Germline mutations in shelterin complex genes are associated with familial chronic lymphocytic leukemia

Helen E. Speedy[1], Ben Kinnersley[1], Daniel Chubb[1], Peter Broderick[1], Philip J. Law[1], Kevin Litchfield[1], Sandrine Jayne[2], Martin J. S. Dyer[2], Claire Dearden[3], George A. Follows[4], Daniel Catovsky[5] and Richard S. Houlston[1,5,*]

[1]Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK; [2]Ernest and Helen Scott Haematological Research Institute, University of Leicester, Leicester, UK; [3]Department of Haemato-Oncology, Royal Marsden Hospital, Sutton, UK; [4]Department of Haematology, Cambridge University Hospitals NHS Foundation Trust, Addenbrooke's Hospital, Cambridge, UK; and [5]Division of Molecular Pathology, The Institute of Cancer Research, London, UK

Running title: *POT1, ACD, TERF2IP* associated with familial CLL

* Correspondence to Richard S. Houlston; Division of Genetics and Epidemiology, Institute of Cancer Research, Sutton, Surrey, SM2 5NG, UK; Tel: +44 (0) 208 722 4175; E-mail: richard.houlston@icr.ac.uk

Abstract word count: 142
Text word count: 3466
Number of Figures: 3 + 5 Supplemental
Number of Tables: 1 + 3 Supplemental
Number of References: 58

Scientific category: Lymphoid neoplasia

Key points

- Germline loss-of-function mutations in shelterin genes occur in a subset of families with CLL
- Telomere dysregulation is further implicated in CLL predisposition

Abstract

Chronic lymphocytic leukemia (CLL) can be familial, however thus far no rare germline disruptive alleles for CLL have been identified. We performed whole-exome sequencing of 66 CLL families, identifying four families where loss-of-function mutations in *POT1* co-segregated with CLL. The p.Tyr36Cys mutation is predicted to disrupt the interaction between POT1 and the telomeric overhang. The c.1164-1G>A splice-site, p.Gln358SerfsTer13 frameshift and p.Gln376Arg missense mutations are likely to impact the interaction between POT1 and ACD, part of the telomere-capping shelterin complex. We also identified mutations in *ACD* (c.752-2A>C) and another shelterin component, *TERF2IP* (p.Ala104Pro and p.Arg133Gln), in three CLL families. In a complementary analysis of 1,083 cases and 5,854 controls, the *POT1* p.Gln376Arg variant, which has a global minor allele frequency of 0.0005, conferred a 3.61-fold increased risk of CLL (P=0.009). This study further highlights telomere dysregulation as a key process in CLL development.

Introduction

Chronic lymphocytic leukemia (CLL; MIM151400) is clinically defined by the presence of a clonal population of B-cell lymphocytes (>5 $\times 10^9$ cells/L) with a characteristic immunophenotype. The disease accounts for approximately 25% of all leukemia and is the most common form of lymphoid malignancy in Western countries, affecting around 16,000 individuals in the United States each year[1]. While the last decade has seen a dramatic evolution in the treatment options for CLL[2-4], it still remains an incurable malignancy. It is anticipated that an increased understanding of CLL pathogenesis will generate further therapeutic targets to either delay or prevent progression of the precursor to frank malignancy.

CLL has one of the highest familial risks of any cancer, with risk being increased eight-fold in relatives of patients[5]. Recent genome-wide association studies (GWAS) have identified common risk single nucleotide polymorphisms (SNPs) at 31 loci associated with sporadic CLL[6-13]. The risk of CLL associated with each of these variants is however modest at best. While families segregating CLL provide evidence for Mendelian susceptibility, no rare alleles of large effect have thus far been discovered. The identification of this class of susceptibility is especially important since mutations are causal and provide direct insight to cancer biology, in contrast to GWAS associations.

Here we report on the whole exome sequencing of familial CLL and establish a key role for rare disruptive mutations in *POT1* and other shelterin complex genes as determinants of susceptibility to CLL. Our findings thus extend the spectrum of cancer types associated with germline mutation in these genes.

Materials and methods

Patient samples and DNA extraction

The families and CLL cases included in this study were recruited through a UK national study of CLL genetics established by the Institute of Cancer Research Divisions of Genetics and Epidemiology and Molecular Pathology in 1996. The diagnosis of CLL and other haematological cancers in family members were established. In all cases the diagnosis of CLL was based on accepted standard clinico-pathological and immunological criteria that are in accordance with current World Health Organization classification guidelines. Informed consent was obtained under MREC 99/1/082. Genomic DNA was extracted from peripheral blood and saliva using standard methods and quantified by PicoGreen (Invitrogen).

Pedigrees and clinical presentation

Supplemental Table 1 details all pedigrees, the number of cases of CLL in each family, and the number of cases that were whole-exome sequenced.

Sequence alignment and analysis

Exon capture was performed using the Nextera Rapid Capture Exome Enrichment kit (Illumina, San Diego, CA, USA). The Illumina HiSeq2000 analyser with 101 base pair reads was used for sequencing. Paired-end fastq files were extracted using CASAVA software (version 1.8.1, Illumina) and aligned to build 37 (hg19) of the human reference genome using Stampy[14] and BWA[15] software. Alignments were processed using the Genome Analysis Tool Kit (GATK) pipeline (version 3.2-2)[16], according to best practices[17,18]. Variants were filtered for positions found in >1 sample from an in-house collection of 1,609 control exomes including; 961 samples from the ICR1000 dataset generated by Professor Nazneen Rahman's Team in the Division of Genetics & Epidemiology at The Institute of Cancer Research, London[19] plus an extra 648 samples from the UK 1958 Birth Cohort (BC)[20], sequenced in-house using Illumina TruSeq exome methodology. We also filtered variants based on frequencies in the 1000 Genomes Project, National Heart, Lung and Blood Institute Exome Sequencing Project (ESP6500) and the Exome Aggregation Consortium (ExAC) catalog. Positions resulting in protein-altering changes were identified using the Ensembl Variant Effect Predictor (version 78) and variants shared between family members were annotated

using custom scripts. The predicted functional consequences of missense variants were assessed using SIFT[21], CADD[22] and SuSPect[23] algorithms.

Sanger Sequencing

Germline verification of variants found by next generation sequencing was performed by Sanger sequencing of mouthwash DNA samples. Primers are listed in Supplemental Table 2.

MaxEntScan scoring of splice acceptor variants

We used the MaxEntScan algorithm[24] to assess the effect of *POT1* g.124481233C>T and *ACD* g.67692984T>G mutations. Scores for the mutated splice acceptor site and wild-type splice site sequence were 5.66, -3.08 and 9.88, 1.84 respectively.

