Supplementary Table 1. Clinical characteristics of the 51 individuals included for whole-exome sequencing.

Clinical selection	Number of individuals
Number of polyps	
Between 5 and 10 polyps	3
Between 10 and 20 polyps	20
Between 20 and 50 polyps	22
More than 50 polyps	6
Type of polyps	
Adenomatous polyps	22
Adenomatous + hyperplastic polyps	22
Adenomatous + serrated polyps	1
Adenomatous + hyperplastic + serrated polyps	6
Also diagnosed with colorectal cancer	
Yes	21
No	30
Family history FDR (first degree relative)	
Polyposis	3
CRC	12
Polyposis + CRC	12
Negative/unknown	24

Nature Genetics: doi:10.1038/ng.3287

Supplementary Table 2. Clinical characteristics of individuals included for exome sequencing.

Subject	Polyp type	Polyp number	Colorectal cancer	FDR with CRC or polyposis	NS BER gene
P07	A	50	Y	Y	*NTHL1* (p.Gln90*)
P30	A	30	Y	Y	
P09	A	20	Y	Y	*SMUG1* (p.Arg124*)
P23	A	20	Y	Y	*NTHL1* (p.Gln90*)
P01[a]	A	15	Y	Y	*NTHL1* (p.Gln90*)
P22	A	15	Y	Y	
P15	A	10	Y	Y	
P08	A	250	Y	N	
P36	A	40	Y	N	
P03	A	25	Y	N	
P02	A	20	Y	N	
P52	A	15	Y	N	
P12	A	10	Y	N	
P42	A	50	N	Y	
P49[a]	A	40	N	Y	*NTHL1* (p.Gln90*)
P25	A	20	N	Y	
P16	A	10	N	Y	
P31	A	30	N	N	
P34	A	30	N	N	
P58	A	16	N	N	
P19	A	15	N	N	
P57	A	15	N	N	*MPG* (p.Arg118*)
P04	A H	25	Y	Y	
P20	A H	15	Y	Y	
P51	A H	10	Y	Y	
P05	A H	30	Y	N	
P28	A H	25	Y	N	
P26	A H	24	Y	N	
P24	A H	20	Y	N	
P38	A H	60	N	Y	
P44	A H	50	N	Y	
P37	A H	45	N	Y	
P32	A H	30	N	Y	
P27	A H	25	N	Y	
P46[b]	A H	17	N	Y	
P18	A H	15	N	Y	
P17	A H	14	N	Y	
P11	A H	10	N	Y	
P47[b]	A H	10	N	Y	
P45[b]	A H	5–10	N	Y	
P54	A H	21	N	N	*OGG1* (p.Arg131*)
P53	A H	15	N	N	
P56	A H	15	N	N	
P41	A H	5–10	N	N	
P06	A H S	30	Y	N	
P43	A H S	50	N	Y	
P62	A H S	9	N	Y	
P35	A H S	30	N	N	
P29	A H S	25	N	N	
P55	A H S	12	N	N	
P14	A S	10	N	N	

List is ranked based on (i) polyp type, (ii) positive history of CRC development in the index patient, (iii) positive history of CRC or polyposis development in a first degree relative (FDR) of the index patients and (iv) the number of polyps present at time of diagnosis. Polyp type: A: Adenomatous polyps, H:Hyperplastic polyps, S:Serrated polyps. Polyp number: amount of polyps present at time of diagnosis. Colorectal cancer: positive (Y) if index patient developed CRC. FDR with CRC or polyposis: first degree relatives diagnosed with CRC and/or polyps; negative (N) if available clinical data of relatives did not mention a positive history of CRC or polyposis. NS BER gene: germline nonsense variant present in one of the base excision repair genes. For two families two (a) and three (b) family members were included.

Supplementary Table 3. Overview of exome sequencing statistics.

Subject	NGS platform	Total bases sequenced (Gb)	Total bases on / near target (Gb)	% bases on target	Average coverage of targets	Median coverage	% regions ≥ 10x coverage	% regions ≥ 20x coverage
P01	SOLID 5500	5.48	4.36	73.72%	73.53	57	93.01%	84.97%
P02	SOLID 5500	6.14	4.91	74.11%	82.56	63.5	93.24%	86.06%
P03	SOLID 5500	6.31	4.98	73.01%	81.93	63	93.35%	86.09%
P04	SOLID 5500	6.78	5.44	74.35%	89.96	68	93.82%	87.19%
P05	SOLID 5500	6.75	5.43	74.60%	89.92	69	93.97%	87.75%
P06	SOLID 5500	6.53	5.09	72.07%	85.16	66	93.65%	87.05%
P07	SOLID 5500	6.91	5.53	74.15%	92.49	71	94.32%	88.22%
P08	SOLID 5500	4.77	3.81	73.99%	63.05	48	91.47%	81.37%
P09	SOLID 5500	6.56	5.23	73.77%	87.14	67	93.56%	86.72%
P11	Illumina Hi Seq	5.95	4.91	64.61%	77.39	75	98.72%	96.44%
P12	Illumina Hi Seq	6.68	5.56	64.71%	88.64	85	98.91%	97.03%
P14	Illumina Hi Seq	6.39	5.27	63.90%	83.95	80	98.89%	96.67%
P15	Illumina Hi Seq	6.82	5.59	63.33%	88.06	84	99.04%	97.10%
P16	Illumina Hi Seq	6.17	5.13	64.99%	78.60	75	98.90%	96.64%
P17	Illumina Hi Seq	6.50	5.32	63.18%	83.07	79.5	98.79%	96.53%
P18	Illumina Hi Seq	6.67	5.54	64.91%	91.96	87	98.98%	97.02%
P19	Illumina Hi Seq	7.33	6.16	66.18%	99.25	95	99.15%	97.55%
P20	Illumina Hi Seq	6.30	5.26	65.48%	80.08	76	98.85%	96.55%
P22	Illumina Hi Seq	7.15	5.96	65.32%	94.83	90.5	99.08%	97.42%
P23	Illumina Hi Seq	6.60	5.56	66.01%	88.32	85	98.84%	96.96%
P24	Illumina Hi Seq	7.30	6.13	66.46%	98.74	94	99.16%	97.56%
P25	Illumina Hi Seq	7.13	5.88	64.80%	93.43	89	99.08%	97.28%
P26	Illumina Hi Seq	5.25	4.35	64.89%	67.49	64	98.44%	95.26%
P27	Illumina Hi Seq	6.70	5.61	66.00%	89.73	86	98.91%	97.03%
P28	Illumina Hi Seq	6.89	5.75	65.95%	95.20	91	98.91%	97.04%
P29	Illumina Hi Seq	5.30	4.43	66.11%	69.76	66	98.52%	95.55%
P30	Illumina Hi Seq	5.93	4.96	65.68%	82.75	79	98.70%	96.42%
P31	Illumina Hi Seq	5.95	4.93	66.27%	82.55	79	98.82%	96.63%
P32	Illumina Hi Seq	6.72	5.57	65.84%	92.52	88	99.01%	97.13%
P34	Illumina Hi Seq	6.69	5.54	65.72%	89.33	85	98.96%	97.03%
P35	Illumina Hi Seq	6.49	5.44	66.12%	89.88	86	98.90%	96.96%
P36	Illumina Hi Seq	6.59	5.51	66.04%	89.65	85	98.95%	96.85%
P37	Illumina Hi Seq	5.90	4.88	65.84%	79.13	75	98.81%	96.33%
P38	Illumina Hi Seq	6.25	5.21	64.97%	85.67	81	98.79%	96.67%
P41	Illumina Hi Seq	6.16	5.15	66.73%	86.48	82.5	98.90%	97.01%
P42	Illumina Hi Seq	5.79	4.83	65.59%	75.93	72	98.81%	96.15%
P43	Illumina Hi Seq	6.37	5.36	67.08%	91.62	87	99.09%	97.26%
P44	Illumina Hi Seq	7.24	6.03	65.50%	98.67	94	98.96%	97.24%
P45	Illumina Hi Seq	5.37	4.44	65.16%	69.94	65	98.56%	94.92%
P46	Illumina Hi Seq	5.41	4.53	66.81%	83.99	79.5	98.93%	96.92%
P47	Illumina Hi Seq	6.65	5.63	67.64%	92.70	88.5	98.95%	97.25%
P49	Illumina Hi Seq	6.42	5.35	66.26%	84.96	76	98.96%	96.54%
P51	Illumina Hi Seq	6.76	5.65	65.61%	91.20	87	99.04%	97.15%
P52	Illumina Hi Seq	6.66	5.54	66.14%	90.33	86	99.08%	97.25%
P53	Illumina Hi Seq	6.76	5.62	65.81%	94.31	90	99.11%	97.33%
P54	Illumina Hi Seq	4.98	4.14	66.10%	64.22	61	98.39%	94.95%
P55	Illumina Hi Seq	5.61	4.69	65.29%	78.16	74	98.72%	96.39%
P56	Illumina Hi Seq	5.92	4.94	65.74%	79.13	75	98.74%	96.45%
P57	Illumina Hi Seq	5.92	4.93	65.71%	81.12	77	98.90%	96.59%
P58	Illumina Hi Seq	6.47	5.37	65.05%	87.48	83	98.93%	96.86%
P62	Illumina Hi Seq	4.87	4.06	66.26%	64.92	62	98.44%	94.89%

