Application Note

Development of an Individual Identity SNP panel and workflow for paternity and forensic applications

Andreas Tillmar[1], Ida Grandell[1], Erik Soderback[2], Ben Turner[2], Cecilie Bøysen[2], Raed Samara[3], Matthew Fosbrink[3], John DiCarlo[3], Eric Lader[3] and Keith Elliott[4].

[1] Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Artillerigatan 12, SE-58758 Linköping, Sweden

[2] QIAGEN Aarhus, Silkeborgvej 2, Prismet, 8000 Aarhus, Denmark

[3] QIAGEN Sciences, Inc. Biological Research Content. 6951 Executive Way, Frederick, 21703, USA

[4] QIAGEN Ltd, Skelton House, Lloyd Street North, Manchester, M15 6SH, UK

Introduction

Next-generation sequencing (NGS), or Massively Parallel Sequencing (MPS), offer significant benefits to forensic genetics and human identity over the established Capillary Electrophoresis (CE) workflow. These benefits include the ability to analyze a much larger number of markers in a single test, thereby increasing discrimination power and the statistical significance of the results obtained. This requirement for increased discrimination is increasingly demanded in complex paternity cases, for example, in cases involving siblings and other closely related individuals, where the standard short tandem repeat (STR) test cannot always provide sufficient statistical values to identify the true relationships with adequate confidence.

In addition, in moving away from the size- and fluorescence-based analysis used in CE, constraints on amplicon length (previously required to view large numbers of markers in the read-window of CE systems) are removed. This, in turn, enables the design of amplicons of much shorter size, making these markers more amenable to forensic testing from challenged or degraded samples.

The National Board of Forensic Medicine (RMV) in Sweden analyzes over fifteen thousand samples a year, including samples for paternity testing, other kinship testing, as well as for missing person identification. To utilize the benefits of NGS, and to address the need for increased discrimination in complex paternity cases and other kinship cases, RMV defined a comprehensive identity SNP panel, encompassing the well-characterized SNPs in the II SNP set defined by Pakstis et al (1) ▷

and the SNPforID set defined by Sanchez et al (2). Using the genomic coordinates for these SNPs, QIAGEN® developed a GeneRead® DNAseq Targeted V2 Panel and provided this to RMV for evaluation. In addition, QIAGEN also provided GeneRead library prep reagents to simplify the front-end pre-analytics and the Biomedical Genomics Workbench analysis software to enable the development of an automated bioinformatics pipeline for the analysis of results. Together, this SNP enrichment panel, along with the library prep kits and software, enabled the development of a streamlined and focused workflow for the generation and analysis of SNP data in complex paternity cases.

Methods

Panel design

The 88 bi-allelic Individual Identity (II) SNPs from the set characterized and published by Pakstis et al were combined with the 52 SNPs characterized and published by Sanchez et al in a 140 SNP panel. Rs numbers for these 140 SNPs were provided by RMV to QIAGEN's Biological Research Content facility, where primer design was conducted and the panel was developed and produced. The final panel design was made up of 280 amplicons to cover every SNP with two amplicons, each being 150 bp long on average.

Samples

Blood samples from 49 unrelated Swedish individuals were selected for analysis. DNA was extracted using a urea extraction method previously described (3). Ten FTA® samples (buccal cells) were selected from routine casework. The FTA samples were analyzed using direct amplification from a 1.2 mm^2 punch after three washing steps with water. For three of the FTA samples, blood samples from the same three individuals were used for comparison. A dilution series of the 2800M Control DNA (Promega) was prepared with the following concentrations 2.5 ng/µl, 0.625 ng/µl, 0.2 ng/µl and 0.025 ng/µl, of which 8 µl was used as template DNA in the analyses.

Target enrichment

Purified DNA underwent target enrichment using the GeneRead DNAseq Targeted V2 Panel and the GeneRead DNAseq PCR Kit V2 (comprising oligonucleotides, enzymes and buffers) according to manufacturer's instructions. For sensitivity studies, two replicates each of 0.2 ng, 1 ng, 5 ng and 20 ng input DNA were used for enrichment of each sample.

