5.84%	7.80%	4.99%
7814	193	153	101		28.72%	27.45%	36.56%	28.66%
7890	228	184	144		8.83%	0.45%	12.12%	9.63%
8395	134	89	55		17.53%	15.72%	24.35%	17.66%
8539	147	87	57		7.34%	8.05%	5.50%	6.81%
8709	224	189	155		7.43%	8.06%	3.94%	7.39%
8730	182	136	116		23.04%	23.94%	13.74%	22.88%
8934	222	183	151		2.65%	3.45%	2.24%	2.51%
8972	225	176	120		5.27%	5.47%	5.41%	5.43%
9167	204	171	132		7.64%	5.59%	4.03%	7.83%
9345	212	166	113		11.04%	8.13%	6.84%	11.13%
9451	208	167	125		4.86%	4.42%	5.84%	5.33%

Table A.2 continued

9628	238	199	148		4.37%	5.18%	6.59%	4.36%
9717	231	199	163		6.05%	5.27%	5.71%	5.19%
9785	211	173	132		2.51%	2.20%	3.06%	2.50%
9981	204	155	129		10.94%	9.54%	6.62%	10.92%

Table A.3 RGB quantitative eye color predictions for all 4 BN models.

Sample	R	G	B	Actual	Pigmentation SNPs only, Error %	All SNPs, Error %	Trait SNPs only, Error %	Ancestry + Pigmentation, Error %
1370	57	29	21		7.65%	8.01%	6.07%	7.96%
1654	41	13	14		3.34%	4.11%	3.29%	3.12%
1736	31	9	10		6.74%	6.90%	5.57%	3.55%
1784	33	10	8		4.42%	3.36%	2.56%	3.19%
1892	60	61	62		2.55%	2.80%	1.93%	2.31%
1902	5	4	4		6.54%	4.41%	6.42%	6.49%
1905	68	60	59		2.21%	3.92%	0.85%	2.31%
2079	52	25	24		8.58%	8.73%	8.49%	9.13%
2093	49	41	44		8.94%	8.51%	7.88%	9.19%
2435	40	38	46		11.63%	11.69%	10.80%	10.39%
3187	48	51	60		4.95%	2.46%	4.90%	5.19%
3471	59	63	73		3.78%	3.60%	3.67%	4.02%
3542	84	54	45		8.81%	8.85%	8.71%	8.08%
4063	14	11	11		9.27%	5.80%	3.74%	10.83%
4069	48	26	23		7.40%	7.59%	4.73%	7.99%
4258	44	37	38		5.97%	6.73%	5.12%	8.38%

Table A.3 continued

4389	42	19	18		11.32%	8.09%	11.22%	11.79%
4635	37	39	40		10.13%	11.02%	10.01%	10.36%
4710	24	15	13		2.39%	1.64%	1.62%	3.70%
4819	49	46	51		7.34%	7.65%	7.57%	7.39%
5168	47	18	19		2.68%	6.15%	1.74%	1.87%
5230	86	68	65		7.71%	7.99%	6.71%	11.74%
6084	52	28	27		5.10%	3.01%	4.83%	1.74%
6149	71	55	48		5.65%	5.60%	4.57%	5.39%
6305	62	36	35		3.84%	3.31%	4.40%	3.89%
6329	34	16	17		2.35%	4.85%	2.12%	3.04%
6347	58	56	58		5.94%	6.77%	4.86%	5.39%
6789	50	35	34		3.39%	3.22%	3.13%	1.85%
7181	36	18	20		3.16%	5.01%	3.92%	0.97%
7263	89	69	56		12.69%	12.69%	12.85%	13.18%
7280	12	3	4		9.95%	12.61%	8.96%	12.28%
7294	67	46	38		6.56%	6.59%	6.29%	6.11%
7482	88	84	83		3.12%	1.77%	8.30%	3.13%
7632	50	41	43		8.96%	4.91%	8.21%	8.40%
7659	47	25	17		4.20%	4.30%	3.03%	5.03%
7814	50	27	27		1.19%	1.64%	7.77%	0.96%
7890	48	23	22		2.89%	4.74%	3.63%	3.83%
8395	15	7	5		4.82%	1.70%	4.17%	5.52%
8539	4	3	3		9.50%	11.11%	10.99%	7.85%
8709	98	79	69		8.21%	11.92%	8.24%	8.05%
8730	83	50	41		10.82%	11.72%	9.17%	11.48%
8934	98	89	92		12.99%	12.37%	13.07%	12.51%

Table A.3 continued

8972	44	44	48		8.29%	8.77%	7.88%	8.75%
9167	71	61	59		10.65%	11.83%	10.71%	9.99%
9345	27	9	12		8.62%	8.56%	9.09%	8.88%
9451	61	31	27		6.41%	11.47%	4.98%	8.19%
9628	63	35	29		4.49%	4.47%	4.76%	4.61%
9717	28	13	12		5.88%	1.13%	0.47%	7.99%
9785	105	103	106		18.90%	18.29%	17.63%	18.93%
9981	91	69	61		13.54%	17.17%	12.73%	13.60%

Table A.4 RGB quantitative hair color predictions for all 4 BN models. Blank rows indicate no hair sample was analyzed.

