The National Cancer Institute's Core Genotyping Facility (CGF): Approaching a gold standard in genotyping assay validation

Robert Welch†, Meredith Yeager, Brian Staats, Michael Beerman, Bernice Packer, Amy Hutchinson, Andrew Bergen, Tabassum Bandey, Salma Chowdhury, Andrew Crenshaw, Sunita Yadavalli, Hugues Sicotte, Edward Miller, Maureen Kiley, and Stephen J. Chanock

Core Genotyping Facility, Advanced Technology Center, National Cancer Institute, 8717 Grovemont Circle, Gaithersburg, MD USA 20892-4605, Intramural Research Support program, SAIC Frederick, NCI-FCRDC, Frederick, MD

Section on Genomic Variation, Pediatric Oncology Branch, NCI, Bethesda, MD. †welchr@mail.nih.gov

Introduction:

The National Cancer Institute's Core Genotyping Facility (CGF) is a high-throughput genotyping laboratory providing genotyping for NCI investigators, primarily of candidate gene single nucleotide polymorphisms (SNPs) nominated for association studies in cancer etiology. SNP genotyping methodologies are highly dependent on the sequence surrounding the SNP of interest and public databases describing sequence variation and sequence context do not always include allele frequencies of the SNP of interest or surrounding polymorphisms. To improve SNP genotype assay performance, the CGF has adopted a defined pipeline for assay validation and subsequent genotyping, including:

1) Annotation -- complete annotation of the SNP of interest and 600 base pairs of surrounding genomic sequence ("mini-contig")

2) Sequencing -- each mini-contig is sequenced in the SNP500Cancer panel (N=102 individual DNA samples), to determine allele frequencies of the SNP of interest and the presence and frequency of surrounding SNPs

3) Genotyping assay development and optimization -- assays are designed and optimized on the SNP500Cancer panel

4) Genotyping assay validation -- For each optimized genotyping assay, the genotyping assay results are compared to the sequencing results per SNP500Cancer individual. There must be 100% concordance between genotypes determined from genotyping assays and from sequencing analysis for a genotyping assay to be considered "validated"

This pipeline has improved the throughput and accuracy of genotyping at the CGF, and has resulted in the discovery and annotation of genetic variants that may have degraded genotyping assay performance if left un-identified.

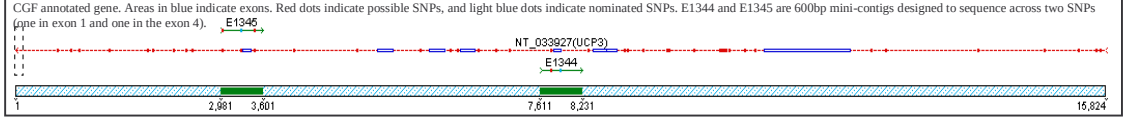
All validated genotyping assays, mini-contig sequencing primers and conditions, and SNP sequence and frequency information, is publicly available at http://snp500cancer.nci.nih.gov

1. Annotation (Front End Bioinformatics)

Genes or sequence variants within these genes are nominated primarily by investigators. Genomic sequence is downloaded from public databases (Genbank, dbSNP, HGVBASE, etc…) and annotated, i.e. the gene structure (exon/intron boundaries, open reading frames, and SNPs) is determined. Candidate SNPs chosen for analysis are derived from public databases and from the literature. For each candidate SNP, a 600bp mini-contig is prepared for sequencing, and contains annotated sequence of approximately 300bp on either side of the nominated SNP. This mini-contig is then used to design PCR primers for PCR-based sequencing. Thus, the sequence information used to design PCR primers contains information from a variety of public sources, which can reduce potential problems with the PCR sequencing. A Sequencher™ (Gene Codes, MI) overview of a completely annotated gene with two mini-contigs is shown in Figure 1 below.

Figure 1.

CGF annotated gene. Areas in blue indicate exons. Red dots indicate possible SNPs, and light blue dots indicate nominated SNPs. E1344 and E1345 are 600bp mini-contigs designed to sequence across two SNPs (one in exon 1 and one in the exon 4).



