nature cancer

Article

https://doi.org/10.1038/s43018-023-00641-9

Somatic evolution of marine transmissible leukemias in the common cockle, *Cerastoderma edule*

Received: 22 November 2022

Accepted: 23 August 2023

Published online: 02 October 2023

Check for updates

Alicia L. Bruzos [1,2,3,29], **Martín Santamarina** [1,2,3,29], **Daniel García-Souto** [1,2,3,4,29], **Seila Díaz**[1,5,29], **Sara Rocha**[6,29], **Jorge Zamora**[1], **Yunah Lee** [7], **Alejandro Viña-Feás**[1,3], **Michael A. Quail** [4], **Iago Otero** [1,2,3], **Ana Pequeño-Valtierra**[1,3], **Javier Temes** [1,3], **Jorge Rodriguez-Castro** [1,3], **Leyre Aramburu** [1], **André Vidal-Capón** [8], **Antonio Villanueva**[9], **Damián Costas**[9], **Rosana Rodríguez**[9], **Tamara Prieto** [6,10,11], **Laura Tomás** [6,10], **Pilar Alvariño** [6,10], **Juana Alonso**[6,10], **Asunción Cao**[12], **David Iglesias** [12], **María J. Carballal**[12], **Ana M. Amaral**[13], **Pablo Balseiro** [14,15], **Ricardo Calado**[5], **Bouchra El Khalfi**[16], **Urtzi Izagirre** [17,18], **Xavier de Montaudouin**[19], **Nicolas G. Pade**[20], **Ian Probert**[21], **Fernando Ricardo**[5], **Pamela Ruiz** [17,18], **Maria Skazina**[22], **Katarzyna Smolarz**[23], **Juan J. Pasantes**[8,9], **Antonio Villalba** [12,17,24], **Zemin Ning**[4], **Young Seok Ju** [7], **David Posada** [6,8,10], **Jonas Demeulemeester** [25,26,27], **Adrian Baez-Ortega** [4,28,30] ✉ & **Jose M. C. Tubio** [1,2,3,30] ✉

Transmissible cancers are malignant cell lineages that spread clonally between individuals. Several such cancers, termed bivalve transmissible neoplasia (BTN), induce leukemia-like disease in marine bivalves. This is the case of BTN lineages affecting the common cockle, *Cerastoderma edule*, which inhabits the Atlantic coasts of Europe and northwest Africa. To investigate the evolution of cockle BTN, we collected 6,854 cockles, diagnosed 390 BTN tumors, generated a reference genome and assessed genomic variation across 61 tumors. Our analyses confirmed the existence of two BTN lineages with hemocytic origins. Mitochondrial variation revealed mitochondrial capture and host co-infection events. Mutational analyses identified lineage-specific signatures, one of which likely reflects DNA alkylation. Cytogenetic and copy number analyses uncovered pervasive genomic instability, with whole-genome duplication, oncogene amplification and alkylation-repair suppression as likely drivers. Satellite DNA distributions suggested ancient clonal origins. Our study illuminates long-term cancer evolution under the sea and reveals tolerance of extreme instability in neoplastic genomes.

Transmissible cancers are clonal somatic cell lineages that spread between individuals via direct transfer of living cancer cells, in a process reminiscent of tumor metastasis[1,2]. Naturally occurring transmissible cancers have been identified in dogs[3–5], Tasmanian devils[6–8] and, more recently, several species of marine bivalve mollusks[9–14]. To date, eight transmissible cancer lineages, collectively known as BTN, have been described in bivalves, probably spreading via transfer of free-floating cells in seawater. BTN infection causes a leukemia-like disease termed

A full list of affiliations appears at the end of the paper. ✉e-mail: ab2324@cam.ac.uk; jose.mc.tubio@usc.es

disseminated neoplasia (DN), in which neoplastic cells proliferate and accumulate in the host's hemolymph and solid tissues[15]. DN is typically diagnosed by cytological or histological methods, as neoplastic cells tend to present a distinctively large, rounded and nonadherent morphology. Although DN is generally fatal, slow progression and remission have been described[16,17]. Due to its propensity for acute epidemic outbreaks, sometimes associated with mass mortalities in bivalve populations[15], this disease also poses an ecological threat to coastal environments and commercial aquaculture.

Among the species affected by DN is the common cockle, *Cerastoderma edule*. This marine bivalve is distributed along the Atlantic coasts of Europe and northwest Africa, being typically found in tidal flats at bays and estuaries[18]. Adult cockles bury themselves in the seabed sediment and use their siphons and gills to filter seawater for sustenance. DN in common cockles was first documented 40 years ago in Ireland[19], and later identified in other European countries[15]. A genetic study recently provided evidence that some cases of DN in *C. edule* are caused by transmissible cancer, and suggested the existence of at least two BTN lineages in this species[10]. Nevertheless, the origins and evolution of cockle BTN remain entirely unexplored.

Here, we present a comprehensive study of the genomes of BTN lineages affecting *C. edule* in Europe. We sampled thousands of common cockle specimens across 11 countries, obtained a chromosome-level reference genome for the species and used it to catalog the genomic variation in 61 BTN tumors identified in these animals. Combining histopathology, cytogenetics and sequencing of whole genomes and transcriptomes, our study illuminates the evolutionary history of the marine leukemias that have colonized cockle populations along the coasts of Europe.

Prevalence of DN in common cockles

To investigate the current prevalence of DN in *C. edule*, we collected 6,854 specimens at 36 locations from 11 countries along the Atlantic coasts of Europe and north Africa between 2016 and 2021 (Fig. 1a and Supplementary Table 1). This included intensive sampling on the coasts of Ireland and Galicia (northwest Spain), two regions where high prevalence of DN has been reported in the past[20–22]. Cytohistological examination of hemolymph and solid tissues revealed that 5.7% (390 of 6,854) of specimens were infected by abnormal circulating cells displaying the features of DN (Fig. 1a and Supplementary Table 1). High overall prevalence was observed in Portugal (17.6%), Ireland (7.4%) and Spain (6.4%), with lower prevalence found in the United Kingdom (3.6%) and France (1.1%); no DN cases were detected in the remaining six countries (Denmark, Germany, Morocco, the Netherlands, Norway, Russia). Twenty percent (77 of 390) of neoplastic specimens presented a severe form of the disease (stage N3), characterized by high levels (>75%) of neoplastic cells in the hemolymph and massive tissue infiltration; 26% (102 of 390) presented an intermediate form (stage N2), distinguished by 15–75% of neoplastic cells in the hemolymph and presence of small infiltration foci in one or more organs; the remaining individuals (53%, 208 of 390) were diagnosed with a mild form (stage N1), where low levels (<15%) of neoplastic cells circulate in the hemolymph and infiltrate solid tissues in small numbers[22] (Extended Data Fig. 1, Supplementary Table 2 and Supplementary Note).

Reference genome and transcriptome of the common cockle

As an initial step in our genomic study of cockle DN, we applied multiplatform DNA sequencing to obtain a reference assembly of the *C. edule* genome. As our reference specimen, we selected a healthy adult male cockle (Fig. 1b) carrying a standard karyotype with 19 chromosome pairs. Hybrid genome assembly yielded a chromosome-level reconstruction of the cockle nuclear genome into 19 scaffolds with N50 = 39.6 megabases (Mb; 50% of the assembly is contained in scaffolds of length N50 or larger; Supplementary Table 3), with an additional 14.9-kilobase (kb) scaffold containing the mitochondrial genome. Haploid genome size was estimated at 790 Mb, with a G + C content of 35.6%. We additionally employed RNA sequencing data from seven tissues to reconstruct a 290-Mb reference transcriptome presenting 98.8% completeness in metazoan gene content (Supplementary Table 3). Gene annotation resulted in a 42-Mb exome with 14,055 protein-coding genes. While this protein-coding exome constitutes 5.3% of the total nuclear genome size, repetitive sequences comprise 46.2% of the genome, with long interspersed nuclear elements (LINEs) being the most frequent type of transposable element (TE) among annotated repeats (Extended Data Fig. 2 and Supplementary Table 4).

Two transmissible cancers propagate through cockle populations

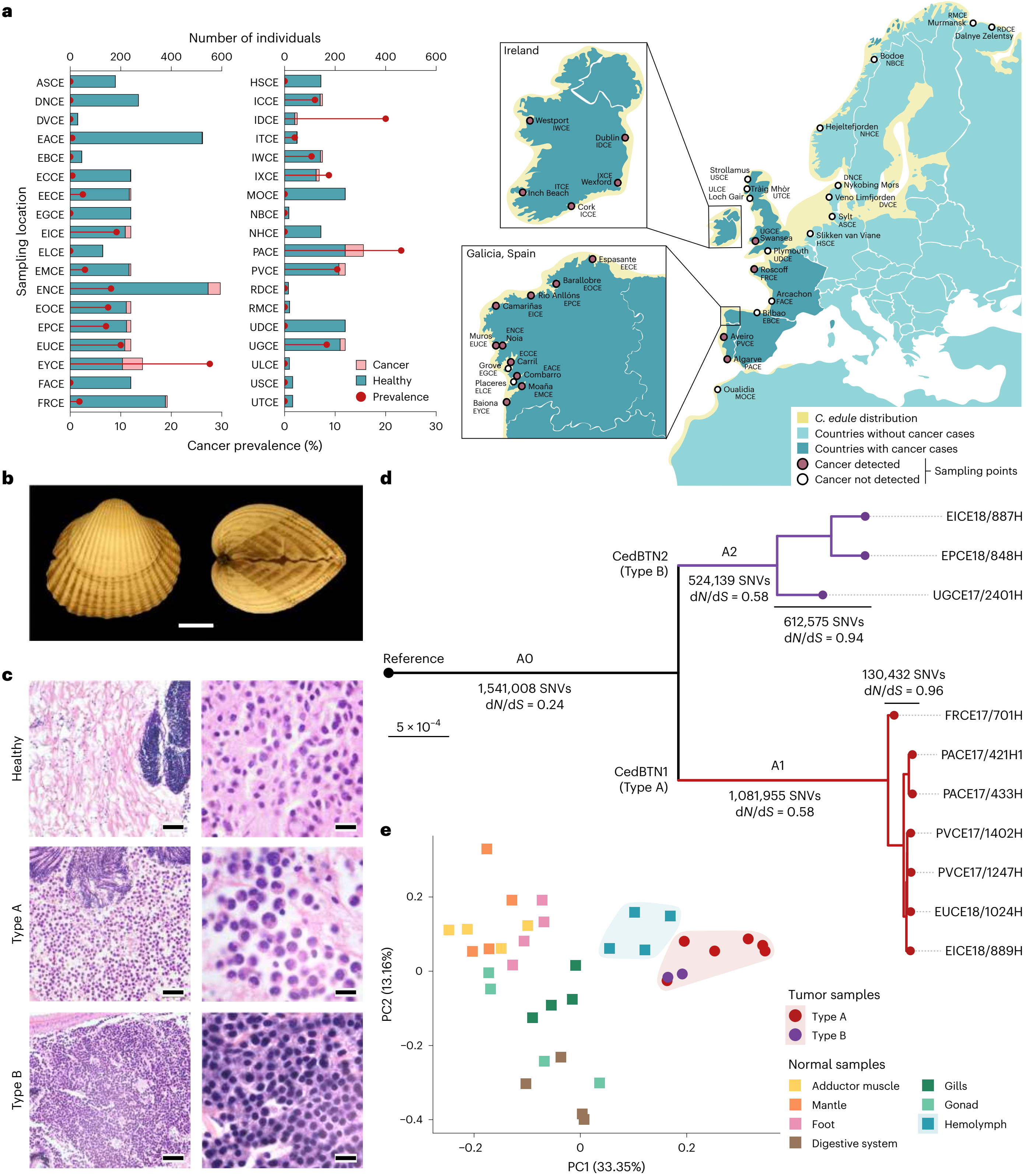
Traditionally, two distinct classes of cockle DN, termed types 'A' and 'B', have been described through cytohistological methods, on the basis of differences in tumor cell size and morphology[21] (Fig. 1c). A previous analysis of microsatellite variation and single-nucleotide variants (SNVs) in both mitochondrial DNA (mtDNA) and one nuclear gene (*EF1α*) provided evidence that these DN types represent two transmissible cancer lineages[10], although it is possible that further lineages exist, as well as nontransmissible cases of DN such as those reported in marine mussels[10,23].

To investigate further the origins and evolution of cockle BTN, we performed whole-genome sequencing of neoplastic hemolymph samples from 61 individuals diagnosed with DN (Supplementary Table 5). Ten of these samples, presenting very high (>97%) tumor purity, were designated as a BTN 'golden set', and used to identify a collection of high-confidence candidate somatic variants. We also sequenced normal tissue samples from 40 host (BTN-infected) individuals and 462 healthy (non-neoplastic) individuals collected across the species' distribution range (Supplementary Table 5). After accounting for host DNA contamination and common germline polymorphisms, we identified a total of 4.3 million SNVs (2.5–3.1 million SNVs per sample) and 0.7 million short insertions and deletions (indels) in BTN samples (Supplementary Table 6). This 'BTN-specific' variant set includes both somatic mutations in each BTN lineage and ancestral germline polymorphisms (from the 'founder' individuals that spawned each lineage) that are absent from our panel of 462 non-neoplastic cockles.

We used BTN-specific SNVs to reconstruct a tumor phylogenetic tree, which split the ten 'golden set' tumors into two divergent lineages

Fig. 1 | Distribution, origins and clonal structure of transmissible neoplasia in common cockles. a, Numbers of healthy and neoplastic *C. edule* cockles collected at each sampling location, with overall cancer prevalence per location for 2016–2021 (left). Map shows sampling locations and geographical distribution of the species. **b**, Photographs of the individual from which the reference *C. edule* genome was assembled. Scale bar, 10 mm. **c**, Micrographs of histological sections from healthy and DN-affected cockle tissues. Images in the left-hand column show healthy connective tissue surrounding the male gonadal follicle (top) and connective tissue heavily infiltrated by type A and type B DN cells. Scale bars, 50 µm. Images in the right-hand column show details of normal hemocytes (top), type A and type B DN cells. Scale bar, 10 µm. Images are representative of 345 independent specimens with similar results. **d**, Phylogenetic tree inferred from BTN-specific SNVs in ten high-purity tumor samples, showing concordance between histological DN types A and B and two clonal transmissible cancer lineages, CedBTN1 and CedBTN2. Numbers of SNVs and d*N*/d*S* ratios are provided for different sections of the tree. All nodes have bootstrap support values of 100 ($n = 1,000$ replicates). Scale bar indicates phylogenetic distance (SNVs per site). **e**, Principal component (PC) analysis of gene expression for genes with tissue-specific expression in normal cockle tissue samples ($n = 4$ per tissue type), type A DN samples ($n = 6$) and type B DN samples ($n = 2$), indicating a clustering of DN (red shading) with healthy hemolymph (blue shading).

