Supplementary Methods

Allele-specific expression (ASE) and loss of heterozygosity (LOH) analysis

PCR fragments encompassing the c.1865T>A change were generated from genomic DNA (gDNA) and cDNA obtained from peripheral blood lymphocytes, or tumor DNA (tDNA) from mutation carriers (Table S1). PCR products were purified using GFX^{TM} PCR DNA and Gel Band Purification kit (GE Healthcare). Allele-specific expression (ASE) of the variant c.1865T>A of *MLH1* at the cDNA level was analyzed by single-nucleotide primer extension (SNuPE) using the SNaPshot kit (Applied Biosystems) with primer 5'-TCTGAAGAAGAAGGCTGAGATGC-3', according to the manufacturer's instructions. Briefly, reactions were performed in a total volume of 10 µL containing 1.5 µL purified PCR product, 5 µL SNaPshot Ready Reaction Mix, and 0.2 µmol/L extension primer. Primer extension thermocycling conditions consisted of 25 cycles of 96ºC for 10 s, 50ºC for 5 s, and 60ºC for 30 s. SNaPshot reaction products were treated with 1 U shrimp alkaline phosphatase (usb) for 60 min at 37ºC and then 15 min at 75ºC. Products were run in an ABI Prism 3130 DNA sequencer and were analyzed by GeneMapper v4.0 (Applied Biosystems). Samples from c.1865T>A carriers showed a profile with two peaks (green and red, representing A and T alleles, respectively). ASE at cDNA level was calculated as the proportion of T allele between cDNA and gDNA (ASE = $f(T)_{cDNA}$/ $f(T)_{gDNA}$), were $f(T)$ was obtained from peak intensities using the formula $f(T)$ = [T allele/(T+A alleles)]. Loss of heterozygosity (LOH) at tumor DNA (tDNA) level was measured as the proportion of T allele between tDNA and gDNA (LOH = $f(T)_{tDNA}$/ $f(T)_{gDNA}$). We used gDNA samples from 8 carriers to establish a range for normal ASE and LOH between 0.89 and 1.11 (mean values ± 3·SD). Experiments were performed in duplicate.

Details on estimation of mutation age

Counting recombination events

For the c.306+5G>A mutation we counted recombination events in the region spanning microsatellites D3S2369 and D3S1298; for c.1865T>A we chose the region spanning D3S1612 and D3S1298. We only counted recombinations on a subset of disease haplotypes corresponding to those in bold in Table 3 (for further discussion see the subsection below on estimating the genealogy

height). For each mutation, all diseased individuals had at least one allele in common in the immediate neighborhood of the disease allele, as would be expected if the mutation occurred on a single shared ancestral haplotype. We chose the most frequently occurring of the distinct disease haplotypes to be the ancestral one and counted the minimum number of recombination events necessary to obtain all of the other haplotypes. The choice of the ancestral disease haplotype does not affect the count of recombination events. When a haplotype was distinct from the ancestral disease haplotype we assumed that it resulted from a recombination, and we assumed that any two identical non-ancestral haplotypes arose from the same recombination event.

Estimating L from the count of recombination events

The length of the genealogy of the sampled copies of a mutation (L) was obtained by dividing the number of recombination events by the per-generation probability of a recombination on the ancestral haplotype. The recombination map length of the c.306+5G>A haplotype was obtained using the Rutgers Map Interpolator (1). We supplied the interpolator with the position of D3S1298, which was included in the Rutgers smoothed map position file, and we interpolated the position of D3S2369, which was not included. For the interpolation we used the physical position of 36,472,209 bp from release 50 of the Ensembl database (2). We used the difference between the sex-averaged map position of D3S2369 and that of D3S1298 to obtain a map length of r=0.9796 cM for the haplotype. The procedure for the c.1865T>A mutation was similar except that both endpoints of the haplotype appeared in the smoothed map position file, and the haplotype length was r=2.2578 cM.

