Development of a Multiplex PCR to MiSeq-Based DNA Fingerprinting Process for Contamination and Sample Swap Detection

JONNA GRIMSBY, YOSSI FARJOUN , MAURA COSTELLO, DANIELLE PERRIN

BROAD INSTITUTE GENOMICS PLATFORM, CAMBRIDGE, MA, USA

Introduction

Despite the fact that we have access to advanced laboratory automation systems with integrated LIMs tracking, incorrect sample identification uploads, sample swaps and contamination still occur.

To address this, we have devised a new multiplex PCR to MiSeq-based DNA fingerprinting assay for checking sample identity. This assay:

- uses genomic DNA remaining from our incoming sample quantification assay to serve as input (no additional sample required)
- uses multiplexed PCR with target-specific primers (95 pairs for SNP genotyping and 1 pair for sex identification) that are 5'-tailed with "adapter" sequence
- uses a second "tailing" PCR with primers containing adapter sequence overlap, molecular indices, and flow cell attachment sequence, generating full sequence-ready constructs
- results in a 384 sample PCR product pool that can be sequenced on a single MiSeq run and analyzed.

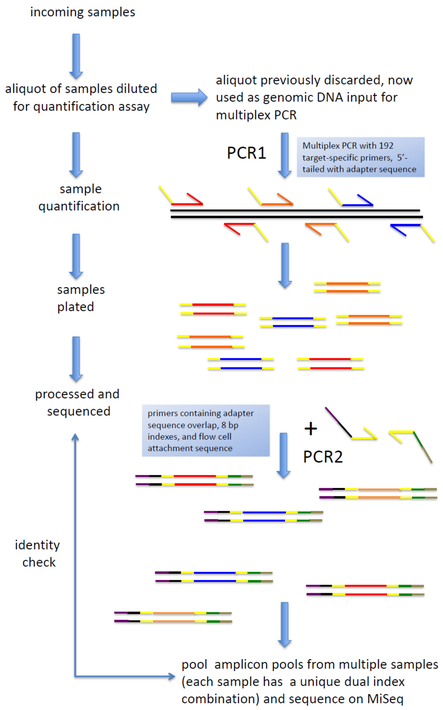

Figure 1. DNA fingerprinting process using multiplex PCR and MiSeq sequencing.

Results: Target Coverage, Sample Types and Input Amount

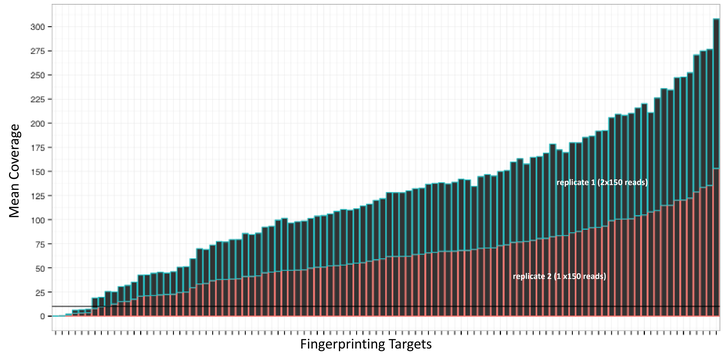

Figure 2. Mean coverage by target for replicate experiments (384 samples).

Coverage of multiplexed fingerprint targets is not even, but 10X is sufficient for calling SNPs (Fig. 2). This assay works for a variety of sample types including FFPE as well as a large range of sample inputs (Figs. 3 and 4).

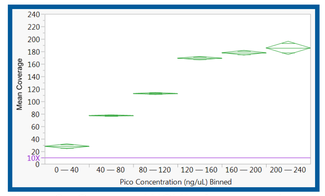

Figure 3. One way ANOVA of coverage by sample concentration (384 samples in pool; samples are diluted 1:10 before multiplex PCR).

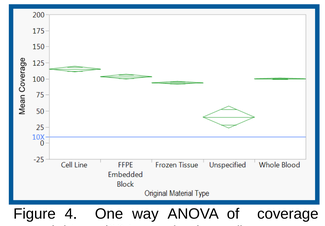

Figure 4. One way ANOVA of coverage by material type (384 samples in pool).

