Chapter 11

APPLICATION OF TAQMAN® CHEMISTRY FOR ALLELIC DISCRIMINATION*

Victor J. Johnson

Toxicology and Molecular Biology Branch, Health Effects Laboratory Division,
National Institute for Occupational Safety and Health,
Centers for Disease Control and Prevention, Morgantown, WV 26505

Contact:
Victor J. Johnson
Toxicology and Molecular Biology Branch,
Health Effects Laboratory Division, NIOSH/CDC,
1095 Willowdale Road, M/S 3014.
Morgantown, WV 26505-2888 USA
Email: VJohnson3@cdc.gov

**The findings and conclusions in this report are those of the author and do not necessarily represent the views of the National Institute for Occupational Safety and Health. Mention of any company name or product does not constitute endorsement by the National Institute for Occupational Safety and Health (NIOSH).*

11. 1. Introduction

Genetic variation in the form of single nucleotide polymorphisms (SNP's) within genes can influence many levels of signaling including gene expression and protein function. Therefore, analysis of such variation can provide valuable insight into disease biology potentially providing important information for the discovery of new therapeutic targets. Traditionally, genotyping technologies used for allelic discrimination were direct sequencing analysis and polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP). Due to the laborious and low-throughput nature of these technologies, they are only amenable to analysis of small numbers of SNP's in relatively small populations. Low sample size results in low statistical power, which contributes to the lack of reproducibility of many SNP-disease association studies[1]. The completion of the human genome sequencing project has been accompanied by an exponential increase in genotyping demands. Larger population sizes are being genotyped for larger number of SNP markers requiring a concomitant increase in the throughput of genotyping techniques and a decrease in the associated costs.

Numerous non-gel-based technologies displaying increased throughput have been developed and are reviewed elsewhere[2,3]. The focus of the present chapter will be on the use of Taqman® chemistry for allelic discrimination. Homogeneous solution hybridization using fluorescence resonance energy transfer (FRET) is one of the newly validated real-time PCR methodologies currently used. It is typified by the Taqman® assay[4,5] and has recently been adapted for use in allelic discrimination[6]. Being homogeneous, all aspects including amplification and detection of genotypes are performed in a single tube. This can greatly reduce technical error and technician time leading to increased genotyping capacity at a reduced cost as well as a reduction in the likelihood of cross contamination and genotyping misidentification. Amplification is performed using a single pair of PCR primers flanking the polymorphic site and detection is achieved by solution hybridization of two dual-labeled fluorescent probes each with a characteristic reporter fluorophore on the 5' end and a quencher molecule on the 3' end[5]. The probes differ by one base at the site of the SNP and thus one probe is specific for each allele present and will bind with high affinity only to that allele. The secondary structure of un-hybridized probes results in quenched fluorescence due to the close physical proximity of the fluorophore and the quencher molecules. Once hybridized to the amplicon, the probe is cleaved by the 5'-3' nuclease activity of *Taq* DNA polymerase resulting in the release of the 5' fluorophore from its 3' quencher molecule (see *Figure 11-1A*). The resulting increase in fluorescent signal is proportional to PCR product accumulation[4]. The probe containing a single mismatch to the allele present at the polymorphic site will be displaced by *Taq* polymerase allowing refolding and quenching of the fluorescence.

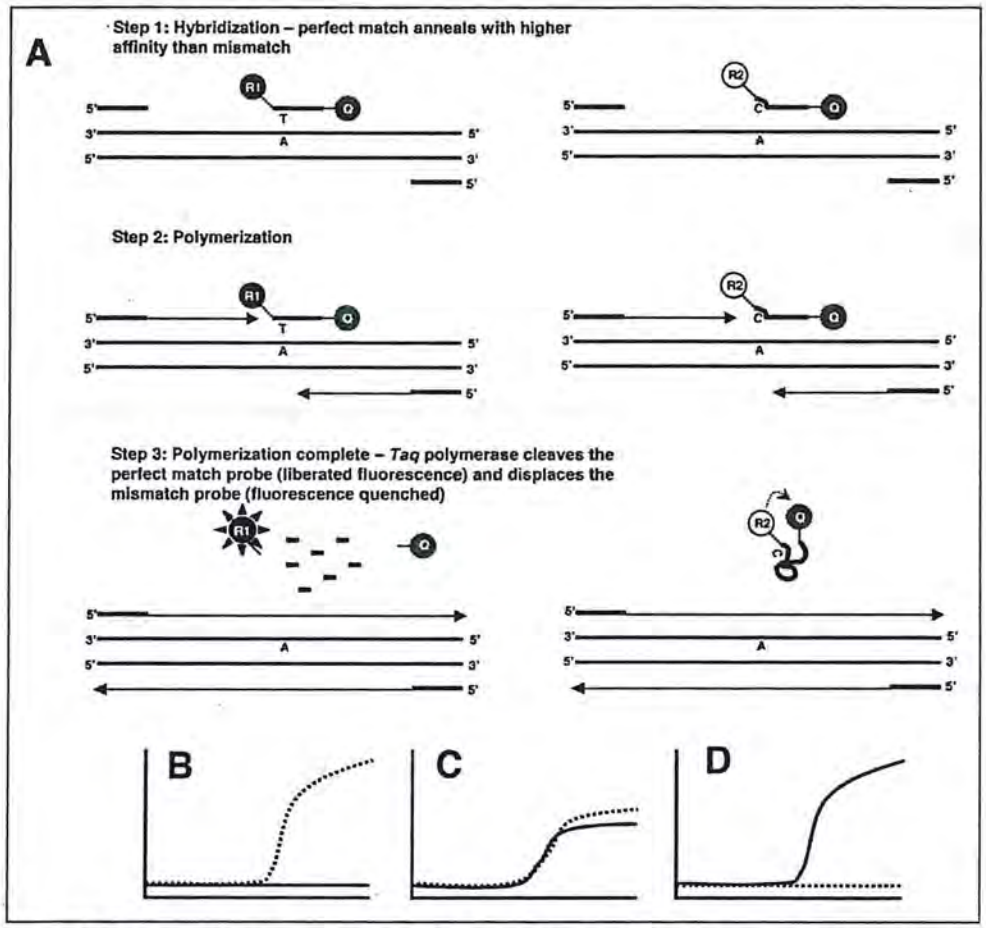

Figure 11-1: **Schematic diagram showing the molecular basis of the Taqman® allelic discrimination assay. A)** Primers and probes hybridize to the genomic template followed by polymerization by *Taq* DNA polymerase. A perfect-match probe is cleaved by the 5'-3' exonuclease activity of *Taq* liberating the fluorophore from its quencher and the signal is proportional to amplicon accumulation, whereas the mismatch probe is displaced and fluorescence is quenched by resonance energy transfer. Three fluorescence patterns are possible: **B)** only "dotted" wavelength (green) indicating homozygous for allele 1, **C)** "dotted" and "solid" wavelength (green and red) indicating heterozygous and **D)** only "solid" wavelength (red) indicating homozygous for allele 2. R1 - reporter fluorophore 1, green (dotted line); R2 - reporter fluorophore 2, red (solid line); Q - quencher.

Therefore, genotypes are determined by measuring fluorescence intensity at two wavelengths; detection of only one wavelength indicating homozygous (*Figure 11-1B and D*) whereas fluorescence at both wavelengths represents heterozygous (*Figure 11-1C*).