Library prep

Library preparation was then conducted on the amplified DNA using QIAGEN's GeneRead library prep workflow designed for sequencing on Illumina® instruments: GeneRead DNA Library I Core Kit, GeneRead DNA I Amp Kit, GeneRead Adapter I Set 12-plex, followed by size selection using the GeneRead Size Selection Kit and QIAquick® PCR. All steps were conducted according to the manufacturer's instructions.

Library quantification and sequencing

The quantification of the libraries was performed using the Qubit® dsDNA BR or HS Assay (Thermo Fisher Scientific) with a Qubit 2.0 Fluorometer (Invitrogen). Agilent high-sensitivity DNA kit was used with the Agilent® 2100 Bioanalyzer to check the average size of the libraries. After quality measurements, libraries were diluted to 4 nM and pooled together. Additional dilution to 10 pM of the pool was performed prior to loading on a MiSeq® Reagent Kit v3 and then sequencing on a MiSeq (Illumina) instrument, according to a standard procedure.

Analysis

Analysis of the raw data was achieved using the Biomedical Genomics Workbench software. This software enables the automation of NGS data analysis from FASTQ file to genotype and facilitates rapid, easy and flexible generation of a sequencing analysis pipeline. Using the Biomedical Genomics Workbench an analysis pipeline was constructed to reproducibly automate the analysis of SNP data from the samples in the study.

SNP genotyping was completely automated. Briefly, a workflow was constructed that maps all reads to hg19 as the reference genome, then genotypes each of the 140 SNP loci using the "Identify Known Mutations from Sample Mappings" tool. The results are presented in a genome browser view for each sample for analyst review.

The minimum coverage for genotype calling was set to 200X. Heterozygote balance was analyzed based on allele read frequency (ARF) values. ARF was defined, for each marker, as the number of reads for a reference allele divided by the total number of reads for the specific marker. ARF for a homozygous genotype should, in theory, either be 0 or 1 and 0.5 for a heterozygous genotype. Unless otherwise stated, ARF between 0 and 0.1 and between 0.9 and 1 were used as thresholds for inclusion of a homozygous genotype and ARF between 0.4 and 0.6 was used as threshold for inclusion of a heterozygous genotype.

Results were evaluated for uniformity of coverage, heterozygous balance, sensitivity, concordance and reproducibility.

Results

Uniformity of coverage

Uniformity of coverage was evaluated for all 140 SNPs. Results are shown in Figure 1. ▷

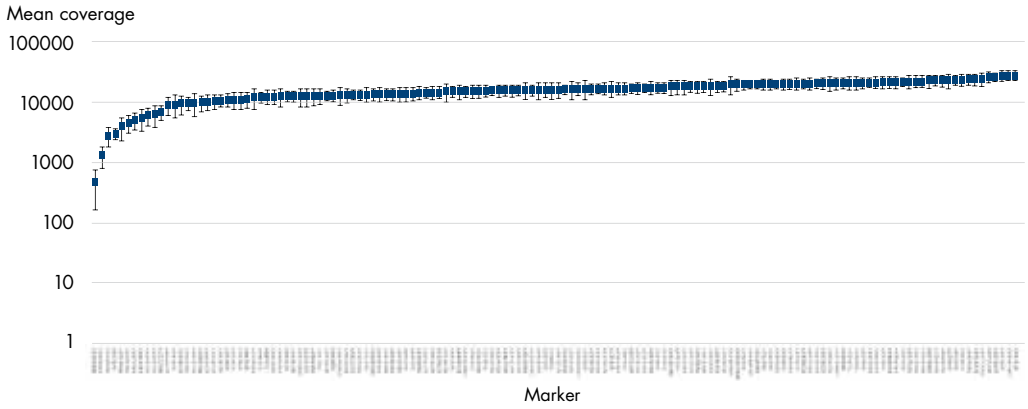

Figure 1. Uniformity of coverage, for the 49 blood samples, for the 140 SNPs in the target enrichment panel sequenced on Illumina MiSeq. All 140 SNPs are shown along the x-axis and a logarithmic scale of coverage on the y-axis. Only 1 SNP was below 1000X coverage (the coverage of this SNP was enough to make accurate calls).