Supplementary Table 4. Potential deleterious variant calls in known cancer predisposing genes.

Gene	Subject	mRNA accession number	mRNA change	Protein change	Mutation type[1]	CRC phenotype[2]	Polyposis phenotype[2]	Co-segregation	Mode of inheritance[3]	Pathogenic[4]
APC	P35	NM_001127510	c.6403A>G	p.Ile2135Val	MSS	Y	Y	ND	AD	N
POLD1	P15	NM_002691	c.371T>C	p.Val124Ala	MSS	Y	Y	ND	AD	N
POLD1	P17	NM_002691	c.961G>A	p.Gly321Ser	MSS	Y	Y	ND	AD	P[5]
POLE	P18	NM_006231	c.665G>A	p.Arg222His	MSS	Y	Y	ND	AR/AD	N
POLE	P35	NM_006231	c.850A>G	p.Lys284Glu	MSS	Y	Y	ND	AR/AD	P[6]
BLM	P51	NM_000057	c.488C>T	p.Ser163Phe	MSS	N	N	ND	AR	N
MSH2	P22	NM_000251	c.1144C>T	p.Arg382Cys	MSS	Y	N	ND	AR/AD	N
MSH6	P06	NM_000179	c.2651C>G	p.Ser884Cys	MSS	Y	N	ND	AR/AD	N
PMS2	P42	NM_000535	c.917T>C	p.Val306Ala	MSS	Y	N	ND	AR/AD	N
PMS2	P05	NM_000535	c.620G>A	p.Gly207Glu	MSS	Y	N	ND	AR/AD	N
ATM	P16	NM_000051	c.5938G>A	p.Gly1980Arg	MSS	N	N	ND	AR/AD	N
ATM	P65	NM_000051	c.5753G>C	p.Arg1918Thr	MSS	N	N	ND	AR/AD	N
ATM	P15	NM_000051	c.1564-1565del	p.Glu522fs	FS	N	N	ND	AR/AD	N
ATM	P20	NM_000051	c.1564-1565del	p.Glu522fs	FS	N	N	ND	AR/AD	N
BRCA2	P54	NM_000059	c.5645C>A	p.Ser1882*	NS	N	N	ND	AR/AD	N[7]
CEBPA	P34	NM_004364	c.827A>G	p.Lys276Arg	MSS	N	N	ND	AD	N
CYLD	P26	NM_015247	c.100C>G	p.Gln34Glu	MSS	N	N	ND	AD	N
DICER1	P15	NM_030621	c.4195A>G	p.Lys1399Glu	MSS	N	N	ND	AD	N
DIS3L2	P01	NM_152383	c.520G>A	p.Asp174Asn	MSS	N	N	N	AR	N
GATA2	P45	NM_001145661	c.82G>A	p.Gly28Ser	MSS	N	N	N	AD	N
KIT	P28	NM_000222	c.1724A>T	p.Gln575Leu	MSS	N	N	ND	AD	N
NF1	P08	NM_001042492	c.529A>G	p.Ile177Val	MSS	N	N	ND	AD	N
PTCH1	P17	NM_000264	c.1993C>T	p.Arg665Cys	MSS	N	N	ND	AD	N
RAD51D	P28	NM_001142571	c.137C>G	p.Ser46Cys	MSS	N	N	ND	AD	N
RHBDF2	P03	NM_024599	c.940G>A	p.Ala314Thr	MSS	N	N	ND	AD	N
TMEM127	P55	NM_001193304	c.433G>C	p.Gly145Arg	MSS	N	N	ND	AD	N
TSC1	P47	NM_000368	c.568C>T	p.Arg190Cys	MSS	N	N	N	AD	N
TSC1	P45	NM_000368	c.568C>T	p.Arg190Cys	MSS	N	N	N	AD	N
TSC2	P20	NM_000548	c.607T>G	p.Cys203Gly	MSS	N	N	ND	AD	N
TSC2	P46	NM_000548	c.1387A>G	p.Ile463Val	MSS	N	N	N	AD	N
WAS	P27	NM_000377	c.794C>G	p.Pro265Arg	MSS	N	N	ND	XR	N

1) FS: frameshift variant, MSS: missense variant and NS: nonsense variant. Only missense variants with a minor allele frequency of <0.1% and phyloP score of >3.0 were selected (see online methods).

2) Presence (Y) or absence (N) of increased risk for the development of CRC or polyposis in individuals with monoallelic (autosomal dominant inheritance) or biallelic (autosomal recessive inheritance) mutations.

3) AD: Autosomal Dominant, AR: Autosomal Recessive, XR: X-linked Recessive.

4) Variant considered to be causative for observed polyposis phenotype, based on *in silico* predictions using SIFT, Align GVGD, and PolyPhen. P: possibly causative, N: not likely causative.

5) Variant located in the exonuclease domain of POLD1 (aa 304–517); SIFT: Deleterious (score: 0), Align GVGD: C55 (GV: 0.00 - GD: 55.27) and PolyPhen2: possibly damaging (score: 0.880).

6) Variant located in the exonuclease domain of POLE (aa 268–471); SIFT: Deleterious (score: 0), Align GVGD: Class C0 (GV: 353.86 - GD: 0.00), and PolyPhen2: probably damaging (score: 1.000).