Sample	R	G	B	Actual	Pigmentation SNPs only, Error %	All SNPs, Error %	Trait SNPs only, Error %	Ancestry + Pigmentation, Error %
1370	103	63	41		4.87%	5.42%	3.35%	5.12%
1654	40	30	30		4.65%	11.46%	5.80%	4.44%
1736	10	11	13		19.45%	20.01%	17.03%	16.65%
1784	16	13	19		11.22%	12.39%	9.63%	10.15%
1892	203	185	165		42.83%	41.96%	42.72%	43.06%
1902	56	34	31		6.47%	9.98%	6.07%	6.51%
1905	125	104	81		12.51%	16.75%	10.06%	12.42%
2079	115	87	65		3.44%	3.30%	2.59%	3.00%
2093	64	32	23		17.16%	17.83%	23.69%	17.35%
2435	167	150	126		28.55%	27.79%	27.15%	29.54%
3187	181	157	122		19.97%	12.64%	23.09%	19.76%

Table A.4 continued

3471	111	81	64		4.19%	2.42%	2.03%	3.65%
3542								
4063	12	10	18		4.50%	7.30%	4.04%	5.67%
4069	83	53	36		8.16%	7.55%	7.84%	8.64%
4258	81	62	50		5.47%	5.31%	2.05%	7.40%
4389								
4635	119	92	71		1.51%	3.73%	2.09%	1.38%
4710	9	9	13		2.71%	2.78%	3.56%	2.73%
4819	102	68	49		5.17%	5.78%	7.31%	5.20%
5168	50	30	27		4.02%	9.02%	3.76%	4.96%
5230	102	64	35		6.81%	5.58%	7.44%	6.29%
6084	31	16	19		23.73%	24.98%	26.15%	18.54%
6149	79	42	41		7.81%	8.87%	14.79%	7.55%
6305	132	82	57		4.13%	4.09%	3.75%	4.13%
6329	12	15	19		2.84%	4.01%	1.78%	2.83%
6347	189	195	191		41.65%	41.18%	38.92%	42.07%
6789	176	154	132		26.10%	27.31%	27.43%	27.40%
7181	15	15	18		15.70%	8.26%	18.20%	13.82%
7263	171	140	110		27.09%	26.64%	27.45%	27.47%
7280	96	64	52		7.98%	14.02%	14.05%	6.22%
7294	64	41	30		8.77%	9.56%	10.74%	9.14%
7482	133	112	96		12.10%	12.04%	13.97%	12.08%
7632	175	160	145		19.59%	19.42%	17.45%	20.05%
7659	130	89	51		10.22%	9.20%	9.08%	9.58%
7814	38	23	23		2.70%	1.26%	3.94%	2.91%
7890								

Table A.4 continued

8395	22	17	20		2.20%	3.30%	0.62%	2.02%
8539	50	37	32		2.37%	6.44%	2.58%	3.57%
8709	63	45	41		11.04%	14.76%	17.13%	11.17%
8730	125	88	61		2.56%	2.99%	4.66%	2.56%
8934	136	105	74		5.16%	4.36%	5.06%	4.80%
8972	97	73	57		8.36%	8.70%	8.98%	8.74%
9167	113	82	60		4.76%	1.75%	5.76%	4.19%
9345								
9451	64	33	21		4.66%	6.20%	4.22%	4.57%
9628	92	55	35		12.48%	13.09%	12.47%	12.62%
9717	9	9	17		7.19%	11.45%	6.01%	8.88%
9785	164	148	136		25.48%	24.66%	23.50%	25.50%
9981	130	94	73		10.69%	7.62%	12.33%	10.73%

APPENDIX F. ALL CANDIDATE CPG SITES

Table A.5 Candidate age CpG sites. Those highlighted in blue had more than one possible CpG site in the mapped read.

Chromosome	CpG Site	NCBI Refseq Gene	r^2	Methylation Pattern	Non-gene Genomic Feature
chr1	4063800	--	0.95	hypo	n/a
chr1	4196025	--	0.89	hyper	AluSc -SINE
chr1	14977413	*KAZN*	0.71	hypo	
chr1	18256473	*IGSF21*	0.90	hypo	
chr1	18442677	--	0.90	hyper	AluJr -SINE
chr1	41891046	*HIVEP3*	0.87	hypo	
chr1	42354692	--	0.91	hyper	AluY -SINE
chr1	50116466	*ELAVL4*	0.82	hypo	
chr1	59107862	--	0.97	hypo	LINC01358
chr1	68242601	--	0.96	hyper	n/a
chr1	88035578	--	0.95	hyper	MER61C -LTR
chr1	930012437	--	0.97	hypo	
chr1	107261767	*NTNG1*	0.98	hyper	AluY-SINE
chr1	112264284	--	0.78	hyper	n/a
chr1	116532957	*CD58*	0.94	hyper	
chr1	143665699	--	0.94	hypo	n/a
chr1	143747286	--	0.90	hypo	AluSz -SINE
chr1	143765020	--	0.88	hypo	MER66C- LTR
chr1	146053356	*LOC101928979*	0.86	hyper	LINC01719; AluYk3-SINE
chr1	153615176	*S100A14*	0.89	hypo	
chr1	153796804	--	0.71	hyper	L2c- LINE
chr1	158790497	--	0.81	hyper	AluY-SINE
chr1	161692317	--	0.98	hyper	AluSx1-SINE
chr1	163416999	--	0.72	hyper	AluY-SINE
chr1	165582202	*LOC400794*	0.78	hyper	
chr1	185020039	--	0.75	hypo	n/a
chr1	192209826	--	0.80	hyper	AluSp-SINE
chr1	195783494	--	0.80	hyper	AluY-SINE
chr1	202705741	*SYT2*	0.85	hypo	AluJo - SINE
chr1	210668395	*HHAT*	0.94	hypo	THE1B-LTR
chr1	217066347	*ESRRG*	0.81	hyper	
chr1	222521710	--	0.97	hypo	LTR8B- LTR

