

Title: Solving Lynch syndrome missing heritability by integrating long-read DNA sequencing and short-read RNA sequencing

Authors: Christopher J. Hatch (Ambry Genetics), Sami Belhadj (Ambry Genetics), John Harting (Ambry Genetics), Cassidy Carraway (Ambry Genetics), Hoda Mirsafian (Ambry Genetics), Stuti Joshi (Ambry Genetics), Kenny Chen (Ambry Genetics), Stephen Tanner (Ambry Genetics), Bojan Losic (Ambry Genetics), Marcy E. Richardson (Ambry Genetics), Seth I. Berger (Ambry Genetics), Rachid Karam (Ambry Genetics)

Learning Objective (146/150 characters): Evaluate the role of long-read DNA sequencing and short-read RNA sequencing to resolve missing heritability in cases suspected for Lynch syndrome.

Abstract: (2490/2500 characters, including spaces)

Lynch Syndrome (LS) is an inherited genetic condition that increases the risk of developing numerous cancers, including colorectal and endometrial cancer, and affects approximately 1:300 individuals. It is caused by pathogenic germline variants in the DNA mismatch repair (MMR) genes *MLH1*, *MSH2*, *MSH6* and *PMS2*; and select deletions in *EPCAM*. Constitutional *MLH1* promoter methylation (hereinafter epimutation), can also cause LS. Many individuals suspected to have LS receive negative clinical testing results potentially due to inherent limitations of conventional testing in detecting structural variants (SVs), regions complicated by pseudogenes, and epimutations. Advancements in long-read sequencing (LRS) and targeted RNA sequencing analysis may hold the potential to close these gaps.

Two cohorts of unexplained MMR-deficient patients were analyzed. Cohort 1 included 96 patients diagnosed with at least 1 LS-associated cancer that underwent targeted LRS (T-LRS). Cohort 2 included 250 individuals, unselected for personal or family cancer history, that underwent allele-specific expression (ASE) analysis to prioritize cases for Long-Read WGS (LR-WGS). For both cohorts, DNA isolated from whole blood, buccal swabs or saliva specimens was available for additional testing.

In Cohort 1, T-LRS did not identify additional causative germline variants affecting MMR genes. Single nucleotide variants (SNV) with a deleterious *in-silico* prediction were already classified as (likely) benign. Intronic SVs showed no significant impact on splicing with short-read RNA sequencing. One sample was identified as a possible epimutant based on ASE of *MLH1*. 54% of samples could not undergo ASE because of a lack of exonic SNPs, but two showed decreased normalized expression levels of *MLH1* suggestive of an epimutation.

Both T-LRS and clinical tests are PCR-based approaches that lose methylation signals. However, LR-WGS can endogenously detect 5-mC methylation directly from native DNA. With this in mind, ASE analysis ran on cohort 2 identified 3/250 cases (1.2%) with complete *MLH1* ASE. Subsequent LR-WGS confirmed *MLH1* epimutations in the silenced allele.

In these cohorts, where most individuals already received paired DNA and RNA clinical testing, the ability to detect epimutants had a greater impact on undiagnosed LS cases than searching for additional SNVs and SVs, suggesting that further investigation is warranted to assess the contribution of epimutations to hereditary cancer risk.

Draft Figures and tables for reference:

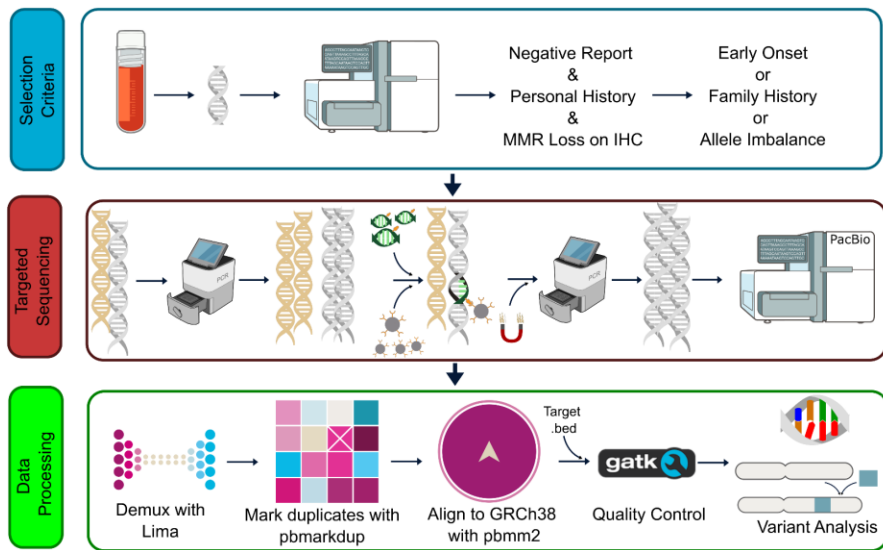


Figure 1. Overview of workflow. Sample selection criteria, an overview of the twist targeted sequencing workflow, and the data processing.