7) Patient diagnosed with breast cancer at young age and has a positive family history of early onset BC. Increased risk for polyposis and CRC is unlikely. Mutation was known prior to exome sequencing.

All variants were observed in a heterozygous state and no signs of compound heterozygosity was found. For applied selection criteria, see main text.

Supplementary Table 5a. Genes with protein-truncating heterozygous variant calls in multiple unrelated individuals.

Gene	Subjects	Number of variants	CPG	mRNA accession number	mRNA change	Protein changes
ATM	P15,P20	1	Y	NM_000051	c.1564-1565del	p.Glu522fs
ATXN3	P34;P54	2	N	NM_004993	c.915_916ins20; c.873del	p.Gly306fs; p.Lys291fs
C14orf37	P06;P20	1	N	NM_001001872	c.698del	p.Thr233fs
CALCR	P03;P38	1	N	NM_001164737	c.1460del	p.Asn487fs
GRIN3B	P32;P36	2	N	NM_138690	c.3085_3086insGA; c.2211dup	p.Pro1029fs; p.Lys738fs
MMP25	P42;P55	1	N	NM_022468	c.352C>T	p.Arg118*
OR4S1	P49;P54	2	N	NM_001004725	c.831_832del; c.50_51insC	p.Pro278fs; p.Glu17fs
PGLS	P02;P11	2	N	NM_012088	c.640-1G>T; c.550C>T	p.?; p.Gln184*
RBFOX3	P15;P24;P25;P29;P44;P55	1	N	NM_001082575	c.26_29del	p.Gln9fs
SSPO	P02;P31	2	N	NM_198455	c.7682G>A; c.2626del	p.Trp2561*; p.Pro876fs
TAS2R46	P18;P20	1	N	NM_176887	c.262del	p.Trp88fs
TLL2	P26;P57	2	N	NM_012465	c.1268-1G>A; c.1201dup	p.?; p.Ser401fs
TTC3	P29;P30	2	N	NM_001001894	c.1066C>T; c.5268_5271dup	p.Arg356*; p.Thr1758fs
WDR63	P07;P26	2	N	NM_145172	c.1917+1G>A; c.1956dup	p.?; p.Asp653fs

CPG: cancer predisposing gene.

Supplementary Table 5b. Homozygous variant calls resulting in protein truncation.

Gene	Subject	mRNA accession number	mRNA change	Protein change	Coverage	%Var. Reads	Mutation Type	CPG	Rec.	Co-segr.
AHCTF1	P24	NM_015446	c.6419-2A>C	p.?	62	100	CSS	N	1	ND
ATXN3	P34	NM_004993	c.916_917insC	p.Gly306fs	120	76.6	FS	N	1	ND
CAMKK2	P43 P45	NM_001270486	c.1614_1615insAAAA	p.Gly539fs	106 69	95.7 100	FS	N	2	N
CDCP2	P33	NM_201546	c.1224_1225insGC	p.Met409fs	64	100	FS	N	1	ND
COL28A1	P07	NM_001037763	c.882+1G>T	p.?	98	74	CSS	N	1	ND
COL6A6	P42	NM_001102608	c.1282+1G>C	p.?	75	73.3	CSS	N	1	ND
CTSA	P15	NM_000308	c.107_108del	p.Leu36fs	68	73.9	FS	N	1	ND
DCHS2	P23	NM_017639	c.6024dup	p.Tyr2009fs	33	76.7	FS	N	1	ND
MROH6	P25 P33 P39	NM_001100878	c.2027dup	p.Cys677fs	11 16 10	72.7 85.7 100	FS	N	3	ND
MRPS18C	P20	NM_016067	c.150+2T>G	p.?	37	81.08	CSS	N	1	ND
MTCH1	P47 P49	NM_001271641	c.30del	p.Trp11fs	16 27	73.3 100	FS	N	2	N
MZT2A	P26 P52	NM_001085365	c.33del	p.Ser12fs	12 11	83.3 100	FS	N	2	ND
NTHL1	P01 P07 P23 P49	NM_002528	c.268C>T	p.Gln90*	26 39 119 168	73.1 84.6 99.2 100	NS	N	4	Y
ORAI1	P12	NM_032790	c.141_142insT	p.Pro47fs	36	76.5	FS	N	1	ND
PIF1	P62	NM_025049	c.145G>T	p.Glu49*	11	72.7	NS	N	1	N
RASAL3	P11	NM_022904	c.1239_1252del	p.Ala414fs	28	73.7	FS	N	1	ND
SAMD1	P22	NM_138352	c.336_337insC	p.Ala113fs	10	90	FS	N	1	ND
ZNF84	P01	NM_003428	c.2192_2193del	p.Arg731fs	14	78.6	FS	N	1	N

CPG: cancer predisposing gene, Rec.: number of times the specific variant was observed in our cohort, Co-segr.: co-segregation of variant with disease in family (ND: not determined, N: no co-segregation, Y: co-segregation confirmed). NS: nonsense variant, FS: frameshift (insertion or deletion) and CSS: canonical splice site. Variant calls with at least 10 reads and >70% variant reads are depicted.

Supplementary Table 6. Frequency of the p.Gln90* variant in *NTHL1* in different ethnic groups.

Population	# Allele $^{p.Gln90*}$	# Allele ^{total}	# Homozygotes	Allele Frequency
European (Dutch)	17	4,658	0	0.003650
European (Finnish)	25	6,748	0	0.003705
European (Non-Finnish)	154	66,850	0	0.002304
Other	1	916	0	0.001092
Latino	5	11,598	0	0.000431
African	2	10,514	0	0.000190
South Asian	2	16,626	0	0.000120
East Asian	0	8,740	0	0
Total	206	126,650	0	0.001627

Supplementary Table 7. Homozygous stretches observed in the exome data of polyposis patients.

Subject	Family	Total length of homozygous segments (Mb)	P-Value[1]	*NTHL1* homozygous segment[2]
P01	A	7,3	0,35	N
P07 *	B	20,2	0,027	Y
P23	C	25,1	0,006	Y
P49	A	9,0	0,28	N

1) Compared to 200 controls to determine possible significant enrichment of homozygous segments in the corresponding exomes; one-sided, right-tailed z-test. (average length of homozygous segments in controls: 4,2Mb; StDev: 8,3).

2) Homozygous region encompassing the genomic locus of the *NTHL1* gene.

* consanguinity confirmed based on family history. For further details: see online methods.

Supplementary Table 8. Somatic base substitutions in comprehensive cancer panel genes in CRCs of individuals with biallelic *NTHL1* mutations.