(Fig. 1d) consistently matching the two histological types of cockle DN (Extended Data Fig. 3). We hereafter refer to these two clonal lineages of *C. edule* BTN, respectively corresponding to DN types A and B, as CedBTN1 and CedBTN2. To assess the quality of our variant set and confirm the independent origins of both BTN lineages, we estimated the ratio of nonsynonymous-to-synonymous mutation rates (d*N*/d*S*)[24,25] along the phylogenetic tree (Fig. 1d). The d*N*/d*S* ratios for variants shared by all ten tumors (ancestral variant set 'A0') and variants shared by all tumors in each lineage (predivergence sets 'A1' and 'A2') strongly suggest that these sets contain a large fraction of germline polymorphisms from two separate founder individuals (d*N*/d*S* = 0.24 for A0, 0.58 for A1, 0.58 for A2). In contrast, the d*N*/d*S* for the terminal branches approximates a neutral value of 1.0 (0.96 for CedBTN1, 0.94 for CedBTN2), as expected for pure sets of somatic mutations in cancer



genomes[25,26]. Accordingly, the d*N*/d*S* of 'private' variants found in only one tumor is 1.00 (Supplementary Table 7).

Additionally, we performed principal component analysis on a set of germline polymorphisms genotyped across the ten 'golden set' tumors and 100 non-neoplastic cockles covering all sampled populations (Extended Data Fig. 4a). This analysis split the tumors into two divergent clusters matching CedBTN1 and CedBTN2, and set apart from two non-neoplastic sample clusters representing relatively divergent groups of cockle populations from northern and southern Europe[27]. This result suggests that CedBTN lineages are highly divergent both from each other and from modern cockle populations, and strongly supports two independent clonal origins. Nevertheless, analysis of sequence mapping data showed that the fractions of sequence reads aligning against the *C. edule* reference genome in BTN tumors (97–98%, 'golden set' samples) are comparable to those for 462 non-neoplastic cockles (interquartile interval, 97–98%) and substantially higher than fractions for cockles of the closest known species, *Cerastoderma glaucum* (48–60%, six samples). This is consistent with both lineages having arisen from *C. edule* founder individuals.

Hemocytic origin of cockle BTN

The ontogeny of bivalve DN is a long-standing question with relevance for the biology and evolution of BTN. The fact that DN cells are observed in the circulatory system and share morphological features with hemocytes has traditionally led to their consideration as neoplastic hemocytes[15]. However, some studies have proposed alternative tissues of origin for these cancers, including gonad follicles, gill epithelium and others[12,15].

To shed light on the origins of CedBTN lineages, we sequenced the transcriptomes of hemolymph samples from eight cockles diagnosed with late-stage DN, and a collection of seven organs or tissues (adductor muscle, mantle, foot, digestive system, gills, gonad and hemolymph) from 28 non-neoplastic animals (Supplementary Table 5). Gene expression analysis of 420 genes with tissue-specific expression (60 genes per tissue type) indicated a consistent transcriptional profile for type A and type B DN samples, which was close to that of non-neoplastic hemolymph samples and divergent from those of all other tissues (Fig. 1e, Extended Data Fig. 4b,c and Supplementary Table 8). While our collection of normal samples does not include every tissue type described in bivalves, our results suggest that cockle BTN lineages are cancers of the hemolymphatic system, derived from somatic hemocytes or hemic progenitor cells. Furthermore, a companion study by Hart et al.[28] also identified normal hemolymph as the tissue with the closest transcriptional similarity to cells from an independent BTN lineage affecting American soft-shell clams. This recurrent cellular origin may reflect a distinctive capability of malignant hemocytes to exploit the transmission opportunities offered by the open circulatory system of bivalves.

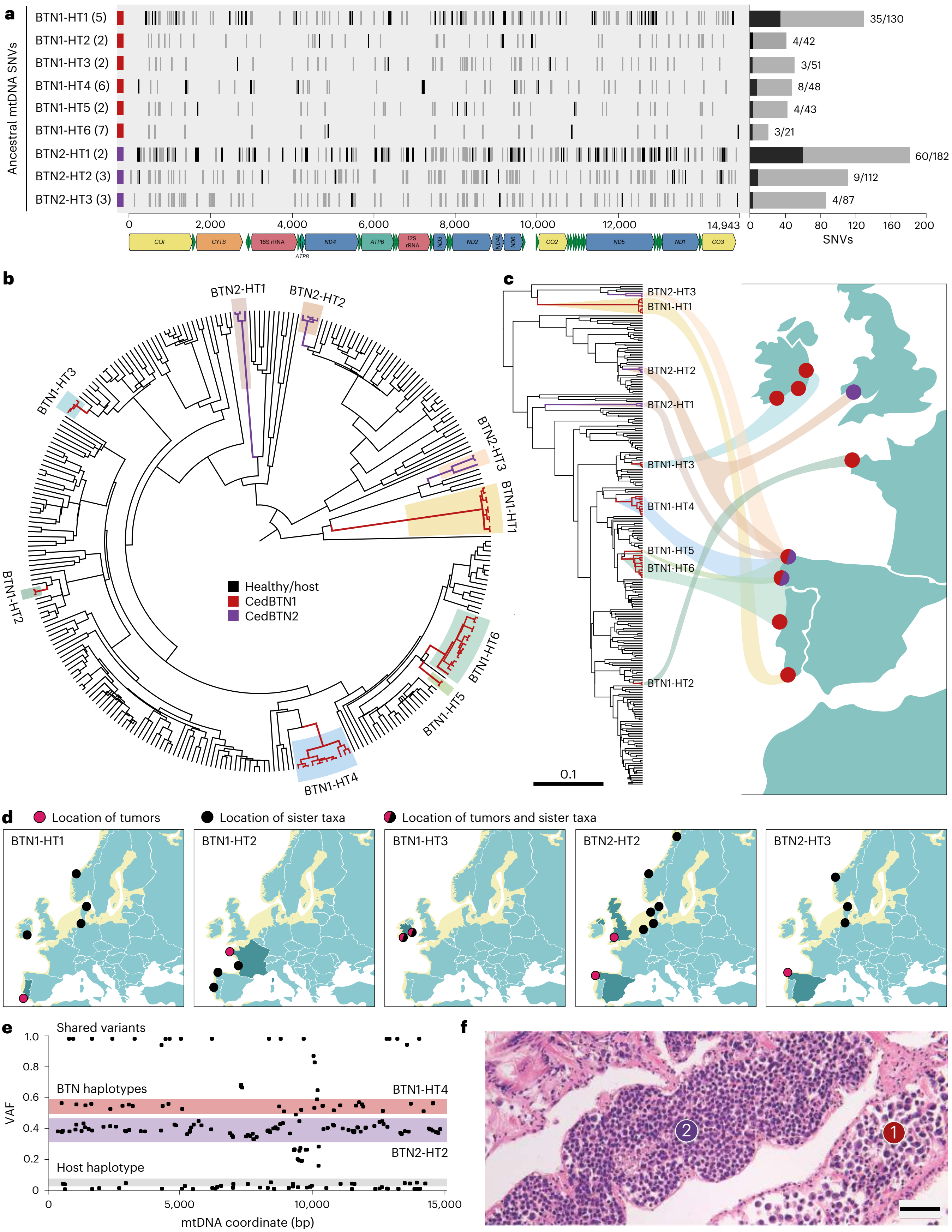
Mitochondrial transfer delineates the clonal structure of CedBTN

To explore the evolutionary history of CedBTN at the mitochondrial level, we identified SNVs in the mtDNA of 51 hemolymph samples from neoplastic cockles, 40 host tissue samples and 168 non-neoplastic cockle samples. In neoplastic animals, sequencing data showed two mtDNA haplotypes at distinct variant allele fractions (VAFs), corresponding to the host and CedBTN mitochondrial genomes. Combining tumor purity and mtDNA VAF information to deconvolute the mtDNA haplotypes within each sample, we identified nine distinct tumor haplotypes (six in CedBTN1 and three in CedBTN2), each distinguished by a specific set of mtDNA variants (Fig. 2a and Supplementary Table 9).

The findings above suggested the existence of nine CedBTN mtDNA lineages. This was confirmed through phylogenetic reconstruction via maximum likelihood (ML) and Bayesian methods (Fig. 2b and Extended Data Fig. 5a). The presence of multiple mtDNA lineages within each CedBTN nuclear clone indicates that mitochondria from transient hosts have repeatedly been acquired by these tumors, as previously described for other transmissible cancers[11,29,30] and for normal and cancer cells in vitro and in vivo[31–33]. We therefore labeled these mtDNA lineages, and their associated haplotypes, after putative mitochondrial horizontal transfer (HT) events (BTN1-HT1 to -HT6 and BTN2-HT1 to -HT3), although it is currently impossible to ascertain whether any of these represent the original mtDNA haplotypes of the CedBTN founder individuals. The correspondence of each nuclear lineage to multiple mtDNA lineages was supported by a phylogenetic tree inferred from the genotypes of nuclear BTN-specific SNVs across the set of 61 sequenced tumors (Extended Data Fig. 5b). Furthermore, tumors from distinct mtDNA lineages within the same CedBTN nuclear lineage presented no evident cytohistological differences (Supplementary Table 10). We evaluated the potentially independent origins of the nine mtDNA lineages using three topology testing methods on the mtDNA phylogenies (Shimodaira–Hasegawa and approximately unbiased tests for the ML tree, posterior odds for the Bayesian tree), which consistently supported independent origins for all the lineages except BTN1-HT5 ($P = 0$ for Shimodaira–Hasegawa, $P < 5 \times 10^{-5}$ for approximately unbiased, posterior odds = 0).

Analyses of the geographical distribution of mtDNA haplotypes from tumors and their sister taxa (defined as non-neoplastic samples derived from the same node in the phylogeny) provided insight into the origins and spread of CedBTN mtDNA lineages. First, although most tumor samples from the same mtDNA lineage are usually found in the same geographical region (for example, BTN1-HT1 in south Portugal, BTN1-HT2 in France, BTN1-HT3 in Ireland), this is not the case for BTN2-HT2, for which tumor specimens were collected in northwest Spain and Wales (Fig. 2c and Extended Data Fig. 6). Second, the geographical ranges of tumors and their sister taxa may be expected to overlap (for example, BTN1-HT3 and sister taxa in Ireland), or at least be proximate (for example, BTN1-HT2 in France and sister taxa in Spain and Portugal), yet we observed four mtDNA lineages (BTN1-HT1, BTN2-HT1, BTN2-HT2, BTN2-HT3) occupying regions distant from the ranges of their sister taxa (Fig. 2d and Extended Data Fig. 6). Two remarkable cases are BTN1-HT1 and BTN2-HT3, for which tumors were found in Portugal and Spain, respectively, while their sister taxa were sampled in Ireland, Germany, Denmark and Norway. Third, the sister taxa of CedBTN2 mtDNA lineages were almost invariably found in northern regions (Denmark, Germany, Norway and the Netherlands), despite

Fig. 2 | mtDNA phylogeny, mtDNA HT and host co-infection in CedBTN. **a**, Ancestral mtDNA haplotypes identified in CedBTN samples, with ancestral SNVs (common to all samples carrying the haplotype) arranged along the reference mtDNA sequence (*x* axis). Potentially somatic SNVs (absent from non-neoplastic samples) are shown in black. Potential mtDNA HT events associated with each haplotype in CedBTN1 (red) and CedBTN2 (purple) are labeled, with the number of samples used to identify ancestral variants given in parentheses. Bar plot presents numbers of potentially somatic (black) and total (gray) ancestral variants per haplotype; numbers are indicated next to each bar. A schematic representation of the mtDNA gene annotation is shown at the bottom. **b**, Bayesian phylogenetic tree of mtDNA haplotypes in normal and CedBTN samples, with identified tumor mtDNA lineages highlighted and labeled. Branch lengths represent phylogenetic distance (scale bar given in **c**). **c**, Correspondence between mtDNA phylogenetic tree and tumor sampling regions; map point colors denote CedBTN nuclear lineages as in **b**. Sampling points in Galicia (northwest Spain) are grouped into northern and southern points. Scale bar indicates phylogenetic distance (SNVs per site). **d**, Maps showing locations of tumors and normal sister taxa for five mtDNA lineages. **e**, VAF plot evidencing co-infection of a host (EICE18/910) by cells from two mtDNA lineages, one from each CedBTN nuclear lineage. Three observed mtDNA haplotypes are shaded in different colors. **f**, Micrograph of histological section of gills from EICE18/910, confirming co-infection by both CedBTN lineages. Dilated efferent vessels are shown; vessels labeled '1' and '2' are mainly infiltrated by type A and type B neoplastic cells, respectively. Scale bar, 50 µm.


a
Ancestral mtDNA SNVs
BTN1-HT1 (5)
BTN1-HT2 (2)
BTN1-HT3 (2)
BTN1-HT4 (6)
BTN1-HT5 (2)
BTN1-HT6 (7)
BTN2-HT1 (2)
BTN2-HT2 (3)
BTN2-HT3 (3)
35/130
4/42
3/51
8/48
4/43
3/21
60/182
9/112
4/87
0
2,000
4,000
6,000
8,000
10,000
12,000
14,943
0
40
80
120
160
200
SNVs
COI
CYTB
16S rRNA
ND4
ATP6
12S rRNA
ND3
ND2
ND4L
ND6
CO2
ND5
ND1
CO3
ATP8
b
BTN2-HT1
BTN2-HT2
BTN1-HT3
BTN2-HT3
BTN1-HT1
BTN1-HT2
BTN1-HT6
BTN1-HT5
BTN1-HT4
Healthy/host
CedBTN1
CedBTN2
c
BTN2-HT3
BTN1-HT1
BTN2-HT2
BTN2-HT1
BTN1-HT3
BTN1-HT4
BTN1-HT5
BTN1-HT6
BTN1-HT2
0.1
d
Location of tumors
Location of sister taxa
Location of tumors and sister taxa
BTN1-HT1
BTN1-HT2
BTN1-HT3
BTN2-HT2
BTN2-HT3
e
Shared variants
BTN haplotypes
BTN1-HT4
BTN2-HT2
Host haplotype
VAF
1.0
0.8
0.6
0.4
0.2
0
0
5,000
10,000
15,000
mtDNA coordinate (bp)
f
2
1

the fact that no CedBTN2 tumors were observed in this range (Fig. 2d and Extended Data Fig. 6). Although we cannot rule out anthropogenic contributions to some of these patterns, the geographical structure of the mtDNA phylogeny suggests that CedBTN lineages have spread over long distances along the Atlantic coasts of Europe, probably through a gradual process of natural colonization. Host mitochondria have been captured by CedBTN cells at different points during this process, potentially to replace somatically mutated incumbent mtDNA[11,29,31,32]. Notably, this phenomenon has not been detected in soft-shell clam BTN[28], possibly due to differences in age (and thus mitochondrial capture opportunity) among BTN lineages, differences in genetic structure between the two host species[27,34], or limitations of sample size and distribution in the study by Hart et al.[28].

In addition to SNVs, inspection of mtDNA sequencing data revealed three independent amplifications spanning the control region of the mtDNA D-loop in CedBTN1, which are absent from healthy cockles (Extended Data Fig. 7a–c). The amplified sequences share a common start motif and overlapping microhomology at the boundaries, which is associated with imperfect DNA break repair[35]. The evolutionary importance of these recurrent amplifications is unclear; they may be neutral changes, or the result of selfish selection at the mitochondrial level[30], or yet confer an advantageous phenotype on BTN cells. Notably, similar D-loop amplifications have been identified in both BTN and non-neoplastic samples from soft-shell clams[28], as well as human cancers[36].