Confirmation of aberrant splicing in an individual carrying the splice acceptor variant 7:g.124481233C>T

RNA extracted from the whole blood of a splice acceptor variant carrier and control was converted to cDNA using Superscript III Reverse Transcriptase (Invitrogen). PCR was then performed to confirm that 7:g.124481233C>T disrupted splicing. The product was visualized on a 2.5% agarose gel. Sanger sequencing was used to confirm the sequence of the product.

Exome array genotyping

1,111 unrelated CLL cases were genotyped for the p.Gln376Arg variant using the Illumina OmniExpress Exome array as previously described[12]. After quality control filtering, genotype data were available for 1,083 CLL cases. For controls, we used publicly accessible data for 5,854 individuals from the 1958 Birth Cohort[20] genotyped using the Illumina HumanExome-12v1 array. These data are available from the European Genome-phenome Archive (EGA) under accession number EGAD00010000234. The chi-squared test was used to determine the significance of the difference in case-control allele counts. Confidence intervals were calculated by the Woolf method.

Protein alignment and structural modelling

Multiple sequence alignments were generated for homologous POT1 and TERF2IP protein sequences using T-Coffee[25,26] to evaluate conservation. POT1 alignments were generated with the following sequences: NP_056265.2, XP_519345.2, NP_001127526.1, XP_009001386.1,

XP_006149256.1, NP_598692.1, XP_002712135.2, XP_010802750.1, XP_005628494.1, XP_001501458.4, XP_006910616.1, XP_010585693.1, XP_004478311.1, XP_007504310.1, XP_001508179.2, NP_996875.1 and NP_001084422.1. TERF2IP alignments were generated with the following sequences: NP_061848.2, NP_001267142.1, XP_003780774.2, XP_008984478.1, XP_006152679.1, NP_065609.2, XP_002711780.1, NP_001068880.1, XP_536776.2, XP_005608497.1, XP_006908867.1, XP_010595146.1, XP_004470975.1, XP_001508762.2, NP_989799.1 and NP_001084428.1. Jalview[27] was used to visualize and format the alignments. The crystal structure of the N-terminal region (OB1 and OB2 domains) of the human POT1 protein (RCSB PDB, 3KJP and 1XJV) was visualized using Chimera (version 1.10.2)[28] and Cn3D (version 4.3.1)[29]. The impact of the p.Tyr36Cys mutation on stability of the POT1:DNA interaction was assessed using mCSM[30]. The effect of missense mutations on protein stability was assessed using INPS[31].

Loss-of-heterozygosity analyses

Loss-of-heterozygosity (LOH) analysis was conducted using ExomeCNV[32] which detects copy number variation (CNV) and LOH events using depth-of-coverage and B-allele frequencies. LOH calls were made by first identifying all heterozygous germline positions. GATK was then used to create BAF files and ExomeCNV used to call LOH at heterozygous positions individually and at combined LOH segments.

Assessment of telomere length

Relative telomere length was determined by two methods: using exome sequencing data and with real-time PCR. Analysis of off-target reads from exome sequencing data was performed essentially as described[33], using a telomeric repeat copy number of k=4. We used data from blood-derived DNA only and also excluded samples with average sequencing depth <20 and with missing covariate data (n=12). Telomere length was adjusted for age at blood draw, sex and sequencing batch by a linear model determined using data from non-carriers only. For the SYBR green real-time PCR, the ratio of telomere repeat units to a single-copy gene (β-globin) for 109 samples, was determined as previously described[34,35]. Primers are listed in Supplementary Table 2. Reactions were performed in triplicate, using 10ng DNA per sample. Each 10µL reaction also contained 5µL of 2X SYBR Green Master Mix (Applied Biosystems) plus either 300 and 700 nmol/L of the control forward and reverse primers, respectively or 100 and 900 nmol/L of the telomere unit forward and reverse primers, respectively. The telomere reaction also included 0.3µL DMSO. Cycling was

performed using an ABI7900HT thermal cycler as previously described[35]. Relative telomere length was calculated using $2^{-\Delta Ct}$ derived from the real-time PCR data and was adjusted for age at blood draw and sex. A Wilcoxon rank-sum test was used to compare the relative adjusted telomere length for *POT1* mutation carriers versus non-carriers of shelterin gene mutations.

Results

Identification of shelterin gene mutations

To maximise the prospects of identifying rare disease-causing variants for CLL we initially focussed our search on the 18 families with the strongest family histories of CLL (Supplemental Table 1) which had been ascertained through an ongoing study[36]. We performed whole exome sequencing on genomic DNA from blood of 45 affected individuals from the 18 families. We excluded variants that were observed more than once in our in-house database of 1,609 healthy individuals from the 1958 BC who had been exome sequenced. We also discounted variants with an allele frequency of greater than 0.1% in large-scale sequencing projects (1000 Genomes Project, ESP6500 or the ExAC catalog). In our first stage analysis we required the filtered variants to be present in all sequenced affecteds within the family.

To further filter the variants identified, we prioritised missense and disruptive variants (nonsense, splice acceptor/donor and frameshift) occurring in genes with a reported cancer association or documented role in cancer predisposition. Analysis of these genes led us to identify pedigree 5047 in which all three affected family members carried a splice acceptor variant in intron 13 (chromosome 7 g.124481233C>T/c.1164-1G>A) of *POT1* (protection of telomeres 1; MIM 606478) (Figure 1 and 2a). We confirmed the mutation by Sanger sequencing in blood and saliva derived DNA in all three cases (Supplemental Figure 1). The mutation was predicted to disrupt splicing by the MaxEntScan algorithm[24] (wild-type score 5.66 versus mutated score -3.08, 154% reduction). This also identified a potential alternative splice acceptor site 43bp downstream (MaxEntScan score=7.66), the use of which would result in a truncated protein product. We confirmed the presence of an aberrant splicing product in a mutation carrier by RT-PCR and validated the use of the predicted alternative splice site using Sanger sequencing (Figure 3, Supplemental Figure 2).

To further investigate the potential role of *POT1* and other members of the shelterin gene complex in familial CLL, we expanded our exome sequencing dataset to include an additional 96 affected relative-pairs from 48 families (Supplemental Table 1). We then looked for shared missense and disruptive variants in the six components of the shelterin complex (*POT1*, *ACD*, *TINF2*, *TERF1*, *TERF2* and *TERF2IP*). Through this analysis we identified three additional families that harbored *POT1* mutations (Figure 1); two missense mutations, p.Tyr36Cys and p.Gln376Arg, occurring at evolutionarily conserved residues (Figure 2b), predicted *in silico* to be damaging by

multiple algorithms, and a frameshift mutation (Table 1). Collectively, we therefore identified mutations in *POT1* in 6% of the CLL families, as compared with the documented frequency of such variants of only 0.9% amongst the 60,706 individuals included in the ExAC catalog (*P*=0.003).