Table A.5 continued

chr1	223885109	--	0.88	hyper	AluY-SINE
chr1	228063253	--	0.79	hypo	n/a
chr1	236123570	--	0.90	hypo	n/a
chr1	238494677	--	0.89	hyper	AluY- SINE
chr1	243952428	*LOC339529*	0.87	hyper	
chr1		*ZNF695*	0.94	hyper	AluY- SINE
chr10	3472563	*LOC105376360*	0.93	hypo	
chr10	3995341	--	0.79	hypo	L1PA8-LINE
chr10	17501725	--	0.92	hypo	AluSq2 - SINE
chr10	21084829	*NEBL*	0.89	hypo	
chr10	27526362	*RAB18*	0.76	hypo	
chr10	43779805	--	0.90	hypo	n/a
chr10	86103582	*GRID1*	0.86	hypo	
chr10	110567780	*SMC3*	0.79	hypo	CpG island
chr10	116785385	--	0.90	hypo	n/a
chr10	119275373	*GRK5*	0.87	hypo	LFSINE_Vert - SINE
chr10	125082211	*CTBP2*	0.77	hypo	
chr10	128642775	--	0.95	hypo	MLT1C - LTR
chr10	131825744	--	0.97	hypo	n/a
chr11	11999029	*DKK3*	0.87	hyper	MIR- SINE
chr11	12724900	*TEAD1*	0.71	hypo	
chr11	19941856	*NAV2*	0.89	hypo	L3 - LINE
chr11	35603992	--	0.85	hyper	AluY- SINE
chr11	3809028	--	0.82	hyper	MER11A - LTR
chr11	39022561	--	0.92	hyper	AluYb8 - SINE
chr11	41410487	*LRRC4C*	0.90	hyper	AluY- SINE
chr11	48429548	--	0.99	hyper	AluSg- SINE
chr11	69008985	*MRGPRF*	0.88	hypo	
chr11	73367989	*ARHGEF17*	0.90	hyper	
chr11	73752429	*RAB6A*	0.96	hypo	AluSx1-SINE
chr11	81009321	--	0.92	hyper	AluSc8-SINE
chr11	93740945	*TAF1D*	0.81	hypo	CpG island
chr11	101448620	--	0.95	hypo	n/a
chr11	105573991	--	0.77	hypo	LTR54- LTR
chr11	118145091	*SCN4B*	0.72	hyper	
chr11	120910116	*GRIK4*	0.71	hypo	MSTD-LTR
chr11	127956596	--	0.78	hypo	AluSx- SINE
chr11	129702075	--	0.89	hypo	n/a
chr12	6337976	*TNFRSF1A*	0.92	hypo	
chr12	6975299	*EMG1*	0.89	hypo	

Table A.5 continued

chr12	14162352	--	0.83	hyper	AluSp-SINE
chr12	16112299	--	0.90	hypo	n/a
chr12	20402242	*PDE3A*	0.95	hypo	AluSz6-SINE
chr12	27244066	*STK38L*	0.96	hypo	CpG island
chr12	42671522	*LOC105369738*	0.80	hyper	MER50-LTR; LINC02451
chr12	43409409	*ADAMTS20*	0.94	hyper	AluYa5- SINE
chr12	52439302	--	0.82	hypo	AluJb-SINE
chr12	78794187	--	0.94	hyper	n/a
chr12	80429596	--	0.90	hyper	AluYc- SINE
chr12	93573848	*SOCS2*	0.91	hypo	
chr12	10020204	*ACTR6*	0.70	hyper	AluSg- SINE
chr12	106113901	*NUAK1*	0.97	hypo	
chr12	107197597	--	0.70	hyper	AluSc- SINE
chr12	110339453	*ATP2A2*	0.95	hypo	
chr12	121906347	*PSMD9*	0.86	hypo	
chr12	125397132	*TMEM132B*	0.82	hypo	AluSx3- SINE
chr12	130286919	--	0.90	hypo	n/a
chr13	29539635	*SLC7A1*	0.84	hypo	
chr13		*COG6*	0.92	hypo	CpG island
chr13	41047562	*ELF1*	0.95	hypo	L1PA5- LINE
chr13	48838575	--	0.84	hypo	n/a
chr13	58859813	--	0.94	hypo	HUERS-P2-int - LTR
chr13	78101976	*RNF219-AS1*	0.97	hyper	AluYb9- SINE
chr13	89483285	*LINC01040*	0.94	hyper	MLT1E2- LTR
chr13	106375143	--	0.79	hyper	n/a
chr13	110129547	--	0.71	hypo	n/a
chr13	112620712	--	0.86	hyper	AluSp-SINE
chr14	19169193	--	0.84	hyper	AluY- SINE
chr14	20601879	*LOC254028*	0.97	hypo	HERVL 18-int LTR
chr14	21525829	*SALL2*	0.98	hypo	
chr14	27726354	--	0.77	hypo	n/a
chr14	34147401	--	0.91	hypo	n/a
chr14	42419944	--	0.93	hypo	L1PA15- LTR
chr14	64760474	*SPTB*	0.94	hypo	
chr14	69728941	*SRSF5*	0.76	hypo	MER41B - LTR
chr14	74892948	*DLST*	0.81	hypo	
chr14	74916943	*RPS6KL1*	0.92	hypo	
chr14	84502166	--	0.99	hypo	HERVH-int - LTR
chr14	93186029	*TMEM251*	0.73	hyper	AluY- SINE