Although mtDNA VAFs were generally consistent with homoplasmy in CedBTN samples, analysis of VAF differences across distinct tissues of the same animal revealed three cases in which two CedBTN mtDNA lineages coexisted within the same host (Extended Data Fig. 7d). In one remarkable animal (EICE18/910), VAF analysis revealed the presence of mtDNA haplotypes from both cancer clones (Fig. 2e), with co-infection by CedBTN1 and CedBTN2 cells being confirmed through histopathological identification of cell morphologies matching DN types A and B (Fig. 2f), as well as through genotyping of BTN-specific nuclear SNVs (99% and 88% of SNVs in the predivergence sets A1 and A2, respectively, were detected in this animal's hemolymph sample). Histopathological re-evaluation of our tumor collection uncovered seven additional cases of co-infection by both types of DN, for which sequencing data are not available (Supplementary Table 2). Thus, we estimated an incidence of 2.6% (10 of 390) for detectable co-infection by distinct tumor lineages, which is probably an underestimate of the overall co-infection rate (including co-infection by cells from multiple tumors that carry the same mtDNA haplotype). This suggests that, in contrast to its extreme rarity in mammalian transmissible cancers, host co-infection is a relatively frequent event in cockle BTN.

Lineage-specific mutational processes in CedBTN

To investigate the processes of DNA damage and repair causing mutations in CedBTN, we examined patterns of SNVs and indels at particular sequence contexts, termed mutational signatures[37]. The mutational spectra of germline cockle polymorphisms and BTN-specific SNVs are broadly similar, the major difference being a higher fraction of cytosine-to-thymine (C>T) substitutions at non-CpG sites in CedBTN relative to the germline (Fig. 3a). We assessed mutational processes across the CedBTN phylogeny by defining six subsets of BTN-specific variants (Fig. 3b): SNVs shared by all samples from each lineage, but not shared between lineages (two predivergence sets, A1 and A2; Fig. 1d); SNVs shared by only some tumors in each lineage (two nonprivate postdivergence sets); and SNVs present in one tumor (two private sets). We also defined two germline sets: ancestral SNVs shared by both CedBTN lineages (ancestral set A0), and SNVs identified in three non-neoplastic cockles. While the two predivergence sets, containing mostly germline variants, present similar mutational spectra, the largely somatic postdivergence sets exhibit notable differences, particularly in the C>T component (Fig. 3b).

With the aim of quantifying the contribution of different mutational processes to these variant sets, we applied a Bayesian approach[38] to infer five mutational signatures de novo from their mutational spectra (Fig. 3c). Three of these signatures (SBS-A, SBS-B, SBS-C) are shared by germline and BTN-specific sets, while the remaining two (SBS-D, SBS-E) are BTN-specific. Most signatures show similarity to human mutational signatures, especially if the latter are corrected for the trinucleotide composition of the human genome. Among the germline signatures, SBS-A probably corresponds to a mixture of human signatures SBS1 (cosine similarity 0.84), caused by spontaneous deamination of 5-methylcytosine at CpG sites[35,39], and SBS5 (0.90), thought to arise from multiple endogenous mutational processes[35,40]; SBS-B resembles human SBS40 (0.79), possibly caused by the same endogenous processes as SBS5 (ref. 40); and SBS-C is similar to SBS8 (0.82), a signature associated with DNA repair and replication errors in human cancers and absent from the human germline[41,42]. Of the BTN-specific signatures, SBS-D resembles both SBS23 (0.86), a signature of unknown etiology described in human myeloid and brain tumors[35], and SBS11 (0.81), associated with the alkylating chemotherapeutic agent temozolomide[37]; the profile of SBS-E has no evident human counterpart, the closest match being SBS40 (0.71).

To explore variation in the activity of mutational processes, we assessed mutational signature exposures across the BTN phylogeny. Signatures SBS-D and SBS-E, while undetectable in germline variant sets, are each predominantly associated with one BTN lineage: whereas SBS-D dominates the spectrum of CedBTN1 postdivergence mutations, SBS-E is mainly active in the CedBTN2 postdivergence set (Fig. 3d and Supplementary Table 11). We note that, while BTN-specific variant sets (including A0) present lower SBS-A exposures relative to the cockle germline, this may reflect disproportionate filtering of variants at CpG sites, which are underrepresented relative to other sequence contexts in the cockle genome. Due to this CpG depletion, independent C>T changes at these sites have a higher probability of being shared between tumor and non-neoplastic samples, and hence being classified as germline variants.

Inspection of indel spectra provided evidence for a variety of mutational processes in germline and BTN-specific sets (Fig. 3e). Although not every observed pattern can be matched to a human signature, germline indels appear to be enriched in signatures ID1 and ID2 (single-nucleotide insertions and deletions at long A/T homopolymers, caused by strand slippage during DNA replication[35]), as well as ID9 and ID14 (single-nucleotide deletions and insertions of unknown etiology). BTN-specific indels present lower contributions from ID1 and ID2 relative to the germline, and appear enriched in ID5 (single-nucleotide deletions at short A/T homopolymers, of unknown etiology) and ID8 (long deletions, possibly caused by repair of DNA double-strand breaks via nonhomologous end-joining[35]). Hence, mutational processes absent from the germline, and possibly linked to genomic instability, appear to have contributed substantial fractions of indels to CedBTN genomes.

Pervasive genomic instability drives CedBTN evolution

Previous cellular studies have shown that cockle DN is distinguished by an unusual, broad continuum of ploidy ranging from 1.3*n* to 9.6*n*, and a variable karyotype marked by an abundance of small chromosomes[43–45]. To investigate further this hallmark of DN in cockle BTN, we performed cytogenetic analysis of 261 metaphase spreads from neoplastic cells in six tumors, three from each CedBTN lineage (Extended Data Fig. 8). This revealed extensive variation in chromosome number and size across tumors, with the median chromosome number per sample varying between 98 and 276 (Supplementary Table 12). Notably, we also observed wide variability in chromosome number within individual tumors. For instance, neoplastic metaphase spreads from sample PACE17/478H contained 11–354 chromosomes of variable size and structure. Fluorescence in situ hybridization (FISH) probes targeting

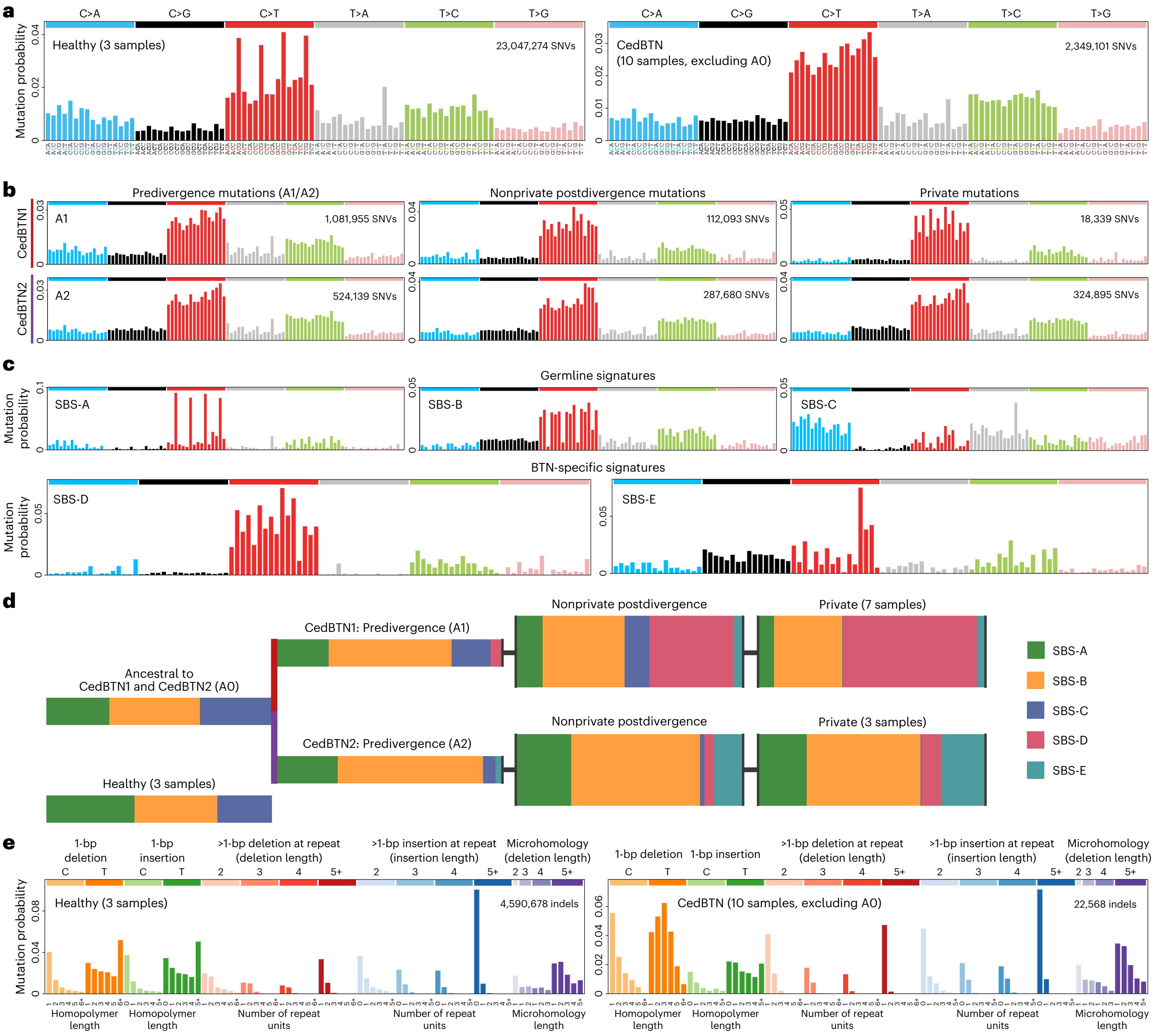


Fig. 3 | Mutational processes in CedBTN. a, Mutational spectra of germline SNVs in three healthy cockle samples (left) and BTN-specific SNVs in ten CedBTN samples (excluding the set of shared ancestral SNVs, A0 in Fig. 1d). The *x* axis presents 96 mutation types in a trinucleotide context, colored by base substitution type[35]; the *y* axis presents mutation probability, normalized to correct for the cockle genome trinucleotide frequencies. **b**, Mutational spectra of subsets of BTN-specific variants in CedBTN1 (top) and CedBTN2, including predivergence variants (left; A1/A2), nonprivate postdivergence variants (center) and private variants. **c**, Germline (top) and BTN-specific mutational signatures inferred from the spectra shown in **a** and **b** (plus the A0 spectrum). **d**, Contribution of each mutational signature to the SNVs in each segment of the CedBTN phylogenetic tree (Fig. 1d) and in healthy samples. Bars for postdivergence variant sets are depicted with greater width to denote collapsing of multiple internal branches of the tree. **e**, Mutational spectra of germline indels in three healthy samples (left) and BTN-specific indels in ten CedBTN samples (excluding the shared ancestral set, A0). The *x* axis presents 83 insertion/deletion types colored by type and length[35]; the *y* axis presents unnormalized mutation probability.

telomeric sequences showed that, despite such karyotypic plasticity, all the chromosomes in CedBTN cells present a canonical structure (Fig. 4a). These results suggest that the shifting karyotypes of CedBTN are probably the outcome of extensive chromosomal reorganization and frequent chromosome missegregation during anaphase.

Next, we inferred copy number (CN) profiles from whole-genome sequencing data for each tumor in our 'golden set'. The profiles were marked by a ubiquitous pattern of highly complex CN alterations along every reference chromosome, with lower CN states visibly underrepresented (Fig. 4b). CN distributions were consistent with a modal CN of 4.0, suggestive of ancestral tetraploidy, except for one tumor (UGCE17/2401H) presenting a modal CN of 5.0. Profiles were loosely conserved across tumors from each lineage, with a combination of shared and sample-specific CN features (Extended Data Fig. 9). Moreover, CN distributions revealed a strong aberrant background of chromosomal regions with additional CN states, which in some cases obscured the expected tetramodal or pentamodal CN profile (Extended Data Fig. 9). The cytogenetic findings above suggest that this aberrant CN background results from persistent chromosome missegregation, generating extensive intra-tumor heterogeneity in