Intriguingly, somatic mutations of residue Tyr36 have previously been reported in CLL (Figure 2a)[37-39]. The p.Gln376Arg variant, identified in pedigree 4013, has a global minor allele frequency (MAF) of 0.0005 in the ESP6500 database and is included on the Illumina Exome array. We therefore initiated a genetic association study of this recurrent variant making use of Illumina exome array data on 1,083 unselected CLL cases and 5,854 1958 BC controls. Six of the cases and nine of the controls were heterozygous for the p.Gln376Arg variant (odds ratio, OR=3.61, 95% confidence interval: 1.28-10.15, *P*=0.009).

In addition to *POT1* mutations, we identified mutations in other shelterin complex genes in families 233, 4092 and 4014. Specifically, the *ACD* (adrenocortical dysplasia protein homolog, MIM 609377) splice site variant c.752-2A>C was carried by both affected siblings in pedigree 233 (Figure 1, Supplemental Figure 3a) and was predicted by the MaxEntScan algorithm to disrupt the exon 7 splice acceptor signal (wild-type score 9.88 versus mutated score 1.84, 81% reduction, Supplemental Figure 3b). In *TERF2IP* (telomeric repeat binding factor 2, interacting protein, MIM 605061), the missense mutation c.398G>A (p.Arg133Gln) was identified in two out of three CLL cases sequenced in family 4014 (Figure 1, Supplemental Figure 4a). This mutation occurs at an evolutionarily conserved site and was predicted to be damaging by multiple methods (Table 1, Supplemental Figure 4b). We also found the c.310G>C (p.Ala104Pro) *TERF2IP* variant in both siblings in family 4092 (Figure 1, Table 1, Supplemental Figure 4a). While this residue is partially conserved, the p.Ala104Pro mutation is not predicted to be damaging by SIFT or SuSPect (Table 1, Supplemental Figure 4b).

Structural predictions

The POT1 N-terminus contains two oligonucleotide/oligosaccharide binding (OB) folds that bind to the single-stranded telomeric overhang (Figure 2) while the C-terminus is responsible for binding to ACD and anchoring the shelterin complex. The crystal structure of human POT1 has been resolved for only the N-terminal OB-folds (Protein Data Bank (PDB) 3KJP and 1XJV). Based upon these structures, Tyr36 is one of 24 residues found at the POT1:telomeric polynucleotide

interface[39] (Figure 2). The p.Tyr36Cys mutation is predicted by mCSM to reduce the POT1:DNA complex affinity (PDB 1XJV, ΔΔG -0.27Kcal/mol; PDB 3KJP, predicted ΔΔG -0.21Kcal/mol).

Since crystal structures for full-length POT1 and TERF2IP are lacking, we used the machine learning algorithm INPS to predict the thermodynamic change in free energy caused by the p.Gln376Arg (POT1), p.Ala104Pro and p.Arg133Gln (TERF2IP) mutations, based upon the protein sequence (Supplemental Table 3). Using this method p.Arg133Gln was predicted to have the largest effect upon protein stability.

Analysis of somatic events

We used ExomeCNV to look for evidence of LOH in the proband of pedigree 5047, comparing exome sequencing data from blood-derived DNA to saliva-derived DNA, finding no evidence of a somatic abnormality at the *POT1* locus. We also looked for deleterious variants identified only in the blood-derived DNA of this case (i.e. absent from the saliva-derived DNA sample and also absent from the other affected individuals in pedigree 5047) and found no somatic inactivating *POT1* mutations. We did however note the presence of a somatic splice donor site mutation affecting the first base of intron 10 of *ATR* (ataxia-telangiectasia and rad3-related) in this case.

Effect of *POT1* mutations on maintenance of telomere length

Given the role of the shelterin complex in telomere length maintenance we examined whether CLL cases from shelterin-mutated pedigrees had telomere lengths that differed from non-carrier CLL cases using exome sequencing and real-time PCR data. We observed no consistent significant difference between the telomere lengths of *POT1* mutation carriers and CLL cases without a mutation in a shelterin complex gene, by exome sequencing or real-time PCR (P=0.03/P=0.57, respectively). The telomere lengths of cases with *ACD* or *TERF2IP* variants also displayed no obvious trend, although the small numbers of cases harbouring these variants precluded a meaningful evaluation of their impact on telomere length.

Discussion

Here we have implemented whole exome sequencing to search for rare disruptive risk alleles for CLL, identifying germline-inactivating shelterin gene mutations in a subset of CLL families. These findings are consistent with the evidence of linkage of familial CLL to chromosomes 7q31.32-q33 and 16q12.2-q23.1 that we previously observed (Supplemental Figure 5)[40].

Germline disruptive variants within shelterin genes have recently been implicated in predisposition to familial melanoma[41,42], cardiac angiosarcoma[43], glioma[44] and colorectal cancer[45], whilst somatic mutations of *POT1* are detectable in 3.5% of all CLL and 9% of *IGHV* (encoding immunoglobulin heavy chain variable)-unmutated CLL[39] and were also identified in 10% of patients with cutaneous T cell lymphoma[46].

POT1-mutated CLL cells have numerous telomeric and chromosomal abnormalities, suggesting that *POT1* mutation facilitates the acquisition of these malignant features[39]. Our observation of germline mutations in *POT1* being associated with familial CLL would concur with this assumption. Our findings also support the proposal that *POT1* mutation is an early event in CLL development[37]
In a CLL GWAS we previously reported an association between the common allele of the *POT1* 3'UTR variant rs17246404 (risk allele frequency=0.75) and increased CLL risk, with a small per allele effect size (OR=1.22)[12]. The recurrent *POT1* coding variant, p.Gln376Arg, identified in the current study is not however in linkage disequilibrium with SNP rs17246404 (r^2=0.00). Therefore, although further studies are required to determine exactly how rs17246404 influences CLL risk, it is plausible that the functional basis of the association is through differential gene expression.

Shelterin is a telomere-specific protein complex composed of six family members, encoded by *POT1, ACD, TERF2IP, TERF1* (encoding telomeric repeat binding factor 1), *TERF2* (encoding telomeric repeat binding factor 2) and *TINF2* (encoding TERF1 interacting nuclear factor 2), that protects the ends of chromosomes. Together, the components of the shelterin complex are necessary for all telomere functions, including the protection of telomeres from degradation, aberrant recombination and incorrect processing by DNA-repair machinery as well as facilitating chromosome capping to mediate telomerase activity[47].

POT1 directly contacts telomeric DNA overhangs[48] and also binds to ACD[49] which connects POT1 to the other shelterin components via its bridge with TINF2[50]. The POT1:ACD interaction enhances the affinity of POT1 for telomeric DNA[49,51]. The *ACD* splice site mutation c.752-2A>C, will disrupt the POT1 binding domain and abolish the TINF2 binding domain, so would therefore be predicted to result in an unformed shelterin complex.