Table A.5 continued

chr14	98000632	--	0.90	hypo	n/a
chr14	101029421	*MIR494*	0.80	hypo	microRNA 494
chr14	105249899	*BRF1; BTBD6*	0.79	hyper	CpG island
chr14		--	0.93	hyper	CpG island
chr15	23965186	--	0.97	hyper	AluY- SINE
chr15	40333450	*C15orf52*	0.75	hyper	AluJb- SINE
chr15	53222767	--	0.96	hyper	n/a
chr15	57897471	--	0.80	hyper	n/a
chr15	65410294	*IGDCC4*	0.73	hypo	
chr15	74049311	--	0.76	hyper	AluSc8-SINE
chr15	79284038	*ANKRD34C*	0.98	hypo	CpG island north shore
chr15	84717135	--	0.93	hypo	n/a
chr15	868928876	*AGBL1*	0.80	hypo	
chr15	87761896	--	0.91	hypo	n/a
chr15	89005093	--	0.88	hypo	n/a
chr15	94157803	--	0.95	hyper	LTR17- LTR
chr15	96367880	*SPATA8*	0.91	hypo	
chr15	101347166	*PCSK6*	0.94	hypo	
chr16		--	0.94	hypo	SVA_D retroposon
chr16	6010345	--	0.99	hypo	LTR16C- LTR
chr16	11757206	*ZC3H7A*	0.95	hypo	AluY- SINE
chr16	13200344	*SHISA9*	0.87	hyper	LTR16E1- LTR
chr16	18373911	--	0.91	hypo	AluSx4- SINE
chr16	18958377	--	0.72	hypo	AluSg-SINE
chr16	23509475	*GGA2*	0.88	hypo	L2c - LINE
chr16	24627310	*LCMT1*	0.83	hyper	
chr16	27485251	*GTF3C1*	0.79	hypo	
chr16	32670516	--	0.97	hyper	n/a
chr16		--	0.75	hypo	AluSx1- SINE
chr16	51767907	*LINC01571*	0.88	hypo	
chr16	57462658	*POLR2C*	0.78	hypo	
chr16	64944282	*CDH11*	0.91	hypo	L2a-LINE
chr16	65117897	*CDH11*	0.81	hyper	
chr16	66065995	--	0.87	hypo	n/a
chr16	71353879	--	0.94	hypo	n/a
chr16	83817394	*HSBP1*	0.92	hypo	
chr16	83918165	*MLYCD*	0.95	hyper	CpG island south shore
chr16	84114686	*MBTPS1*	0.74	hyper	AluSx1-SINE
chr16	86670506	--	0.79	hyper	n/a

Table A.5 continued

chr16	87661630	*JPH3*	0.96	hypo	
chr16	89614823	*DPEP1*	0.85	hypo	
chr17	3911939	*P2RX1*	0.99	hypo	MLT1K -LINE
chr17	7477472	*ZBTB4*	0.70	hyper	
chr17	9786768	*DHRS7C*	0.87	hypo	MIR-SINE
chr17		--	0.83	hyper	AluY- SINE
chr17	18270847	--	0.97	hypo	n/a
chr17	20064834	*SPECC1*	0.78	hyper	
chr17	20708532	*DHRS7B*	0.76	hyper	AluSx1- SINE
chr17	30390004	*CPD*	0.90	hypo	
chr17	36396863	--	0.70	hyper	AluY- SINE
chr17	39151146	*PLXDC1*	0.96	hypo	CpG island north shore
chr17	43734456	--	0.70	hyper	n/a
chr17	45949280	*MAPT*	0.98	hypo	CpG island
chr17		--	0.92	hyper	n/a
chr17	50365019	--	0.83	hypo	n/a
chr17	56831991	*C17orf67*	0.73	hyper	AluSg7- SINE
chr17	59107008	*TRIM37*	0.96	hypo	CpG island
chr17	61371775	*BCAS3*	0.73	hyper	
chr17	72364500	--	0.81	hyper	n/a
chr17		--	0.92	hyper	AluY- SINE
chr17	79670160	--	0.99	hypo	n/a
chr17	80662935	*RPTOR*	0.98	hyper	LTR8B- LTR
chr18	5028355	--	0.85	hyper	AluY- SINE
chr18	7501277	--	0.92	hyper	n/a
chr18	10472966	*APCDD1*	0.80	hypo	
chr18	11284112	--	0.98	hypo	n/a
chr18	14962783	*LINC01443*	0.78	hypo	
chr18	15315932	--	0.89	hyper	HERVIP10F-int - LTR
chr18	16280392	--	0.71	hyper	ALR/Alpha- Satellite
chr18	45782822	*SLC14A1*	0.91	hypo	
chr18	64765447	--	0.93	hyper	n/a
chr19	3054479	*AES*	0.86	hypo	
chr19	5517943	--	0.71	hyper	n/a
chr19	13059944	*NFIX*	0.83	hypo	
chr19	16121759	*RAB8A*	0.97	hypo	
chr19	16570110	*SLC35E1*	0.90	hypo	
chr19	18292644	--	0.80	hypo	MIR3- SINE; CpG island south shore
chr19	28946814	--	0.75	hypo	n/a