The use of Taqman® chemistry for the detection of allelic variation has been validated by demonstrating comparable results with PCR-RFLP[7], mutagenically separated, allele specific PCR[8], PCR with sequence-specific primers[9,10], and sequencing[11]. These assays have also been compared with other FRET probe systems including molecular beacons[12]. Taqman® assays are now being used as a tool for comparison and validation of many genotyping technologies[13-15]. The purpose of the following discussion is to present basic guidelines for designing, performing and interpreting allelic discrimination assays based on Taqman® technology. Subsequently, some of the potential uses of these assays will be discussed.

11.2. Methodology

11.2.1. Primer/Probe Design

The design of primers and probes for real-time PCR genotyping is one of the most critical aspects of this technology. It is at this stage of assay development that effort should be directed towards designing primers and probes that can utilize standardized hybridization and amplification parameters. Standardized conditions help to minimize technical error as well as increase throughput. Selecting and verifying the sequence surrounding the SNP of interest is vital to the successful design of a Taqman® genotyping assay. The sequence should be between 300 and 600 base pairs (bp) with the SNP close to the mid-point in order to maximize the number of potential primer-probe combinations that can be generated. Once a suitable region of DNA is selected, it is important to determine the sequence quality. The most rigorous method to determine the quality is to re-sequence using several independent genome samples. This will allow for detection of sequencing errors and also for identification of novel SNP's in the vicinity of the target SNP. Identification of novel SNP's is critical to assay success since an unidentified mismatch located in the complimentary regions will affect primer and/or probe hybridization independent of the SNP of interest. When novel SNP's are identified, an N should be substituted for the base to inform the primer/probe design software to avoid that site. Another method to determine the presence of potentially interfering SNP's is to perform a BLAST search using the NCBI dbSNP database (http://www.ncbi.nlm.nih.gov/SNP/snp_blastByOrg.cgi). A sequence BLAST® search should also be performed to eliminate the possibility of the primers or probes binding to another region of the human genome. Non-unique sequences should be masked with N to avoid these regions. The high quality sequence can then be entered into a design software package such as Primer Express® (Applied Biosystems) or Beacon Designer® (Bio-Rad Laboratories) in order to generate potential primer/probe combinations as outlined below.

Our laboratory has used Primer Express® to design assays since this software has the advantage of incorporating a minor grove binder (MGB) at the 3' end of the probe. Placing a MGB at the 3' end of the probe increases the binding affinity for DNA allowing for the use of shorter probes effectively increasing the destabilizing effect of a single mismatch. This results in a greater difference in melting temperature between the perfect match and the mismatch probe, thus improving genotype discrimination[16,17]. Probes should be designed as close as possible to the following guidelines;

- The 5' end of the probe should not contain guanine residues
- Probe melting temperature should be approximately 5-10°C above primer melting temperature
- Nucleotide runs should be avoided especially guanines which can stabilize probe binding independent of the SNP
- Probes for both alleles should be specific for the same strand, either sense or complement
- Melting temperature of both probes should differ by no more than 1°C
- SNP site should be close to the center of the probe

The allele-specific probes are dual labeled with a MGB and a non-fluorescent quencher at the 3' end and a unique fluorescent molecule at the 5' end. Primers are designed around the probe so that the amplicon is between 50 and 100 bp in length. The melting temperature of the primers should be 60°C. Optimization of candidate primers using standardized PCR conditions (see section 11.2.2.) should be conducted prior to the addition of probes. Primer pairs generating a high yield of specific product in the absence of non-specific bands and primer-dimers should be selected. High efficacy primers can then be used to optimize probe concentrations and the ability to detect specific genotypes.

11.2.2. PCR Conditions

The genotyping assays that our laboratory has designed, as well as those available commercially (Applied Biosystems) and in public databases (SNP500Cancer database,[11]), have been optimized to run under universal PCR conditions. Our conditions include a 25µl reaction volume consisting of 10ng of genomic DNA, 900nM of each primer, 200nM of each probe, and 12.5µl of 2x Taqman® Universal PCR Master

Mix (contains PCR buffer, passive reference dye ROX, deoxynucleotides, uridine, Uracil-N-glycosylase and AmpliTaq Gold® DNA polymerase; Perkin-Elmer, Applied Biosystems, Foster City, CA). Other universal PCR master mixes can be substituted although empirical testing should be conducted to verify performance of the primer/probe sets. Amplification can be performed using a real-time thermal cycler with cycling conditions of 2 min at 50°C; 10 min at 95°C; 50 cycles of 92°C for 30 sec; 60°C for 1 min. Real-time fluorescence detection is performed during the 60°C annealing/extension step of each cycle. The use of this technology for genotyping is not limited to laboratories with real-time PCR instrumentation. Amplification can be performed in a regular thermal cycler followed by end-point detection of fluorescence intensity using a fluorescent plate reader[16,7]. The cycling conditions would be the same as for real-time instruments although optimization of temperatures may be needed due to subtle variation in temperature profiles between instruments.

11.2.3. Signal detection

Genotype identification is based on the detection of the two unique fluorophores that are conjugated to the allele-specific probes, each of which fluoresce at distinct wavelengths. As shown in *Figure 11-1*, the 5'-3' exonuclease activity of Taq polymerase results in the liberation of the fluorescent molecule from the quencher only when a perfect match between probe and genomic template exists. Therefore, amplification of DNA samples homozygous for either allele results in fluorescence detected in one channel while heterozygous samples will fluoresce in both channels (*see Figure 11-1B-D*). The fluorescent signals can be detected during amplification using real-time PCR instrumentation or at the end of amplification using a fluorescent plate reader. Maximum separation between genotypes is achieved when the detection instrument is capable of performing spectral compensation. The fluorescent signal generated by one fluorophore may bleed into the detection channel of the second fluorophore and spectral compensation is used to minimize such effects. The coefficient of variation (CV) for the intensity of each genotype grouping will be wider in the absence of compensation, which may complicate genotype identification. We observe 100% concordance for genotypes derived from signals detected using real-time PCR instrumentation or fluorescence plate readers[7].

11.2.4. Genotype "Calling"

The data generated by Taqman® genotyping assays are plotted on a two-color bivariate plot using threshold cycle (Ct) values or endpoint fluorescence. The ability to discriminate alleles based on endpoint (final cycle) fluorescence intensity makes this technology available to laboratories that do not have real-time PCR instrumentation. Individual samples cluster into 4 quadrants on the bivariate plot.

No template controls and un-amplifiable samples are located in the lower left quadrant, samples homozygous for allele 1 in the lower right quadrant, samples homozygous for allele 2 in the upper left quadrant and heterozygous samples in the upper right quadrant (*see Figure 11-2C*).