In silico predictions suggest that the germline p.Tyr36Cys mutation identified in pedigree 162 is likely to disrupt the interaction between POT1 and the single-stranded telomeric DNA overhang. The *POT1* frameshift and splice site mutation are likely to result in truncated protein products, impairing their interaction with ACD. The p.Gln376Arg variant, though not predicted *in silico* to impact protein stability, alters an evolutionarily constrained residue thus implying functional importance.

Previous experiments have shown that when the ACD/POT1 subunit is inhibited, the telomerase complex increases telomere length[48,50]. We observed no significant differences between the telomere lengths of CLL cases with a *POT1* mutation and those who did not harbour a shelterin gene mutation. This observation is comparable to that in tumor cells derived directly from CLL cases with a somatic *POT1* mutation versus matched cases with no *POT1* mutation[39] and may reflect the numerous unmeasured variables that can influence telomere length in human populations. We also acknowledge that our telomere length measurements are based on blood-derived DNA and therefore could be subject to the effects of uncharacterized somatic mutations.
In this regard, we note that the proband of pedigree 5047 harbored a somatic splice site mutation in ATR, a gene also know to play a key role in telomere maintenance[52]. Furthermore, whilst GWAS have identified SNPs at loci including other telomere maintenance genes that are associated with telomere length, there has been no such association reported for a *POT1* SNP[12,53,54].

TERF2IP associates with the shelterin complex via its C-terminus to a central region of TERF2, forming a stable 1:1 complex. TERF2IP, as part of the shelterin complex, is vital for the repression of homology-directed repair of double strand chromosomal break at the telomere. While the novel missense variant p.Arg133Gln is predicted to be pathogenic, markedly reducing the stability of the protein, p.Ala104Pro is less well conserved and is thus more likely to be tolerated[40].

Germline disruptive mutations in *POT1* have previously been associated with susceptibility to melanoma in nine families[41,42] and glioma in three families[44]. Furthermore, recent studies have identified *POT1* p.Arg117Cys in four Li-Fraumeni-like syndrome families[43], and *ACD* and *TERF2IP* mutations in eight melanoma families[55]. None of the mutation carriers in the melanoma families featured cases of glioma or CLL. Similarly, the glioma families did not feature cases of melanoma or CLL and the only case of melanoma was seen in one of the Li-Fraumeni-like syndrome families. Collectively these data, and the fact that none of our families segregated glioma or melanoma, suggest that the penetrance associated with rare shelterin complex mutations is modest. Such an assertion is supported by our observation that the predicted deleterious p.Arg133Gln *TERF2IP* variant was identified in only two out of three CLL cases sequenced in family 4014. Additionally, in our case-control analysis, the *POT1* p.Gln376Arg mutation was shown to confer a modest 3.6-fold increase in risk of CLL. Furthermore, the absence of significant loss-of-heterozygosity in the tumors of carriers, when examined[43], suggests mutations in the shelterin complex genes do not function as high penetrance tumor suppressors, but rather moderate penetrance alleles.

Early age of onset in cancer can be indicative of inherited predisposition and it is noteworthy that in this study, mutation carriers were diagnosed with CLL much younger than the population average (59 years as compared to 71 years). Since seven of the 66 CLL families were carriers of shelterin mutations, this translates to 11% of familial CLL being ascribed to mutations in this class of genes (95% confidence interval, 4%–21%). However, we acknowledge that our analyses were based only on families ascertained in the UK and therefore the impact of such mutations on familial CLL could vary depending on ethnicity. Moreover, it remains to be established, through additional studies, whether other CLL families are the consequence of polygenic susceptibility or as yet unidentified higher impact disease-causing mutations.

In summary, the *POT1, ACD* and *TERF2IP* loss-of-function mutations we report here suggest that multiple components of the shelterin complex play a role in CLL predisposition. Moreover, they extend the spectrum of cancer associated with inherited mutations in these genes. It is however, likely that shelterin complex gene mutations confer cancer risks analogous to those associated with *ATM* heterozygosity[56] or *CHEK2*[57] for breast cancer. Nevertheless, since the dysregulation of telomere protection has been identified as a target for potential therapeutic intervention in CLL, it may be possible that early identification of mutation carriers will facilitate improvements in future disease management.

Acknowledgements

Bloodwise provided principal funding for the study (LRF05001, LRF06002, LRF13044). We acknowledge support from Cancer Research UK (C1298/A8362 supported by the Bobby Moore Fund), the Arbib Fund and the Leicester Experimental Cancer Medicine Centre (C325/A15575 Cancer Research UK/UK Department of Health). BK received a PhD studentship from The Institute of Cancer Research, supported by the Sir John Fisher Foundation. The study made use of genotyping data on the 1958 Birth Cohort; a full list of the investigators who contributed to the generation of these data is available at http://www.wtccc.org.uk/. We are grateful to all investigators and all the patients and individuals for their participation.

Author contributions

HES and RSH drafted the manuscript; HES performed project management, sequencing and bioinformatic analysis; BK, DC, PJL and KL performed bioinformatic analysis; PB performed sample preparation; SJ performed sample database management; CD, MJSD, GAF and DC performed sample recruitment; RSH obtained financial support.

Conflict of interest disclosure statement

The authors declare no competing financial interest.

