Long Read Sequencing Elucidates Complex Germline Variants in Individuals Undergoing Hereditary Gastrointestinal Cancer Testing

Sami Belhadj, PhD[1*]; Felicia Hernandez, PhD[1]; Hoda Mirsafian[1], PhD; Ashley PL Marsh, PhD[1]; Colin Young, PhD[1]; Jesus Ramirez Castaño, BS[1]; Cassidy Carraway, BS[1]; Jessica Grzybowski, MS-CGC[1]; Carrie Horton, MS-CGC[1]; Marcy Richardson, PhD[1]; Matthew Schultz, PhD[1]; Rachid Karam, MD, PhD[1]

[1] Ambry Genetics. 1 Enterprise, Aliso Viejo, CA 92656 | * Email: sbelhadj@ambrygen.com | samibelhadj@outlook.com

BACKGROUND

Long read sequencing (LRS) is emerging as a key technology to elucidate complex variants that are missed and/or not fully characterized by conventional short read (SR) sequencing approaches. Among these, structural variants (SV) and mobile elements are particularly challenging to resolve. We sought to apply LRS to explore missing heritability in cases suspected to have Lynch Syndrome (LS; **Results I**) and to resolve complex SVs involving *PMS2* and *APC* (**Results II**).

METHODS

DNA samples from patients with inconclusive/negative results from multigene panel testing underwent multiple LRS-based strategies including long read genome sequencing (LR-GS) and targeted DNA sequencing (T-LRS) using a custom Twist panel on the PacBio Sequel IIe / Revio instruments. Short read DNA and RNA sequencing data, as well as MLPA data, was available for all studied cases as part of clinical testing.

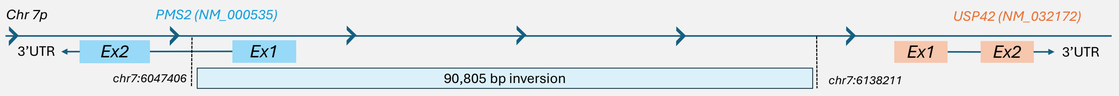
RESULTS I: Using LRS to explore Lynch Syndrome missing heritability

A. Cohort 1: Manually curated – Clinical history prioritization (n=17)

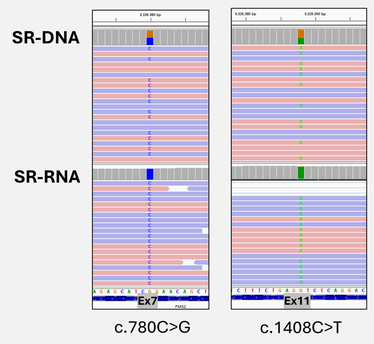
- 15/17 diagnosed with CRC/Endometrial cancer; 3/17 diagnosed with other LS-related cancer types (Dx age 38-77, median 50)
- 13/17 with ≥1 FDR and 4/17 with ≥2 SDRs diagnosed with LS-related cancers
- All cases with negative genetic testing and abnormal Immunohistochemistry (IHC) results (MMR proteins)

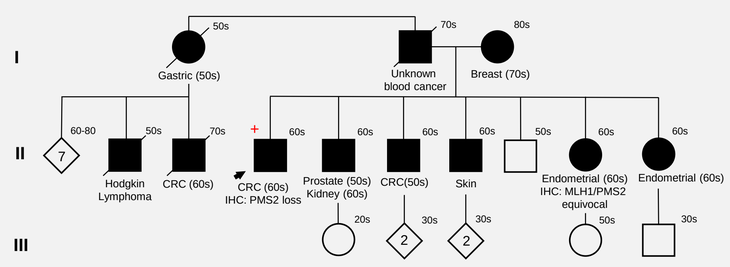
T-LRS *(Hybrid capture) MLH1, MSH2, MSH6, PMS2 and EPCAM*

- No alterations identified in 16/17 cases.
- *PMS2* Ex1 Inversion identified in 1/17 (5.8%) cases.



- RNA data shows complete Allele Specific Expression (ASE), Consistent with the inversion disrupting the transcription start site.





Cancer diagnosis ages are noted after cancer type. Current age or age at passing denoted in top-right corner.
Note: Proband is also a carrier of *NBN* p.R43*.