Most real-time PCR instruments will automatically plot the data and assign genotypes based on the location of homozygous and heterozygous positive control samples. Genotypes can also be called manually from the location of the sample on the bivariate plot or the data can be plotted using Microsoft Excel and genotypes called

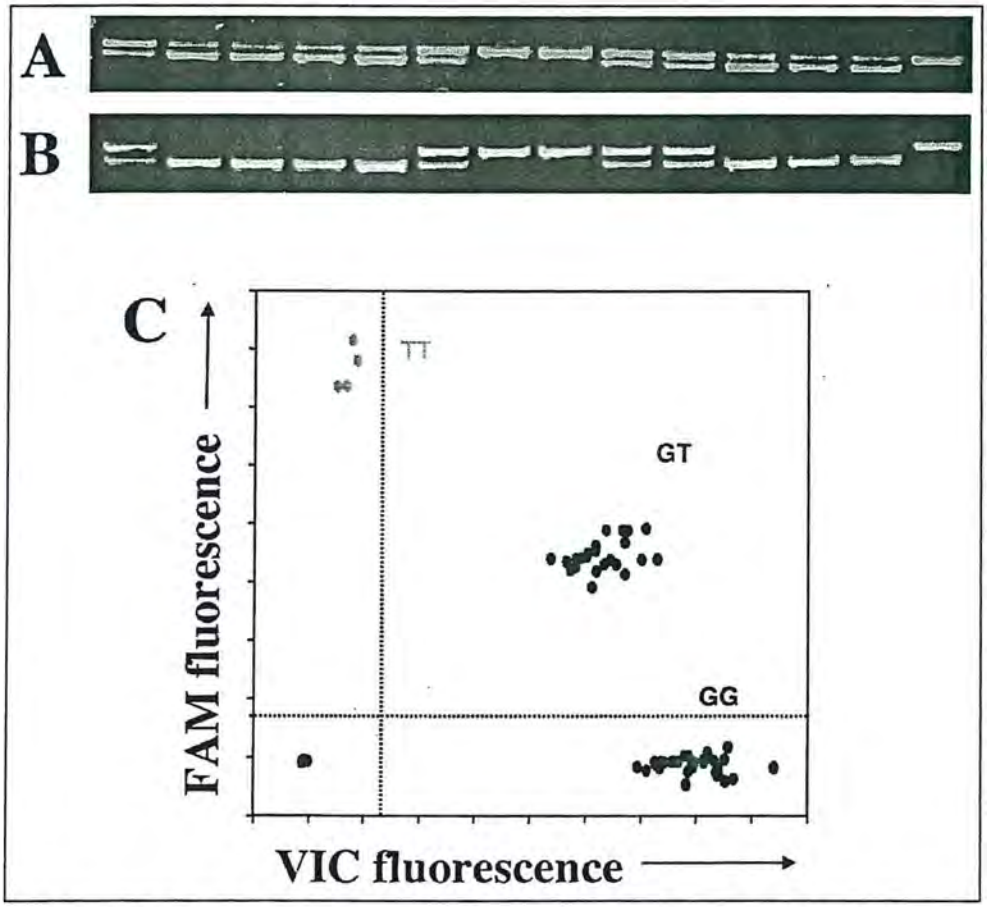

Figure 11-2: **The effect of enzyme concentration and incubation time on TGFB1(-509) amplicon digestion.** PCR-RFLP samples were digested with **A**) 10 U Eco 81 I or **B**) 20 U Eco 81 I at 37°C for 18 hours followed by gel detection of banding patterns. It is clearly evident that 10 U of enzyme resulted in incomplete digestion and misidentification of genotypes. **C**) The Taqman® assay did not produce any ambiguous results.

automatically using Genotyper-Endpoint, an Excel macro created by Max Myakishev (NCI, Bethesda, MD). The data can also be analyzed using statistical approaches including K-means clustering to define genotypes[18,19]. Manual examination of the bivariate plot is important for the identification of outliers that cannot be assigned definitively to one genotype population. Fluctuation in the amount of input genomic template between samples is a substantial cause of wide CV's and outliers. Therefore, it is important to accurately determine the DNA concentration of each sample and to normalize to a constant concentration. Sample degradation is the other main cause of outliers and increased CV's, an issue that becomes important when genotyping archival samples (see section 11.3.3.).

11.2.5. Quality Control/Quality Assurance

Quality control/quality assurance (QC/QA) practice is critical for high throughput genotyping based on Taqman® technology. Proper adherence to an established QC/QA program will minimize genotype misclassification and decrease genotyping error rate. The following are some guidelines for development of an effective QC/QA program. During the assay development phase, successful assays should be used to genotype a panel of genomic DNA samples (around 100 independent samples) to determine genotype frequencies. It is advisable that the minor allele frequency be greater than 5%, otherwise the population size required to maintain adequate statistical power may become prohibitive for most association studies. Genotype frequencies should be in accordance with those in the published literature for the ethnic population under study. In addition, genotype frequencies should be tested for compliance with Hardy-Weinberg equilibrium. Assays meeting these requirements can be used for genotyping.

Each genotyping plate should include the following controls; 1) no template control (PCR master mix only), 2) positive sample homozygous for allele 1, 3) positive sample homozygous for allele 2, and 4) positive sample heterozygous for both alleles. These positive and negative samples will define genotype locations on the bivariate plot and all unknown samples should form clusters around one of the controls samples. If any of the controls do not amplify as expected or the no template control shows amplification, all of the results for that plate are suspect and should be repeated. In addition to the positive and negative controls suggested above, the bivariate plot should be examined following each plate run to identify samples that do not clearly belong to a specific cluster. Normalizing the DNA concentration of each sample will help to minimize the CV for each cluster by ensuring equal amounts of genomic template. The DNA concentration and quality of ambiguous samples should be checked and adjusted if necessary, and these samples should be repeated using the Taqman® assay. Under certain circumstances such as when using formalin-

fixed archival tissue samples as the genomic DNA source, it may be more effective to determine the concentration of amplifiable DNA in the samples as opposed to the actual DNA concentration. This can be done by using a quantitative Taqman® assay to measure the amplifiable concentration of a non-polymorphic region of the genome, as validated for the BRCA1 gene[20]. If the sample remains ambiguous, another genotyping technique such as PCR-RFLP should be considered to confirm the genotype otherwise no genotype should be assigned. The inclusion of a few blinded duplicate samples for each run will help discover misidentification and sample contamination. Upon completion of the study samples, genotype frequencies should be tested for Hardy-Weinberg equilibrium and 10% of samples should be randomly selected and repeated to further avoid misclassification and estimate error rate. Acceptable levels of overall genotyping error rate (includes genotype identification errors and technical error) vary depending on the minor allele frequency such that the cost of genotyping errors greatly increases as the minor allele frequency approaches zero[21,22]. Therefore, QC/QA procedures should strive to maintain genotyping error rate < 1%. Other QC/QA measures generally applicable to the use of PCR should also be followed, including cleaning and running diagnostic procedures on thermal cyclers and detection instrumentation, complete physical separation of pre-PCR and post-PCR reaction components, the use of Uracil-N-glycosylase and UTP to degrade PCR amplicons from previous reaction if contamination occurs, and the use of barrier pipet tips and other engineering controls to prevent contamination.

11.3. Results, Interpretation and Applications

11.3.1. Cytokine Polymorphism Genotyping and Validation with PCR-RFLP

We have designed and validated Taqman® genotyping assays for numerous SNP's in cytokine genes[7] and antioxidant genes[23]. These assays have been validated based on parallel genotyping by PCR-RFLP[7]. Based on genotyping results for 800 independent DNA samples by Taqman® and PCR-RFLP, the calculated concordance rate was 99.4% (5 mismatches/800 genotypes). Re-analysis of the mismatching genotypes confirmed that the original real-time PCR result was correct. Inspection of the original RFLP gel results showed that discrepant samples produced poor restriction patterns thus accounting for the misidentification. The genotyping error rate also decreased from 3.2% for PCR-RFLP to < 1% for the Taqman® assays[7], similar to the Taqman® error rate reported by Ranade *et al.*[19].