Sample Identity

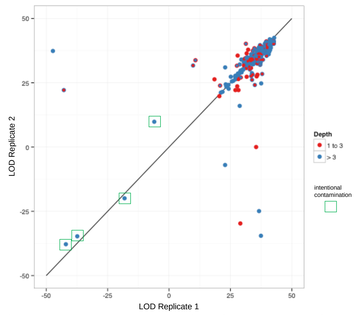

Figure 5. LOD score comparison between experiment replicates. (Larger data point = replicate 1, smaller data point = replicate 2)

LOD scores (representing matched sample identity to existing fingerprint profiles if LOD > -3) were acquired from the Picard fingerprinting tool. This plot indicates that LOD scores are generally correlated between assay replicates. Cases where high depth samples have mismatching LOD scores between replicates can be explained by contamination in one replicate. A few samples were intentionally contaminated for testing purposes, and these are highlighted (Fig. 5).

Sex

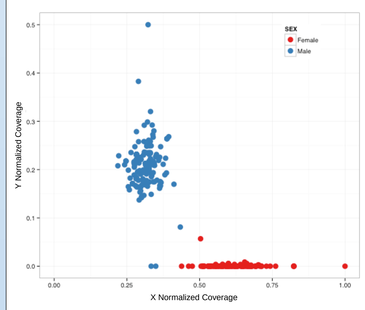

Figure 6. Inference of sex based on normalized AMELX/Y coverage (X vs. Y reads per kilobase per million mapped reads).

The assay can successfully distinguish between males and females using a single primer pair that amplifies a small region of AMELX and AMELY (Amelogenin X-linked and Y-linked). The algorithm for inferring sex from this coverage data is still in design (Fig. 6).

Contamination

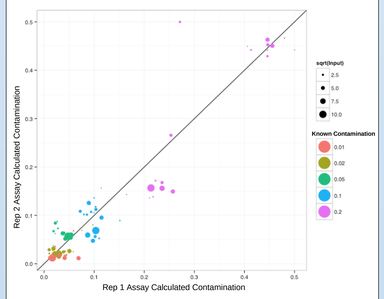

Figure 7. Contamination detection and calculation test (replicates) with varying reaction input amounts and known contamination levels.

DNA from the CEPH/Utah trio (Corielle NA12878 and her parents) was mixed with unrelated NA12877. We tested a range of inputs into PCR1 and a range of known contamination levels. In general, our assay contamination calculation is not accurate for 20% contamination. The calculation is sensitive to low input amount when both it and the known contamination level are low, in that it tends to over-estimate contamination (Fig. 7). We can successfully detect the presence of contamination, but estimation has ~50% error.

Conclusions

- This new assay provides adequate coverage of SNP targets to correctly generate a fingerprint profile and determine sex, and works for a variety of sample types at a range of concentrations.
- Contamination detection is possible using this assay, but estimation of actual contamination levels is error-prone.
- This new method will enable us to fingerprint all samples coming through the door without utilizing any additional sample material.
- We will be able to check incoming sample sex against collaborator-provided meta-data and check DNA fingerprint concordance of multiple samples from the same patient source (such as tumor/normal pairs). This alone is expected to reduce our sample swaps by 25%.
- Our ability to detect swaps or contamination in newly received samples will allow us to pull samples before sending them downstream for costly lab processing and data analysis.
- Most importantly, checks of incoming fingerprints with outgoing sequence data will ensure that correct data is associated with the correct sample.

In the future, we plan to:

- fully automate this process
- develop and incorporate a genomic DNA quality- score into this assay based on a ratio of large to small PCR fragment coverage
- use fingerprint sequencing coverage data to roughly estimate genomic DNA concentration

These tools, in combination with the DNA fingerprint, would provide an all-encompassing sample quality assessment and identity profile that can be linked to all samples as they enter our facility.

Acknowledgements

We thank Brendan Blumenstiel, Matthew DeFelice, Thomas Howd, Kristin Anderka, Maria Baco, Michelle Cipicchio, Justin Abreu, Sheli McDonough, Cole Walsh, Scott Anderson and the Broad Institute Genomics Platform.