- ASE analysis with phASER[1] (short read DNA + RNA data) confirms the allelic imbalance observed in IGV for two informative variants:

Case ID	NM_ID	gene	aCount	bCount	totalCount	log2_aFC	variants(hg19)	Variant name	variant_location	binom_p	binom_q
AMB395	NM_000535	*PMS2*	7	342	349	-5.61	chr7_6026988_G_A	c.1408C>T	Exon11	2.10E-91	8.60E-90
	NM_000535	*PMS2*	138	0	138	Inf	chr7_6036980_G_C	c.780C>G	Exon7	5.74E-42	1.18E-40

B. Cohort 2: Allele-Specific Expression-prioritized cases (n=250)

- 250 probands with abnormal IHC (MMR proteins) and DNA still available.
- Unselected for personal and family cancer history.

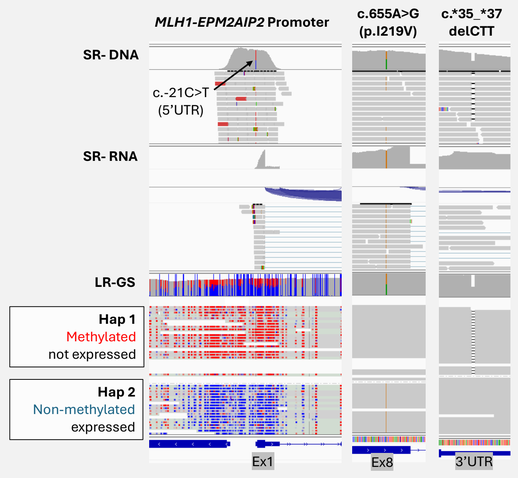
ASE analysis with phASER[1] identified 3/250 (1.2%) cases with complete *MLH1* ASE.

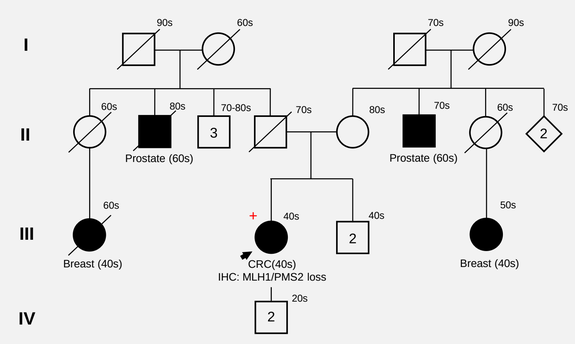
Case ID	NM_ID	gene	aCount	bCount	totalCount	log2_aFC	variants (hg19 location)	Variant name	binom_p	binom_q
AMB259*	NM_000249	*MLH1*	0	299	299	Inf	chr3_37053568_A_G	c.655A>G	2E-90	1E-87
	NM_000249	*MLH1*	111	0	111	Inf	chr3_37092173_GTTC_G	c.*35_*37delCTT	8E-34	1E-31
AMB790	NM_000249	*MLH1*	95	1	96	6.57	chr3_37092173_GTTC_G	c.*35_*37delCTT	2E-27	2E-25
AMB918	NM_000249	*MLH1*	36	4	40	3.17	chr3_37092173_GTTC_G, chr3_37092301_GGATT_G	c.*35_*37delCTT, c.*162_*165delGATT	2E-07	3E-06

*Note: Case AMB259 is a carrier of *MLH1* c.-21C>T (VUS; MAF < 0.001%; gnomAD v2.1)

T-LRS *(Hybrid capture) MLH1, MSH2, MSH6, PMS2 and EPCAM* → No alterations were identified →

- LR-GS in 1/3 candidates identified constitutional *MLH1* promoter methylation.





Cancer diagnosis ages are noted after cancer type. Current age or age at passing denoted in top-right corner.

RESULTS II: Using T-LRS to resolve challenging germline variants

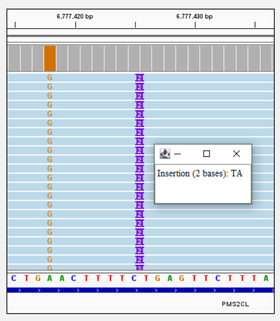
A. Refuting a *PMS2* Ex11 duplication identified by MLPA

We performed T-LRS and separated PMS2/PMS2CL haplotypes using Paraphase[2]. This revealed a gene conversion event that made the *PMS2*–specific probe #364 (SALSA MLPA Probemix P008) bind to *PMS2CL* resulting in a false positive MLPA signal.