Taqman® assays are capable of overcoming some technical issues that are associated with PCR-RFLP. For example, incomplete enzyme digestion is a problem inherent to the PCR-RFLP technique for genotyping. The *TGFβ1(-509)* cytokine SNP can serve to illustrate this issue. The digestion products for the CC, CT, and TT genotypes

are 429+26, 455+429+26, and 455 bp, respectively as illustrated in *Figure 11-2* (26 bp band is not visible). Incomplete digestion of a CC genotype will result in the appearance of bands at 455, 429 and 26 bp resulting in a false positive for the heterozygous CT genotype. The RFLP products shown in *Figure 11-2A* are the result of digestion with 10 U *Eco* 81 I at 37°C for 18 hours. The same samples digested with twice the enzyme concentration (*Figure 11-2B*) demonstrated the errant identification of the CC genotypes as heterozygous due to incomplete digestion. Similar findings have been observed for the *IL-6(-174)* cytokine SNP and the authors suggested that a non-mutation dependent cleavage site be introduced into one primer to monitor completion of digestion[8]. Incomplete digestion can conceivably contribute to errors in samples that yield greater amounts of PCR product. Such samples may require additional enzyme or incubation time to fully digest indicating that optimization of enzyme concentration and digestion time are critical to the success of PCR-RFLP based genotyping. This problem is overcome with the Taqman® allelic discrimination assay since the technique does not rely on any post-PCR manipulations. The Taqman® assay for *TGFβ1(-509)* did not produce any ambiguous samples (*Figure 11-2C*).

Another advantage of using Taqman® assays is that increased sensitivity of detection can be achieved compared to PCR-RFLP. The limit of detection for the Taqman® assay developed for *IL-1β(-511)* SNP was 0.5 ng of genomic DNA per reaction, whereas the lowest amount of input genomic DNA that resulted in unambiguous restriction patterns for the same SNP was 3.9ng of DNA/reaction (*Figure 11-3*). Therefore, the Taqman® assay would lead to an 8-fold reduction in DNA consumption, which is an important step in conserving and maximizing the use of patient DNA samples. The limits of detection for these two techniques will vary for individual SNP's and should be tested empirically.

Another advantage of the Taqman® genotyping assay is ease of performance compared to PCR-RFLP. The homogeneous nature of these assays means that the entire test, including amplification and genotype identification is performed in a single closed tube. This leads to reduced hands-on time by eliminating post-PCR manipulations, thus increasing the throughput and reducing the cost. The throughput for our lab has increased from approximately 6 - 24 hours/80 genotypes by PCR-RFLP to less than 2 hours/80 genotypes by real-time PCR assay with only 20 minutes hands-on-time. The reagent costs for the RFLP assays used in our laboratory are estimated at $3.50/sample, whereas reagents for the Taqman® assays are approximately $0.85/sample at the time of this writing. This indicates that real-time PCR genotyping can reduce costs up to 75% over traditional RFLP methods. The use of plate formats with more wells (i.e. 384 wells) or the use of microfluidics and nanoliter reaction volumes can also dramatically increase assay throughput as well as further decrease the time and cost associated with genotyping large populations.

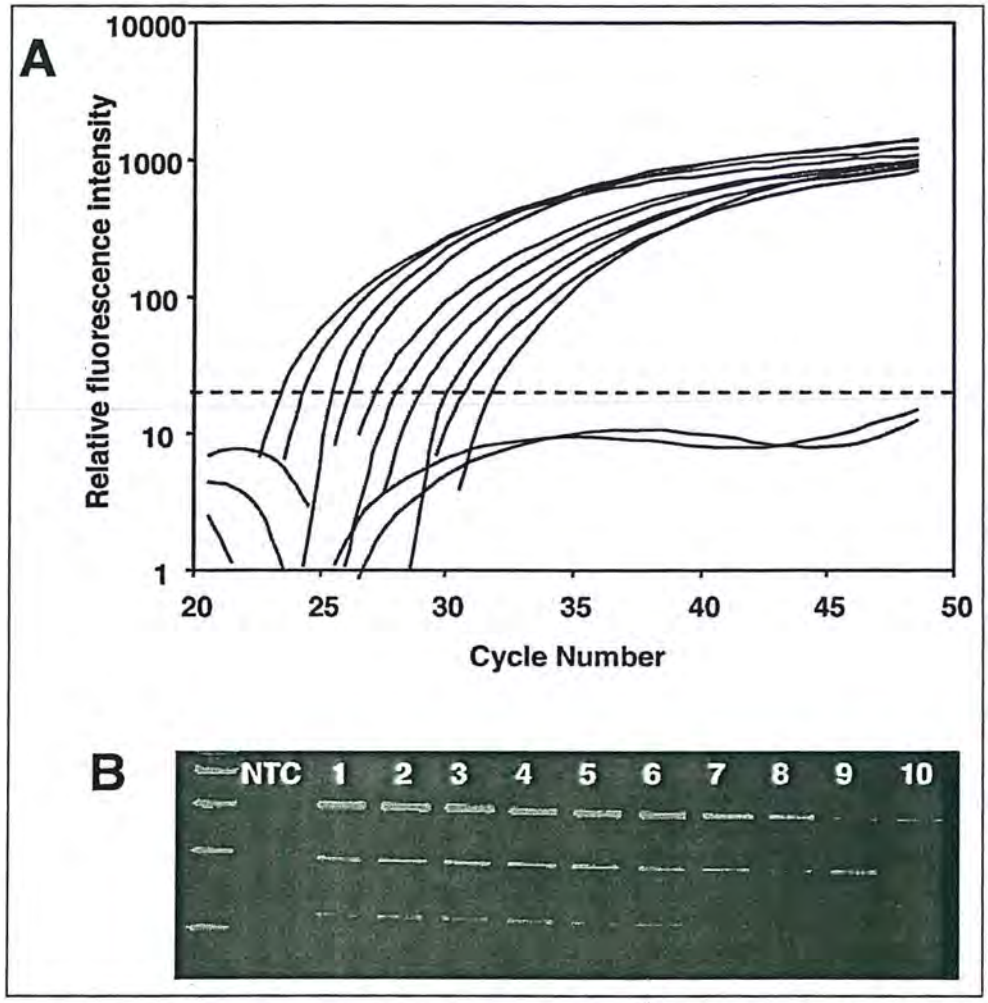

Figure 11-3: **Detection sensitivity for Taqman® versus PCR-RFLP assays for *IL-1β(-511)*.** Ten serial 2-fold dilutions between 500 and 0.5ng of genomic DNA were amplified by both techniques. **A)** All samples containing genomic template clearly amplified for the Taqman® assay, whereas no amplification was evident in the two no-template controls (below the dashed threshold line). **B)** Ambiguous results were obtained by PCR-RFLP starting at dilution number 7 corresponding to 3.9ng of input DNA template. NTC = no template control.

11.3.2. Insertion/Deletion Mutations

Taqman® assays are not restricted to the detection of SNP mutations. We have developed Taqman®-based assays for the detection of deletion mutations in glutathione-S-transferase *(GST)* family of genes[23]. Traditionally, deletions in *GSTM* and

GSTT isoforms have been identified using traditional PCR with primers that are within the deleted stretch of DNA. The patient is considered *GSTM1* or *GSTT1* null when no band results in the PCR. A control for the performance of the PCR reaction is usually included by amplifying a region in another gene. A Taqman® primer/probe set was designed for *GSTM1* and *GSTT1*, each containing a single 6FAM-labeled probe that specifically hybridizes within the deleted fragment of the respective gene. The primers were designed to be within the deleted region also. An internal control primer/probe set was designed to amplify and detect the presence of the *GAPDH* gene using a VIC labeled probe[23]. Since the primer/probe sets for the deletion mutations and the internal control were designed to perform under standardized conditions, both assays can be run in the same tube and a two color bivariate plot can be used for genotype identification (*Figure 11-4*). Samples with intact *GSTM1* or *GSTT1* are located in the upper right quadrant and the true null genotypes (those that did not amplify for the *GST* gene but were positive for *GAPDH*) are located in the lower right quadrant. Samples that do not amplify for *GAPDH* (samples in the lower left quadrant with the no template controls) should be repeated. Recently, similar Taqman® assays have been developed that are capable of distinguishing samples that contain two functional copies (homozygous normal) from those with only one functional copy (heterozygous) by using the rate of amplification for *GSTM1* or *GSTT1* to determine active copy number[24].