References

1. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2012. *CA Cancer J Clin*. 2012;62(1):10-29.
2. Byrd JC, Brown JR, O'Brien S, et al. Ibrutinib versus ofatumumab in previously treated chronic lymphoid leukemia. *N Engl J Med*. 2014;371(3):213-223.
3. Cartron G, de Guibert S, Dilhuydy MS, et al. Obinutuzumab (GA101) in relapsed/refractory chronic lymphocytic leukemia: final data from the phase 1/2 GAUGUIN study. *Blood*. 2014;124(14):2196-2202.
4. Furman RR, Sharman JP, Coutre SE, et al. Idelalisib and rituximab in relapsed chronic lymphocytic leukemia. *N Engl J Med*. 2014;370(11):997-1007.
5. Goldin LR, Bjorkholm M, Kristinsson SY, Turesson I, Landgren O. Elevated risk of chronic lymphocytic leukemia and other indolent non-Hodgkin's lymphomas among relatives of patients with chronic lymphocytic leukemia. *Haematologica*. 2009;94(5):647-653.
6. Berndt SI, Skibola CF, Joseph V, et al. Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. *Nat Genet*. 2013;45(8):868-876.
7. Crowther-Swanepoel D, Broderick P, Di Bernardo MC, et al. Common variants at 2q37.3, 8q24.21, 15q21.3 and 16q24.1 influence chronic lymphocytic leukemia risk. *Nat Genet*. 2010;42(2):132-136.
8. Di Bernardo MC, Crowther-Swanepoel D, Broderick P, et al. A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. *Nat Genet*. 2008;40(10):1204-1210.
9. Sava GP, Speedy HE, Di Bernardo MC, et al. Common variation at 12q24.13 (OAS3) influences chronic lymphocytic leukemia risk. *Leukemia*. 2015;29(3):748-751.
10. Slager SL, Rabe KG, Achenbach SJ, et al. Genome-wide association study identifies a novel susceptibility locus at 6p21.3 among familial CLL. *Blood*. 2011;117(6):1911-1916.
11. Slager SL, Skibola CF, Di Bernardo MC, et al. Common variation at 6p21.31 (BAK1) influences the risk of chronic lymphocytic leukemia. *Blood*. 2012;120(4):843-846.
12. Speedy HE, Di Bernardo MC, Sava GP, et al. A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia. *Nat Genet*. 2014;46(1):56-60.
13. Berndt SI, Camp NJ, Skibola CF, et al. Meta-analysis of genome-wide association studies discovers multiple loci for chronic lymphocytic leukemia. *Nat Commun*. 2016;7:10933.
14. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res*. 2011;21(6):936-939.
15. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760.
16. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303.
17. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43(5):491-498.
18. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;43:11.10.11-11.10.33.
19. Ruark E, Münz M, Renwick A, et al. The ICR1000 UK exome series: a resource of gene variation in an outbred population. *F1000 Research*. 2015(4).
20. Power C, Elliott J. Cohort profile: 1958 British birth cohort (National Child Development Study). *Int J Epidemiol*. 2006;35(1):34-41.

21. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*. 2003;31(13):3812-3814.
22. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46(3):310-315.
23. Yates CM, Filippis I, Kelley LA, Sternberg MJ. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. *J Mol Biol*. 2014;426(14):2692-2701.
24. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol*. 2004;11(2-3):377-394.
25. Di Tommaso P, Moretti S, Xenarios I, et al. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res*. 2011;39(Web Server issue):W13-17.
26. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*. 2000;302(1):205-217.
27. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics*. 2009;25(9):1189-1191.
28. Pettersen EF, Goddard TD, Huang CC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem*. 2004;25(13):1605-1612.
29. Wang Y, Geer LY, Chappey C, Kans JA, Bryant SH. Cn3D: sequence and structure views for Entrez. *Trends Biochem Sci*. 2000;25(6):300-302.
30. Pires DE, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*. 2014;30(3):335-342.
31. Fariselli P, Martelli PL, Savojardo C, Casadio R. INPS: predicting the impact of non-synonymous variations on protein stability from sequence. *Bioinformatics*. 2015;31(17):2816-2821.
32. Sathirapongsasuti JF, Lee H, Horst BA, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics*. 2011;27(19):2648-2654.
33. Ding Z, Mangino M, Aviv A, Spector T, Durbin R, Consortium UK. Estimating telomere length from whole genome sequence data. *Nucleic Acids Res*. 2014;42(9):e75.
34. Cawthon RM. Telomere measurement by quantitative PCR. *Nucleic Acids Res*. 2002;30(10):e47.
35. Pooley KA, Sandhu MS, Tyrer J, et al. Telomere length in prospective and retrospective cancer case-control studies. *Cancer Res*. 2010;70(8):3170-3176.
36. Sellick GS, Webb EL, Allinson R, et al. A high-density SNP genomewide linkage scan for chronic lymphocytic leukemia-susceptibility loci. *Am J Hum Genet*. 2005;77(3):420-429.
37. Landau DA, Carter SL, Stojanov P, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell*. 2013;152(4):714-726.
38. Landau DA, Tausch E, Taylor-Weiner AN, et al. Mutations driving CLL and their evolution in progression and relapse. *Nature*. 2015.
39. Ramsay AJ, Quesada V, Foronda M, et al. POT1 mutations cause telomere dysfunction in chronic lymphocytic leukemia. *Nat Genet*. 2013;45(5):526-530.
40. Sellick GS, Goldin LR, Wild RW, et al. A high-density SNP genome-wide linkage search of 206 families identifies susceptibility loci for chronic lymphocytic leukemia. *Blood*. 2007;110(9):3326-3333.
41. Robles-Espinoza CD, Harland M, Ramsay AJ, et al. POT1 loss-of-function variants predispose to familial melanoma. *Nat Genet*. 2014;46(5):478-481.
42. Shi J, Yang XR, Ballew B, et al. Rare missense variants in POT1 predispose to familial cutaneous malignant melanoma. *Nat Genet*. 2014;46(5):482-486.
43. Calvete O, Martinez P, Garcia-Pavia P, et al. A mutation in the POT1 gene is responsible for cardiac angiosarcoma in TP53-negative Li-Fraumeni-like families. *Nat Commun*. 2015;6:8383.

44. Bainbridge MN, Armstrong GN, Gramatges MM, et al. Germline mutations in shelterin complex genes are associated with familial glioma. *J Natl Cancer Inst*. 2015;107(1):384.
45. Chubb D, Broderick P, Dobbins SE, et al. Rare disruptive mutations and their contribution to the heritable risk of colorectal cancer. *Nat Commun*. 2016;7:11883.
46. Pinzaru AM, Hom RA, Beal A, et al. Telomere Replication Stress Induced by POT1 Inactivation Accelerates Tumorigenesis. *Cell Rep*. 2016;15(10):2170-2184.
47. de Lange T. Shelterin: the protein complex that shapes and safeguards human telomeres. *Genes Dev*. 2005;19(18):2100-2110.
48. Loayza D, De Lange T. POT1 as a terminal transducer of TRF1 telomere length control. *Nature*. 2003;423(6943):1013-1018.
49. Xin H, Liu D, Wan M, et al. TPP1 is a homologue of ciliate TEBP-beta and interacts with POT1 to recruit telomerase. *Nature*. 2007;445(7127):559-562.
50. Ye JZ, Hockemeyer D, Krutchinsky AN, et al. POT1-interacting protein PIP1: a telomere length regulator that recruits POT1 to the TIN2/TRF1 complex. *Genes Dev*. 2004;18(14):1649-1654.
51. Wang F, Podell ER, Zaug AJ, et al. The POT1-TPP1 telomere complex is a telomerase processivity factor. *Nature*. 2007;445(7127):506-510.
52. Tong AS, Stern JL, Sfeir A, et al. ATM and ATR Signaling Regulate the Recruitment of Human Telomerase to Telomeres. *Cell Rep*. 2015;13(8):1633-1646.
53. Codd V, Nelson CP, Albrecht E, et al. Identification of seven loci affecting mean telomere length and their association with disease. *Nat Genet*. 2013;45(4):422-427.
54. Pooley KA, Bojesen SE, Weischer M, et al. A genome-wide association scan (GWAS) for mean telomere length within the COGS project: identified loci show little association with hormone-related cancer risk. *Hum Mol Genet*. 2013;22(24):5056-5064.
55. Aoude LG, Pritchard AL, Robles-Espinoza CD, et al. Nonsense mutations in the shelterin complex genes ACD and TERF2IP in familial melanoma. *J Natl Cancer Inst*. 2015;107(2):pii: dju408.
56. Renwick A, Thompson D, Seal S, et al. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat Genet*. 2006;38(8):873-875.
57. Meijers-Heijboer H, van den Ouweland A, Klijn J, et al. Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat Genet*. 2002;31(1):55-59.
58. Puente XS, Bea S, Valdes-Mas R, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature*. 2015;526(7574):519-524.