CTGGACTTTTCTATGAGTTCTTTA — *Probe #364 PMS2 binding site (24nt) –* ***PMS2 probe binds***

CTGAACTTTTC--TGAGTTCTTTA — *Analogous region for probe #364 in PMS2CL (22nt) –* ***probe does not bind***

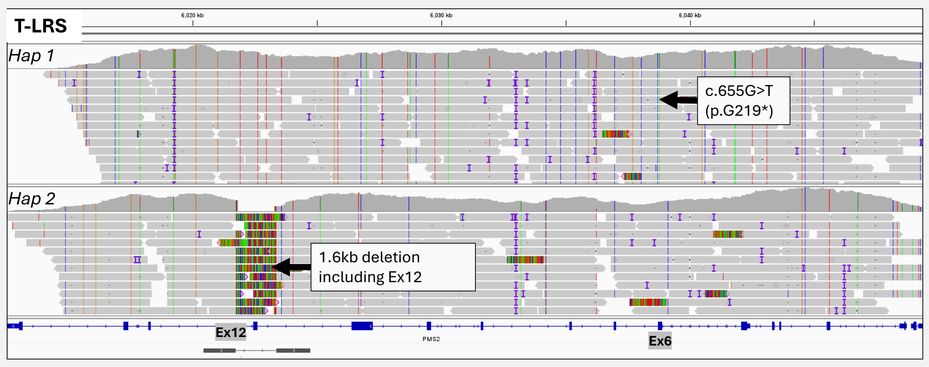
CTGGACTTTTCTATGAGTTCTTTA — *PMS2CL sequence after conversion event (24nt) –****probe binds***



Paraphase: *PMS2CL* reads supporting the gene conversion event

B. Uncovering a pathogenic *PMS2* Ex12 deletion in a Constitutional Mismatch Repair Deficiency (CMMRD) case

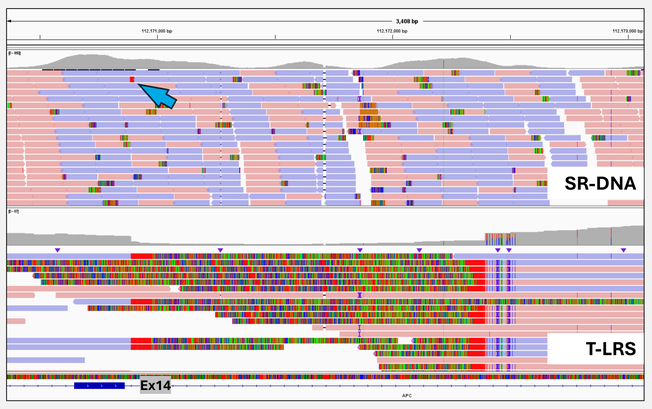
Conventional clinical testing (SR DNA+RNA and MLPA) identified a single pathogenic *PMS2* variant (p.G219*) in a CMMRD case. By using T-LRS, we identified a 1.6kb deletion affecting *PMS2* Ex12, *in-trans* with the already reported nonsense variant.
T-LRS data unequivocally assigned the deletion to *PMS2,* which was subsequently confirmed by long-range PCR and sequencing of *PMS2* Ex10-15. MLPA failed to identify this deletion.



C. *APC* Alu-driven recombination event resulting in Exon 14 skipping

Several unexplained FAP cases presented with *APC* Ex14 skipping (RNA data), where the underlying DNA variant couldn't be fully resolved by SR sequencing approaches. Using T-LRS, we identified a previously described Alu-driven recombination event[3]. The event resulted in a 1.5kb deletion with the 5' breakpoint within ~23bp from Ex14 splice donor site.

Despite coverage in Intron14, SR data fails to resolve this variant. Only <1% of reads supported the 5' breakpoint (blue arrow).



TAKE HOME POINTS

1. LRS increases the accuracy of germline genetic testing by providing a better characterization of complex variants and identifying alterations that could have been previously missed.
2. LRS is a useful complement to conventional assays such as short read DNA and RNA sequencing.
3. T-LRS applications are more cost effective, allow sample multiplexing and yield increased read depth.

REFERENCES

1. Castel SE et al. Nat. Comm. 2016 (PMID: 27605262)
2. https://github.com/PacificBiosciences/paraphase
3. Tuohy TMF et al. Hum. Genetics 2010 (PMID: 20033212)