Table A.5 continued

chr19	30204514	--	0.94	hyper	n/a
chr19	32719581	--	0.89	hyper	CpG island
chr19	38791549	*LGALS7B*	0.86	hypo	CpG island
chr19	44209881	--	0.74	hyper	AluY- SINE
chr19	46777016	*SLC1A5*	0.98	hyper	
chr19	47919710	--	0.82	hypo	AluSx1- SINE
chr19		*SLC6A16*	0.89	hypo	AluY-SINE
chr19	50099738	--	0.74	hypo	AluSx1- SINE
chr2	355785	--	0.71	hyper	n/a
chr2	1813951	*MYT1L*	0.88	hypo	AluY- SINE
chr2	3422322	*TRAPPC12*	0.96	hypo	
chr2	3559211	*RNASEH1-AS1*	0.93	hyper	AluSg- SINE
chr2		--	0.86	hypo	n/a
chr2	5407160	--	0.88	hyper	AluY- SINE
chr2	7472236	--	0.97	hyper	AluY- SINE
chr2	8224700	*LINC00299*	0.87	hyper	L2b- LINE
chr2	9797707	--	0.98	hypo	L2b- LINE
chr2	11598412	*LPIN1*	0.73	hypo	
chr2		*TRIB2*	0.98	hypo	CpG island
chr2	17909503	*KCNS3*	0.92	hypo	L1PA5- LINE
chr2	18048445	--	0.97	hypo	n/a
chr2	22105966	--	0.81	hyper	AluYc3- SINE
chr2	27293431	*TRIM54*	0.77	hypo	
chr2	49198224	--	0.83	hypo	n/a
chr2	57085962	--	0.75	hyper	MER50- LTR
chr2	59065629	--	0.97	hyper	AluYa5- SINE
chr2	60352130	--	0.90	hypo	n/a
chr2	64312183	--	0.97	hypo	n/a
chr2	70303874	--	0.78	hypo	n/a
chr2	73720245	--	1.00	hypo	n/a
chr2	81822755	--	0.97	hypo	AluY- SINE
chr2	84407393	--	0.84	hypo	n/a
chr2	84915328	--	0.89	hyper	AluSq2- SINE
chr2	91883623	--	0.97	hyper	n/a
chr2	94943394	*LOC442028*	0.86	hyper	AluY- SINE
chr2	96505661	*NEURL3*	0.83	hypo	
chr2	96639541	*KANSL3*	0.83	hyper	CpG island south shore; AluSg- SINE
chr2	103495113	--	0.84	hyper	n/a
chr2	115261238	*DPP10*	0.78	hyper	

Table A.5 continued

chr2	119478869	*SCTR*	0.99	hypo	
chr2	121069189	--	0.86	hypo	n/a
chr2	123245386	--	0.82	hypo	L1PA10- LINE
chr2	128238962	--	0.78	hyper	AluSx1- SINE
chr2	130574299	*TISP43*	0.92	hypo	
chr2	133567301	*NCKAP5*	0.77	hypo	
chr2	137420985	*THSD7B*	0.82	hypo	MLT1D - LTR
chr2		--	0.99	hypo	L1P2- LINE
chr2	138931502	--	0.91	hyper	L1PA8A- LINE
chr2	153202145	--	0.76	hyper	AluYb8- SINE
chr2	162996026	--	0.94	hypo	n/a
chr2	171153761	*TLK1*	0.98	hyper	
chr2	186962582	--	0.85	hyper	LTR12C- LTR
chr2	189204054	--	0.81	hyper	AluY- SINE
chr2	197165374	*ANKRD44*	0.71	hyper	CpG island south shore
chr2	206252430	*GPR1-AS*	0.77	hyper	AluSc8- SINE
chr2	227872705	*DAW1*	0.89	hyper	
chr2	233473583	--	0.92	hypo	n/a
chr2	242070223	*LINC01237*	0.94	hypo	
chr20	2042748	--	0.71	hypo	MIR3- SINE
chr20	2103419	*STK35*	0.86	hypo	CpG island
chr20	4199373	--	0.95	hyper	AluY- SINE
chr20	5508262	*LOC101929207*	0.98	hypo	LINC01729
chr20	17055088	--	0.74	hyper	L1PA12- LINE
chr20	21020228	--	0.86	hypo	CpG island
chr20	21990770	--	0.93	hyper	n/a
chr20	24134912	--	0.72	hyper	n/a
chr20	28520706	--	0.79	hyper	ALR/Alpha- Satellite
chr20	29816832	--	0.74	hypo	L1P1- LINE
chr20	30169025	--	0.98	hyper	SST1- satellite
chr20	35269494	*MMP24*	0.94	hypo	
chr20	41176226	*PLCG1*	0.91	hypo	
chr20	52791043	--	0.89	hypo	LTR33- LTR
chr20	57464328	--	0.77	hypo	AluSx3- SINE
chr20	60481072	--	0.79	hyper	n/a
chr20	62768607	--	0.91	hypo	n/a
chr20	63475434	--	0.94	hypo	n/a
chr21	8238040	--	0.95	hyper	n/a
chr21	8421064	--	0.98	hyper	CpG island south shelf; CpG island north shore