Sample	Gene	mRNA accession number	mRNA change	AA Change	Base Subst.
P07	*APC*	NM_001127510	c.2269C>T	p.Gln757*	C:G>T:A
P07	*APC*	NM_001127510	c.4285C>T	p.Gln1429*	C:G>T:A
P07	*ERBB3*	NM_001982	c.695C>T	p.Ala232Val	C:G>T:A
P07	*ERBB4*	NM_005235	c.1183G>T	p.Val395Phe	C:G>A:T
P07	*FANCG*	NM_004629	c.1402G>A	p.Ala468Thr	C:G>T:A
P07	*KRAS*	NM_033360	c.35G>A	p.Gly12Asp	C:G>T:A
P07	*PIK3CA*	NM_006218	c.2176G>A	p.Glu726Lys	C:G>T:A
P07	*PIK3CA*	NM_006218	c.353G>A	p.Gly118Asp	C:G>T:A
P07	*RNF213*	NM_001256071	c.10853G>A	p.Gly3618Asp	C:G>T:A
P07	*SEPT9*	NM_001113491	c.1321C>T	p.Pro441Ser	C:G>T:A
P07	*SYNE1*	NM_033071	c.16285A>G	p.Lys5429Glu	T:A>C:G
P07	*TP53*	NM_000546	c.853G>A	p.Glu285Lys	C:G>T:A
P07	*ZNF521*	NM_015461	c.1490G>A	p.Arg497Gln	C:G>T:A
P07	*NF1*	NM_001042492	c.6678G>A	p.Leu2226Leu	C:G>T:A
P07	*RNF213*	NM_001256071	c.13704G>A	p.Val4568Val	C:G>T:A
P07	*SMARCA4*	NM_001128845	c.4038C>T	p.Tyr1346Tyr	C:G>T:A
P07	*SOX11*	NM_003108	c.693C>T	p.Asp231Asp	C:G>T:A
P23	*APC*	NM_001127510	c.1909G>A	p.Gly637Arg	C:G>T:A
P23	*APC*	NM_001127510	c.4405C>T	p.Gln1469*	C:G>T:A
P23	*APC*	NM_001127510	c.3268C>T	p.Gln1090*	C:G>T:A
P23	*ARID1A*	NM_006015	c.6208C>T	p.Gln2070*	C:G>T:A
P23	*BRAF*	NM_004333	c.1780G>A	p.Asp594Asn	C:G>T:A
P23	*KRAS*	NM_033360	c.38G>A	p.Gly13Asp	C:G>T:A
P23	*PIK3CA*	NM_006218	c.1624G>A	p.Glu542Lys	C:G>T:A
P23	*PTCH1*	NM_001083602	c.3219G>A	p.Met1073Ile	C:G>T:A
P23	*SMAD4*	NM_005359	c.1255G>A	p.Gly419Arg	C:G>T:A
P23	*TP53*	NM_000546	c.722C>T	p.Ser241Phe	C:G>T:A
P23	*ATR*	NM_001184	c.2850G>A	p.Pro950Pro	C:G>T:A
P23	*PTPRT*	NM_133170	c.3486G>A	p.Ala1162Ala	C:G>T:A
P23	*RALGDS*	NM_006266	c.225C>T	p.Val75Val	C:G>T:A
P23	*TCF3*	NM_003200	c.1743C>T	p.Asn581Asn	C:G>T:A
P23	*TRIM33*	NM_015906	c.1914C>T	p.Thr638Thr	C:G>T:A
P69	*APC*	NM_001127510	c.4285C>T	p.Gln1429*	C:G>T:A
P69	*KRAS*	NM_033360	c.35G>T	p.Gly12Val	C:G>A:T
P69	*LPHN3*	NM_015236	c.252T>G	p.Tyr84*	T:A>G:C
P69	*PIK3CA*	NM_006218	c.1624G>A	p.Glu542Lys	C:G>T:A
P69	*PIK3CG*	NM_001282427	c.1703C>A	p.Thr568Lys	C:G>A:T
P69	*ROS1*	NM_002944	c.5104G>T	p.Val1702Phe	C:G>A:T
P69	*SETD2*	NM_014159	c.535G>A	p.Glu179Lys	C:G>T:A
P69	*TCF3*	NM_003200	c.1670G>A	p.Arg557Gln	C:G>T:A
P69	*TP53*	NM_000546	c.743G>A	p.Arg248Gln	C:G>T:A
P69	*BCL9*	NM_004326	c.660+1G>A	p.?	C:G>T:A
P69	*CSMD3*	NM_198123	c.8778T>A	p.Pro2926Pro	T:A>A:T
P69	*ITGA9*	NM_002207	c.1299C>T	p.Gly433Gly	C:G>T:A
P69	*PRKDC*	NM_006904	c.5532G>A	p.Val1844Val	C:G>T:A

Supplementary Table 9. Somatic mutations in *APC* in colorectal adenomas of individuals with biallelic *NTHL1* mutations.

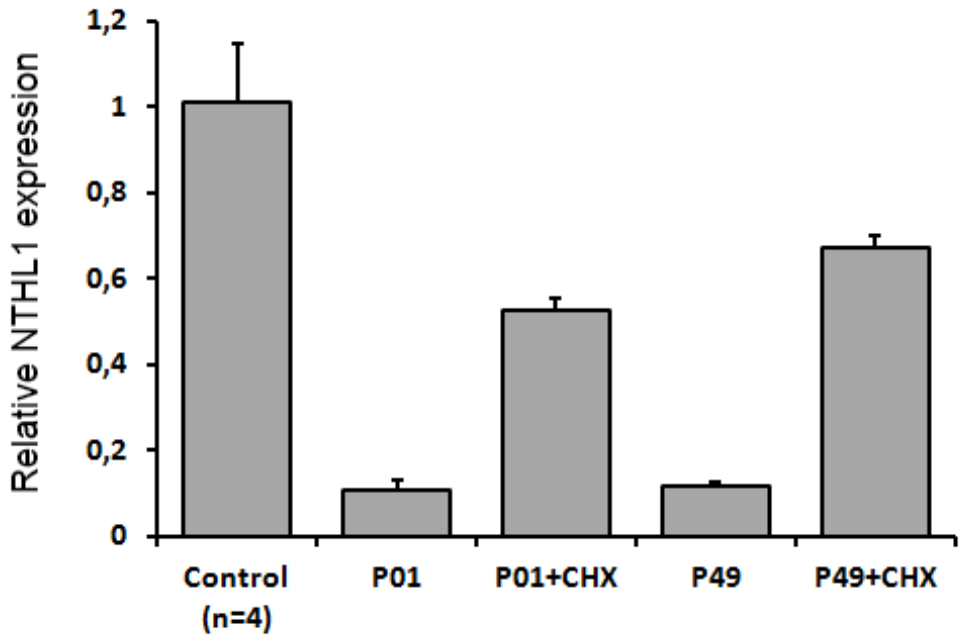
Sample	mRNA change*	AA Change	# reads Fw	# reads Rv	%Var. Reads Fw	%Var. Reads Rv	Base Subst.
P49 (2)	c.1417C>T	p.Gln473*	7,463	9,506	13.1	13.6	C:G>T:A
P49 (1)	c.2097G>A	p.Trp699*	6,011	1,791	24.6	34.2	C:G>T:A
P49 (3)	c.3688C>T	p.Gln1230*	6,609	6,544	42.2	42.0	C:G>T:A
P49 (3)	c.3880C>T	p.Gln1294*	2,306	2,218	32.8	54.7	C:G>T:A
P71 (2)	c.739C>T	p.Gln247*	4,087	4,101	18.7	18.6	C:G>T:A
P71 (1)	c.2995C>T	p.Gln999*	1,748	4,768	16.1	10.1	C:G>T:A

*NM_001127510. Numbers between brackets represent corresponding polyp sample.

Supplementary Table 10. Somatic base substitutions in comprehensive cancer panel genes in CRCs of individuals with biallelic *MUTYH* mutations.

