A PUTATIVE SNP IN THE FRAGILE X GENE NEITHER AFFECTS PCR-BASED GENOTYPING NOR IS VERIFIED BY SANGER SEQUENCING: PRACTICAL IMPLICATIONS FOR GENOME DATABASE ACCURACY

Homero L Rey, Andrew G Hadd, Stephen Morales, Alison Anderson, and Gary J Latham
Asuragen, Inc., Austin, TX

SUMMARY

- A single nucleotide polymorphism (SNP), detected through next-generation sequencing (NGS), could potentially interfere with PCR-based genotyping methods within the 5′ UTR of the *FMR1* gene.
- Utilizing a cohort of samples identified within the 1000 Genomes database as containing the SNP (rs111485627), no interference with *FMR1* fragment analysis was observed.
- Using Sanger sequencing, we were unable to verify the presence of the putative SNP, making it likely that the SNP in question is actually an NGS artifact.
- Furthermore, utilizing synthetic DNA constructs engineered to contain the specific SNP, we saw no discernible impact on *FMR1* genotyping using the AmplideX® PCR/CE *FMR1* Kit (RUO).

INTRODUCTION

Availability of next-generation sequencing (NGS) data from thousands of individuals has substantially increased our understanding of human DNA variation across different populations. However, many SNPs in publicly-available databases lack stringent peer-reviewed validation and may thus reflect intrinsic errors generated by large-scale NGS studies. This is particularly evident in highly repetitive and/or highly GC-rich regions of the genome. One such SNP, rs111485627, reportedly overlaps the region targeted by the FAM-labeled reverse primer of the AmplideX PCR/CE *FMR1* Kit[1], a commonly used reagent set for *FMR1* genotyping. This study addresses primary and secondary PCR methods, Sanger sequencing confirmation and assessment of plasmids containing this SNP as they pertain to *FMR1* genotyping. The results address the need for secondary confirmation and recognition of NGS-specific errors in populating SNP databases.

MATERIALS AND METHODS

A total of 20 unique cell-line genomic DNA samples (13 female, 7 male) were obtained from the Coriell Cell Repositories. These samples were identified as containing SNP rs111485627 in either Phase I or Phase III of the 1000 Genomes Project[2,3,4,5]. DNA from these samples was analyzed using the AmplideX PCR/CE *FMR1* Kit (Asuragen) using standard primers as well as a single-base-offset primer designed to avoid the putative SNP. Samples were tested in duplicate following manufacturer instructions. The primer binding region for each cell-line DNA was independently assessed using Sanger sequencing for 6 of the samples following purification and cycle sequencing with BigDye® 3.1 terminators (Thermo Fisher). A map of the *FMR1* locus region, sequence context and primer locations are shown in Figure 1. Finally, a synthetic construct of the CGG region, designed to contain the SNP, was prepared, sequenced verified and analyzed using the AmplideX kit as shown in Figures 4 and 5.

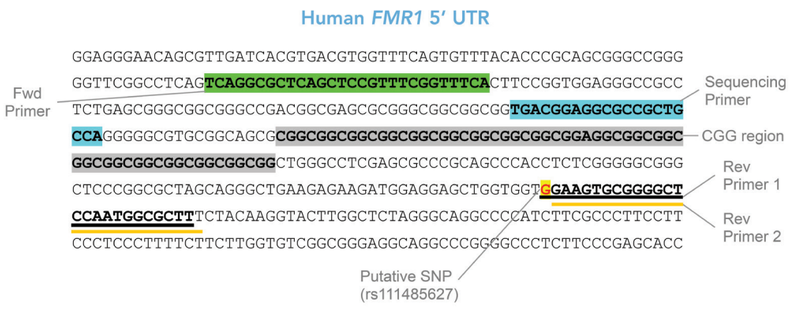

Figure 1. The putative SNP is identified within a highly GC-rich and polymorphic region of the X chromosome. The 5′ UTR of the *FMR1* gene region (X: 147911851-147912330, GRCh37/HG19) is shown with relative placements of the forward (Fwd) primer, sequencing primer and two reverse (Rev) primers in relation to the CGG repeat region and the putative SNP. Rev Primer 1 corresponds to the standard primer and Rev Primer 2 to a single-nucleotide offset. The sequencing primer is shown in the forward direction.