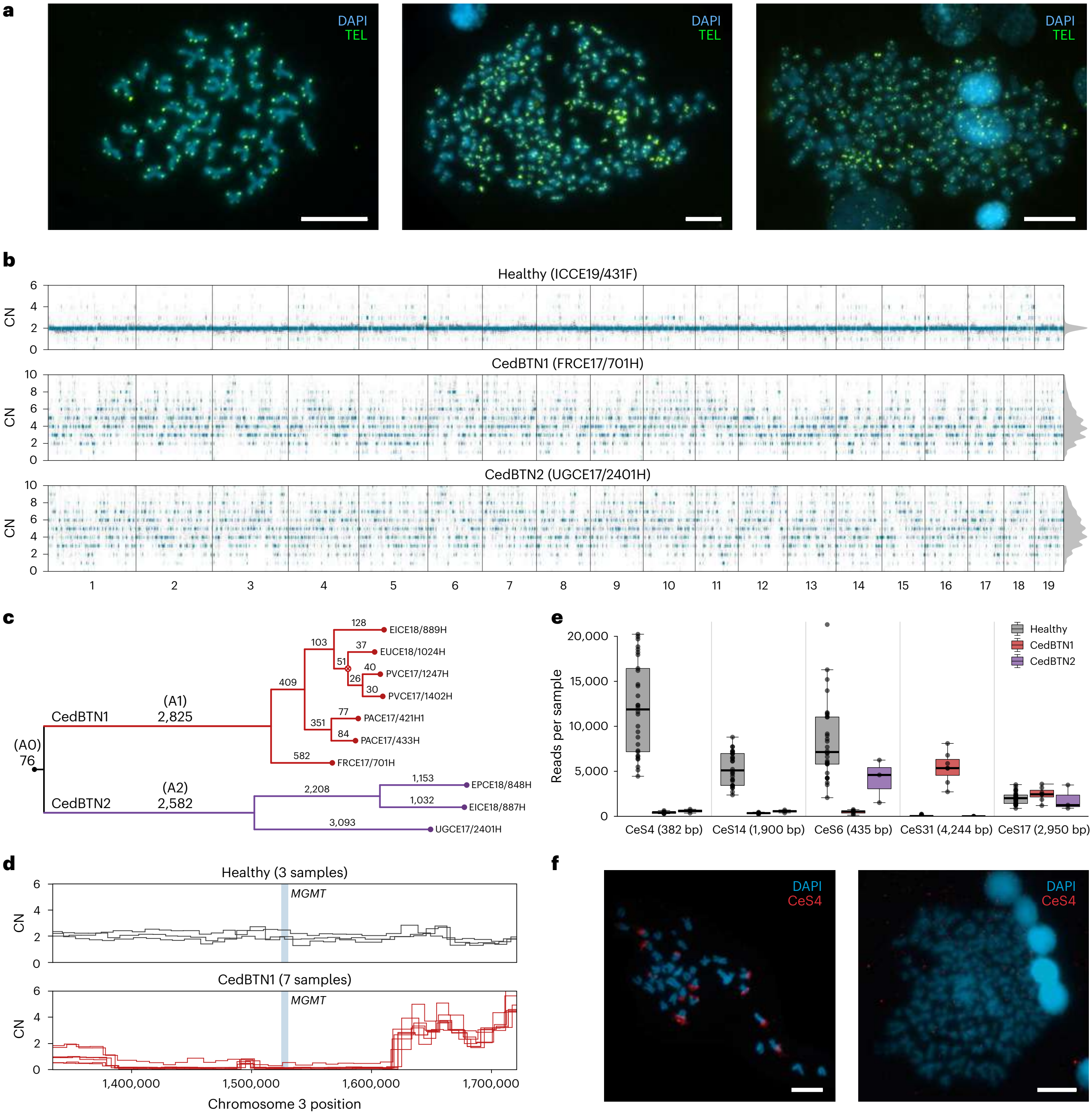


Fig. 4 | Chromosomal, CN and structural variation in cockle BTN. **a**, FISH of telomeric peptide nucleic acid probes (TEL, shown in green) onto healthy (left) and example type A (center) and type B (right) neoplastic metaphase spreads. All chromosomes, including the smallest neoplastic chromosomes, hold telomeric signals on all chromatid ends. Scale bars, 10 µm. Images are representative of ten independent experiments with similar results. **b**, CN profiles of representative healthy, CedBTN1 and CedBTN2 samples. Gray dots represent estimates of unrounded CN for 10-kb windows along the reference genome (*x* axis); blue segments indicate inferred segments of integer CN. Distributions of unrounded CN are shown on the right margin. Profiles are representative of eight healthy, seven CedBTN1 and three CedBTN2 samples with similar results. **c**, Phylogenetic tree inferred from BTN-specific SVs. The number of SVs per branch is indicated, and branches corresponding to sets of ancestral or predivergence variants (A0, A1, A2) are labeled. Bootstrap support values ($n = 1{,}000$ replicates) are ≥99.9 for all nodes except that marked with symbol ⊗ (91.6). **d**, CN profiles in a 500-kb region around the *MGMT* gene locus in healthy ($n = 3$) and CedBTN1 ($n = 7$) samples. Each sample is represented by a line. The highest CedBTN1 CN estimate at the gene locus (CN = 0.4) corresponds to sample EICE18/889H. **e**, Numbers of sequence reads aligning to five satellite DNA elements identified in a diverse set of healthy cockles ($n = 30$) and CedBTN1 ($n = 7$) and CedBTN2 ($n = 3$) tumors. Each dot represents a sample. Boxes represent first and third quartiles; middle line within each box denotes the median; whiskers indicate values within 1.5 × interquartile range from the first and third quartiles. Monomer size is provided for each satellite. **f**, FISH of DNA probe for satellite CeS4 (red) onto representative metaphases of healthy (left) and neoplastic specimens. The red channel in the second image is oversaturated to verify absence of CeS4 from neoplastic chromosomes. Scale bars, 10 µm. Images are representative of ten independent experiments with similar results.

CN. Such heterogeneity is most likely amplified by the effect of cell transmission bottlenecks to produce the observed inter-tumor CN variability. Overall, our analyses indicate that both CedBTN clones are highly aneuploid lineages that underwent at least one whole-genome duplication event in early tumorigenesis, leading to a likely tetraploid state that, in the case of CedBTN2, later developed further CN gains in the UGCE17/2401H branch. Due to the inability to discriminate completely between germline and early somatic variants in each lineage, however, it is currently not possible to date these genome duplication events with greater precision.

To characterize further the landscape of somatic alterations in cockle BTN, we applied multiple established algorithms to call structural variants (SVs) in the ten 'golden set' tumors. We then removed potentially germline events by genotyping these variants on 455 non-neoplastic samples. This approach yielded a conservative set of 18,272 high-confidence SVs (7,347 in CedBTN1, 11,356 in CedBTN2), with deletions being the most frequent type of event (80%, 14,589 of 18,272; Extended Data Fig. 10a,b). A maximum parsimony phylogenetic tree reconstructed from these variants confirmed the CedBTN nuclear phylogeny inferred from SNVs, supporting two divergent lineages with a minimal fraction of shared structural variation (Fig. 4c).

The combination of gene CN data and nonsynonymous predivergence mutations in each lineage did not reveal any high-confidence candidate cancer-driver events (Supplementary Table 13). Similarly, d*N*/d*S* ratios yielded no evidence of positive selection for postdivergence SNVs or indels in either lineage. However, the availability of CN data offered an additional opportunity to identify potential early driver CN alterations. We systematically screened for gains and losses of regions containing oncogenes and tumor suppressor genes (TSGs), respectively. This analysis identified likely ancestral amplifications involving two canonical oncogenes in CedBTN1: *MDM2* (10–13 copies; mean CN = 10.9; gene CN percentile = 98.3), encoding the principal cellular antagonist of the p53 protein, and *CCND3* (8–18 copies; mean CN = 10.7; gene CN percentile = 98.2), encoding a cyclin that promotes G1/S cell cycle transition (Supplementary Table 14). Recurrent amplification of these genes has been observed in multiple cancer types, and is thought to prevent cell cycle arrest and apoptosis under conditions of genomic instability[46–49]. In CedBTN2, we found evidence of a likely ancestral *MYC* amplification (7–11 copies; mean CN = 9.2; gene CN percentile = 96.3; Supplementary Table 14). Interestingly, *MYC* activation has also been proposed as an early driver of a mammalian transmissible cancer[1].

Notably, we also identified an ancestral homozygous deletion of *MGMT* in CedBTN1 (Fig. 4d and Supplementary Table 14). The enzyme encoded by this gene, O^6-methylguanine-DNA methyltransferase, is essential for repair of alkylated DNA bases, and its inactivation results in hypersensitivity to the toxic and mutagenic effects of alkylating agents[50,51]. Given the cumulative and virtually lineage-specific activity of signature SBS-D (Fig. 3d), and its similarity to human signature SBS11 (caused by the alkylating agent temozolomide[37]), SBS-D most likely reflects unrepaired alkylation of DNA bases due to loss of *MGMT*. The resemblance between SBS-D and SBS23 further suggests that SBS23 may arise from deficient DNA alkylation repair in human cancers.

We examined gene expression estimates for these candidate early drivers, and found evidence of increased expression of amplified genes *CCND3*, *MDM2* and *MYC* in the relevant CedBTN lineage relative to normal tissues, as well as absence of *MGMT* expression in CedBTN1 (Extended Data Fig. 10c and Supplementary Table 14). Remarkably, we also observed overexpression of *MDM2* in CedBTN2 relative to normal tissues, perhaps related to moderate amplification of this gene in CedBTN2 (3–8 copies). These findings support the conclusion that CN alterations of *CCND3*, *MDM2*, *MGMT* and *MYC* are likely drivers of early CedBTN evolution, and raise the possibility that upregulation of *MDM2* has been independently selected in both cancer lineages. Despite the high gene content completeness of our *C. edule* genome assembly (Supplementary Table 3), we cannot exclude the possibility that further early driver events have escaped detection due to lack of homology between the sets of cancer genes in humans and bivalves.

Satellite DNA expansions illuminate the emergence of CedBTN

Finally, we applied a computational method to examine the repetitive complement of the *C. edule* genome, with a focus on satellite DNA. These repetitive sequences are relevant for genome stability, exhibiting long-term conservation and propensity for rapid CN changes[52]. Our method identified 34 satellite DNA candidates in the common cockle reference genome (Supplementary Table 15), four of which varied in frequency between non-neoplastic and BTN genomes, providing further insight into the origins of cockle BTN (Fig. 4e). Two satellites, named CeS4 and CeS14, were found at high frequency in all samples from a genetically diverse cohort of non-neoplastic cockles, yet were entirely absent from both BTN lineages. We designed FISH probes to target satellite CeS4, which confirmed the results obtained from sequencing data (Fig. 4f). This finding suggests that both CedBTN1 and CedBTN2 may be ancient cancer lineages that diverged from the cockle population before the emergence and expansion of CeS4 and CeS14 in the *C. edule* germline. Another satellite, CeS6, was found in cockle populations and CedBTN2 samples, while absent from CedBTN1 (Fig. 4e). Lastly, despite satellite CeS31 being exclusive to CedBTN1, our data did not support exclusive presence of any satellite DNA in CedBTN2 samples. Although we cannot exclude other explanations, these observations suggest that CedBTN2 possibly diverged from the cockle population more recently than CedBTN1.

Discussion

Despite several BTN lineages having been newly described in recent years[9–14,23], to our knowledge no analyses of whole BTN genomes have yet been reported. Combining a range of approaches, our study provides an expansive outlook into the genomes of these singular marine leukemias in European common cockles, complementing the work of Hart et al.[28] on American soft-shell clams. Both studies reveal neoplastic genomes marked by aneuploidy, pervasive genomic instability and lineage-specific mutational processes. In the case of cockle BTN, we find evidence for sustained chromosomal instability, most likely activated by early whole-genome duplication[53,54] and fueled by recurrent chromosome missegregation during mitosis[55,56]. Moreover, the likely upregulation of MDM2 and cyclin D by means of ancestral gene amplification suggests that BTN lineages may evolve tolerance of chromosomal instability through disruption of p53-dependent responses against aneuploidy[57,58]. Interestingly, suppression of p53 via cytoplasmic sequestration has been reported in the BTN lineage affecting soft-shell clams[59], raising the possibility that BTN lineages in different bivalve species may have evolved distinct adaptations in response to common evolutionary pressures.

The extreme chromosomal instability of CedBTN genomes contrasts with the quiescent karyotypes of transmissible cancers in dogs and Tasmanian devils[1,5,60], challenging the notion that a stable genomic architecture is required for long-term survival of cancer lineages. Although our data do not allow estimation of precise ages for cockle BTN, multiple lines of evidence suggest that these cancers may have emerged centuries or millennia ago. These include the broad geographical distribution of tumors, the marked genetic divergence between tumors and modern cockles, the recurrent capture of host mitochondria by tumors (not observed in soft-shell clam BTN[28]) and the absence in tumors of satellite DNA elements that are vastly expanded in the cockle germline. Furthermore, Hart et al. estimate an age of ~423 years for the soft-shell clam BTN lineage[28], demonstrating the potential for long-term survival of marine transmissible cancers. Taken together, our findings suggest that CedBTN lineages have undergone a long history of sustained genomic instability. Studying the mechanisms by which BTN cells overcome the effects of such instability promises to broaden our understanding of the conditions required for tumors to survive and adapt over the long term.

Methods

This research complied with all relevant ethical regulations. Animal samples were obtained under the approval of the Standing Committee on Conflict of Interest, Scientific Misconduct and Ethical Issues (CoIME) of the European Research Council, and under regional licenses for mollusk extractions and trading authorizations. Our institutional facilities conformed to safety requirements. Seawater was subjected to disinfection protocols and laboratory personnel possessed the required experimental work certifications.

Statistics and reproducibility

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation or diagnosis during experiments and outcome assessment. A fraction of the data were excluded from certain analyses for quality reasons; exclusion criteria for particular analyses are detailed in the sections below.

To ensure reproducibility of cytological and histological diagnosis of disease stage (Extended Data Fig. 1b–p) and type of neoplasia (Figs. 1c and 2f and Extended Data Fig. 3), we performed diagnosis independently on five tissues for each individual: hemolymph (cytology), foot, gonad, gills and digestive gland (histology). To ensure reproducibility of cytogenetic analyses, we conducted independent FISH experiments on metaphases using different probes and conditions, including CedBTN1, CedBTN2 and healthy specimens. The number of metaphases/experiments for each probe and condition were as follows. Telomeric probes (Fig. 4a): CedBTN1 (51/3); CedBTN2 (27/2); healthy (52/5). Satellite DNA (Fig. 4f): CedBTN1 (42/2); CedBTN2 (19/2); healthy (64/6). Histone genes and ribosomal DNA (Extended Data Fig. 8): CedBTN1 (153/10); CedBTN2 (71/9); healthy (49/10). Consistent results were obtained in all cases.

Sample collection, processing and diagnosis

Between 2016 and 2021, 6,854 *C. edule* specimens were collected from seabeds of 11 countries covering the species' geographical range (Supplementary Table 1). Cockles were maintained in closed-circuit seawater tanks for 48 h.

For information on sample processing and diagnosis, see the Supplementary Note.

Karyotyping

Mitotic chromosomes were obtained following standard protocols[61]. Neoplastic animals received an in vivo colchicine treatment (0.005%, 8 h), a hypotonic treatment and fixation in ethanol-acetic acid (3:1). Fixed gills were disaggregated in acetic acid (60%), dropped onto pre-heated glass slides, stained with 4′,6-diamidino-2-phenylindole (DAPI: 0.14 μg ml^{-1} in 2 × SSC buffer) and mounted with Antifade (Vectashield, Vector). Metaphase visualization was performed with a Nikon Eclipse E800 microscope and a DS-Qi1Mc CCD camera using the Nikon NIS-Elements software (v.5.42.01). Image processing was performed with Adobe Photoshop CS6 (v.13.1.3).

DNA isolation and sequencing

DNA was isolated using a QIAamp DNA Mini Kit (Qiagen), with an additional precipitation step with SDS/CH_3COOH (70 °C, 10 min). Samples presenting insufficient DNA yields were whole-genome-amplified using a REPLI-g Mini Kit (Qiagen) (Supplementary Table 5). DNA libraries were prepared using Illumina whole-genome protocols, multiplexed and sequenced on an Illumina NovaSeq 6000 platform to generate 150-base pair (bp) paired-end (PE) reads. Sequencing depth ranged between ~20× and 150×, depending on the type and purpose of each sample (Supplementary Table 5).

RNA isolation and sequencing

Total RNA was isolated using the RNeasy Mini Kit (Qiagen) from normal tissue samples (adductor muscle, gills, digestive system, mantle, foot, gonad and hemolymph) of 28 healthy cockles, and hemolymph samples from eight neoplastic cockles. RNA libraries were prepared using the Illumina TruSeq RNA library kit with the Illumina Ribo-Zero ribosomal RNA removal kit, and sequenced on an Illumina NovaSeq 6000 platform to generate 150-bp PE reads (insert size 250 bp, 100 million reads per sample).

Sequence read alignment

DNA reads were aligned to the reference genome assembly using BWA (v.0.7.17)[62] with default settings, and processed using samtools (v.1.9)[63] and bammarkduplicates (v.2.0.87). RNA reads were mapped to the reference genome using STAR (v.2.7.3a)[64]. Before alignment, 5 of 13 alignment parameters were optimized for one healthy (ENCE17_H_Pool) and one cancer sample (PACE17_656H). Default values were used for all parameters except the following: 'outFilterMismatchNmax = 33, seedSearchStartLmax = 50, AlignSJoverhangMin = 5, AlignSJDBoverhangMin = 3, outFilterType = BySJout'. Aligned reads were quantified with RSEM (v.1.3.1)[65] to produce tables of read counts and transcripts per million. A total of 14,067 genes were captured.