Table 1. Germline mutations in shelterin complex genes identified in CLL pedigrees

	Mutation position						Effect predictions			
Gene	Genomic (hg19)	cDNA[a]	Protein	Variant type	Pedigree	Carriers[b]	CADD[c]	SIFT	SuSPect[d]	GERP
POT1	7:g.124481233C>T	c.1164-1G>A	na	Splice acceptor	5047	3 / 3 / 3	17.16	na	na	4.67
POT1	7:g.124532337T>C	c.107A>G	p.Tyr36Cys	Missense	162	2 / 2 / 2	18.9	Deleterious	71	5.71
POT1	7:g.124482952_124482953insA	c.1071_1072insT	p.Gln358SerfsTer13	Frameshift	4029	2 / 2 / 4	na	na	na	4.71
POT1	7:g.124482897T>C	c.1127A>G	p.Gln376Arg	Missense	4013	2 / 2 / 3	23	Deleterious	50	5.55
ACD	16:g.67692984T>G	c.752-2A>C	na	Splice acceptor	233	2 / 2 / 2	19.35	na	na	5.15
TERF2IP	16:g.75682090G>C	c.310G>C	p.Ala104Pro	Missense	4092	2 / 2 / 3	10.15	Tolerated	10	2
TERF2IP	16:g.75682178G>A	c.398G>A	p.Arg133Gln	Missense	4014	2 / 3 / 3	14.88	Deleterious	73	5.34

na = not applicable; GERP= Genomic Evolutionary Rate Profiling score

[a] *POT1* reference transcript is NM_015450 and *ACD* reference transcript is NM_001082486

[b] Carriers given as number of familial cases with mutation / number of cases in family exome sequenced / total number of CLL cases in family

[c] CADD Phred-like score

[d] Scores of 50 and above considered to indicate deleterious mutations

Figure legends

Figure 1. Rare *POT1*, *ACD* and *TERF2IP* mutations in chronic lymphocytic leukemia families. Black filled symbols indicate CLL cases, other cancers are indicated by a blue filled symbol and an unfilled symbol indicates an individual with no known cancer. These symbols have a central dot to indicate cases who were exome sequenced. A central orange dot denotes a shelterin gene mutation carrier; a pink dot denotes a wild-type individual. A line through a symbol indicates that an individual is deceased. Age of diagnosis (in years) is listed for CLL cases. Splice acceptor variants are numbered relative to *POT1* transcript NM_015450 and *ACD* transcript NM_001082486. NHL= non-Hodgkin lymphoma.

Figure 2. Impact of rare familial mutations on POT1 protein

(A) Schematic showing position of germline POT1 mutations identified in CLL families relative to OB domains (orange) and ACD binding region (blue). Also shown are somatic POT1 mutations identified in previous studies of CLL patients[39,58] (unshaded background) and germline mutations found in familial cutaneous melanoma[41,42] (yellow background).

(B) Cross-species conversion of POT1 amino acids subject to missense mutation in CLL families.

(C) Schematic of the crystal structure of human POT1 N-terminal OB-domains bound to a telomeric DNA sequence (Protein Data Bank 3KJP), illustrating the proximity of tyrosine 36 to the DNA strand. OB-domains are shown in grey, DNA is illustrated in blue and Tyr.36 is highlighted in pink.

Figure 3. Impact of *POT1* splice acceptor site mutation on splicing

(A) Splice acceptor site consensus scores predicted by MaxEntScan[24] for each base from the natural *POT1* intron 13/exon 14 boundary across exon 14. For clarity, only part of the intron (lower case text above black line) and exon (upper case text above black box) are shown. The predicted score for the unmutated natural splice site (pink bar) is also labelled. Positive scores are otherwise marked in green, negative scores in orange.

(B) MaxEntScan splice acceptor consensus scores for the same region based upon the sequence of c.1164-1G>A *POT1* mutation carriers. The scores of the mutated natural splice acceptor (pink bar) and the predicted alternative splice site with the highest MaxEntScan score (43bp downstream) are labelled. The part of exon 14 that would be removed by use of this splice site is indicated by a grey box.

(C) Abnormal splicing product detected by RT-PCR using cDNA from a CLL case (Ca) carrying the c.1164-1G>A mutation. This product is absent from control (Co) cDNA. (NT=no template reaction, L=ladder, bp=base pairs).