Table 1. Cohort Summary

Tier	Personal Cancer Hx (CRC/Ute)	Family Hx*	MMR IHC	In case of MLH1 IHC loss		Final
				Somatic MLH1 mutation status	Alternative: BRAF V600E	
AAA	EO (<50)	At least 1 FDR no prostate	Any	Discarded	Discarded/Unknown	4
AAA-	EO (<50)	At least 1 FDR prostate	Any	Discarded	Discarded/Unknown	1
AA	EO (<50)	At least 1 FDR	Any	Unknown	Unknown	1
A	Any age	At least 1 FDR	Any	Discarded	Discarded/Unknown	11
A-	Any age	At least 1 FDR / Prostate	MLH1	Unknown	Unknown	26
BBB	EO (<50)	At least 2 SDRs	Any	Discarded	Discarded/Unknown	6
BBB-	Any age	At least 2 SDRs	Any	Discarded	Discarded/unknown	7
BB-EO	EO (<50)	at least 1 SDR	any	discarded	Discarded/Unknown	10
BB	Any age	At least 1 SDR	Any	Discarded/unknown	Discarded/unknown	8
B	Any age	At least 2 TDRs	Any	Discarded/Unknown	Discarded/unknown	2
CCC	EO (<50)	Absent/No data	Any	Discarded	Discarded/unknown	16
CCC-	EO (<50)	Absent/No data	MLH1	unknown	unknown	2
CC	Any age	Absent/No data	MSH2/MSH6/PMS2	NA	NA	
C	Any age	Absent/No data	MLH1	Unknown	Unknown	2
						96

Table 2. SNV findings in MMR genes

SYMBOL	ambry_dot	ambry_pdot	Ambry Class	CHROM	POS	REF	ALT	samples	genotypes	AG	AL	DG	DL	CLIN_SIG	REVEL	BayesDel	MAX_AF	EXON	INTRON	Inclusion	
MSH6	MSH6 c.242C>T	MSH6 p.Ala81Val	VLB	chr2	47783475	C	T	('25_106595',)	('0 1',)	0	0	0.32	0.01	Conflicting	0.22	-0.294	0.00015	1/10	.	SpliceAI	
MSH2	MSH2 c.2201C>G	MSH2 p.Ser734Cys	VLB	chr2	47476562	C	G	('25_260109',)	('1 0',)	0	0	0	0	VUS/Conflict	0.74	0.088	8.99E-07	13/16	.	ClinVar	
MLH1	MLH1 c.1853A>C	MLH1 p.Lys618Thr	B	chr3	37047640	A	C	('25_213292', '25_265924',)	('0 1', '0 1',)	0	0	0	0	LB/B	0.96	0.509	0.0099	16/19	.	REVEL/BayesDel	
MLH1	MLH1 c.1852A>G	MLH1 p.Lys618Glu	B	chr3	37047639	A	G	('25_213292', '25_265924',)	('0 1', '0 1',)	0	0	0	0	LB/B	0.87	0.315	0.0099	16/19	.	REVEL/BayesDel	
MLH1	MLH1 c.1321G>A	MLH1 p.Ala441Thr	B	chr3	37025919	G	A	('25_203231',)	('0 1',)	0	0	0	0	VUS/LB/B	0.54	0.128	0.00141	6	12/19	.	REVEL/BayesDel
MSH6	MSH6 c.3961A>G	MSH6 p.Arg1321Gly	VLB	chr2	47806611	A	G	('25_213292',)	('0 1',)	0	0	0	0	LB	0.67	0.236	0.001	9/10	.	REVEL/BayesDel	