C. edule reference genome and transcriptome

Sampling, histopathology and cytogenetics. A large male *C. edule* specimen (weight 19.15 g, length 40 mm, height 37 mm), collected from Noia, Spain (42° 47′ 35.1″ N, 8° 54′ 42.5″ W) in November 2017, was selected as the reference animal. Histological examination confirmed absence of parasites or evident pathologies, and absence of cytogenetic aberrations was confirmed by surface spreading of synaptonemal complexes[66], stained as described above. Tissue samples from hemolymph, foot, gill, mantle, adductor muscle, digestive system, gonad and siphons were preserved in RNAlater (Qiagen), flash-frozen and stored at −80 °C.

Genome sequencing and assembly. Sequencing: A multiplatform approach was applied, combining short- and long-read sequencing. Illumina sequencing comprised PE libraries with insert sizes of 350, 550 and 850 bp, and mate-paired libraries with insert sizes of 2.5, 5, 8 and 10 kb, prepared using the Illumina TruSeq PCR-Free DNA (350-bp inserts) and Illumina TruSeq DNA library kits (550-bp and 850-bp inserts), and sequenced on Illumina NovaSeq 6000 and HiSeq 4000 platforms. Long-read sequencing was performed with Oxford Nanopore Technologies (ONT). Following end-repairing and dA-tailing (NEBNext End Repair/dA-tailing module, New England Biolabs), we constructed whole-genome libraries from unsheared DNA (SQK-LSK109, ONT) and sequenced them in MinION R9.4 flowcells (FLO-MIN106, ONT) controlled by the MinKNOW software (v.18.12.09). Base calling was performed using Guppy (v.2.3.1). Hi-C sequencing was performed using the Arima Genomics Hi-C v1 kit and one 150-bp PE library with the NEBNext Ultra II DNA library kit (New England Biolabs). Genome size, heterozygosity and GC content: GenomeScope (v.1.0.0) and wtdgb2 (v.2.5) were employed to estimate cockle genome size from short and long reads, respectively. The initial *k*-mer counting required by GenomeScope was assessed using Jellyfish (v.2.2.10) on 630 million PE reads (read size 100 bp, insert size 500 bp) from the reference animal. We ran GenomeScope using default parameters, estimating a haploid genome size of 812 Mb and 1.86% heterozygosity. The wtdgb2 assembly, using 50 gigabases (Gb) of ONT data with a minimum read length of 10 kb (ref. 67), estimated a haploid genome size of 840 Mb. The G + C content of the genome was 35.6%. Genome assembly: MaSuRCA (v.3.2.4)[68] was run on 50 Gb of ONT reads (depth 60×, minimum read length 10 kb) and 180 Gb (depth 143×) of Illumina 100-bp reads from five libraries: PE reads (depth 93×, insert size 550 bp) and mate-paired libraries (insert sizes 2.5, 5, 8 and 10 kb, total depth 50×). The resulting set of contigs was close to the theoretical diploid genome size (1.76 Gb). Homologous contigs were purged with purge_haplotigs (v.1.1.0) and HaploMerger2 (v.3.4), masked with WindowMasker (v.1.0.0) and fed back to HaploMerger2. Haplotig removal efficiency was assessed using the KAT toolkit (v.2.3.2). The resulting haploid contig set had a size of 793 Mb, N50 = 1.28 Mb (Supplementary Table 3) and BUSCO (v.3.0.2) completeness (using 'metazoa' dataset with '--long' option) of 95.2%. Scaffolding with Arima Hi-C reads was conducted using three rounds of 3D-DNA (v.180922). Hi-C reads were aligned to the scaffolds using BWA-MEM (v.0.7.17) with the '-5SP' setting. The output file was converted into a contact map and visualized using PretextMap and PretextView (v.0.0.2). The scaffolded genome had N50 = 39.6 Mb, with 95% of the genome contained in 19 chromosomal scaffolds. For genome polishing, we first ran GATK HaplotypeCaller (v.4.1.6.0)[69] to call SNVs and indels using 630 million PE reads (length 100 bp, insert size 500 bp). Then, we replaced reference alleles with alternate alleles presenting VAF ≥ 0.75, using varibase (v.1.0).

Transcriptome sequencing and assembly. RNA libraries were prepared using the Illumina TruSeq RNA kit with the Illumina Ribo-Zero rRNA removal kit, and sequenced on an Illumina HiSeq 2500 platform to generate 100-bp PE reads (insert size 250 bp). Reads were aligned to the reference assembly using HISAT2 (v.2.1.0)[70]. Alignments were assembled and merged into a nonredundant transcript set using StringTie (v.2.1.1). Final transcriptome size was 290 Mb, presenting 98.8% completeness on the BUSCO metazoan dataset (v.3.0.2) (Supplementary Table 3).

Genome annotation. TE sequences were identified with RepeatModeler (v.1.0.11) and used to locate TEs on the primary assembly with RepeatMasker (v.4.1.0). The last 6.3 Mb of chromosome 11 was masked with WindowMasker (v.1.0) to account for TE overrepresentation (Extended Data Fig. 2). This annotation approach yielded >2 million repetitive elements (48.7% of the genome; Supplementary Table 4). Gene predictions were obtained using two Maker2 runs (v.3.01.03), the first by supplying *C. edule* transcripts and proteins from bivalves *Mizuhopecten yessoensis* (GCA_002113885), *Crassostrea gigas* (GCA_000297895), *Crassostrea virginica* (GCA_002022765) and *Mytilus galloprovincialis* (GCA_001676915). The outcome of this round was used to train SNAP (v.0.15), and its output was fed into the second Maker2 round. *C. edule* TE sequences were also supplied to mask the genome. This approach identified 14,055 protein-coding genes. We performed Gene Ontology annotation using Blast2GO (v.1.4.5) against the BLAST 'nr' database and InterProScan2 (v.2); explored the metabolic pathways of these proteins with KEGG Automatic Annotation Server (v.2.1), using GHOSTX with bi-directional best hit against *Lottia gigantea*, *Pomacea canaliculata*, *Crassostrea gigas*, *Mizuhopecten yessoensis*, *Octopus bimaculoides*; and transferred functional orthology information using EggNOG (v.4.5.1).

Calling and filtering of mitochondrial SNVs and indels

Calling of SNVs and indels in mtDNA was performed using GATK MuTect2 (v.4.1.6.0)[71] in 'mitochondria mode' (option '-L MT'). A maximum of 100 reads were retained per alignment start position, and filtering of duplicates was disabled. Sites with median mapping quality >50 were omitted. An orientation bias model was used to filter the calls, and multi-nucleotide-variant calling was disabled. A median autosomal coverage of 50 was assumed to filter potential polymorphic nuclear mtDNA (NUMT) integrations; the autosomal coverage was estimated using samtools (v.1.9) on nuclear sequence data. The minimum number of supporting reads required on each strand was set to 1. Biallelic SNVs were filtered as follows: (1) For healthy specimens, for which all variants typically presented VAF ≈ 1, variants with 0.5 ≤ VAF < 1 had their VAF converted to 1, while variants with 0 < VAF < 0.5 had their VAF converted to 0. The case 0.5 ≤ VAF < 1 may be explained by read-mapping or coverage issues, unidentified CN variants or high-frequency heteroplasmy; while the case 0 < VAF < 0.5 probably corresponds to false positives and low-frequency heteroplasmic positions. (2) For tumor samples with a matched-host sample, we compared the mtDNA alignments between both samples and removed variants found exclusively in one sample (usually at low frequency). (3) For tumor samples without a matched-host sample, VAF distributions were visually inspected to determine a VAF threshold for variant acceptance. Biallelic indels were discarded, as they were almost exclusively found at low frequency in whole-genome-amplified samples, strongly suggesting their being artefacts. Multiallelic positions were individually examined across all samples, and labeled as true or false positives on the basis of concordance between their VAFs and those of most mtDNA variants.

Deconvolution of mtDNA haplotypes and co-occurrence analysis

Deconvolution of mtDNA haplotypes was performed by directly inspecting sample-specific VAF ranges, together with the estimated purity (tumor cell fraction). For a set of 51 neoplastic hemolymph samples, 42 matched-host tissue samples and 168 non-neoplastic cockle samples, this method allowed identification of tumor and host mtDNA haplotypes within each tumor and matched-host sample. The modal VAF of the tumor mtDNA haplotype in a given hemolymph sample was generally consistent with its tumor cell fraction. Those cases where host and tumor alleles could not be confidently assigned were excluded from the analysis. Mitochondrial genomes present in each sample were reconstructed and used to produce a multiple-sequence alignment. A small number of sample pairs showed more than two mtDNA haplotypes present in both sequenced tissues; these were interpreted to reflect co-occurrence of cells from two distinct tumor mtDNA lineages. However, samples presenting evidence for more than two haplotypes were conservatively discarded if they met any of the following conditions: (1) the third haplotype did not appear in both tissues of the individual; (2) the third haplotype appeared at very low frequency; (3) the third haplotype originated a long branch in the mtDNA phylogenetic tree, suggesting artefactual variants.

Phylogenetic inference from mtDNA variants

The alignment of deconvoluted mtDNA sequences was visually inspected using Genious Prime (v.11.03) to check correctness of reading frames across coding genes and basic alignment statistics. Region MT:9018–10168 was excluded due to existence of amplifications in some genomes, yielding an alignment length of 13,792 bp. As the mean divergence among sequences was low (~1%), preliminary neighbor-joining trees were used to examine the placement of uncertain haplotypes (see 'Deconvolution of mtDNA haplotypes and co-occurrence analysis').

ModelTest-NG (v.0.1.6)[72] was used to select the best-fitting nucleotide substitution model for the dataset. Models were estimated for each gene or region separately (30 regions; some regions overlapping transfer RNAs or intergenic sequences were merged), as well as for the complete dataset and for a three-partitioned dataset (coding regions, rRNAs and tRNAs). The best model in each case was selected according to the Bayesian Information Criterion. Phylogenetic relationships were inferred using ML and Bayesian inference. For ML, we used RAxML-NG (v.0.8.1) with ten starting parsimony trees and 1,000 bootstrap replicates. Partitioned analyses were implemented using the 30 partitions described above; exploratory analyses yielded identical results using one and three partitions. Bayesian inference analyses were conducted with BEAST (v.2.6.2), again implementing different models for the 30 a priori established partitions. Runs were implemented with a single or three partitions (coding regions, rRNAs and tRNAs), further partitioning being avoided to reduce bias on node ages[73]. Linked clock models and tree topology were used, with both coalescent and Yule priors on the tree topology. Multiple independent

MCMC chains were run for 200 million iterations, sampling every 20,000 iterations. At least two runs were performed. Convergence was checked with Tracer (v.1.7.1), and TreeAnnotator (v.2.6.2) was used to summarize posterior estimates.

For information on additional phylogenetic analyses, see the Supplementary Note.

Selection of high-purity tumor set

The purity (tumor cell fraction) of each sample was estimated by a combination of approaches: (1) manual counting of neoplastic hemocytes in cell monolayers (Supplementary Note); (2) assessment of changes in mtDNA VAF between tumor and matched-host samples; and (3) assessment of the placement of tumor mtDNA haplotypes in the mtDNA phylogenetic tree, to identify cases of host co-infection (see 'Deconvolution of mtDNA haplotypes and co-occurrence analysis'). A set of high-purity tumor samples (hereafter, the 'golden set') was defined by selecting samples that had sequencing depth ≥90 Gb, had purity estimates >97%, had not undergone whole-genome amplification and showed no evidence of co-infection by distinct tumor mtDNA lineages. This subset comprised ten samples: seven CedBTN1 samples (EICE18_889H, EUCE18_1024H, FRCE17_701H, PACE17_433H, PVCE17_1247H, PVCE17_1402H, PACE17_421H1; diagnosed as type A DN) and three CedBTN2 samples (EICE18_887H, EPCE18_848H, UGCE17_2401H; diagnosed as type B DN). The discrepancy in the number of samples from each lineage reflects the overall difference in prevalence across sampling locations (Supplementary Table 1).

Calling, filtering and annotation of nuclear SNVs and indels

Calling of SNVs and indels in the 'golden set' of tumor genomes was performed using GATK MuTect2 (v.4.1.6.0)[71] in 'tumor-only' mode with default settings. Variant calling in samples from healthy (non-neoplastic) cockles was performed using Platypus (v.0.8.1)[74] with default settings. Our cockle genome assembly was used as the reference sequence for all samples. MuTect2 calls were first filtered by assigning filter tags using the FilterMutectCalls tool in GATK (v.4.1.6.0), and then selecting calls showing only filter tags 'PASS' or 'clustered_events'. To isolate potentially somatic variants and filter contaminating germline variation from the hosts, we identified likely germline variants from the sets of tumor variants by comparing them against the combined set of 'PASS'-tagged variant calls obtained by Platypus across the 462 healthy samples in our 'panel of normals'. This approach was required for two reasons: (1) matched-host samples were found to contain substantial fractions of cancer cells, and were therefore unsuitable for filtering of host contamination in tumor samples; (2) because BTN is an allogeneic transplant, tumor cells are genetically unrelated to hosts, and thus the germline variation from the matched host does not capture the germline variation from the 'founder' animal that spawned the cancer lineage.

For information on filtering and annotation of SNVs and indels, see the Supplementary Note.

SV calling and filtering

SVs were called in high-purity tumors using a combination of three algorithms: DELLY (v.0.7.9)[75], LUMPY (v.0.2.13)[76] and Manta (v.1.6.0)[77]. DELLY was run in tumor-only mode with stringent read-filtering criteria (options '-q 20 -s 15') and an exclusion file containing annotated repeat coordinates ('-x'). LUMPY was run in tumor-only mode with discordant and split read pairs pre-extracted with samtools (v.1.9). Manta (v.1.6.0) was run in tumor-only mode ('--minEdgeObservations = 3', '--minCandidateSpanningCount = 3'). To limit false positives, we considered only candidate events with base-level breakpoint resolution, and belonging to the following SV categories: deletions, duplications, inversions and breakends (or BNDs, including translocations). We integrated SV calls using svimmer (v.0.1), requiring events to have been called by at least Manta and one other caller. We genotyped all candidate SVs using GraphTyper (v.2.0) with default settings.

For information on filtering and annotation of SVs, see the Supplementary Note.

CN inference

CN calling was performed with DELLY (v.1.0.3) using our cockle genome assembly to correct for read mappability and GC content. The minimum CN alteration size was set to 10 kb. Read counts were obtained for variable-size bins with 10-kb uniquely mappable (mapping quality ≥ 10) sequence. These bins were constructed by first simulating 2 × 150-bp PE reads from the reference genome using dicey (v.0.1.8) with otherwise default parameters. Simulated reads were aligned back to the reference genome using BWA-MEM (v.0.7.17), sorted, converted into BAM format and indexed with samtools (v.1.9). The final mappability map was generated using the 'dicey mappability2' tool with default parameters. CN segments were called for a range of ploidy values between 2*n* and 6*n*, and the most likely ploidy was then selected for each sample as the value providing the best fit between expected and observed CN modes. The most likely ploidy was found to be 4*n* for all samples except UGCE17/2401H (best fit by 5*n*), PACE17/421H1 and PACE17/433H (for both of which 4*n* was assumed, as their CN distributions were uninformative).