Conflict of Interest Disclosure
All authors have the financial relationship to disclose: Employment by Asuragen.
Research Use Only – Not For Use In Diagnostic Procedures
Presented at AMP 2016 - G57

RESULTS

A cohort of HapMap and Coriell repository DNA samples were analyzed using *FMR1* genotyping and Sanger sequencing. The cohort of 20 samples was analyzed in separate groups based on reports from 1000 Genomes Phase 1[2,3] and confirmation of samples in 1000 Genomes Phase 3[4,5] as shown in Table 1.

Phase 1*	Phase 3**		
HG01521	NA12761***	HG02252	HG01746
NA19923	HG00127	HG00345	HG00308
NA20534	HG01768	HG00360	HG01507
NA18626	NA11832	HG01777	HG02090
NA18978	HG00324	NA20514	NA20819

*Samples identified in 1000 Genomes Phase I (03/2012)
**Secondary cohort identified in 1000 Genomes Phase 3 (05/2013)
***NA12761 was reported to contain the SNP in both published databases

Table 1. 1000 Genomes sample cohort identified with putative SNP rs111485627. This SNP is seen at a frequency of 1.29% within the 1000 Genomes cohort and at 1.92% within the ESP cohort.

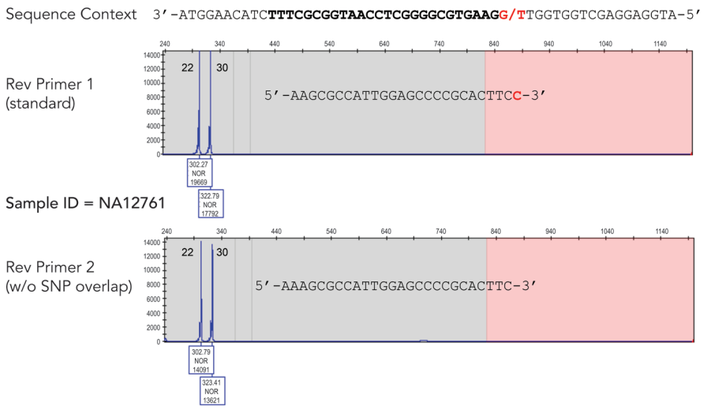

Figure 2. A sample (NA12761) with the putative SNP was equivalently genotyped with two primer sets. Primers of equivalent length overlapping the putative region (Rev Primer 1, overlapping base shown in red) or omitting the SNP (Rev Primer 2) were used for *FMR1* genotyping without impact to repeat size determination (22, 30 CGG) or PCR yield. Primer sequences are shown as insets relative to reported sequence context .

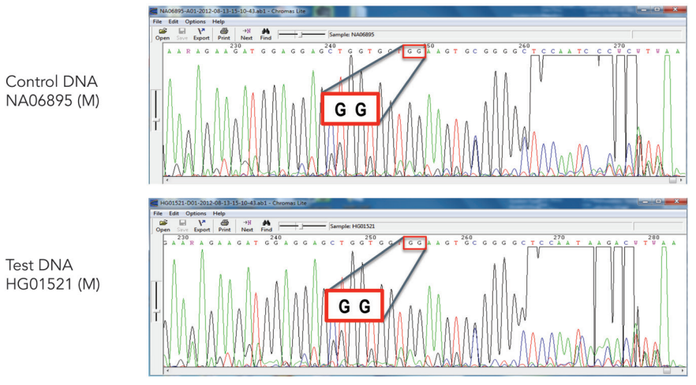

Figure 3. Sanger sequence analysis does not verify presence of putative SNP. Sequence analysis comparing two CGG size-matched cell line DNA samples shows that both contained a "—GG—" sequence within the primer binding site instead of the NGS-reported "—TG—" polymorphism.