Table 3. SV findings in MMR genes

SV chrom	SV_start	SV_end	SV length	SV type	Gene	Location	cols_0_1	cols_1_1	IHC	QUAL	RNA Data
2	47433224	47433351	127	DEL	MSH2	Intron7	['25_265444']	[]	MLH1/PMS2	999	T
2	47433395	47433471	76	DEL	MSH2	Intron7	[]	['25_239190']	MSH2/MSH6	17	T
3	37000665	37000736	71	DEL	MLH1	Intron2	['25_315816']	[]	MLH1/PMS2	999	T
2	47448799	47450299	1500	DEL	MSH2	Intron8	['25_256799', '25_281368', '25-229905', '25-118954', '25_291940', '25_296604']	[]	MSH2/MSH6, PMS2, MLH1/PMS2...	999	T
2	47448801	47450435	1634	DEL	MSH2	Intron8	['25_256799']	[]	MSH2/MSH6	269	F
2	47426741	47426741	120	INS	MSH2	Intron6	['25-309533', '24-595731', '25_265924']	[]	MLH1/PMS2	412	T
2	47683321	47683321	0	TRA	MSH2	[chr14:56926953] AAAAAA	['25_127258']	[]	MSH2/MSH6	58	T
2	47683321	47683321	0	TRA	MSH2	[2:47683321] TTTTTA	['25_127258']	[]	MHS2/MSH6	58	T
2	47788249	47788249	254	INS	MSH6	Intron1	['24-595731']	[]	MLH1/PMS2	115	T
7	5980877	5980877	120	INS	PMS2	Intron3	['25_306857', '25_265444']	[]	MLH1/PMS2, MLH1/PMS2	535	T
7	5980919	5980919	120	INS	PMS2	Intron3	['25_170244']	[]	MSH2/MSH6	999	T
7	5980943	5980943	96	INS	PMS2	Intron3	['25_239190', '25_265444']	[]	MSH2/MSH6	999	T

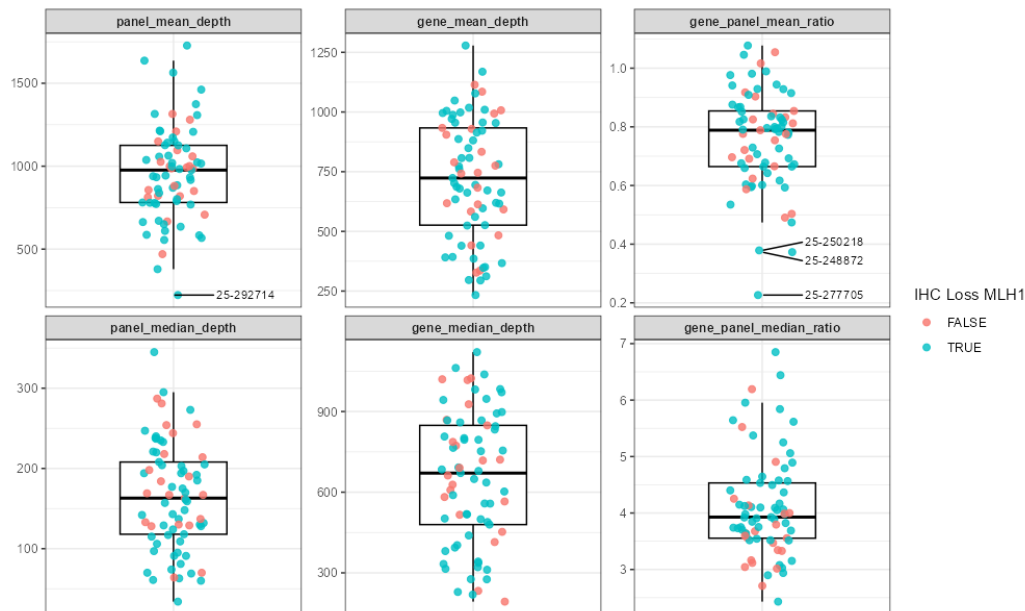


Figure 2. RNA coverage for *MLH1* compared to other genes in the ambry panel. The mean ratio for RNA coverage identifies 3 outliers that all have IHC loss of *MLH1* (row 1, column 3).

Table 4. ASE on flagged samples from *MLH1* coverage supports 1 sample, but is inconclusive for remaining 2.

sample	panel mean depth	panel median depth	gene mean depth	gene median depth	Mean ratio	Median Ratio	IHC	Exonic Het SNP Ref/Alt	RNA SNP Ref/Alt
All	947 +/- 266	168 +/- 65	737 +/- 243	669 +/- 237	0.78 +/- 0.14	4.12 +/- 0.88			
25-248872	790	74	295	276	0.37	3.73	Deficient	5': 242/175 (58/42%)	5': 2/3 (40/60%)
25-250218	917	137	347	333	0.38	2.43	MLH1/ PMS2	Ex8: 218/210 (51/49%) In14: 181/126 (59/41%)	Ex8: 42/453 (8/92%) In14: 6/59 (9/91%)
25-277705	1727	60	391	315	0.23	5.25	MLH1	No Het SNPs	