Phylogenetic inference from nuclear variants

Tumor phylogenetic trees were estimated from nuclear sets of BTN-specific SNVs and SVs in the ten high-purity tumors. For SNVs, variant alleles were concatenated into an alignment containing 4,340,713 sites and 3,724 different site patterns. ML trees were estimated with RAxML (v.8.2.12) using a single partition for the whole nuclear genome. A GTRGAMMA substitution model was assumed, given the low number of sequences and high number of sites. Stamatakis ascertainment bias correction was applied, incorporating exact nucleotide frequencies of invariable sites along the partition. A hundred trees were generated by optimizing alternative parsimony starting trees, and the tree with the best Gamma-based likelihood was selected. Tree consistency was evaluated using nonparametric bootstrap analysis with 1,000 replicates. This tree was rooted using the reference sequence as an outgroup.

For SVs, binary genotypes derived from GraphTyper (v.0.2) were concatenated into an alignment using functions from the phangorn (v.2.8.1) R package. Heuristic parsimony tree searches were performed with the implementation of the parsimony ratchet[78] in phangorn. To evaluate the level of homoplasy, tree consistency indexes (CI) were calculated for the alternative phylogenies estimated from different types of SVs: deletions (CI = 0.81), duplications (CI = 0.91), inversions (CI = 0.79) and breakends (CI = 0.77). A maximum parsimony tree search was performed using PAUP* (v.4.0a168). An alignment of 18,272 SV binary genotypes was analyzed, encoding the variants as unordered reversible characters with equal weights, and an exhaustive parsimony tree search was performed. Rooting was done using a user-specified outgroup corresponding to the reference sequence. Consistency of the tree was evaluated using nonparametric bootstrap analysis with 1,000 replicates.

For information on additional phylogenetic analyses, see the Supplementary Note.

Mutational signature analysis

Mutational signatures were inferred from sets of BTN-specific and germline variants using the sigfit (v.2.2.0)[38] R package. First, mutational catalogs were produced ('build_catalogues' function) from eight nonoverlapping SNV sets: (i) germline variants from three normal samples (BNg14, ENCE17_3575F, ICCE19_366F_HC); (ii) variants ancestral to both CedBTN clones (that is, present in all ten tumor samples); (iii) predivergence variants in CedBTN1 (present in all CedBTN1 and no CedBTN2 samples); (iv) predivergence variants in CedBTN2; (v) nonprivate postdivergence variants in CedBTN1 (present in at least two, but not all, CedBTN1 samples, and no CedBTN2 samples); (vi) nonprivate postdivergence variants in CedBTN2; (vii) private variants in CedBTN1 (present in exactly one CedBTN1 and no CedBTN2 samples); and (viii) private variants in CedBTN2. Mutational catalogs were corrected by the trinucleotide context frequencies of the reference genome using the 'convert_signatures' function, and then multiplied by the total mutation counts in the original catalogs. To prevent large mutation count differences, catalogs with >100,000 mutations were downsampled to this number.

Inference of mutational signatures was performed in three stages. First, sets of 2–4 signatures were extracted ('extract_signatures' function) from the mutational catalogs obtained from variant sets (i)–(iv) (germline and predivergence). The number of signatures yielding the cleanest signature deconvolution (based on goodness-of-fit, low redundancy and orthogonality of signatures) was $N = 3$ (signatures SBS-A, SBS-B, SBS-C). Next, the sigfit 'Fit-Ext' model[38] was used to fit these three signatures to the mutational catalogs from variant sets (v)–(viii) (postdivergence), while simultaneously extracting 1–3 additional signatures ('fit_extract_signatures' function). In this case, the best-supported number of additional signatures was $M = 2$ (signatures SBS-D, SBS-E), resulting in a total of five inferred signatures. Finally, the five signatures were fitted to all eight mutational catalogs ('fit_signatures' function) to estimate signature exposures. Signatures SBS-D and SBS-E were found to have nonsignificant exposures in variant sets (i) and (ii) (germline); therefore, more accurate exposures were obtained for these two sets by re-fitting signatures SBS-A to SBS-C only. Comparison of the inferred signatures against human mutational signatures in the COSMIC database (v.3.2) by means of cosine similarity yielded the following correspondence for SBS-A to SBS-E: SBS1 (similarity 0.84), SBS5 (0.86), SBS8 (0.80), SBS23 (0.81), SBS40 (0.65). Because COSMIC signatures are relative to the sequence composition of the human genome, whereas signatures SBS-A to SBS-E were inferred from genome-independent catalogs, COSMIC signatures were also transformed to a genome-independent representation ('convert_signatures' function), which led to the following correspondence for SBS-A to SBS-E: SBS5 (0.90), SBS40 (0.79), SBS8 (0.82), SBS23 (0.86), SBS40 (0.71). Mutational spectra of indels obtained from the variant sets described above were generated using the 'indel.spectrum' function in the Indelwald tool (version 24/09/2021; github.com/Maximilian-Stammnitz/Indelwald) and compared against human indel signatures in the COSMIC database (v.3.2).

Selection analyses

Evidence of selection for somatic mutations in protein-coding genes was assessed using normalized nonsynonymous-to-synonymous substitution ratios (d*N*/d*S*) for BTN-specific variants. dNdScv (v.0.0.1.0)[25] was used to estimate d*N*/d*S* ratios for somatic missense and truncating substitutions (SNVs) and indels. A reference CDS database (RefCDS) was built from the gene annotation for the reference genome assembly using the 'buildref' function in dNdScv. The 'dndscv' function was applied to two subsets of BTN-specific SNVs and indels: (1) 'postdivergence' variants in either clone, defined as those present in any sample from either CedBTN1 or CedBTN2, but not present in all samples from the same clone, nor in any sample from the other clone; and (2) all nonshared variants, defined as those variants present in only one clone. Variants shared by both clones were excluded, as these are likely germline. 'dndscv' was run with options 'max_coding_muts_per_sample = Inf', 'max_muts_per_gene_per_sample = Inf', 'cv = NULL', 'refdb = RefCDS'. No genes with d*N*/d*S* ratios significantly different from 1.0 were identified for any mutation type.

For information on additional d*N*/d*S* analyses related to the CedBTN phylogeny, see the Supplementary Note.

Identification of candidate driver mutations

To identify candidate early cancer-driver mutations in CedBTN, a set of cancer gene orthologs was first defined. The COSMIC Cancer Gene Census database of genes causally involved in human cancer (COSMIC v.95) was retrieved. EggNOG gene identifiers were used to find *C. edule* orthologs, rendering 226 putative cancer genes across the cockle reference genome. A screen for potential early driver mutations in CedBTN1 and CBTN2 was conducted by searching for tumor-specific SNVs, indels and SVs satisfying the following criteria: (1) the variant occurred predivergence; (2) the variant affects the coding sequence of a cancer gene ortholog; (3) the variant is nonsynonymous; and (4) the type of mutation matches one of the mutation types listed for the overlapping gene in the Cancer Gene Census. This search yielded the list of early mutations in cancer genes reported in Supplementary Table 13. However, the combination of mutation consequence and gene CN did not provide sufficient evidence that these events had affected the genes in a manner consistent with biological knowledge, and therefore none was considered a high-confidence candidate early driver mutation.

Identification of candidate driver genes in CedBTN1 and CedBTN2 was also performed through the detection of CN changes associated with ancestral inactivation of a TSG, or ancestral amplification of an oncogene. CN estimation for each cancer gene was obtained through the intersection of unrounded CN segments with cancer gene chromosomal coordinates. Gene-wise CN was set as the average from intersected segments, normalized by their relative size. Gene-wise CN estimates were used to search for driver candidates in CedBTN1 and CedBTN2. Potential candidate driver oncogenes were defined as those with ancestral amplifications with CN > 6. Potential candidate driver TSGs were defined as those with ancestral losses with CN < 2. Candidate driver TSGs with multiple genomic copies, or with average CN status decrease indicative of hemizygous deletion, were further inspected in search of additional deleterious variants disrupting remaining alleles (homozygous inactivation). Candidate driver genes with the strongest support were assessed for changes in gene expression relative to normal cockle tissues, using the results of a differential gene expression analysis performed as described below (Extended Data Fig. 10c).

Cancer histogenesis determination via gene expression analysis

Raw RNA sequence read counts were normalized via regularized log transformation using the DESeq2 (v.1.34.0) R package. Genes that were significantly upregulated in a specific tissue type were identified by differential gene expression analysis, using the 'DESeq' function, with a design including contrasts between each tissue type and all other types combined. The top 60 genes for each of these contrasts, defined as those with the lowest adjusted *P* values, were selected as genes with 'tissue-specific expression'. Both unsupervised hierarchical clustering and principal component analysis were performed on the set of normalized tissue-specific gene expression values, using the functions 'dist' and 'prcomp' in R. A heatmap of gene expression values was produced using the ComplexHeatmap (v.2.10.0) R package, with clustering based on Pearson's correlation.

Satellite DNA identification and analysis

Repetitive elements were recovered with RepeatExplorer (v.2.3.8.1) from Illumina PE data from representative healthy and neoplastic specimens. Subsequently, a comparative analysis was performed on a larger dataset including 30 healthy cockles (uniformly representing all sampled populations) and the ten high-purity tumors (200,000 reads per sample). Reads were aligned to the repetitive elements using BWA-MEM (v.0.7.17), and filtered (mapping quality (MAPQ) ≥ 60 and alignment score (AS) > 70) to assess the relative abundance of each repetitive element in healthy and neoplastic genomes. We generated DNA probes of satellite CeS4 labeled with digoxigenin-11-dUTP (10 × DIG Labeling Mix, Roche) by PCR with primers TACATTTTTGTGACGTTGAGAGGC and GGAGTTAGACAAAAACTATTGCTC. FISH experiments for this satellite and other gene families (28S and 5S rDNAs and histone H3) were performed following published protocols[79,80]. Telomeric repeats were detected with a commercial telomeric (C3TA2)3 probe (Applied Biosystems).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The reference genome sequence for *Cerastoderma edule* and sequencing data supporting the findings of this study have been deposited in the European Nucleotide Archive (www.ebi.ac.uk/ena) under overarching accession code PRJEB58149. Human mutational signatures were retrieved from the COSMIC v.3.2 database (cancer.sanger.ac.uk). Source data are provided with this paper. All other data supporting the findings of this study are available from the corresponding authors on reasonable request.

Code availability

Computer code used for data analyses is available on GitLab (gitlab.com/mobilegenomesgroup/scuba_cancers).

References

1. Murchison, E. P. Clonally transmissible cancers in dogs and Tasmanian devils. *Oncogene* **27**, S19–S30 (2009).
2. Metzger, M. J. & Goff, S. P. A sixth modality of infectious disease: contagious cancer from devils to clams and beyond. *PLoS Pathog.* **12**, e1005904 (2016).
3. Cohen, D. The canine transmissible venereal tumor: a unique result of tumor progression. *Adv. Cancer Res.* **43**, 75–112 (1985).
4. Murgia, C., Pritchard, J. K., Kim, S. Y., Fassati, A. & Weiss, R. A. Clonal origin and evolution of a transmissible cancer. *Cell* **126**, 477–487 (2006).
5. Murchison, E. P. et al. Transmissible dog cancer genome reveals the origin and history of an ancient cell lineage. *Science* **343**, 437–440 (2014).
6. Pearse, A.-M. & Swift, K. Transmission of devil facial-tumour disease. *Nature* **439**, 549–549 (2006).
7. Murchison, E. P. et al. Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer. *Cell* **148**, 780–791 (2012).
8. Stammnitz, M. R. et al. The evolution of two transmissible cancers in Tasmanian devils. *Science* **380**, 283–293 (2023).
9. Metzger, M. J., Reinisch, C., Sherry, J. & Goff, S. P. Horizontal transmission of clonal cancer cells causes leukemia in soft-shell clams. *Cell* **161**, 255–263 (2015).
10. Metzger, M. J. et al. Widespread transmission of independent cancer lineages within multiple bivalve species. *Nature* **534**, 705–709 (2016).
11. Yonemitsu, M. A. et al. A single clonal lineage of transmissible cancer identified in two marine mussel species in South America and Europe. *eLife* **8**, e47788 (2019).
12. Skazina, M. et al. First description of a widespread *Mytilus trossulus*-derived bivalve transmissible cancer lineage in *M. trossulus* itself. *Sci. Rep.* **11**, 5809 (2021).
13. Garcia-Souto, D. et al. Mitochondrial genome sequencing of marine leukaemias reveals cancer contagion between clam species in the Seas of Southern Europe. *eLife* **11**, e66946 (2022).
14. Michnowska, A., Hart, S. F. M., Smolarz, K., Hallmann, A. & Metzger, M. J. Horizontal transmission of disseminated neoplasia in the widespread clam *Macoma balthica* from the Southern Baltic Sea. *Mol. Ecol.* **31**, 3128–3136 (2022).
15. Carballal, M. J., Barber, B. J., Iglesias, D. & Villalba, A. Neoplastic diseases of marine bivalves. *J. Invertebr. Pathol.* **131**, 83–106 (2015).
16. Elston, R. A., Kent, M. & Drum, A. Progression, lethality and remission of hemic neoplasia in the bay mussel *Mytilus edulis*. *Dis. Aquat. Organ.* **4**, 135–142 (1988).
17. Burioli, E. A. V. et al. Implementation of various approaches to study the prevalence, incidence and progression of disseminated neoplasia in mussel stocks. *J. Invertebr. Pathol.* **168**, 107271 (2019).
18. Hayward, P. J. & Ryland, J. S. *Handbook of the Marine Fauna of North-West Europe* (Oxford Univ. Press, 2017).
19. Twomey, E. & Mulcahy, M. F. A proliferative disorder of possible hemic origin in the common cockle, *Cerastoderma edule*. *J. Invertebr. Pathol.* **44**, 109–111 (1984).
20. Villalba, A., Carballal, M. J. & López, C. Disseminated neoplasia and large foci indicating heavy haemocytic infiltration in cockles *Cerastoderma edule* from Galicia (NW Spain). *Dis. Aquat. Organ.* **46**, 213–216 (2001).
21. Carballal, M. J., Iglesias, D., Santamarina, J., Ferro-Soto, B. & Villalba, A. Parasites and pathologic conditions of the cockle *Cerastoderma edule* populations of the coast of Galicia (NW Spain). *J. Invertebr. Pathol.* **78**, 87–97 (2001).
22. Díaz, S., Iglesias, D., Villalba, A. & Carballal, M. J. Long-term epidemiological study of disseminated neoplasia of cockles in Galicia (NW Spain): temporal patterns at individual and population levels, influence of environmental and cockle-based factors and lethality. *J. Fish Dis.* **39**, 1027–1042 (2016).
23. Hammel, M. et al. Prevalence and polymorphism of a mussel transmissible cancer in Europe. *Mol. Ecol.* **31**, 736–751 (2022).
24. Miyata, T. & Yasunaga, T. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J. Mol. Evol.* **16**, 23–36 (1980).
25. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).
26. Baez-Ortega, A. et al. Somatic evolution and global expansion of an ancient transmissible cancer lineage. *Science* **365**, eaau9923 (2019).
27. Martínez, L., Freire, R., Arias-Pérez, A., Méndez, J. & Insua, A. Patterns of genetic variation across the distribution range of the cockle *Cerastoderma edule* inferred from microsatellites and mitochondrial DNA. *Mar. Biol.* **162**, 1393–1406 (2015).
28. Hart, S. F. M. et al. Centuries of genome instability and evolution in soft-shell clam, *Mya arenaria*, bivalve transmissible neoplasia. *Nat. Cancer* https://doi.org/10.1038/s43018-023-00643-7 (2023).
29. Rebbeck, C. A., Leroi, A. M. & Burt, A. Mitochondrial capture by a transmissible cancer. *Science* **331**, 303–303 (2011).
30. Strakova, A. et al. Recurrent horizontal transfer identifies mitochondrial positive selection in a transmissible cancer. *Nat. Commun.* **11**, 3059 (2020).
31. Spees, J. L., Olson, S. D., Whitney, M. J. & Prockop, D. J. Mitochondrial transfer between cells can rescue aerobic respiration. *Proc. Natl Acad. Sci. USA* **103**, 1283–1288 (2006).
32. Tan, A. S. et al. Mitochondrial genome acquisition restores respiratory function and tumorigenic potential of cancer cells without mitochondrial DNA. *Cell Metab.* **21**, 81–94 (2015).
33. Saha, T. et al. Intercellular nanotubes mediate mitochondrial trafficking between cancer and immune cells. *Nat. Nanotechnol.* **17**, 98–106 (2022).
34. Cross, M. E. et al. Genetic evidence supports recolonisation by *Mya arenaria* of western Europe from North America. *Mar. Ecol. Prog. Ser.* **549**, 99–112 (2016).
35. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
36. Yuan, Y. et al. Comprehensive molecular characterization of mitochondrial genomes in human cancers. *Nat. Genet.* **52**, 342–352 (2020).

37. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
38. Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational signatures. Preprint at bioRxiv https://doi.org/10.1101/372896 (2020).
39. Lindahl, T. & Nyberg, B. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry* **13**, 3405–3410 (1974).
40. Zou, X. et al. A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. *Nat. Cancer* **2**, 643–657 (2021).
41. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
42. Singh, V. K., Rastogi, A., Hu, X., Wang, Y. & De, S. Mutational signature SBS8 predominantly arises due to late replication errors in cancer. *Commun. Biol.* **3**, 1–10 (2020).
43. Le Grand, F. et al. Prevalence, intensity, and aneuploidy patterns of disseminated neoplasia in cockles (*Cerastoderma edule*) from Arcachon Bay: seasonal variation and position in sediment. *J. Invertebr. Pathol.* **104**, 110–118 (2010).
44. Díaz, S. et al. Disseminated neoplasia causes changes in ploidy and apoptosis frequency in cockles *Cerastoderma edule*. *J. Invertebr. Pathol.* **113**, 214–219 (2013).
45. Matias, A. M. et al. Karyotype variation in neoplastic cells associated to severity of disseminated neoplasia in the cockle *Cerastoderma edule*. *Aquaculture* **428–429**, 223–225 (2014).
46. Oliner, J. D., Saiki, A. Y. & Caenepeel, S. The role of MDM2 amplification and overexpression in tumorigenesis. *Cold Spring Harb. Perspect. Med.* **6**, a026336 (2016).
47. Kato, S. et al. Analysis of MDM2 amplification: next-generation sequencing of patients with diverse malignancies. *JCO Precis. Oncol.* https://doi.org/10.1200/PO.17.00235 (2018).
48. Büschges, R. et al. Amplification and expression of cyclin D genes (CCND1 CCND2 and CCND3) in human malignant gliomas. *Brain Pathol.* **9**, 435–442 (1999).
49. Kasugai, Y. et al. Identification of CCND3 and BYSL as candidate targets for the 6p21 amplification in diffuse large B-cell lymphoma. *Clin. Cancer Res.* **11**, 8265–8272 (2005).
50. Pegg, A. E., Dolan, M. E. & Moschel, R. C. in *Progress in Nucleic Acid Research and Molecular Biology*, Vol. 51 (eds Cohn, W. E. & Moldave, K.) 167–223 (Academic Press, 1995).
51. Shiraishi, A., Sakumi, K. & Sekiguchi, M. Increased susceptibility to chemotherapeutic alkylating agents of mice deficient in DNA repair methyltransferase. *Carcinogenesis* **21**, 1879–1883 (2000).
52. Lower, S. S., McGurk, M. P., Clark, A. G. & Barbash, D. A. Satellite DNA evolution: old ideas, new approaches. *Curr. Opin. Genet. Dev.* **49**, 70–78 (2018).
53. Davoli, T. & de Lange, T. The causes and consequences of polyploidy in normal development and cancer. *Annu. Rev. Cell Dev. Biol.* **27**, 585–610 (2011).
54. Gemble, S. et al. Genetic instability from a single S phase after whole-genome duplication. *Nature* **604**, 146–151 (2022).
55. Santaguida, S. & Amon, A. Short- and long-term effects of chromosome mis-segregation and aneuploidy. *Nat. Rev. Mol. Cell Biol.* **16**, 473–485 (2015).
56. Ly, P. et al. Chromosome segregation errors generate a diverse spectrum of simple and complex genomic rearrangements. *Nat. Genet.* **51**, 705–715 (2019).
57. Crockford, A. et al. Cyclin D mediates tolerance of genome-doubling in cancers with functional p53. *Ann. Oncol.* **28**, 149–p156 (2017).
58. Matsuo, H. et al. Recurrent CCND3 mutations in MLL-rearranged acute myeloid leukemia. *Blood Adv.* **2**, 2879–2889 (2018).
59. Walker, C., Böttger, S. & Low, B. Mortalin-based cytoplasmic sequestration of p53 in a nonmammalian cancer model. *Am. J. Pathol.* **168**, 1526–1530 (2006).
60. Pye, R. J. et al. A second transmissible cancer in Tasmanian devils. *Proc. Natl Acad. Sci. USA* **113**, 374–379 (2016).
61. García-Souto, D. et al. Methylation profile of a satellite DNA constituting the intercalary G+C-rich heterochromatin of the cut trough shell *Spisula subtruncata* (Bivalvia, Mactridae). *Sci. Rep.* **7**, 6930 (2017).
62. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]* (2013).
63. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
64. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
65. Parrish, N., Hormozdiari, F. & Eskin, E. Assembly of non-unique insertion content using next-generation sequencing. *BMC Bioinformatics* **12**, S3 (2011).
66. Hurtado, N. S. & Pasantes, J. J. Surface spreading of synaptonemal complexes in the clam *Dosinia exoleta* (Mollusca, Bivalvia). *Chromosome Res.* **13**, 575–580 (2005).
67. Lu, H., Giordano, F. & Ning, Z. Oxford Nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinformatics* **14**, 265–279 (2016).
68. Zimin, A. V. et al. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res.* **27**, 787–792 (2017).
69. Poplin, R. et al. Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at bioRxiv https://doi.org/10.1101/201178 (2017).
70. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
71. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
72. Darriba, D. et al. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. *Mol. Biol. Evol.* **37**, 291–294 (2020).
73. Jin, Y. & Brown, R. P. Partition number, rate priors and unreliable divergence times in Bayesian phylogenetic dating. *Cladistics* **34**, 568–573 (2018).
74. Rimmer, A. et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).
75. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
76. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
77. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
78. Nixon, K. C. The Parsimony Ratchet, a new method for rapid parsimony analysis. *Cladistics* **15**, 407–414 (1999).
79. García-Souto, D., Pérez-García, C., Morán, P. & Pasantes, J. J. Divergent evolutionary behavior of H3 histone gene and rDNA clusters in venerid clams. *Mol. Cytogenet.* **8**, 40 (2015).
80. Insua, A., Freire, R. & Méndez, J. The 5S rDNA of the bivalve Cerastoderma edule: nucleotide sequence of the repeat unit and chromosomal location relative to 18S–28S rDNA. *Genet. Sel. Evol.* **31**, 509 (1999).

Acknowledgements

This research was funded by the European Research Council (ERC) Starting Grant no. 716290 ('SCUBA CANCERS'), awarded to J.M.C.T. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. Sampling research was carried out mainly at the Universidade de Vigo's Centro de Investigación Mariña, supported by the 'Excellence in Research (INUGA)' Program from the Regional Council of Culture, Education and Universities, and co-funded by the European Union through the ERDF Operational Program Galicia 2014–2020 'A way to make Europe'. Molecular biology and bioinformatics research was carried out mainly at the Centre for Research in Molecular Medicine and Chronic Diseases (CiMUS), supported by the European Regional Development Fund 'A way to make Europe' and the Research Centre of the Galician University System (2019–2022). A.L.B. was supported by a predoctoral fellowship from the Spanish Ministry of Economy, Industry, and Competitiveness (grant no. BES2016/078166); received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 730984 ASSEMBLE PLUS Transnational Access; and received a travel grant from Boehringer Ingelheim Fonds. M. Santamarina was supported by a predoctoral fellowship from the Spanish regional government of Xunta de Galicia (grant no. ED481A-2017/299). D.G.-S. was supported by postdoctoral contracts from Xunta de Galicia (grant nos. ED481B-2018/091 and ED481D 2022/001). S.D. received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 730984 ASSEMBLE PLUS Transnational Access. I.O. was supported by a predoctoral fellowship from the Spanish regional government of Xunta de Galicia Consellería de Cultura, Educación y Universidad (grant no. ED481A 2021/096). J.R.-C. was partially supported by the program to structure and improve research centers (Centros Singulares 2019, CiMUS). T.P. was supported by a predoctoral fellowship from the Spanish Government (grant no. FPU15/03709) and a predoctoral fellowship from the Spanish regional government of Xunta de Galicia (grant no. ED481A-2015/083). L.T. was supported by a predoctoral fellowship from the Spanish regional government of Xunta de Galicia (grant no. ED481A-2018/303). A.M.A. received funding from Portuguese national funds FCT (Foundation for Science and Technology) through project nos. UIDB/04326/2020, UIDP/04326/2020 and LA/P/0101/2020; and from the operational programs CRESC Algarve 2020 and COMPETE 2020 through project no. EMBRC-PT ALG-01-0145-FEDER-022121. R.C. was supported by FCT/MCTES (grant nos. UIDP/50017/2020, UIDB/50017/2020, LA/P/0094/2020), through national Portuguese funds. M. Skazina was supported by the Russian Science Foundation, grant no. 19-74-20024. N.G.P. provided biological resources supplied by EMBRC-ERIC. K.S. was supported by the National Centre of Science (Poland) with grant no. UMO-2017/26/M/NZ8/00478. J.J.P. was supported by the Spanish regional government of Xunta de Galicia (grant no. ED431C 2020/05), and Fondos Feder 'Unha maneira de facer Europa'. Z.N. was supported by Wellcome grant no. WT098051. Y.S.J. was supported by a grant from the National Research Foundation of Korea funded by the Korean Government (grant no. NRF-2020R1A3B2078973). D.P. was supported by the Spanish Ministry of Science and Innovation, MICINN (grant no. PID2019-106247GB-I00 awarded to D.P.), the ERC (grant no. ERC-617457-PHYLOCANCER awarded to D.P.) and the Spanish regional government of Xunta de Galicia. J.D. was supported by a postdoctoral fellowship from the Belgian Research Foundation, Flanders (FWO; grant no. 12J6921N), and the Flemish Institute for Biotechnology (VIB). We thank the staff of the fishermen's associations ('cofradías') of Galicia for their advice and assistance with sampling, especially L. Solís, J. Alfaya and A. Simón; and the Galicia Supercomputing Centre (CESGA) for the availability of informatic resources. We thank L. F. Møller (National Institute of Aquatic Resources, Denmark), B. Hussel (Alfred Wegener Institute), M. Wolowicz (University of Gdansk), T. Verstraeten (Ghent University), M. L. Martínez and A. M. Insua Pombo (Universidade da Coruña), C. García de Leaniz (Swansea University), A. Smith and A. Harvey (Marine Biological Association, UK), R. Parks (Centre for Environment, Fisheries and Aquaculture Science, UK) and T. Magnesen (Universitetet i Bergen) for providing samples for this project. We thank C. Canchaya (Universidade de Vigo), M. Rey, J. Quinteiro, M. Hermida and P. Martínez (Universidade de Santiago de Compostela) for advice on genome assembly and annotation; M. Rodríguez (Universidade de Vigo) for administrative support; and E. P. Murchison (University of Cambridge) for discussion and critical reading of the manuscript.

Author contributions

A. Villalba, D.P. and J.M.C.T. designed the project. A.L.B., M. Santamarina, D.G.-S., S.D., S.R., J.Z., M.A.Q., I.O., J.J.P., J.D. and A.B.-O. developed methods. A.L.B., M. Santamarina, D.G.-S., S.D., S.R., J.Z., Y.L., I.O., J.T., Y.S.J., J.D. and A.B.-O. performed computational analyses. T.P., L.T., J.A., Z.N. and D.P. assisted with analyses. A.L.B., M. Santamarina, D.G.-S., S.D., L.A., A.V.-C., A. Villanueva, A.P.-V., A.V.-F., J.T., J.R.-C., P.A., J.A. and J.J.P. performed laboratory work. A.L.B., D.G.-S., S.D., M.A.Q., A.P.-V., J.T. and J.R.-C. performed sequencing methods. A.L.B., D.G.-S., S.D., A.V.-F., A. Villanueva, D.C., R.R., J.A., A.M.A., P.B., R.C., B.E.K., U.I., X.M., N.G.P., I.P., F.R., P.R., M. Skazina and K.S. provided samples. A.L.B., M. Santamarina, D.G.-S., S.D., S.R., J.Z., Y.L., J.J.P., Y.S.J., D.P., J.D. and A.B.-O. helped with interpretation of results. A.L.B., D.G.-S., S.D., A.P.-V., J.R.-C., A. Villanueva, P.A. and J.A. performed sample management. A.C., D.I., M.J.C., A. Villalba, Z.N. and D.P. provided technical advice. A.L.B., M. Santamarina, D.G.-S., S.D., Y.L., J.Z., J.D. and A.B.-O. generated figures. A.B.-O. and J.M.C.T. wrote the manuscript with contributions from all other authors. A.L.B., M. Santamarina, D.G.-S., S.D. and S.R. contributed equally. J.Z., Y.L. and A.V.-F. contributed equally.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s43018-023-00641-9.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43018-023-00641-9.

