Accurate detection of copy-number changes in the highly conserved regions of PMS2 using NGS

Vince Ramey, Michael Kennemer, Rachel Harte, Dan Kvitek, Karen Ouyang, Steve Lincoln, Sarah McCalmon, Nila Patil, and Swaroop Aradhya

Invitae Corporation, San Francisco, CA

Disclosure statement: All authors are employees and stockholders of Invitae Corporation.

Introduction

Lynch syndrome is caused by inherited mutations in PMS2 in 4-11% of cases. However, testing for PMS2 is hampered by the presence of a pseudogene, PMS2CL, which is nearly identical to PMS2 in both intronic and exonic regions of exons 12–15. Ordinary NGS reads cannot be unambiguously aligned to PMS2 or PMS2CL, and gene conversion between the 3' end of PMS2 and PMS2CL further complicates this issue. In 2011, scientists at MRC Holland and ARUP introduced a method using MLPA followed by LR-PCR and sequencing to determine the variant location[1]. While innovative and a significant improvement over previously existing methods, this approach relies heavily on MLPA assays designed to target specific paralogous sequence variants, or PSVs, which may often be uninformative. Incorporating NGS data into this method, buttressing the MLPA targeted PSV loci with additional high confidence variant calls is shown to significantly increase accuracy.

Methods

The first step of PMS2 analysis is a bioinformatics screen to detect potential copy-number changes in the 3' end of PMS2. This is done by baiting and using Illumina short reads to sequence the full gene. Reads from PMS2CL are also included due to their near complete homology. PMS2CL is then masked from the reference genome and copy-number normal is set to 4 copies for exons 12-15 of PMS2. Copy number is called by an in-house algorithm, CNVitae. Non-normal copy-number calls are then confirmed using MLPA, with no loss of sensitivity[2], followed by LR-PCR and sequencing for location inference[3].

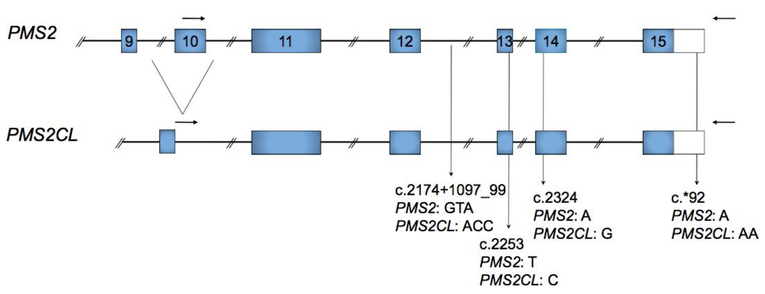

Figure 1: From Vaughn et al (2011): "Representation of homologous regions of PMS2 and PMS2CL encompassing exons 9 and 11–15. The locations of paralogous sequence variants (PSVs) under MLPA probes are noted for intron 12 and exons 13–15. Nucleotide numbering is with respect to reference sequence NM_000535.5 and arrows indicate the location of primers used for LR-PCR."[1]

MLPA PSV targets, where PMS2CL allele is described in HGVS as a variant in PMS2.
Exon 12: NC_000007.13:g.6021356_6021358delinsGGT (NM_000535.5:c.2174+1097_2174+1099delinsACC)
Exon 13: NC_000007.13:g.6018249A>G (NM_000535.5:c.2253T>C)
Exon 14: NC_000007.13:g.6017340T>C (NM_000535.5:c.2324A>G)
Exon 15: NC_000007.13:g.6012938dupT (NM_000535.5:c.*92dupA)

If the MLPA-targeted PSV loci are not informative, NGS variant calls made with high depth of coverage and allele balance ratios that correspond to the MLPA confirmed copy number are used. To be efficient, we preferentially use variants that are covered by existing sequencing primer designs. But will design and order additional sequencing primers, as necessary. This expands the loci used in location inference from 1 per exon to several hundred base pairs covered at depth by NGS in our assay. To date we have found 122 readthrough polymorphisms, where the alternate allele has been observed in more than 1 case, that could potentially be used to discern the location of a copy number alteration.