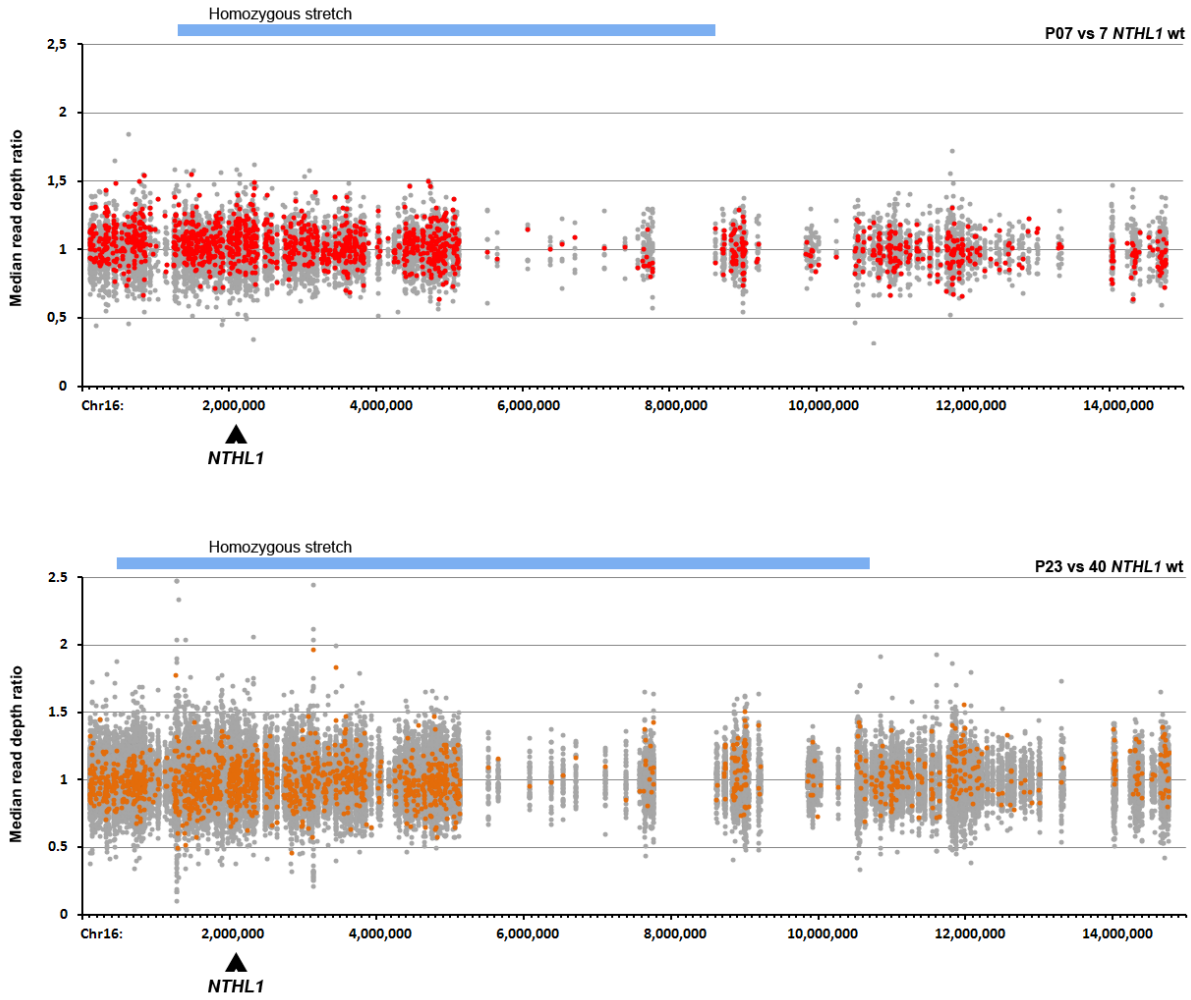
Sample	Gene	mRNA accession number	mRNA change	AA Change	Base Subst.
MAP1	*ADAMTS20*	NM_025003	c.1723G>A	p.Gly575Arg	C:G>T:A
MAP1	*PIK3CG*	NM_001282427	c.1912G>T	p.Glu638*	C:G>A:T
MAP1	*SMAD4*	NM_005359	c.1572G>T	p.Trp524Cys	C:G>A:T
MAP1	*SMARCA4*	NM_001128845	c.4296G>T	p.Met1432Ile	C:G>A:T
MAP2	*APC*	NM_001127510	c.4120G>T	p.Glu1374*	C:G>A:T
MAP2	*KRAS*	NM_033360	c.436G>A	p.Ala146Thr	C:G>T:A
MAP2	*NLRP1*	NM_033004	c.2415G>T	p.Lys805Asn	C:G>A:T

MUTYH germline mutations. MAP1: p.Pro405Leu (homozygous) and MAP2: p.Pro405Leu (heterozygous) and p.Asp161His (heterozygous).