2. Sequencing determination and analysis

Primer3 (http://www.genome.wi.mit.edu/genome_software/other/primer3.html) is used to design PCR primers for the fully annotated "mini-contig" to amplify a 400-500 base pair fragment flanking the SNP of interest. 5' M13 forward and reverse universal sequencing tags are added to the corresponding primer for high-throughput sequencing. PCR conditions for each primer pair are optimized using a thermal gradient matrix of differing salt concentrations. After primer pairs are optimized, PCR-based fluorescent DNA sequencing is performed on the SNP500Cancer panel. The SNP500Cancer panel consists of N=102 DNA samples from cell lines derived from individuals with self-reported ethnicity and is publicly available from Coriell Cell Repositories (http://locus.umdnj.edu/). The distribution of self-reported ethnicity is: African/African American, N=24; Caucasian, N=31; Hispanic, N=23; Pacific Rim, N=24.

Figure 2. ABI 3700's

7 ABI 3700's perform all sequencing. Each sequencer can generate ~1536 sequences a day covering ~50 contigs a week and ~2600 contigs a year.

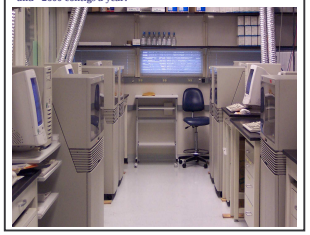
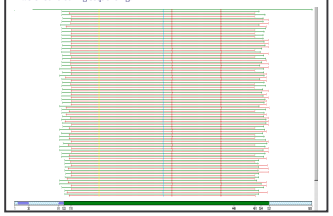


Figure 3. Mini-Contig Assembly

Shows all a sub set of the 102 individuals assembled with the "mini-contig" on the top. SNPs are indicated in the sequence by colored dots. Blue SNPs are those that were nominated, red SNPs are those which were annotated on the "mini-contig" but not nominated, and yellow SNPs are those which were found during sequencing.

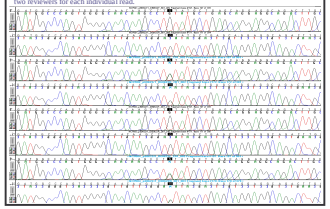


Sequence analysis utilizes Big Dye v3 sequencing chemistry on the ABI 3700 (Applied Biosystems, Foster City CA) (Figure 2). Sequences are aligned using Sequencher™ (Gene Codes, MI). 178 of 204 sequence reads (88%) must align on the first pass for alignment criteria (Minimum Match Percentage is 85 and the Minimum Overlap is 20). If the alignment is less than 88%, the sequencing is rejected and reasons for its rejection investigated, which sometimes leads to the redesign of the PCR primers. If the alignment is greater than 88%, all previously annotated SNPs are genotyped for all 102 individuals with legible sequence by two independent reviewers. A third reviewer is available to examine differences in the two reviewers' genotype calls if these are not completely concordant. A full contig assembly is shown in Figure 3, and some of its corresponding sequence chromatograms in Figure 4. Unannotated variants identified during sequencing analysis are genotyped in the same manner.

Figure 4.

Sequencing Chromatograms

Shows the individual chromatograms that are used to determine bases calls (call genotypes) by the two reviewers for each individual read.



Post-Sequencing Analysis:

Upon completion of sequence determination and analysis, all sequence variants are evaluated for obvious problems (e.g. too many genotypes), for Hardy-Weinberg Equilibrium (HWE), SNP frequency by ethnic grouping, and type of mutation (synonymous versus non-synonymous, etc.). SNP location and coverage in the gene and proximity to other SNPs are evaluated prior to selection of SNP for genotype assay design. SNPs determined to be suitable for assay design are then submitted to the appropriate genotyping platform design pipeline.

3. Genotyping assay development and optimization

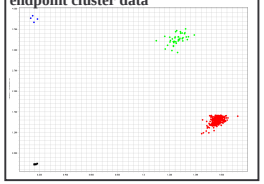
Genotyping assays are designed using the sequence that has been determined in the CGF sequencing pipeline. Any SNPs that have a minor allele frequency > 0.0 are indicated on the sequence that is used to design a genotyping assay. The CGF utilizes three major genotyping platforms for high-throughput genotyping. These are TaqMan™ – 5' exonuclease, MGB Eclipse™ – 3' fluorescent hybridization, and Sequenom™ MassARRAY – primer extension detected by matrix-assisted laser desorption ionization time of flight (MALDI-TOF). While the design and optimization protocol for each platform varies, the validation requirement remains 100% concordance between genotypes derived from specific genotyping assays and resequencing.