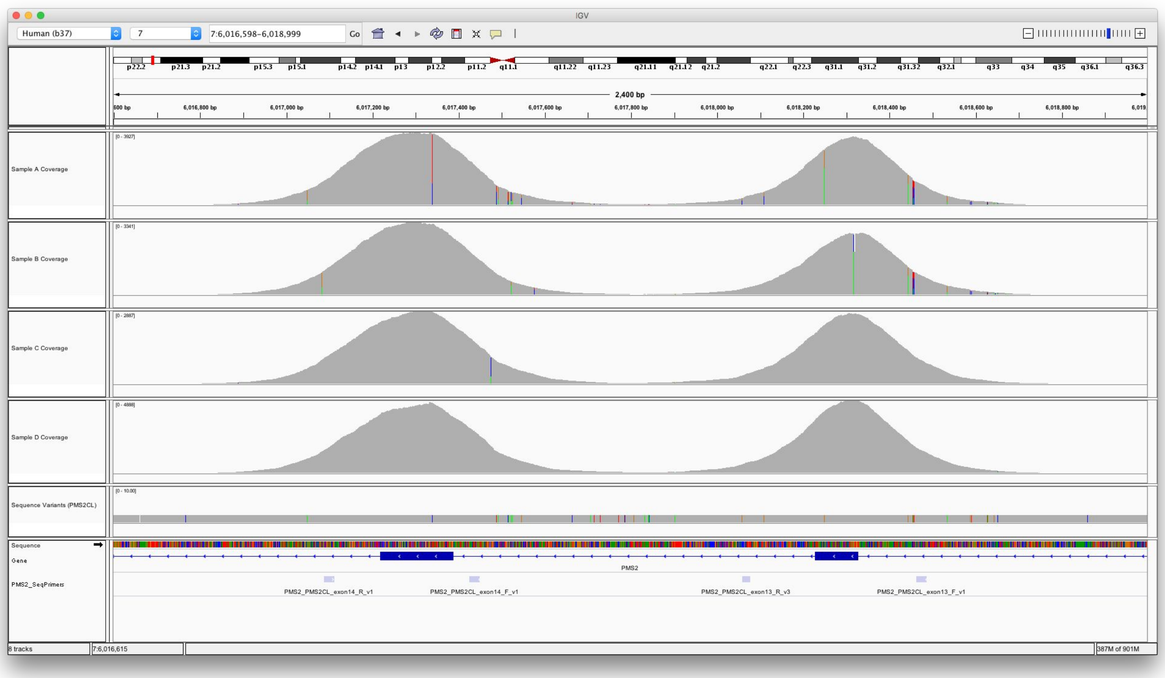

Figure 2: IGV of 4 Samples and seq. primers with putative PMS2 exon 13 and 14 Deletions, where PMS2CL sequence is masked out and all PMS2/PMS2CL-derived NGS reads align to PMS2. (Sample A) Can be disambiguated using MLPA + LR-PCR/Sanger method. Both the exon 13 and 14 PSV loci (c.2253T>C and c.2324A>G) are heterozygous. (B) Neither PSV is present. Alternate variant locus in exon 13, and between the standard sequencing primers, can be used in place of a PSV. (C) Neither PSV is present. Alternate variant locus in exon 14 can be used instead, but a custom pair of sequencing primers would be required. (D) Neither PSV is present. Could not identify an alternate variant locus in NGS data.

Results

Using our NGS read-through variant call rate at the MLPA PSV loci, the rough rate of gene conversion and the expected frequency that these loci are informative can be estimated. We have sequenced over 30,000 clinical samples on a targeted NGS panel that includes PMS2. While a majority of these tests are for hereditary cancer, an indication of Lynch syndrome in this test set is relatively uncommon. For exon 13, we have observed both PSV alleles in 82.95% of samples.

While each additional exon adds information content, the PSV loci in exons 13 and 14 are only 909bp apart. As a result, the alleles are frequently gene converted or inherited alongside each other. This is observed in our samples, as only 0.06% of cases are heterozygous on only one of the PSV loci of exon 13 or 14.