Heterozygous balance

Results for heterozygous balance are shown in Figure 2. The three clearly defined bands of data represent: samples homozygous for the reference genotype (with scores between 1 and 0.9), heterozygous samples (scores between 0.6 and 0.4) and samples homozygous for the non-reference genotype (scores between 0 and 0.1).

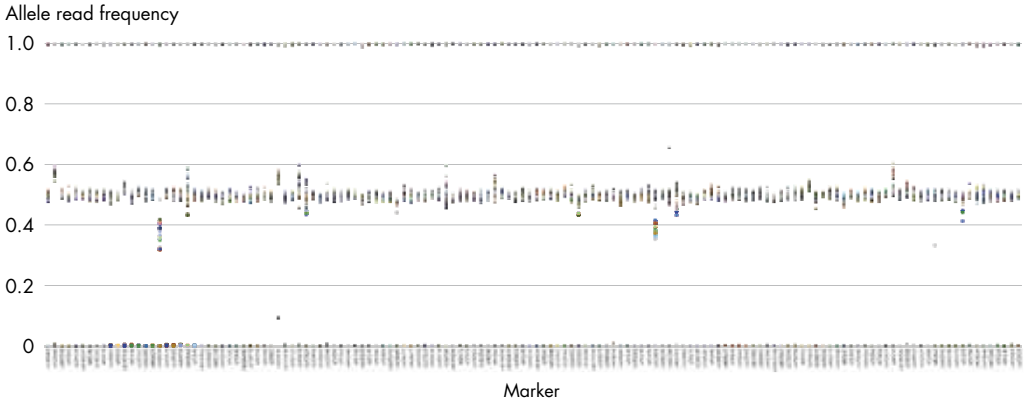

Figure 2. Heterozygous balance data (for the 49 blood samples) for all 140 SNPs. The y-axis shows the number of reads for the reference genotype divided by the total number of reads for each SNP.

Sensitivity

Sensitivity results are shown in Figure 3, where the data points represent 0.2 ng to 20 ng input DNA (analyzed in duplicate) for each SNP. At the low level of DNA input (0.2 ng) no clearly defined thresholds for heterozygous balance are discernible. The minimum amount of required DNA input using the standard protocol is therefore 1 ng. If thresholds for heterozygous balance are increased, the genotypes are, however, still correctly called.

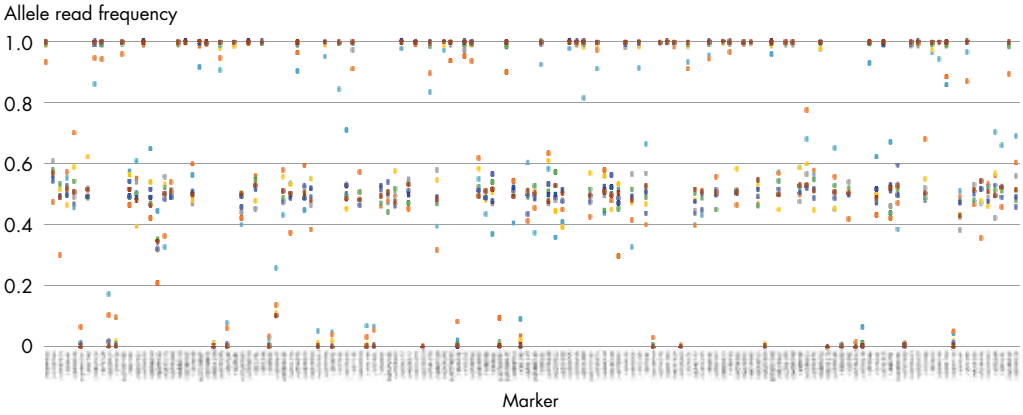

Figure 3. Sensitivity data for all 140 SNPs. Two data points for each SNP are overlaid on to the data for heterozygous balance, showing that imbalance occurs at 0.2 ng input DNA.