Example: TaqMan™ Design and Optimization

DNA sequence is submitted to Applied Biosystems' "Assay by Design" service. Assays that are successfully designed by Applied Biosystems are optimized by genotyping in the SNP500Cancer panel (N=102 DNA samples) in duplicate. Genotype calls are made using endpoint reads on an ABI 7900HT (Applied Biosystems, CA). The CGF has two ABI 7900HTs as shown in Figure 5. A technician manually assigns genotype calls by cluster analysis using Sequence Detection Systems 2.0 software (Applied Biosystems, CA). An example of the type of cluster analysis is shown in Figure 5. Successful optimization is achieved by modifying the annealing temperature of the 5' nuclease reaction until clear separation of the clusters is evident and concordance exists between the duplicated samples. Sometimes it is also necessary to modify the amounts of primer and probe present in a reaction in order to successfully optimize an assay. Once a genotyping assay is optimized, genotype calls are exported to an Oracle database and compared to sequencing calls using Perl scripts.

Figure 5. ABI 7900HT's

Figure 6. Optimized TaqMan endpoint cluster data



4. Genotyping assay validation

SNP genotypes determined by optimized genotyping assays are compared to sequence data. In order for any SNP genotyping assay to be validated and used for high-throughput genotyping, there must be 100% concordance between the assay results and sequencing-based analysis for that SNP. Where genotypes derived from sequencing and one or more genotyping assays are discordant, the sequence traces are evaluated for possible genotype calling errors by a careful visual evaluation. Discordances can be due to poor sequence reads, but there also may be cases where a genotyping assay is not correctly making genotype calls due to unannotated SNPs affecting a sequencing or genotyping primer or probe's performance. In these cases, the genotyping assay appeared to have been performing correctly (i.e., optimized), but proved to be miss-calling alleles when compared to another method. Examples of genotyping assays and sequencing assays which can miss-call alleles are shown in Figure 7.

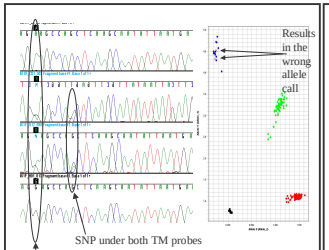


Figure 7A. Example of non described SNP under TM probe resulting in wrong TM genotype call, and correct sequencing call.

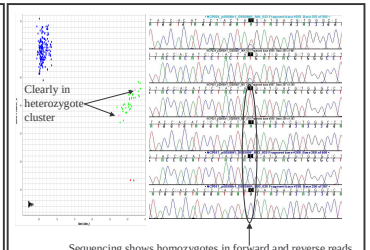


Figure 7B. Example of allele specific sequencing PCR resulting in wrong sequencing genotype call, and correct TM call.

Summary

Genotyping laboratories responding to genotyping requests for particular SNPs will need to develop a workflow or pipeline that evaluates sequence variation surrounding the SNP of interest in silico and in DNA samples derived from populations similar to those populations being genotyped. The CGF's investigator-nominated candidate gene SNP molecular genetic analyses are enabled by a comprehensive bioinformatics analysis unit that performs sequence assembly and SNP annotation, a sequencing validation workflow, and analysis of the resulting sequence and sequence variation in a group of publicly available DNA samples with self-reported ethnicity (N=102, http://snp500cancer.nci.nih.gov). SNPs are verified, and sequence flanking the SNP of interest is determined by sequencing both strands (400-500 base pairs). Genotyping assays using one or more genotyping methods are developed for sequence-verified SNPs with appreciable minor allele frequency (> 5%). The SNP500Cancer panel DNAs are then genotyped to validate both the sequencing and genotyping platform assays for particular SNPs.

Sequencing has long been considered to be the gold-standard for identifying and characterizing sequence variation. Without using multiple approaches to SNP validation (e.g. sequencing and TaqMan™), sequence variants flanking the SNP of interest would have gone undetected. Since these undetected sequence variants can effectively change the resulting genotype call for a particular molecular genetic analysis platform, genotyping datasets in disease association studies without validated assays can lead to loss of statistical power. While these problems can go undetected in large datasets due to the low frequency of individuals affected by unknown sequence variation, if large datasets are subsequently dissected into sub-phenotypes, the percentage of individual genotypes affected may increase stochastically, creating a greater potential for false positive or false negative association results.