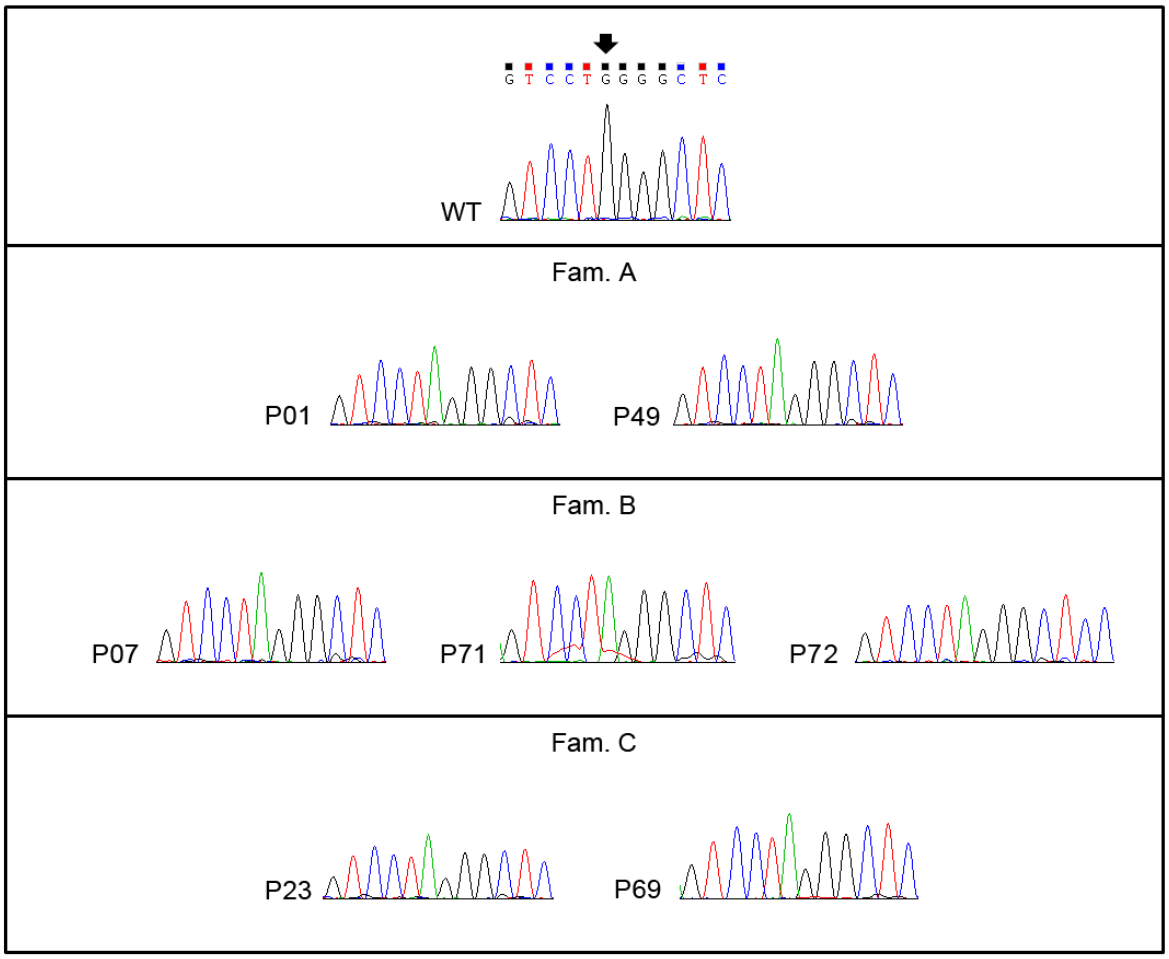


Supplementary Figure 1. Loss of *NTHL1* RNA expression in homozygous p.Gln90* carriers due to nonsense-mediated decay. EBV-transformed B-lymphocytes derived from two homozygous p.Gln90* carriers (P01 and P49, family A) show approximately ten-fold reduced *NTHL1* RNA expression compared to EBV-transformed B-lymphocytes derived from controls. Partial rescue of *NTHL1* expression is obtained after treatment with cyclohexamide (CHX) which indicates that expression is reduced due to nonsense-mediated decay. Data were normalized using the expression of the housekeeping gene *HPRT*. Experiments were performed in duplo and depicted results reflect the data obtained from two independent experiments; error bars represent the standard deviation based on the relative expression levels of *NTHL1* in all included samples.

Nature Genetics: doi:10.1038/ng.3287

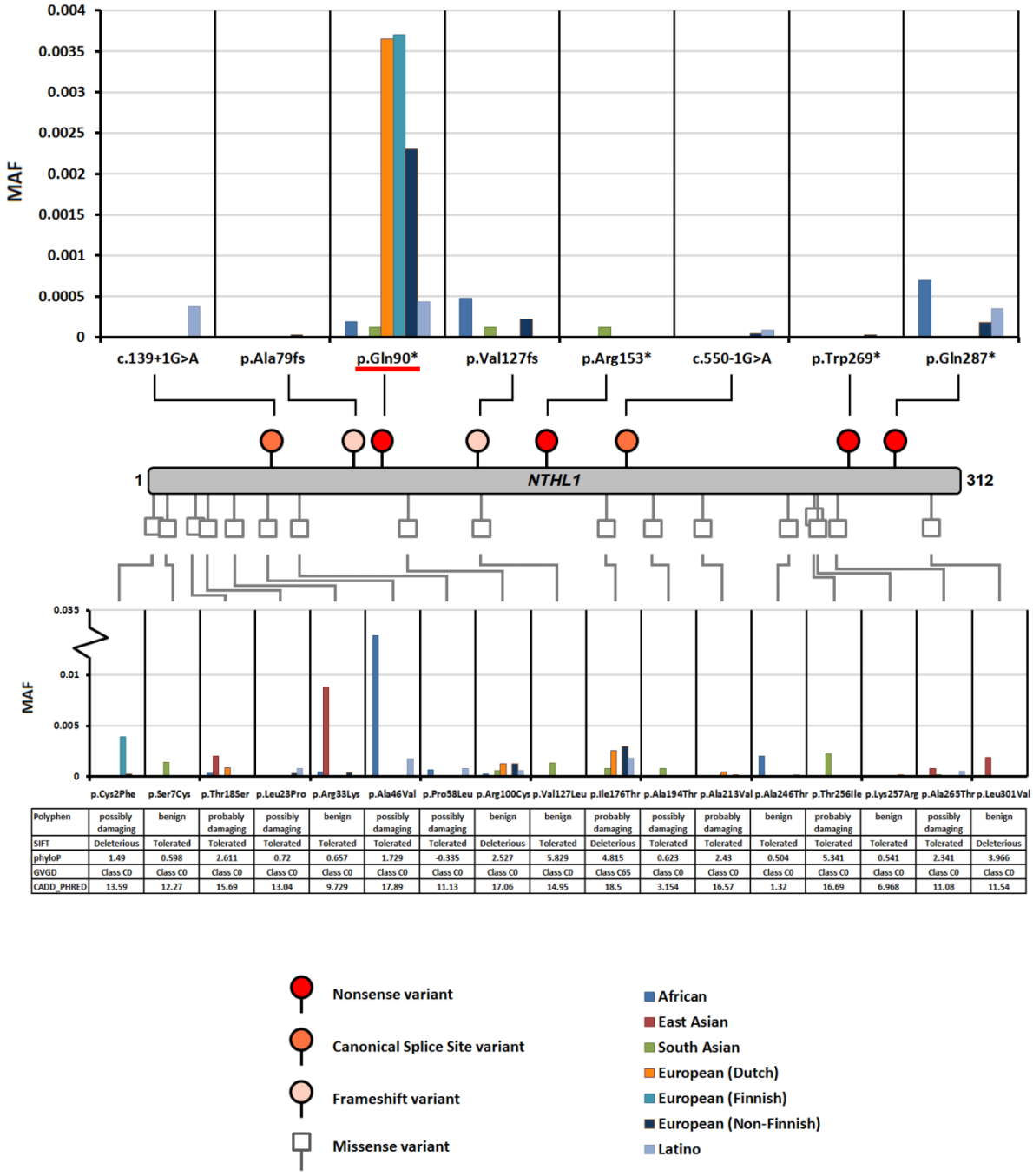


Supplementary Figure 2. Read depth analysis of exome data. Read depth of P07 (Top panel, red dots, SOLiD exome sequencing data) and P23 (Bottom panel, orange dots, Illumina exome sequencing data) compared to 7 and 40 *NTHL1* wild type samples, respectively (grey dots). Both P07 and P23 are homozygous around the *NTHL1* locus (blue bars; **Supplementary Fig. 6**), but median read-depth analysis of chromosome 16 shows that large genomic deletions encompassing this locus are not present.