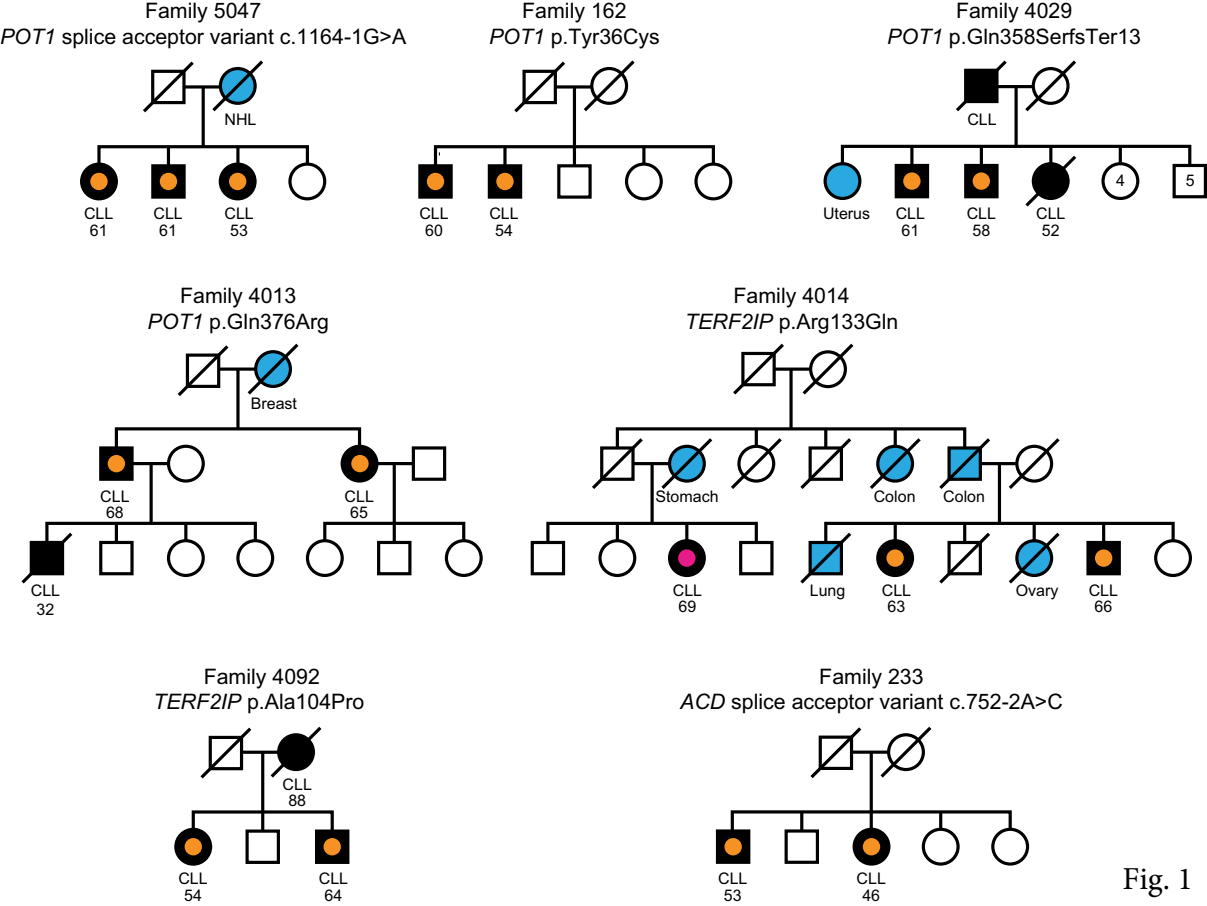
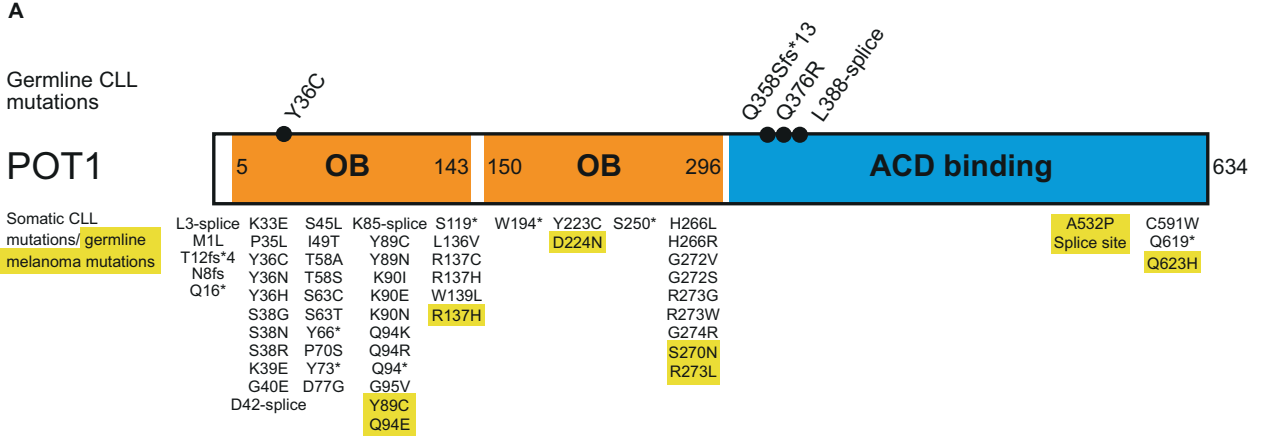
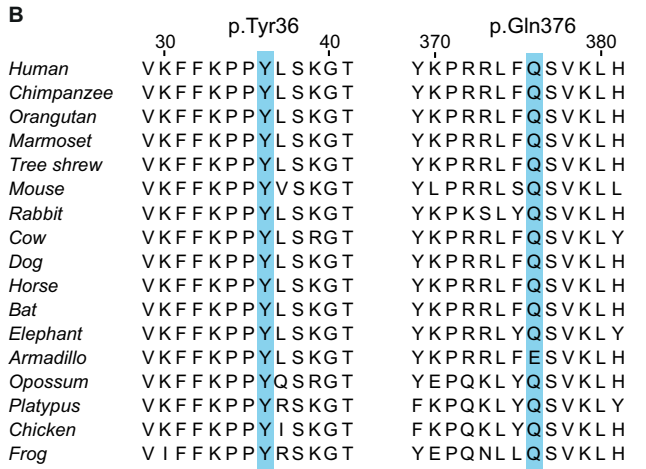

Family 5047
POT1 splice acceptor variant c.1164-1G>A
NHL
CLL 61
CLL 61
CLL 53
Family 162
POT1 p.Tyr36Cys
CLL 60
CLL 54
Family 4029
POT1 p.Gln358SerfsTer13
CLL
Uterus
CLL 61
CLL 58
CLL 52
4
5
Family 4013
POT1 p.Gln376Arg
Breast
CLL 68
CLL 65
CLL 32
Family 4014
TERF2IP p.Arg133Gln
Stomach
Colon
Colon
CLL 69
Lung
CLL 63
Ovary
CLL 66
Family 4092
TERF2IP p.Ala104Pro
CLL 88
CLL 54
CLL 64
Family 233
ACD splice acceptor variant c.752-2A>C
CLL 53
CLL 46

Fig. 1


A
Germline CLL mutations
Y36C
Q358Sfs*13
Q376R
L388-splice
POT1
5
OB
143
150
OB
296
ACD binding
634
Somatic CLL mutations/ germline melanoma mutations
L3-splice
M1L
T12fs*4
N8fs
Q16*
K33E
P35L
Y36C
Y36N
Y36H
S38G
S38N
S38R
K39E
G40E
D42-splice
S45L
I49T
T58A
T58S
S63C
S63T
Y66*
P70S
Y73*
D77G
K85-splice
Y89C
Y89N
K90I
K90E
K90N
Q94K
Q94R
Q94*
G95V
Y89C
Q94E
S119*
L136V
R137C
R137H
W139L
R137H
W194*
Y223C
D224N
S250*
H266L
H266R
G272V
G272S
R273G
R273W
G274R
S270N
R273L
A532P
Splice site
C591W
Q619*
Q623H