Table A.5 continued

chr21	9015849	--	0.70	hyper	n/a
chr21	13052629	*ANKRD30BP2*	0.78	hyper	HERVIP10F-int - LTR
chr21		--	0.70	hyper	AluSz- SINE
chr21	22190657	--	0.95	hyper	AluY- SINE
chr21	31175084	*TIAM1*	0.92	hyper	AluJb- SINE
chr21	33720267	*ITSN1*	0.88	hypo	
chr21	38775976	--	0.97	hypo	n/a
chr21	41715789	*LINC00479*	0.97	hypo	LINC 47
chr21	42234622	*ABCG1*	0.89	hypo	CpG island north shore
chr21	45421692	*COL18A1*	0.87	hypo	
chr21	45950327	--	0.87	hypo	n/a
chr22	15854768	--	0.76	hyper	n/a
chr22	25995622	*MYO18B*	0.95	hypo	
chr22	31805593	*DEPDC5*	0.86	hypo	AluSx1- SINE
chr22	32702402	*SYN3*	0.78	hyper	AluY- SINE
chr22	36550935	--	0.70	hypo	n/a
chr22	41495548	*ACO2*	0.79	hyper	
chr22	43659516	*EFCAB6*	0.85	hypo	AluJb-SINE
chr22	44232700	--	0.96	hyper	AluSp-SINE
chr22	47551087	--	0.83	hypo	MER31-int - LTR
chr22		*C22orf34*	0.87	hypo	n/a
chr3	6323747	--	0.84	hyper	AluY-SINE
chr3	15378337	--	0.81	hypo	MER66C-LTR
chr3	20662005	--	0.89	hyper	AluYf1-SINE
chr3	21010301	--	0.83	hypo	MLT2A2- LTR
chr3	26671108	*LRRC3B*	0.87	hyper	AluYa5- SINE
chr3	28014819	--	0.85	hyper	AluSz- SINE
chr3	32296344	*CMTM8*	0.77	hyper	L2c- LINE
chr3	37736642	*ITGA9*	0.96	hypo	
chr3	38958253	--	0.84	hypo	LTR33- LTR
chr3	40868938	--	0.96	hypo	n/a
chr3	42159397	*TRAK1*	0.96	hypo	
chr3	46276720	--	0.95	hypo	n/a
chr3	46567076	*LRRC2*	0.79	hyper	AluSx1- SINE
chr3	50243396	*GNA12*	0.98	hypo	
chr3	50359681	*CYB561D2*	0.70	hypo	CpG island
chr3	54626604	*CACNA2D3*	0.71	hyper	
chr3	60377246	*FHIT*	0.84	hyper	AluSc- SINE
chr3	82608153	--	0.96	hypo	L1PA4- LINE
chr3	83160968	--	0.75	hyper	AluSx- SINE

Table A.5 continued

chr3	94345443	--	0.81	hyper	LTR17- LTR
chr3	98732314	*ST3GAL6*	0.94	hypo	
chr3	109776215	--	0.73	hyper	HERVL-int- LTR
chr3	122284610	*CASR*	0.95	hyper	
chr3	127392674	--	0.94	hyper	MIRb- SINE
chr3	127652072	*PODXL2*	0.98	hypo	
chr3	128686010	--	0.98	hypo	AluSp- SINE
chr3	141127667	*SPSB4*	0.91	hypo	
chr3	143374349	*SLC9A9*	0.90	hyper	AluY- SINE
chr3	170826454	--	0.96	hypo	n/a
chr3	173585557	*NLGN1*	0.94	hypo	
chr3	173820103	*NLGN1*	0.74	hyper	AluY- SINE
chr3	178037440	--	0.93	hyper	AluSp- SINE
chr3	179681694	*USP13*	0.93	hyper	LTR6A- LTR
chr3		--	0.77	hypo	AluSq2- SINE
chr3	185446348	*MAP3K13*	0.71	hyper	AluYh3- SINE
chr3	185560278	--	0.74	hyper	AluY- SINE
chr4	46241	--	0.75	hyper	AluSq2-SINE
chr4	4736422	--	0.99	hypo	n/a
chr4	7309145	*SORCS2*	0.88	hypo	
chr4	7973393	*ABLIM2*	0.73	hypo	AluSg-SINE
chr4	12783904	--	0.93	hyper	AluY-SINE
chr4	13626245	*BOD1L1*	0.77	hyper	AluSc- SINE
chr4	24401179	--	0.88	hyper	AluY- SINE
chr4	32679934	--	0.92	hyper	AluY- SINE
chr4	37375661	*NWD2*	0.91	hyper	AluY- SINE
chr4		*PGM2*	0.83	hypo	AluSx1- SINE
chr4	38462903	*LINC01258*	0.92	hypo	
chr4	41467606	*LIMCH1*	0.94	hypo	
chr4	53734272	*LOC100506444*	0.95	hyper	L1PA6- LINE
chr4	63219060	--	0.98	hypo	L1PA11- LINE
chr4	66561899	--	0.94	hyper	LTR12C- LTR
chr4	89979495	--	0.83	hypo	AluSg4-SINE
chr4	91267014	*CCSER1*	0.99	hyper	
chr4	117576099	*LINC01378*	0.96	hyper	
chr4	129091174	*SCLT1*	0.80	hyper	
chr4	130042473	--	0.71	hyper	AluSz- SINE
chr4	130433474	--	0.79	hyper	MER50- LTR
chr4	133852268	--	0.71	hypo	THE1D- LTR
chr4	135835605	--	0.96	hyper	AluSx1- SINE