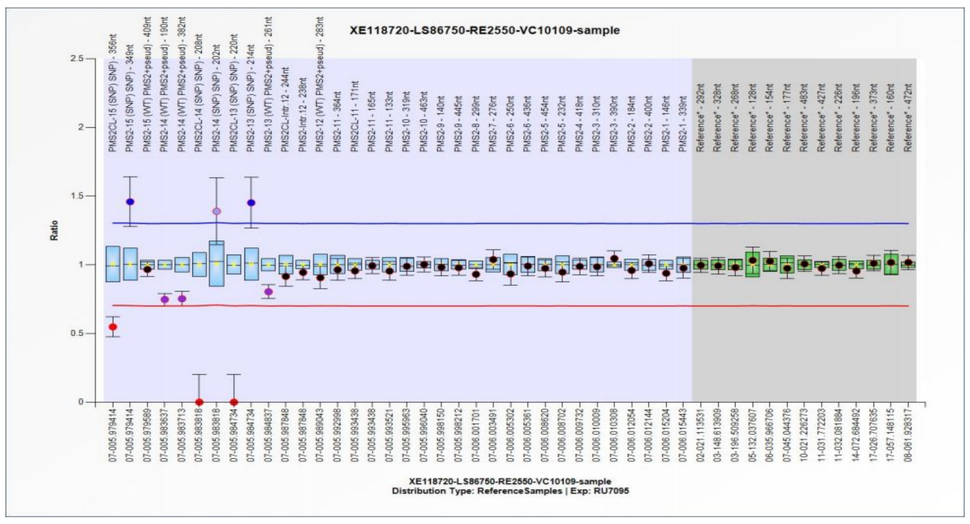

Figure 3: MLPA Coffalyser output for an exon 13 and 14 Deletion where both PSV loci are uninformative as a result of being homozygous for the PMS2 (SNP) reference allele.

Recently, we have had 100 clinical cases of a copy number variant in the 3' end of PMS2 that were confirmed by MLPA. Though likely skewed by testing families, by far the most commonly observed putative copy number change has been an exon 13-14 Deletion. One or both MLPA PSV loci were informative in 36 of 49 exon 13-14 cases, or 73%. In 12 of 49 calls, or 26%, NGS variants were necessary to determine the location of Deletion. Incorporating NGS allowed for location to be inferred in all but 1 exon 13-14 Deletion call (Fig. 2, Sample D).

While gene conversion is more rare in Exon 12, there is an additional risk that the CNV breakpoint may be between the exon and the intronic PSV target. This can also be mitigated by NGS data. Overall, we have been able to determine the location of 97.8% of Deletions.

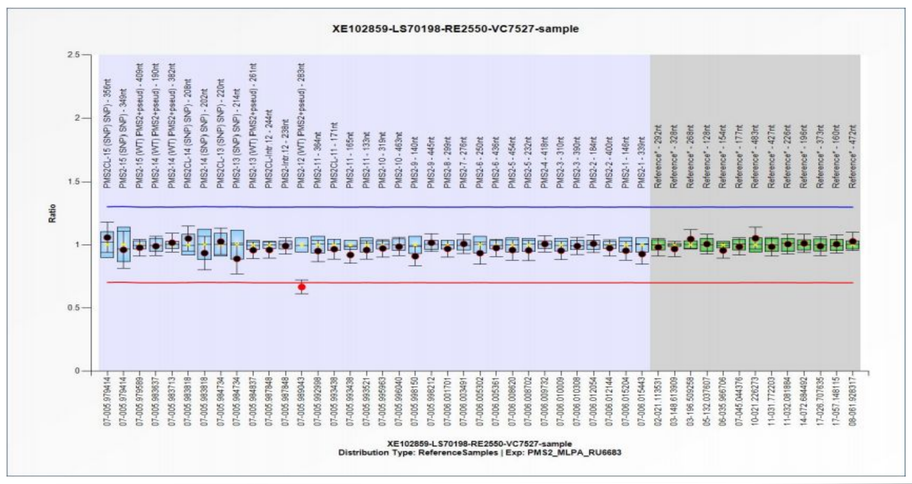

Figure 4: MLPA Coffalyser output for an exon 12 Deletion where intronic PSV locus is uninformative. The Deletion event does not extend far enough into the intron.

Conclusions

Accurate copy-number determination in clinical diagnostic testing is both critical and technically challenging. A mixture of techniques, including microarray, MLPA, ddPCR and NGS, may be necessary to efficiently call structural variants. In the specific case of PMS2/PMS2CL the innovative and recently developed inference method[1], combining MLPA with targeted sequencing, made copy-number calling in the 3' end of PMS2 possible. Our approach is an extension that incorporates NGS data into the process in a way that reduces cost while improving the ability to localize variants in Lynch syndrome cases.

References

1. Vaughn et al., "Avoidance of pseudogene interference in the detection of 3' deletions in PMS2" Human Mutation, 2011
2. Jacobs et al., "Accurate detection of small and large copy number events from targeted next generation sequence data" ASHG presentation 2013
3. http://blog.invitae.com/full-pms2-testing-at-invitae-weve-got-you-covered/