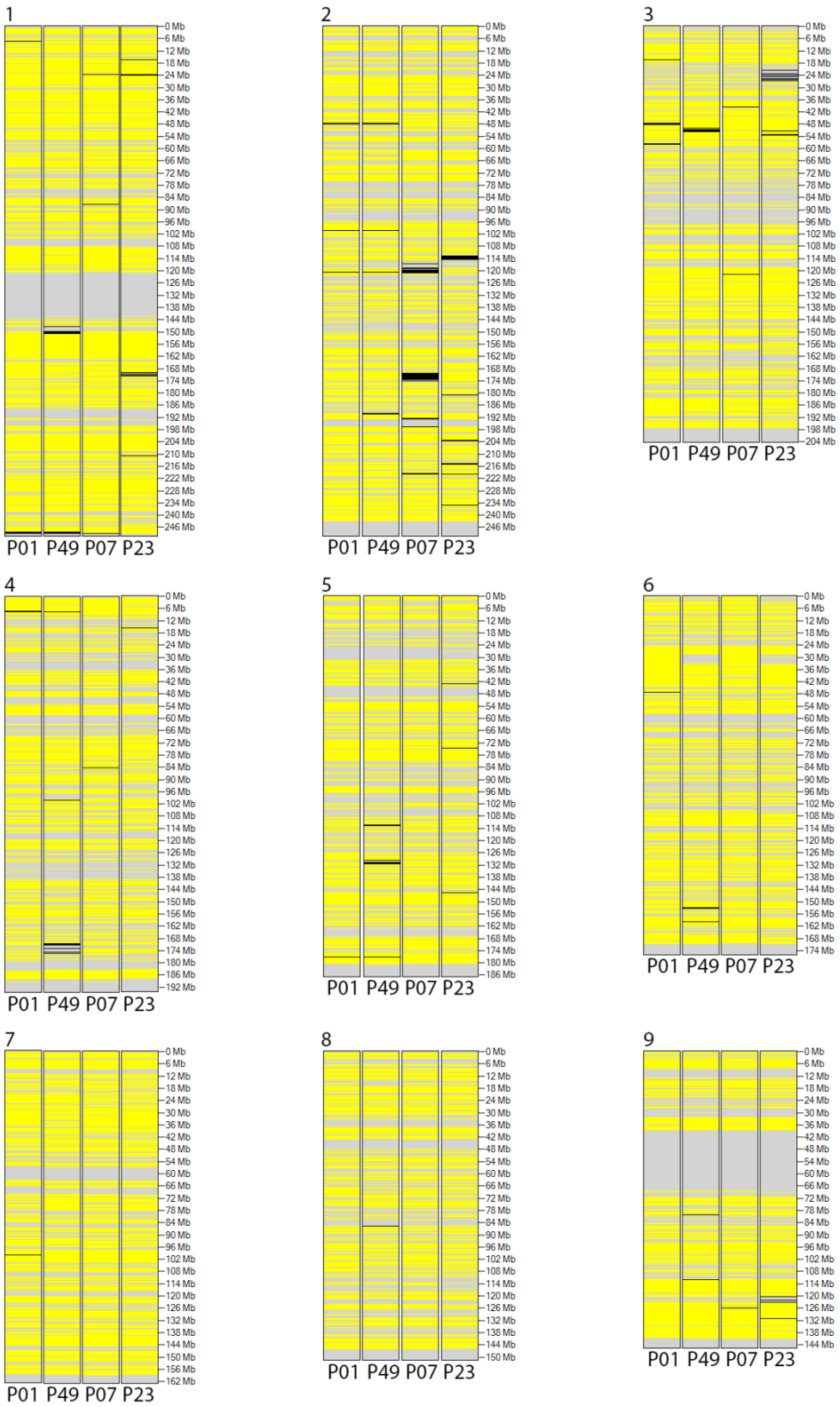
Supplementary Figure 3. Validation and co-segregation analysis by Sanger sequencing. The c.268C>T (p.Gln90*) variant in *NTHL1* was present in all clinically affected members of families A (P01 and P49), B (P07, P71 and P72), and C (p23 and P69). WT: wild type control negative for the c.268C>T variant in the *NTHL1* gene. Arrow: the change of guanine to adenine results in the introduction of a stop codon.

Nature Genetics: doi:10.1038/ng.3287

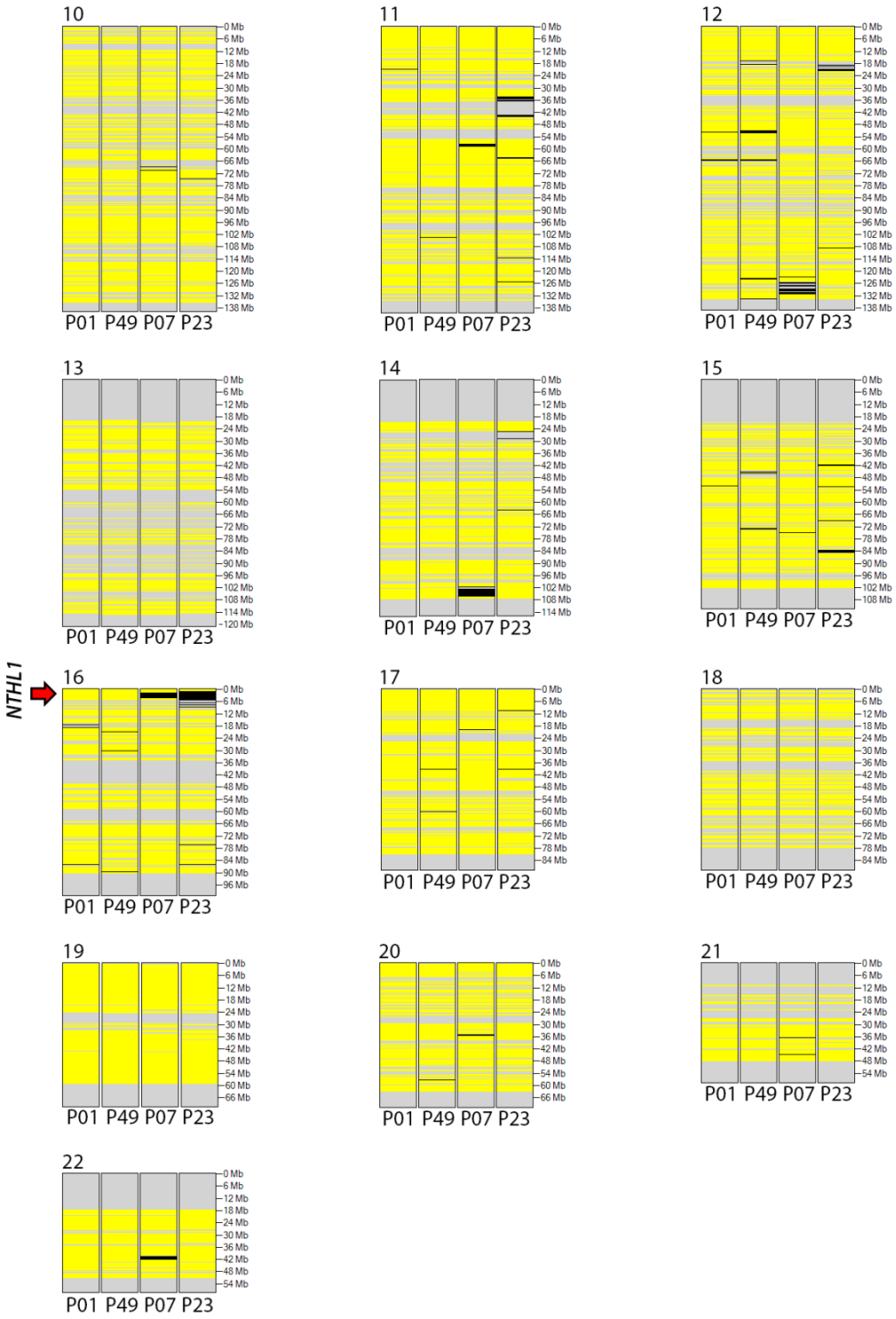


Polyphen	possibly damaging	benign	probably damaging	possibly damaging	benign	possibly damaging	possibly damaging	benign	benign	probably damaging	possibly damaging	probably damaging	benign	probably damaging	benign	possibly damaging	benign
SIFT	Deleterious	Tolerated	Tolerated	Tolerated	Tolerated	Tolerated	Tolerated	Deleterious	Tolerated	Deleterious	Tolerated	Tolerated	Tolerated	Tolerated	Tolerated	Tolerated	Deleterious
phyloP	1.49	0.598	2.611	0.72	0.657	1.729	-0.335	2.527	5.829	4.815	0.623	2.43	0.504	5.341	0.541	2.341	3.966
GVGD	Class C0	Class C0	Class C0	Class C0	Class C0	Class C0	Class C0	Class C0	Class C0	Class C65	Class C0	Class C0	Class C0	Class C0	Class C0	Class C0	Class C0
CADD_PHRED	13.59	12.27	15.69	13.04	9.729	17.89	11.13	17.06	14.95	18.5	3.154	16.57	1.32	16.69	6.968	11.08	11.54