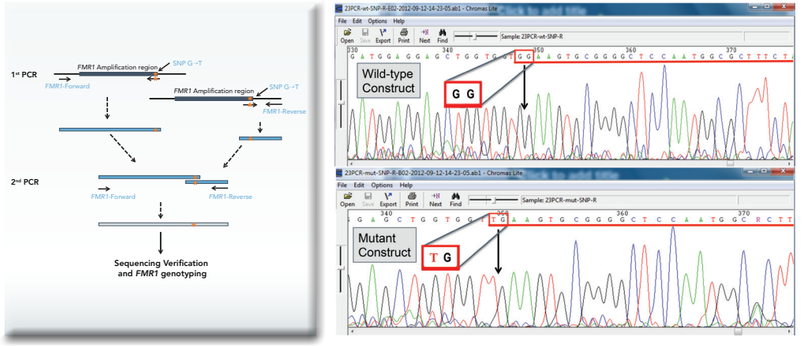

Figure 4. SNP mutant construction and SNP identity confirmation to generate controls for determining impact of the SNP. Two plasmid constructs were made via site-mediated two-step PCR to produce the GG and GT variants in the primer binding site of the 5′ UTR of the *FMR1* gene region (workflow shown in left panel). PCR products were purified, quantified and verified for identity using Sanger sequencing (electropherogram traces shown in right panel).

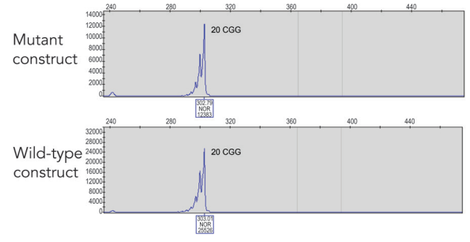

Figure 5. Equivalent genotyping observed for DNA constructs with and without a SNP. PCR/CE analysis of wild-type and mutant constructs produces similar fragment peak profiles (20 CGG), resulting in identical genotype categorization. The apparent 2-fold difference in peak intensity has no effect on the peak sizing and subsequent genotype call (Normal). Constructs were diluted to ~40,000 copies per reaction prior to testing with the AmplideX *FMR1* PCR/CE assay.

CONCLUSIONS

- The reported SNP (rs111485627) has no impact on genotyping analysis of the 5′ UTR region of the *FMR1* gene using the AmplideX *FMR1* PCR/CE Kit (RUO), and its presence cannot be verified via Sanger sequencing.
- This putative SNP may be a consequence of inherent or systematic errors due to NGS, particularly in the context of GC-rich and/or homopolymer regions wherein the technology's reliance on algorithm-based correction for error-prone sequencing are heavily challenged[6].
- These observations suggest the need for caution when interpreting high-homology genetic data derived from NGS analyses, and further support the "Gold Standard" nature of Sanger-based and fragment-based analysis for highly repetitive nucleic acid sequences.

References

1. Filipovic-Sadic S, et al. "A Novel *FMR1* PCR Method That Reproducibly Amplifies Fragile X Full Mutations in Concordance with Southern Blotting and Reliably Detects Low Abundance Expanded Alleles." Clinical chemistry 56.3 (2010): 399–408.
2. The 1000 Genomes Project Consortium. "A Map of Human Genome Variation from Population Scale Sequencing." Nature 467.7319 (2010): 1061–1073.
3. The 1000 Genomes Project Consortium. "An Integrated Map of Genetic Variation from 1,092 Human Genomes." Nature 491.7422 (2012): 56–65.
4. The 1000 Genomes Project Consortium. "A Global Reference for Human Genetic Variation." Nature 526.7571 (2015): 68–74.
5. Sudmant, Peter H. et al. "An Integrated Map of Structural Variation in 2,504 Human Genomes." Nature 526.7571 (2015): 75–81.
6. Wall, Jeffrey D. et al. "Estimating Genotype Error Rates from High-Coverage next-Generation Sequence Data." Genome Research 24.11 (2014): 1734–1739.

Asuragen®
www.asuragen.com