Table A.5 continued

chr4	139950667	*MAML3*	0.95	hypo	
chr4	155072381	--	0.98	hypo	MLT2B5- LTR
chr4	156435901	--	0.87	hyper	AluSc8- SINE
chr4	156943171	*PDGFC*	0.99	hypo	
chr4	156972103	--	0.79	hypo	CpG island
chr4	161875569	*FSTL5*	0.73	hyper	MER51A- LTR
chr4	172122671	*GALNTL6*	0.92	hypo	
chr4	184989517	--	0.91	hyper	LTR12C- LTR
chr4	185065844	--	0.97	hyper	AluYe5- SINE
chr4	185821121	*SORBS2*	0.93	hyper	
chr5	320741	*AHRR*	0.96	hypo	CpG island
chr5	9374929	*SEMA5A*	0.96	hypo	
chr5	16181842	--	0.79	hyper	n/a
chr5	17237268	*BASP1*	0.97	hyper	AluY- SINE
chr5	20295611	*CDH18*	0.77	hyper	AluSc- SINE
chr5	29481037	--	0.71	hypo	L1PA15-16- LINE
chr5	34293457	--	0.94	hypo	n/a
chr5	35760219	*SPEF2*	0.82	hyper	AluY- SINE
chr5	39575351	--	0.78	hyper	AluY- SINE
chr5	41958227	--	0.99	hyper	AluY- SINE
chr5	44860573	--	0.85	hypo	n/a
chr5	63940282	--	0.98	hypo	n/a
chr5	67165616	*MAST4*	0.92	hypo	
chr5	76396503	--	0.95	hypo	AluSq2- SINE
chr5	80487492	--	0.91	hyper	AluSz- SINE
chr5	83938549	--	0.94	hyper	AluSz- SINE
chr5	96309881	*LOC101929710*	0.89	hyper	L1PA7- LINE
chr5	123099609	*PRDM6*	0.85	hypo	CpG island
chr5	124597019	--	0.82	hypo	n/a
chr5	126057037	--	0.92	hypo	AluSx- SINE
chr5	131267215	*CDC42SE2*	0.70	hyper	AluSc- SINE
chr5	135996517	--	0.99	hypo	n/a
chr5	139365068	*PAIP2*	0.99	hypo	AluSp- SINE
chr5	142753415	*LOC101926975*	0.95	hypo	LINC01844
chr5	142819987	*ARHGAP26*	0.95	hypo	
chr5	149779845	*PPARGC1B*	0.77	hypo	
chr5	149810882	*PPARGC1B*	0.84	hyper	
chr5	151504961	*FAT2*	0.96	hypo	L2a- LINE
chr5	155757656	--	0.95	hypo	n/a
chr5	158690335	--	0.81	hypo	n/a

Table A.5 continued

chr5	170603504	*KCNIP1*	0.83	hypo	
chr5	172836036	*ERGIC1*	0.73	hyper	
chr5	173304859	--	0.96	hypo	n/a
chr5	174022730	--	0.84	hypo	n/a
chr5	181429458	--	0.97	hypo	n/a
chr6	1554788	--	0.93	hypo	n/a
chr6	11607556	--	0.81	hypo	n/a
chr6	15897287	--	1.00	hypo	n/a
chr6	27545064	--	0.86	hypo	n/a
chr6	36789906	*CPNE5*	0.97	hypo	
chr6	54737066	--	0.72	hyper	L1PA14- LINE
chr6	56954118	*DST*	0.82	hypo	CpG island
chr6	60758588	--	0.83	hyper	AluYk3- SINE
chr6	61299572	--	0.83	hyper	ALR/Alpha- Satellite
chr6	65838577	--	0.99	hyper	n/a
chr6	71485562	--	0.90	hypo	AluSz- SINE
chr6	77173762	--	0.90	hypo	L1PA15-16 - LINE
chr6	84227467	*CEP162*	0.93	hypo	CpG island
chr6	88588292	--	0.86	hyper	n/a
chr6	125346778	--	0.94	hyper	n/a
chr6	143455353	*PEX3*	0.74	hyper	AluY- SINE
chr6	159253354	*FNDC1*	0.90	hypo	MamRTE1- LINE
chr6	160616504	*LPA*	0.74	hyper	
chr6	161308483	--	0.83	hypo	LTR9B- LTR
chr6	161476111	*PARK2*	0.91	hyper	AluY- SINE
chr6	163740403	--	0.78	hypo	n/a
chr6		--	0.94	hypo	n/a
chr7	1784596	--	0.90	hyper	AluSp- SINE
chr7	6139834	*USP42*	0.93	hypo	
chr7	10610635	--	0.88	hyper	AluY- SINE
chr7	19548524	--	0.92	hypo	L1PA7- LINE
chr7	23701332	*FAM221A*	0.94	hyper	AluY- SINE
chr7	51079767	*COBL*	0.94	hypo	
chr7		--	0.90	hypo	AluY- SINE
chr7	665002780	--	0.86	hypo	MLT1F1- LTR
chr7	72397228	*CALN1*	0.94	hyper	
chr7	76999555	*1-UPK3BP1-PMS2P11*	0.77	hyper	pseudogene
chr7	98956156	*TRRAP*	0.89	hypo	
chr7	108171197	*NRCAM*	0.97	hyper	