Insertion/deletion (*insdel*) polymorphisms with clearly defined and stable break points provide a unique opportunity to use Taqman® technology to provide all possible genotypes without the requirement for an internal control primer/probe set. One set of primers and a single probe (6FAM label) can be constructed to only give a product when the region is deleted, the primers being on the outside of the deleted fragment and the probe hybridizing across the two deletion endpoints. Even though the primers will produce a product when no deletion has occurred, specific hybridization of the probe is disrupted and no signal is generated. The other primer/single probe (VIC label) set would be specific for the deleted fragment with both primers and probes being inside the break points. Standardized reaction conditions allow for both primer/probe sets to be run in the same reaction. Plotting the fluorescence intensities on a bivariate plot results in clustering of the *ins/ins* population in the VIC-only quadrant, the *ins/del* population in the 6FAM-VIC quadrant and the *del/del* population in the 6FAM-only quadrant. This concept has recently been developed for genotyping a 276 bp *insdel* polymorphism on chromosome 22q13 resulting in more rapid, cost-effective and less laborious genotyping[25].

11.3.3. Archival DNA Samples

Archival tissues that have been formalin-fixed and paraffin embedded represent an invaluable resource for studying the association between genetic variation and

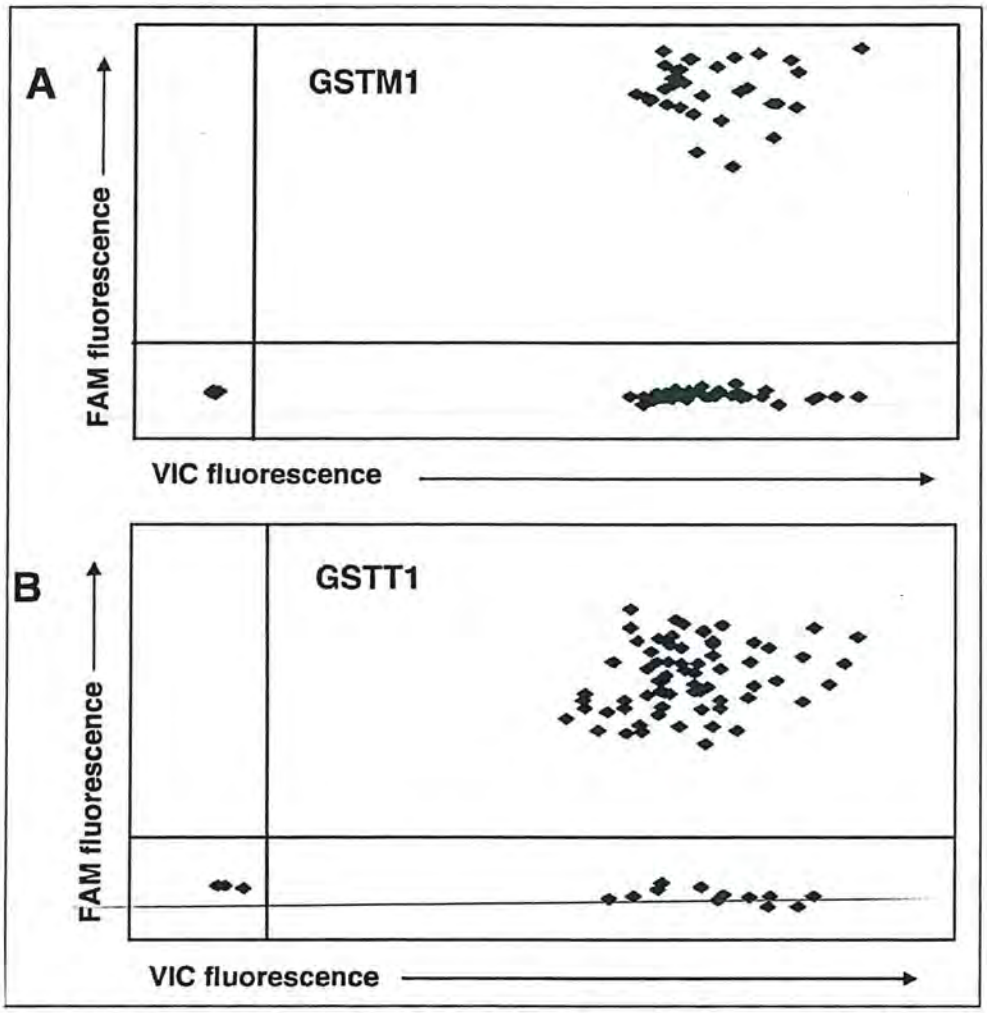

Figure 11-4: **Bivariate allelic discrimination plots of Taqman® assays for A) *GSTM1* and B) *GSTT1*.** *GST* null genotypes are located in the lower right quadrant and *GST* replete genotypes in the upper right quadrant. No template controls are located in the lower left quadrant.

disease, especially when collected as part of autopsy studies. Formalin fixation is known to result in degradation of genomic DNA and the extent of degradation increases with time from fixation. Due to the progressive DNA degradation, the success of a PCR amplification reaction decreases with the size of the desired product and the age of the tissue sample. The degree of degradation may also be gene-specific due to differences in chromosome location and histone association. Amplification products < 180 bp have been recommended as optimal for archival

specimens and will greatly reduce the number of non-amplifiable samples in a population for most genes[26]. This guideline makes the Taqman® genotyping approach ideally suited for archival samples since the target amplicon size is 50-100 bp.

Recently, Taqman® assays were used to study the association between genetic variation in antioxidant genes and the incidence of progressive massive fibrosis in a population of 700 coal workers[23]. The source of genomic DNA was formalin-fixed and paraffin-embedded lung tissues that were collected at the time of autopsy between 1972 and 1996. Success rate for the SNP's examined ranged from 87.7 to 90.5%. Attempts to genotype this population for the *TGFβ1(-509)* SNP were unsuccessful using PCR-RFLP. In contrast, the Taqman® assay amplified successfully for 87.7% of the samples (Yucesoy et al., unpublished results). This is not surprising given that the amplicon for the PCR-RFLP assay was 455 bp increasing the chances of damage within the target region due to the fixation and sample aging processes. The Taqman® assay was able to overcome this problem likely due to the amplification of a 100 bp product. Similar success rates were reported for a panel of Taqman® assays performed on DNA extracted from archived breast tissue[27]. These studies suggest that Taqman® genotyping is a viable strategy when studying genetic variation within population especially when archival samples are the source of genomic DNA as it is capable of amplifying relatively poor integrity DNA.

11.4. Summary and Future Direction

Taqman® chemistry has been successfully adapted for use as a tool to discriminate allelic variation in human study populations. This technique has been validated against a number of well-accepted techniques including PCR-RFLP, sequencing, and allele-specific PCR. The Taqman® assays are now being used as the basis for validation of newer genotyping strategies indicating its growing acceptance in the scientific literature. These assays are amenable to the high-throughput genotyping demands of the post-human genome sequencing era and are associated with less cost and technical error compared to more traditional approaches. The application of microfluidics and nanoliter volumes will greatly further the high-throughput/low cost attributes of this genotyping platform.