Concordance

Of the 140 SNPs used in the enrichment panel, 95 are also present in the Illumina ForenSeq™ DNA Signature Prep Kit enrichment panel. Results exhibited full concordance with the ForenSeq panel data (data not shown) when analyzing the 2800M Control DNA sample for both methods.

The analyses of the FTA samples displayed congruent results, as for the blood samples, both for coverage and ARF values.

Reproducibility

All replicate analyses of the 2800M Control DNA (the dilution series and 6 additional analyses (from 6 different runs) produced consistent and correct genotypes.

Five FTA samples were analyzed twice with consistent genotypes, and samples (blood and buccal cells on FTA) from three different individuals were analyzed and gave consistent genotypes.

Conclusion

We have described the development and evaluation of an Identity SNP panel and workflow for use in paternity and forensic applications. In order to be effective for routine paternity use, any such target enrichment panel and associated workflow needs to demonstrate: high uniformity of coverage to enable cost effective and reliable sequencing of all markers; well defined heterozygous balance thresholds to ensure the ability to safely call homozygous and heterozygous samples with confidence; high sensitivity to prevent problems where input DNA is limited; and, good reproducibility to enable reliable interpretation of guidelines to be set in a high-throughput environment. Furthermore concordance with other commercially available enrichment panels is essential to provide assurance that results are correct and compatible with results generated with other panels and workflows. ▷

Data presented here demonstrate extremely high uniformity of coverage, with 139 of the 140 SNPs exhibiting coverage between 1,000X and 10,000X and all producing coverage on average over 100X. One SNP with coverage sometimes below 200X has been recorded as performing poorly with other NGS enrichment panels (C. Phillips, personal communication). These results compare favorably with all enrichment panels on the market.

The results demonstrate well-defined heterozygous balance for samples down to 1 ng of input DNA, enabling clear calling thresholds to be defined for paternity samples. Such thresholds are vital for the operational use of such a panel where the genotypes of samples are unknown and the presence of contaminating DNA is a possibility.

Results down to 0.2 ng input DNA show that, while correct results are obtained down to this level, heterozygous balance is suboptimal. This is unlikely to be a problem for paternity samples where reference samples typically contain higher levels of starting DNA. However, for future adoption of this panel workflow in forensic casework, further optimization of the panel is required in order to obtain reliable, balanced results with such low input DNA.

Summary of advantages using QIAGEN's universal workflow with custom enrichment panel:

- No "fixed" marker panel. Easy to set up/design custom made marker panel. Thus the workflow offers possibilities to combine panels for different purposes in an easy fashion.
- Easy to modify the workflow (e.g., the different steps with cleaning and removal of unwanted DNA fragments).
- Robust (as shown with the tests using different levels (quant/quality) of DNA extractions). Works well even with FTA cards without any control of the DNA (no quant/no quality measurement).
- Available workflows for different sequencers (MiSeq and Ion Torrent™).

References

1. Pakstis, A. J., et al. (2010) SNPs for a universal individual identification panel. Hum Genet **127 (3)**, 315–324. doi:10.1007/s00439-009-0771-1
2. Sanchez, J. J., et al. (2006) A multiplex assay with 52 single nucleotide polymorphisms for human identification. Electrophoresis **27 (9)**, 1713–1724. doi:10.1002/elps.200500671
3. Lindblom, B., Holmlund, G. (1988) Rapid DNA purification for restriction fragment length polymorphism analysis. Gene analysis techniques **5 (5)**, 97–101

For up-to-date licensing information and product-specific disclaimers, see the respective QIAGEN kit handbook or user manual. QIAGEN kit handbooks and user manuals are available at **www.qiagen.com** or can be requested from QIAGEN Technical Services or your local distributor.

Discover more at **www.qiagen.com**.