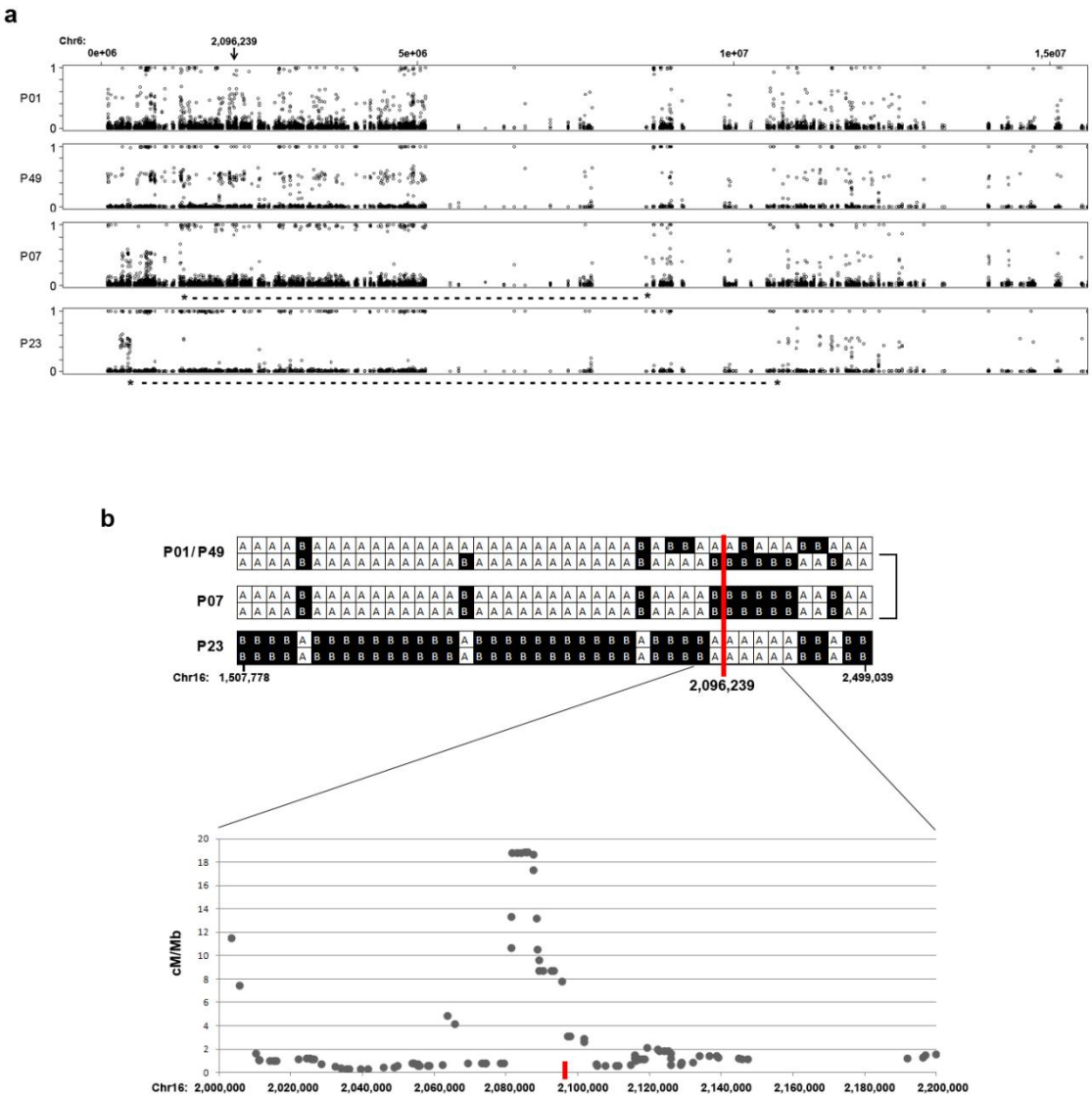
Supplementary Figure 4. Recurrent truncating and missense variants in the *NTHL1* gene and ethnic distributions. Upper part: eight truncating variants are recurrently (≥2 calls) encountered in our in-house database, or the database from the Exome Aggregation Consortium [Nov., 2014 accessed]. All variants were encountered in a heterozygous state. The p.Gln90* variant (red underlined) is present in subjects from different descent, but significantly enriched in European cohorts. Lower panel: seventeen missense variants were recurrently (≥10 calls) encountered in our in-house- or ExAC databases. Frequencies of these variants are depicted for the different populations and, based on in-silico predictions, most of these missense variants are predicted to be benign polymorphisms.



Supplementary Figure 5. Continued on next page.



Supplementary Figure 5. Genome-wide autozygosity mapping. Homozygosity/autozygosity mapping based on exome data from four patients (P01, P07, P23 and P49) from three families confirms homozygous stretches (black regions) encompassing the NTHL1 locus in P07 and P23 (red arrow). Overall, homozygous stretches were more abundant in P07 and P23 compared to P01 and P49 due to consanguinity (family B) or predicted consanguinity (family C) in these families (see supplementary table 7), but overlap of homozygous regions between P07 and P23 was only observed at the NTHL1 locus. Yellow regions represent heterozygous regions, numbers correspond to chromosomal numbering and y-axis defines the genomic position on the corresponding chromosome. For details, see online methods.



Supplementary Figure 6. Haplotype analysis around the *NTHL1* locus. (a) B-allele frequency plots extracted from the exome sequencing data of P01 and P49 (family A), P07 (family B), and P23 (family C). Both individuals from family A encompass heterozygosity around the *NTHL1* locus, whereas homozygous stretches encompassing the *NTHL1* locus were present in P07 and P23. Boundaries of the homozygous regions in P07 (~7 Mb) and P23 (~10 Mb) are indicated by asterisks connected by dashed lines. **(b)** Haplotype analysis of a 1-Mb genomic locus encompassing the *NTHL1* gene (red line). At least three different haplotypes were observed in four carriers of the p.Gln90* variant in the *NTHL1* gene. The presence of a recombination hotspot (recombination rate: 18.9 cM/Mb [17]) upstream of the *NTHL1* p.Gln90* locus (chr16:2,096,239) may explain the observed variability in haplotypes. Blocks represent informative SNPs in close proximity (1 Mb region) of the *NTHL1* p.Gln90* locus; (A) and (B) represent major and minor alleles, respectively.