Acknowledgements

The author would like to thank Drs. Berran Yucesoy and Michael Luster for their critical review of this manuscript during its preparation. The author would like to acknowledge support from the NIOSH-NIEHS Interagency Agreement (Y1-ES0001-06). All experimentation complies with the mandates of the NIOSH Institutional Review Board and the current laws of the United States of America.

References

1. ***Ioannidis JP, Trikalinos TA, Ntzani EE, Contopoulos-Ioannidis DG:*** (2003). Genetic associations in large versus small studies: an empirical assessment. *Lancet* 361, 567-71.

2. ***Chen X, Sullivan PF:*** (2003). Single nucleotide polymorphism genotyping: biochemistry, protocol, cost and throughput. *Pharmacogenomics J* 3, 77-96.

3. ***Shi MM:*** (2001). Enabling large-scale pharmacogenetic studies by high-throughput mutation detection and genotyping technologies. *Clin Chem* 47, 164-72.

4. ***Holland PM, Abramson RD, Watson R, Gelfand DH:*** (1991). Detection of specific polymerase chain reaction product by utilizing the 5'——3' exonuclease activity of Thermus aquaticus DNA polymerase. *Proc Natl Acad Sci U S A* 88, 7276-80.

5. ***Livak KJ, Flood SJ, Marmaro J, Giusti W, Deetz K:*** (1995). Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. *PCR Methods Appl* 4, 357-62.

6. ***Livak KJ:*** (1999). Allelic discrimination using fluorogenic probes and the 5' nuclease assay. *Genet Anal* 14, 143-9.

7. ***Johnson VJ, Yucesoy B, Luster MI:*** (2004). Genotyping of single nucleotide polymorphisms in cytokine genes using real-time PCR allelic discrimination technology. *Cytokine* 27, 135-41.

8. ***Haring D, Endler G, Hsieh K, Mannhalter C:*** (2002). Comparison of methods for determination of interleukin 6 G(-174)C promoter polymorphism. *Biotechniques* 33, 1114-7.

9. ***Ficko T, Galvani V, Rupreht R, Dovc T, Rozman P:*** (2004). Real-time PCR genotyping of human platelet alloantigens HPA-1, HPA-2, HPA-3 and HPA-5 is superior to the standard PCR-SSP method. *Transfus Med* 14, 425-32.

10. ***Sylvain K, Aurelie H, Marc M, Christophe R:*** (2004). Rapid screening for HLA-B27 by a TaqMan-PCR assay using sequence-specific primers and a minor groove binder probe, a novel type of TaqMan trade mark probe. *J Immunol Methods* 287, 179-86.

11. ***Packer B R, Yeager M, Staats B, Welch R, Crenshaw A, Kiley M, Eckert A, Beerman M, Miller E, Bergen A, Rothman N, Strausberg R, Chanock SJ:*** (2004). SNP500Cancer: a public resource for sequence validation and assay development for genetic variation in candidate genes. *Nucleic Acids Res* 32, D528-32.

12. ***Tapp I, Malmberg L, Rennel E, Wik M, Syvanen AC:*** (2000). Homogeneous scoring of single-nucleotide polymorphisms: comparison of the 5'-nuclease Taqman® assay and Molecular Beacon probes. *Biotechniques* 28, 732-8.

13. ***Armstrong B, Stewart M, Mazumder A:*** (2000). Suspension arrays for high throughput, multiplexed single nucleotide polymorphism genotyping. *Cytometry* 40, 102-8.

14. ***Bray MS, Boerwinkle E, Doris PA:*** (2001). High-throughput multiplex SNP genotyping with MALDI-TOF mass spectrometry: practice, problems and promise. *Hum Mutat* 17, 296-304.

15. ***Nordfors L, Jansson M, Sandberg G, Lavebratt C, Sengul S, Schalling M, Arner P:*** (2002). Large-scale genotyping of single nucleotide polymorphisms by Pyrosequencingtrade mark and validation against the 5'nuclease (Taqman®) assay. *Hum Mutat* 19, 395-401.

16. ***de Kok JB, Wiegerinck ET, Giesendorf BA, Swinkels DW:*** (2002). Rapid genotyping of single nucleotide polymorphisms using novel minor groove binding DNA oligonucleotides (MGB probes). *Hum Mutat* 19, 554-9.

17. ***Kutyavin IV, Afonina IA, Mills A, Gorn VV, Lukhtanov, EA, Belousov ES, Singer MJ, Walburger DK, Lokhov SG, Gall AA, Dempcy R, Reed MW, Meyer RB, Hedgpeth J:*** (2000). 3'-minor groove binder-DNA probes increase sequence specificity at PCR extension temperatures. *Nucleic Acids Res* 28, 655-61.

18. ***Olivier M, Chuang LM, Chang MS, Chen YT, Pei D, Ranade K, de Witte A, Allen J, Tran N, Curb D, Pratt R, Neefs H, de Arruda Indig M, Law S, Neri B, Wang L, Cox DR:*** (2002). High-throughput genotyping of single nucleotide polymorphisms using new biplex invader technology. *Nucleic Acids Res* 30, e53.

19. ***Ranade K, Chang MS, Ting CT, Pei D, Hsiao CF, Olivier M, Pesich R, Hebert J, Chen YD, Dzau VJ, Curb D, Olshen R, Risch N, Cox DR, Botstein D:*** (2001). High-throughput genotyping with single nucleotide polymorphisms. *Genome Res* 11, 1262-8.

20. ***Haque KA, Pfeiffer RM, Beerman MB, Struewing JP, Chanock SJ, Bergen AW:*** (2003). Performance of high-throughput DNA quantification methods. *BMC Biotechnol* 3, 20.

21. ***Kang SJ, Finch SJ, Haynes C, Gordon D:*** (2004). Quantifying the percent increase in minimum sample size for SNP genotyping errors in genetic model-based association studies. *Hum Hered* 58, 139-44.

22. ***Kang SJ, Gordon D, Finch SJ:*** (2004). What SNP genotyping errors are most costly for genetic association studies? *Genet Epidemiol* 26, 132-41.

23. ***Yucesoy B, Johnson VJ, Kashon ML, Fluharty K, Vallyathan V, Luster MI:*** (2005) Lack of association between antioxidant gene polymorphisms and susceptibility to progressive massive fibrosis in coal miners. *Thorax* 60, 492-5.

24. ***Covault J, Abreu C, Kranzler H, Oncken C:*** (2003). Quantitative real-time PCR for gene dosage determinations in microdeletion genotypes. *Biotechniques* 35, 594-6, 598.

25. ***Robledo R, Beggs W, Bender P:*** (2003). A simple and cost-effective method for rapid genotyping of insertion/deletion polymorphisms. *Genomics* 82, 580-2.

26. ***Farrand K, Jovanovic L, Delahunt B, McIver B, Hay ID, Eberhardt NL, Grebe SK:*** (2002). Loss of heterozygosity studies revisited: prior quantification of the amplifiable DNA content of archival samples improves efficiency and reliability. *J Mol Diagn* 4, 150-8.

27. ***Jaremko M, Justenhoven C, Abraham BK, Schroth W, Fritz P, Brod S, Vollmert C, Illig T, Brauch H:*** (2005). MALDI-TOF MS and Taqman® assisted SNP genotyping of DNA isolated from formalin-fixed and paraffin-embedded tissues (*FFPET*). *Hum Mutat* 25, 232-8.