B
p.Tyr36
30
40
p.Gln376
370
380
Human V K F F K P P Y L S K G T Y K P R R L F Q S V K L H
Chimpanzee V K F F K P P Y L S K G T Y K P R R L F Q S V K L H
Orangutan V K F F K P P Y L S K G T Y K P R R L F Q S V K L H
Marmoset V K F F K P P Y L S K G T Y K P R R L F Q S V K L H
Tree shrew V K F F K P P Y L S K G T Y K P R R L F Q S V K L H
Mouse V K F F K P P Y V S K G T Y L P R R L S Q S V K L L
Rabbit V K F F K P P Y L S K G T Y K P K S L Y Q S V K L H
Cow V K F F K P P Y L S R G T Y K P R R L F Q S V K L Y
Dog V K F F K P P Y L S K G T Y K P R R L F Q S V K L H
Horse V K F F K P P Y L S K G T Y K P R R L F Q S V K L H
Bat V K F F K P P Y L S K G T Y K P R R L F Q S V K L H
Elephant V K F F K P P Y L S K G T Y K P R R L Y Q S V K L Y
Armadillo V K F F K P P Y L S K G T Y K P R R L F E S V K L H
Opossum V K F F K P P Y Q S R G T Y E P Q K L Y Q S V K L H
Platypus V K F F K P P Y R S K G T F K P Q K L Y Q S V K L Y
Chicken V K F F K P P Y I S K G T F K P Q K L Y Q S V K L H
Frog V I F F K P P Y R S K G T Y E P Q N L L Q S V K L H

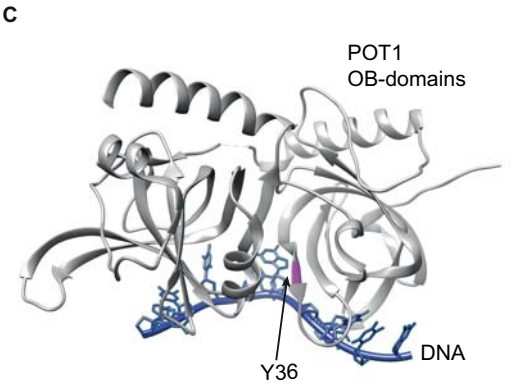

C
POT1
OB-domains
DNA
Y36

Figure 2

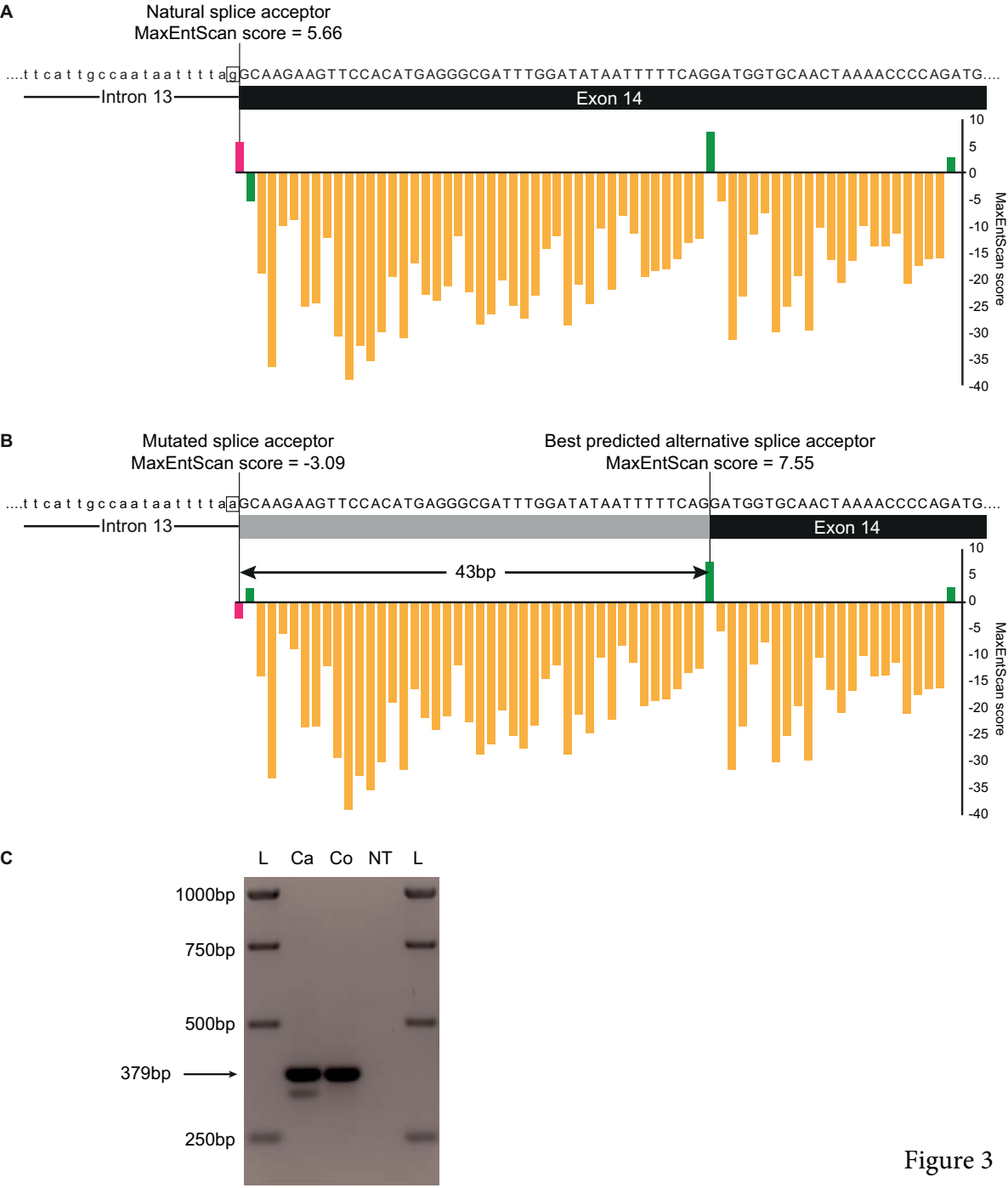

A
Natural splice acceptor
MaxEntScan score = 5.66
....t t c a t t g c c a a t a a t t t t a g GCAAGAAGTTCCACATGAGGGCGATTTGGATATAATTTTTCAGGATGGTGCAACTAAAACCCCAGATG....
Intron 13
Exon 14
10
5
0
-5
-10
-15
-20
-25
-30
-35
-40
MaxEntScan score
B
Mutated splice acceptor
MaxEntScan score = -3.09
Best predicted alternative splice acceptor
MaxEntScan score = 7.55
....t t c a t t g c c a a t a a t t t t a a GCAAGAAGTTCCACATGAGGGCGATTTGGATATAATTTTTCAGGATGGTGCAACTAAAACCCCAGATG....
Intron 13
Exon 14
43bp
10
5
0
-5
-10
-15
-20
-25
-30
-35
-40
MaxEntScan score
C
L
Ca
Co
NT
L
1000bp
750bp
500bp
379bp
250bp

Figure 3