Estimating the height of the genealogy of the sampled copies of each mutant allele

To estimate TMRCA given the estimated genealogy length $\hat{L}$, we simulated the joint probability density of genealogy heights and lengths, $f(H,L;N,n)$. For a constant-sized population under the coalescent, this joint density is determined by the effective population size N and the number n of sampled copies of the mutation (3). Although N is unknown, the ratio of the most likely tree length L_m and the most likely height H_m given this length depends only on n and can be expressed as $L_m/H_m=1/c(n)$ for some function c. Assuming that our genealogy

length estimate $\hat{L}$ is the most likely length under the true demographic history, we estimate TMRCA as $\hat{H} = c(n)\hat{L}$, where $c(n)$ was determined from coalescent simulations.

Our estimation procedure for $c(n)$ used 5×10^7 genealogies simulated under a constant population model with n=17 sampled lineages for the c.306+5G>A mutation and n=12 lineages for c.1865T>A. Simulations were carried out using the program ms (4). We estimated L_m from a histogram of the simulated tree lengths with 1,000 bins. Using a moving window of width 5 bins and step-size 1, we smoothed the histogram by averaging the values within each window, producing a vector of length 996. We then found the maximal element in this vector and averaged the centers of the bins corresponding to that element in order to estimate the location of the maximum of the distribution. Using this method, the estimated most likely gene tree length was $L_{m(\text{Ebro})}$ = 2.8031 in units of $4N$ generations ($L_{m(\text{Jaén})}$=2.3974).

To find H_m, we selected all simulated trees whose lengths fell within a window around L_m, [2.8005, 2.8057] for n=17 and [2.3947, 2.4001] for n=12. Each window was obtained by incrementally increasing the size of a symmetric window around L_m until the number of simulated genealogies with lengths in the window reached or exceeded 1,000. For each mutation the size of the increment was 2×10^{-4} units of $4N$ generations (10^{-4} units of $4N$ generations in each direction). To estimate H_m from these trees, we employed the same procedure used for estimating L_m, except that because of the smaller number of trees, the histogram for H used 100 bins and the moving window had width 3 and step-size 1. The most likely gene tree height for the given gene tree length was estimated to be $H_{m(\text{Ebro})}$=0.6514 in units of $4N$ generations ($H_{m(\text{Jaén})}$ = 0.6159). Dividing H_m by L_m yielded $\hat{c}_{\text{Ebro}}$=0.2324 and $\hat{c}_{\text{Jaén}}$=0.2569.

Note that although the number of sampled copies of the c.306+5G>A mutation is 42 (40 for the c.1865T>A mutation), because no two of the 17 families (12 families for c.1865T>A) are believed to share a common individual within the last five generations, in the simulations we used n=17 (n=12 for c.1865T>A). Thus, to estimate the TMRCA of the sampled copies of a mutation we chose one haplotype per family to represent the haplotype in the fifth generation from that family, estimated the height of the genealogy that relates these disease

haplotypes, and added four generations to the estimated height. The haplotype chosen was the one common to all family members, except in the single case of an intrafamilial recombination in the region considered. In this case, all individuals had a common ancestor in the fourth generation and the disease haplotype of this individual, assumed to be that of her two children, was chosen.

Confidence intervals for TMRCA

To obtain a 95% confidence interval for TMRCA for the c.306+5G>A mutation we simulated 100,000 trees under a coalescent of constant size with n=17 lineages (n=12 for c.1865T>A) and recorded the height H and length L of each tree. The height of each tree was then estimated from its length by H_{est}= cL and the ratio $h=H/H_{est}$ was computed for each tree. We found the 2.5 and 97.5 percentiles of these 100,000 ratios by ordering them from smallest to largest $h_{(i)}$ (i =1…100,000) and taking $h_{(2,501)}$ and $h_{(97,500)}$. The lower bound of the 95% confidence interval for the estimate of the TMRCA was computed as $h_{(2,501)}$ĉ$\hat{L}$+4 and the upper bound was computed as $h_{(97,500)}$ĉ$\hat{L}$+4.

Supplementary References

1. Matise TC, Chen F, Chen W*, et al.* A second-generation combined linkage physical map of the human genome. Genome Res 2007;17:1783-6.
2. Flicek P, Aken BL, Beal K*, et al.* Ensembl 2008. Nucleic Acids Res 2008;36:D707-14.
3. Wakeley J. Coalescent Theory: An Introduction. Greenwood Village, Colorado: Roberts & Company Publishers; 2008.
4. Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 2002;18:337-8.