Table A.5 continued

chr7	131044567	*LINC-PINT*	0.94	hypo	AluY- SINE
chr7	135721405	*SLC13A4*	0.89	hyper	
chr7	140017247	*TBXAS1*	0.82	hypo	
chr7	142457036	--	0.94	hyper	n/a
chr7	148597916	*C7orf33*	0.74	hyper	AluY- SINE
chr7	150720791	*GIMAP1*	0.87	hyper	CpG island
chr7	152427666	*KMT2C*	0.94	hypo	
chr7	158586396	*PTPRN2*	0.78	hyper	
chr7	130734372	*KLF14*	0.43	hyper	CpG island south shore
chr8	5977722	--	0.92	hypo	n/a
chr8	7091379	--	0.87	hyper	L2b- LINE
chr8	7596631	--	0.85	hyper	AluSx3- SINE
chr8	10041145	--	0.85	hypo	n/a
chr8	12568233	*LOC729732*	0.86	hyper	uncharacterized
chr8	30028038	--	0.88	hyper	n/a
chr8	32928270	--	0.91	hypo	n/a
chr8		--	0.97	hypo	CpG island
chr8	40813656	*ZMAT4*	0.94	hypo	
chr8	54187430	--	0.94	hypo	n/a
chr8	54558490	--	0.96	hypo	n/a
chr8	58991982	*TOX*	0.76	hyper	
chr8	64499252	--	0.97	hyper	AluY- SINE
chr8	66626477	--	0.82	hyper	AluSg- SINE
chr8	68505655	*C8orf34*	0.74	hyper	
chr8	74006322	*LY96*	0.87	hypo	AluSx1- SINE
chr8	91214003	*LRRC69; SLC26A7*	0.91	hypo	
chr8	95008667	*NDUFAF6*	0.88	hypo	AluSq- SINE
chr8	95082870	--	0.86	hyper	AluYa5- SINE
chr8	96452284	--	0.93	hyper	LTR12C- LTR
chr8	100560334	--	0.74	hyper	AluSx3- SINE
chr8	107047420	--	0.79	hyper	THE1D- LTR
chr8	107459396	*ANGPT1*	0.90	hyper	AluSz- SINE
chr8	111470502	--	0.74	hyper	MER52A- LTR
chr8	114011598	--	0.96	hypo	n/a
chr8	11434279	--	0.87	hyper	MER50B-LTR
chr8	116146678	*LINC00536*	0.82	hyper	HERVL-int - LTR
chr8	116949780	--	0.97	hyper	LTR12C- LTR
chr8	119207197	*MAL2*	0.99	hypo	
chr8	124426306	--	0.92	hypo	n/a

Table A.5 continued

chr8	131457944	--	0.97	hyper	AluY- SINE
chr8	132672809	*LRRC6*	0.88	hyper	MIR3-SINE
chr8	139493896	--	0.78	hyper	AluY- SINE
chr8	139561834	--	0.76	hyper	AluSg- SINE
chr8	144183050	*MROH1*	0.81	hypo	AluSz6- SINE
chr9	4496182	*SLC1A1*	0.72	hyper	
chr9	38672611	--	0.93	hypo	n/a
chr9	38769009	--	0.73	hyper	n/a
chr9	39666078	--	0.92	hyper	AluYc- SINE
chr9	39807190	*GLIDR*	0.94	hyper	AluSc- SINE
chr9	40522056	*LOC102724580*	0.99	hyper	AluSg4- SINE
chr9	41016947	*PGM5P2*	1.00	hyper	
chr9	60526557	--	0.92	hyper	CER- satellite
chr9	65285785	--	0.75	hyper	AluYh3- SINE
chr9	78125693	--	0.82	hypo	n/a
chr9		--	0.84	hyper	n/a
chr9	92618787	*IPPK*	0.99	hypo	
chr9	94780865	*C9orf3*	0.86	hyper	n/a
chr9	104258663	--	0.75	hypo	MER4A1- LTR
chr9	108385902	--	0.99	hyper	AluYk11- SINE
chr9	111068348	--	0.93	hyper	AluSg- SINE
chr9	113085081	--	0.97	hypo	L1MA7- LINE
chr9	121491473	*GGTA1P*	0.91	hypo	L2b- LINE
chr9	131256702	--	0.95	hypo	n/a
chr9	133793132	*VAV2*	0.96	hypo	
chr9	137858603	--	0.83	hyper	AluY- SINE