Supplementary information S1 (box)

To provide a conservative assessment of narrative potential in a typical exome we used whole-exome sequence data from a 'healthy' control female with no reported disease diagnosis. In summary, the exome sequence data obtained 46.5-fold average coverage using the Agilent Human All Exon 50Mb Kit, on the Illumina HiSeq 2000 machine. To maintain a conservative illustration, all variants were additionally cross-examined in independent whole-genome sequence data generated for the same sample (42.9-fold average coverage). BWA (v0.5.10) was used to align all sequencing reads to the 1,000 Genomes Human reference sequence library (GRCh37p4, the rCRS mitochondrial sequence, Human herpesvirus 4 type 1, and decoy sequence derived from HuRef, Human Bacand Fosmid clones and NA12878), and the variant calls were produced using both samtools (v1.17) and GATK (v1.6-11). GATK was used for quality score recalibration and realignment [1]. To obtain the primary variant calling dataset, SNVs were called from the realigned exome kit capture HiSeq data using GATK with the whole-genome sequence data adopted as an independent resource to exclude potential sequencing SNV artifacts from the whole-exome sequence data. The same whole-exome sequence data was variant-called by samtools to exclude potential technical artifacts arising from alignment or variant calling algorithms. That is, while the GATK whole-exome sequence data was adopted as the primary variant calling data, only the consensus SNVs consistently detected across the three datasets (GATK-exome, samtools-exome, and GATK-genome) were considered for the "narrative potential" illustration. While indel calling has been notably improved using GATK indel realignment, due to the persisting inherent difficulties obtaining high positive predictive values for indel variant calls, we restrict our "disease potential" illustration to SNV data.

Sequence-based Quality Control Filters:

To maintain a conserved "narrative potential", we restricted the illustration to SNVs based on a battery of quality-control filters including: at least 10-fold coverage, with a phred-scaled probability that the variant call is not due to error of 30 (representing 1 in 1000 chance of error), and assigned a "PASS" status by GATK UnifiedGenotyper, which further facilitates filtering for machine and sequence artifacts. Some additional hard-call filters were enforced to ensure remaining SNVs had (a) a quality score versus depth filter, depth-normalized discovery confidence of at least 2, (b) a homopolymer run score less than six, (c) a low strand bias score, thus indicative of non-preferential strand bias for the alternative allele on one of the two possible read orientations, (d) haplotype scores less than 13, (e) mapping quality of at least 40, (f) mapping quality rank sum and read position rank sum test scores greater than -12.5 and -8, respectively. Moreover, as previously described, we restricted illustration to SNVs that were called in both independent whole-exome and whole-genome sequence data, and further SNVs had to be independently called by both samtools and GATK alignment and variant calling platforms. During this step we also applied an indel filter to exclude SNVs called within 3bp flanking ends of putative indel calls. As a final sequence-based quality control step, we leveraged off the NHLBI-Exome Sequencing Project (ESP) to exclude variants flagged as potentially problematic variant calls (SVM or indel5) in the NHLBI-Exome Sequencing Project (ESP).

Annotation-based Quality Control Filters:

Functional annotation of SNVs was performed using ensembl Ve!P 2.5 (release 67). Only annotations of likely-gene-disrupting nature (stop gain/loss and splice acceptor/donor sites), and missense (non-synonymous) were considered. Annotations for predicted NMD transcripts or non-coding genes were excluded for our illustrative purposes. Moreover, for this narrative potential illustration, we restricted our investigation to public CCDS transcripts (Release 9).[3] Variants were excluded if they were homozygous alternative allele for a base where the reference allele was observed at a reference allele frequency rate <1% in the internal control cohort.

Of the ‘qualifying’ nonsense (AF≤1% - Supplementary Table 1) and missense (AF≤1% and meeting damaging score in at least three of the four algorithms used for this narrative-potential illustration – Supplementary Table 2), none were situated within UCSC Genome Browser defined repeat masker regions of reported LINEs, SINEs, LTRs, DNA repeat elements, micro-satellites, low complexity repeats, satellite repeats, or other reported repeat types. Moreover, we compared all the listed sites in supplementary tables 1 and 2 to the chimpanzee genome and observed that the chimpanzee reference allele is consistent with the hg19 reference allele at the listed sites. Thus, none of the alternative alleles represent the expected ancestral allele.

1. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491-498 (2011).

2. Exome Variant Server, NHLBI Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/) August 2012

3. Pruitt, K. D. *et al*. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.* **19**, 1316-1323 (2009)