GENETIC VARIANCE DETECTION TECHNOLOGIES FOR PHARMACOGENOMICS

Karl Heinz Hecker, Ph.D.
Editor

DNA Press™

GENETIC VARIANCE DETECTION-TECHNOLOGIES FOR PHARMACOGENOMICS

ISBN 0-9748765-5-0

Printed in the United States of America on acid-free paper.
Library of Congress Cataloging-in-Publication Data

Genetic variance detection : technologies for pharmacogenomics / Karl H. Hecker, editor.— 1st ed.
p. ; cm.
Includes bibliographical references and index.
ISBN 0-9748765-5-0
1. Molecular genetics—Laboratory manuals. 2. Pharmacogenomics.
[DNLM: 1. Pharmacogenetics—methods. 2. Sequence Analysis, DNA—methods. 3. Chromatography—methods. 4. Oligonucleotide Array Sequence Analysis—methods. 5. Spectrum Analysis, Mass—methods. QU 450 G328 2005] I. Hecker, Karl Heinz.
QH440.5.G45 2005
572.8—dc22

2005014469

DNA Press, LLC
P.O. BOX 572
Eagleville, PA 19408, USA
www.dnapress.com
editors@dnapress.com

Publisher: DNA Press, LLC
Art Direction: Alex Nartea
Cover art & book layout: Studio N Vision (www.studionvision.com)

Summary of chapters

SECTION I: Sequencing Genetic Variations

Chapter 1: Viral Genotyping with the TRUGENE® Platform

The OpenGene® DNA Sequencing System has been applied to a diverse range of variance detection applications including HIV-1 drug resistance testing, HLA tissue typing, Hepatitis B surface antigen and polymerase sequencing, and Hepatitis C genotyping. The detection platform has a modular, scaleable design that allows small and large batch processing. The TRUGENE® HIV-1 Genotyping Kit detects genomic mutations in protease and reverse transcriptase regions of Human Immunodeficiency Virus Type 1 that confer resistance to antiretroviral drugs. It is used as an aid in monitoring and treating HIV infection. This test was the first of its type to be cleared by the FDA for in vitro diagnostic use and the first widely accepted application of pharmacogenetics.

SECTION II: Genetic Variance Detection Based on Heteroduplex Analysis

Chapter 2: Genetic Variance Detection Using Temperature Gradient Capillary Electrophoresis (TGCE)

The temperature gradient capillary electrophoresis (TGCE) technique is a highly sensitive, fast, and cost-effective technique for high throughput mutation/SNP detection. TGCE is based on the heteroduplex analysis of a heterozygous DNA sample with two different alleles. The method usually starts with a heterozygous PCR product. The product is denatured and slowly annealed to form two homoduplexes of original strands of DNA and two heteroduplexes each with a mismatch at the mutation or SNP site. These four species of DNA molecules have different melting temperature profiles. The resultant homoduplex and heteroduplex samples are electrophoresed in capillaries filled with a sieving matrix containing an intercalating dye under a temporal temperature gradient. During electrophoresis two heteroduplexes will be partially denatured, produce a bulky structure at the mismatched site, and thus exhibit a retarded mobility in the matrix. This process results in separation of heteroduplexes from homoduplexes. When these separated species pass through the laser detection window, a CCD camera records the signal. A heterozygous sample will migrate as more than one peak, while its homozygous control will migrate as a single peak. The peak patterns generated by the mutant and control samples are then compared and scored automatically by software. TGCE eliminates the need for calibration runs to determine the optimum temperature for any particular sample, and TGCE with a high throughput capillary electrophoresis system allows simultaneous screening of a large number of samples with different Tm's in a single run. The throughput capability can be further increased by multiplexing different sizes of amplicons or by pooling amplicons amplified from many individual samples from the same DNA

locus in a single capillary. This chapter will focus on discussing the TGCE technology and its related applications implemented on SpectruMedix's instrument platforms.

Chapter 3: The Use of Chemical Cleavage of Mismatch (CCM) for Mutation Detection

The chemical cleavage of mismatch (CCM) is a method for the detection of point mutations in heteroduplex DNA, which is based on the piperidine-mediated cleavage of chemically modified, mismatched bases. First described by Cotton and colleagues in 1988, CCM was adapted from the Maxam-Gilbert sequencing principles that employ a series of reagents to selectively modify nucleotide bases. CCM accurately detects and localizes mismatches in heteroduplex DNA fragments up to 2,000 base pairs in length, allowing the rapid screening of mutations in large genomic segments. Several modifications have been described since the initial development of the method, promoting its widespread use for the detection of sequence variants.

Chapter 4: Genetic Variance Detection Using Surveyor Nuclease

Discovery and mapping of mutations and polymorphisms with members of the CEL I family of plant DNA endonucleases is simple, sensitive, and flexible. These enzymes efficiently cleave DNA with high specificity at the 3' end of any mismatch site in both DNA strands. Use of the enzymes involves four steps: (i) PCR to amplify target DNA from both mutant and wild-type reference DNA; (ii) hybridization to form heteroduplexes between mutant and wild-type reference DNA; (iii) treatment of annealed DNA with nuclease to cleave heteroduplexes; and (iv) analysis of digested DNA products using the detection/separation platform of choice. The use of members of this family of enzymes is described in various applications, including detection of mutations in labeled and unlabeled DNA, in large DNA fragments, in DNA fragments with multiple mutations, and in DNA populations where the mutant is a minor component.

Chapter 5: Experience Applying LightTyper™ Methodology to Human SNP's Relevant to Growth and Cardiovascular Risk

Single nucleotide polymorphisms (SNP's) represent a numerically large fraction of the polymorphic sites in the human genome. Research is proceeding at two main levels in molecular genetic epidemiology and pharmacogenetics. The first is to build genome-wide assays of hundreds of thousands of SNP's in parallel on very high throughput platforms, such as those developed by Illumina, Parallele, and Affymetrix. The second is to undertake focused studies based on smaller numbers of SNP's examined on many thousands of subjects in parallel. A number of 96- and 384-well microplate-based platforms exist for this purpose. In this chapter we describe our experience with the Roche LightTyper™ system (originally beta tested as the Idaho Technologies Odyssey). This system measures fluorescence change while delivering a real time thermal ramp to melt a binding oligonucleotide off an SNP-containing single strand of an asymmetric PCR product. Each scan takes a few minutes so that one LightTyper™ can examine the output plates from many PCR machines per day. There is also flexibility of available fluorescence chemistries. These features permit medium throughput analyses (many thousands of SNP calls per day per worker) at reasonably low capital and consumable costs and with rapid setup time

and turnaround of assays. The approach can, with attention to detail, also be applied to 'microhaplotypes' of adjacent SNP's, to polymorphisms in closely related gene families and other atypical genotyping situations. Brief comparisons are made with other liquid and gel based technologies.

Chapter 6: MADGE-Based Technologies for Identification of Unknown Mutations at the Population Level: MeltMADGE and EndoVII-MADGE

'MADGE' (microplate array diagonal gel electrophoresis), enables the simple setup of polyacrylamide or agarose gel electrophoresis in an open-faced 96-well microplate compatible format. The granted patent hinges on the use of a diagonal angle to obtain reasonable track length. For allele-specific PCR analysis, PCR checking, and PCR-RFLP analysis, 96-well, 192-, 384-, and 768-well versions have been used. 96-well gels can also be used for micro satellite analysis. More recently, for the identification of unknown mutations in large numbers of amplicons, we have developed two new approaches. One, meltMADGE, a physicochemical method of mutation scanning, combines MADGE with thermal ramp electrophoresis to attain an interrogation of PCR products similar with that achieved by denaturing gradient gel electrophoresis. The other, endoVII-MADGE, a biological method of mutation scanning, combines denaturing MADGE gels with single strand cleavage at heteroduplexes of fluorescently end-labeled PCR products. For each method, ten 96-well gels can be run in 1-2 hours in a 2-liter tank, giving economical mutation scanning approaches for unknown single gene mutations in large population studies. The earlier methods are briefly reviewed and the mutation scanning approaches are covered in more detail. These latter should enable the deeper sampling of rarer sequence diversity in case collections with less stringent selection criteria, in population samples and in large trial designs.

SECTION III: Novel Approaches for Genetic Variance Detection

Chapter 7: The Invader® Assay with Highly Multiplexed PCR for SNP Genotyping

Large-scale SNP genotyping projects require quantities of genomic DNA that may be difficult to obtain from standard biological samples. One approach to addressing this issue relies on PCR-based target amplification. In principle, PCR amplification can be carried out in a multiplex format in which multiple loci are amplified in the same tube. In practice, however, this approach can result in highly variable yields of individual amplified products due to PCR bias. We investigated the application of multiplexed PCR as a target pre-amplification step for use in conjunction with the Invader® assay for SNP genotyping. Using the quantitative Invader assay, we studied the kinetics of multiplexed PCR amplification and optimized conditions to minimize bias. Using this approach, we designed 192 pairs of PCR primers for 192 randomly selected SNP's and executed a 192-plex PCR using eight genomic DNA samples. The amplified samples were then genotyped using 192 biplex Invader assays. Of the 192 assays, 161 assays produced successful results. Of the 31 assays that demonstrated poor performance, only 6 assays failed due to the 192-plex PCR conditions. The remaining 25 assays apparently failed due to factors unrelated to the multiplexed PCR amplification, such as high background in the Invader assay due to presence of repeat sequences for some of the 192 SNP's. Analysis

of target amplification carried out using these optimized conditions suggests that it should be possible to amplify 1,000 targets in a single PCR in quantities sufficient for genotyping using a signal amplification method, such as the Invader® assay.

Chapter 8: Highly Multiplexed Approaches to Comprehensive Discovery and Scoring of Genetic Variation

Genetic biomarkers are polymorphisms that are associated with disease susceptibility or individual variation in drug response. The identification and scoring (genotyping) of a vast number of DNA polymorphisms in large populations is of high importance for disease gene identification, pharmacogenetics, and population-based studies. Among the genetic variation being analyzed, single nucleotide polymorphisms seem to be the most powerful in large-scale genetic analysis. This manuscript describe two technologies: one for large-scale variation discovery using a highly multiplexed (>1,000) in vivo mismatch repair detection (MRD) assay and the other for large-scale polymorphism scoring using a very highly multiplexed (>10,000) molecular inversion probe (MIP) assay.

Chapter 9: Multiplex Genotyping of 384 to 1536 SNP Loci on Universal Arrays

We describe a flexible, accurate, and high-throughput SNP genotyping system for large-scale genetic analysis. It includes a miniaturized bead-based platform, a high-resolution confocal scanner (BeadArray Reader), and a highly multiplexed genotyping assay (GoldenGate® assay). Genotyping assay oligos corresponding to 384 - 1,536 specific SNP sequences are each linked to a unique sequence (address) that can hybridize to its complementary strand on universal arrays. The arrays are made of beads randomly assembled into microwells etched into the ends of either optical fiber bundles or the surface of silicon slides. The optical fiber bundles are assembled into an 8 x 12 array matrix (Sentrix® Array Matrix) that matches a 96-well microtiter plate. The arrays on the silicon slides (Sentrix® BeadChip) are multi-channel pipette compatible for loading multiple samples onto a single silicon slide. These formats allow simultaneous processing of 16 - 96 samples, which enables investigators to generate approximately 300,000 genotypes per day with minimal equipment requirements and greater than 1.6 million genotypes per day in a robotics-assisted process.

Chapter 10: Whole Genome Genotyping on BeadChips with Illumina's Infinium™ Assay

We have developed an array-based whole genome genotyping (WGG) assay (Infinium™) using our BeadChip platform that effectively enables unlimited multiplexing from a single sample preparation. Genomic loci of interest are captured by specific hybridization of picomolar concentrations of whole genome amplified gDNA to 50-mer probe arrays. After target capture, SNP's are genotyped on the array by allele-specific primer extension in the presence of biotin-labeled nucleotides. The resultant signal is amplified during staining and the array is read out on a high-resolution confocal scanner. We have employed our high-density BeadChips supporting up to 288,000 bead types to create an array that can query over 100,000 SNP's using the Infinium™ assay. In addition, we have developed an automated BeadChip processing platform using Tecan's GenePaint™ slide processing system. Hybridization, washing, array-based primer exten-

sion, and staining are performed directly in Tecan's capillary gap Te-Flow Through Chambers. This automation process greatly increases assay robustness, throughput, and enables LIMS control of sample tracking.

Chapter 11: Application of Taqman® Chemistry for Allelic Discrimination

Taqman® technology is a high-throughput method for genotyping single nucleotide polymorphisms and other forms of genetic variation including insertion/deletion mutation. This technique is a homogeneous solution hybridization assay using fluorescence resonance energy transfer. Allelic discrimination is achieved by the use of two allele specific dual-labeled fluorescent probes that differ by a single base at the polymorphism site. Probes are constructed so that they are a perfect match to one allele and contain a single mismatch to the variant allele. A perfect match probe is cleaved by 5'_3' exonuclease activity of *Taq* DNA polymerase liberating the fluorophore from its quencher whereas a mismatch destabilizes hybridization resulting in probe displacement without cleavage and quenching of the fluorescence. The detection of two unique wavelengths allows for determination of genotypes with homozygous samples displaying fluorescence at one wavelength and heterozygous samples at both wavelengths. The purpose of this chapter is to provide the reader with a conceptual overview of the methodology and application of Taqman® genotyping. The first section will provide essential details of the techniques utilized for assay design, performance, detection and genotyping discrimination. Guidelines for the development of a quality control/quality assurance program are also provided. The final section illustrates the diversity of application for Taqman® genotyping and highlights several advantages over more traditional genotyping approaches such as restriction fragment length polymorphism analysis.

Chapter 12: Discovery and Genotyping of Single Nucleotide Polymorphisms using Mass Spectrometry

Technologies for the discovery and analysis of SNP's have become important tools to foster our understanding of genomics. One of those tools, the MassARRAY system, based on matrix-assisted laser desorption/ionization (MALDI) time-of-flight (TOF) mass spectrometry (MS) and its use for the analysis of SNP's and beyond is presented here. The MassARRAY system is an integrated multi-application platform for genetic analysis. The tools for SNP genotyping are based on primer-extension biochemistry that generates allele-specific extension products that are automatically analyzed in real-time on a MALDI-TOF mass spectrometer. In addition to SNP genotyping this tool can be applied to the analysis of DNA pools and due to the intrinsic sensitivity and quantitative nature of the detection can also be expanded to gene expression analysis. A different biochemistry based on the generation of base-specific cleavage products is used to screen and analyze SNP's in amplicons up to 1000 base pairs. Beyond the discovery of SNP's this tool can be applied to typing of organisms such as bacteria and the same tool, when used on bisulfite treated DNA, is employed for the analysis of DNA methylation, for example in promoter regions. These ranges of applications makes the MassARRAY platform a versatile and flexible tool to follow-up and validate hits generated through whole-genome screening approaches and to functionally characterize the targets in an integrated and comprehensive approach.