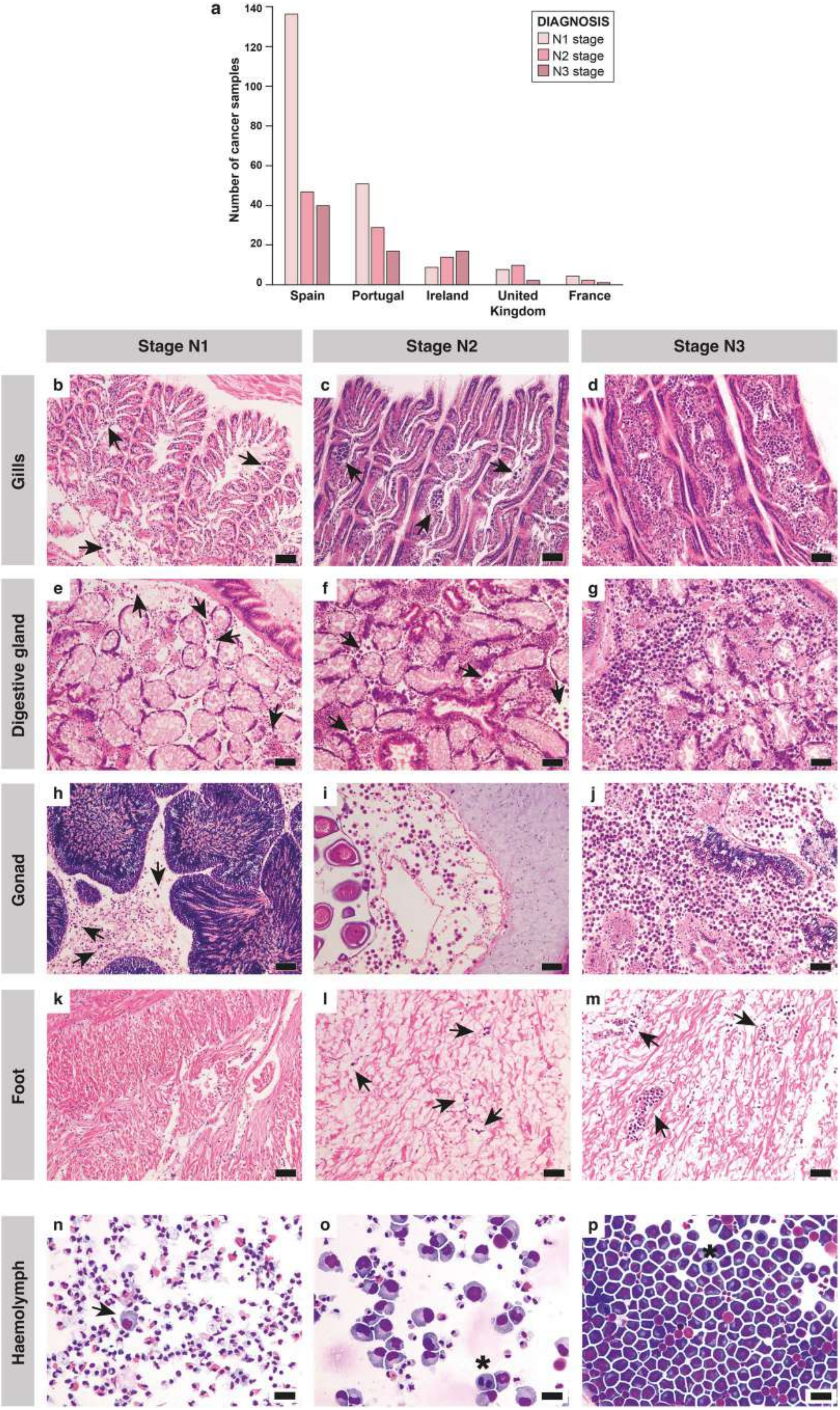
Correspondence and requests for materials should be addressed to Adrian Baez-Ortega or Jose M. C. Tubio.

Peer review information *Nature Cancer* thanks Andreas Bergthaler, Kelley Thomas and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

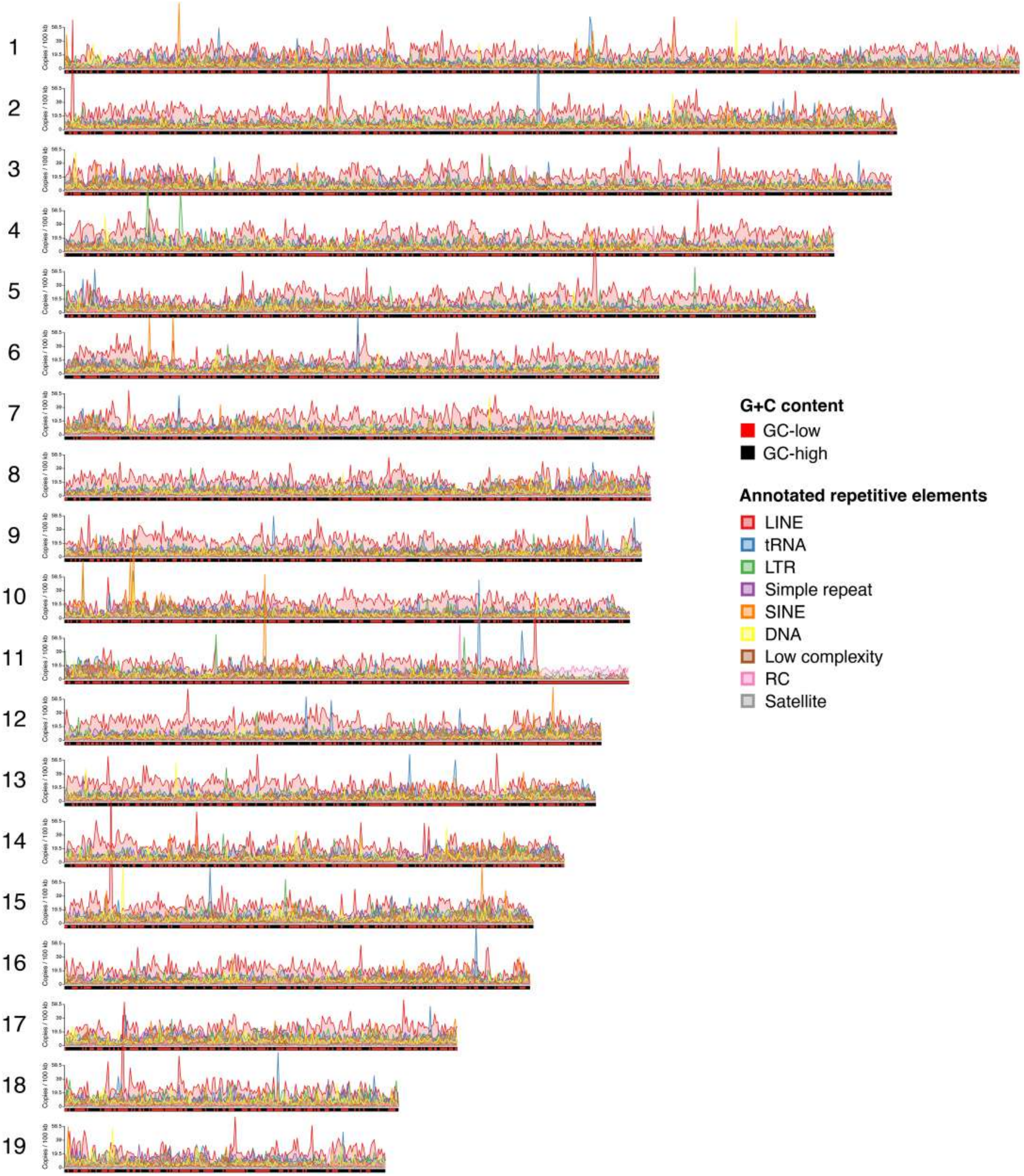
[1]Genomes and Disease, Centre for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain. [2]Department of Zoology, Genetics and Physical Anthropology, Universidade de Santiago de Compostela, Santiago de Compostela, Spain. [3]Instituto de Investigaciones Sanitarias de Santiago de Compostela (IDIS), Santiago de Compostela, Spain. [4]Wellcome Sanger Institute, Hinxton, UK. [5]ECOMARE, Centre for Environmental and Marine Studies (CESAM) & Department of Biology, University of Aveiro, Aveiro, Portugal. [6]CINBIO, Universidade de Vigo, Vigo, Spain. [7]Graduate School of Medical Science and Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea. [8]Department of Biochemistry, Genetics and Immunology, Universidade de Vigo, Vigo, Spain. [9]Centro de Investigación Mariña (CIM-ECIMAT), Universidade de Vigo, Vigo, Spain. [10]Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, Vigo, Spain. [11]New York Genome Center, New York, NY, USA. [12]Centro de Investigacións Mariñas (CIMA), Consellería do Mar, Xunta de Galicia, Vilanova de Arousa, Spain. [13]Centro de Ciencias do Mar do Algarve (CCMAR), Universidade do Algarve, Faro, Portugal. [14]Department of Biological Sciences, University of Bergen, Bergen, Norway. [15]NORCE AS, Bergen, Norway. [16]Laboratory of Physiopathology, Molecular Genetics & Biotechnology, Faculty of Sciences Ain Chock, Health and Biotechnology Research Centre, Hassan II University of Casablanca, Casablanca, Morocco. [17]Research Centre for Experimental Marine Biology and Biotechnology (PiE-UPV/EHU), University of the Basque Country (UPV/EHU), Plenzia-Bitzkaia, Spain. [18]Cell Biology in Environmental Toxicology Research Group, University of the Basque Country (UPV/EHU), Leioa-Bizkaia, Spain. [19]UMR EPOC5805, University of Bordeaux, Arcachon, France. [20]European Marine Biology Resources Centre (EMBRC-ERIC), Paris, France. [21]FR2424 Station Biologique de Roscoff, Sorbonne University/CNRS, Roscoff, France. [22]Department of Applied Ecology, St Petersburg State University, St Petersburg, Russia. [23]Department of Marine Ecosystem Functioning, University of Gdańsk, Gdynia, Poland. [24]Department of Life Sciences, Universidad de Alcalá, Alcalá de Henares, Spain. [25]VIB–KU Leuven Center for Cancer Biology, Leuven, Belgium. [26]Department of Oncology, KU Leuven, Leuven, Belgium. [27]The Francis Crick Institute, London, UK. [28]Magdalene College, University of Cambridge, Cambridge, UK. [29]These authors contributed equally: Alicia L. Bruzos, Martín Santamarina, Daniel García-Souto, Seila Díaz, Sara Rocha. [30]These authors jointly supervised this work: Adrian Baez-Ortega, Jose M. C. Tubio. ✉e-mail: ab2324@cam.ac.uk; jose.mc.tubio@usc.es

https://doi.org/10.1038/s43018-023-00641-9

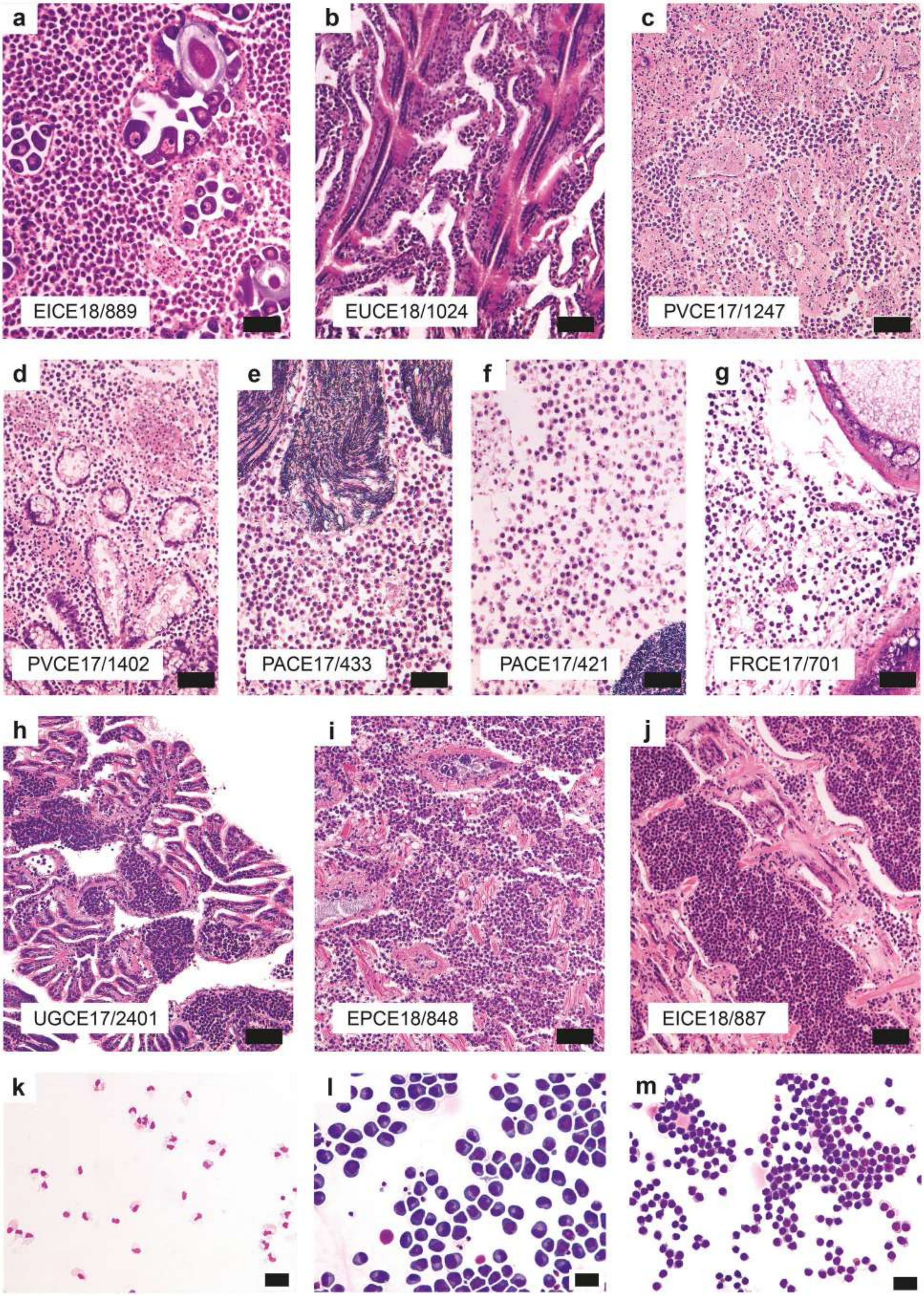


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Frequency and progression stages of disseminated neoplasia in *C. edule*. a, Numbers of individuals diagnosed with each stage of cockle DN (early or N1, intermediate or N2, and late or N3) in each country where DN was detected. **b**–**m**, Micrographs of histological sections of cockle DN at different stages of progression: early stage, N1 (**b**, **e**, **h**, **k**); intermediate stage, N2 (**c**, **f**, **i**, **l**); and late stage, N3 (**d**, **g**, **j**, **m**). Histological sections show the gills (**b**–**d**), digestive gland (**e**–**g**), gonad (**h**–**j**) and foot (**k**–**m**). **n**–**p**, Hemolymph cell monolayers of cockle DN at stages N1 (**n**), N2 (**o**) and N3 (**p**). Arrows indicate neoplastic cells; asterisks mark mitotic phases of neoplastic cells.

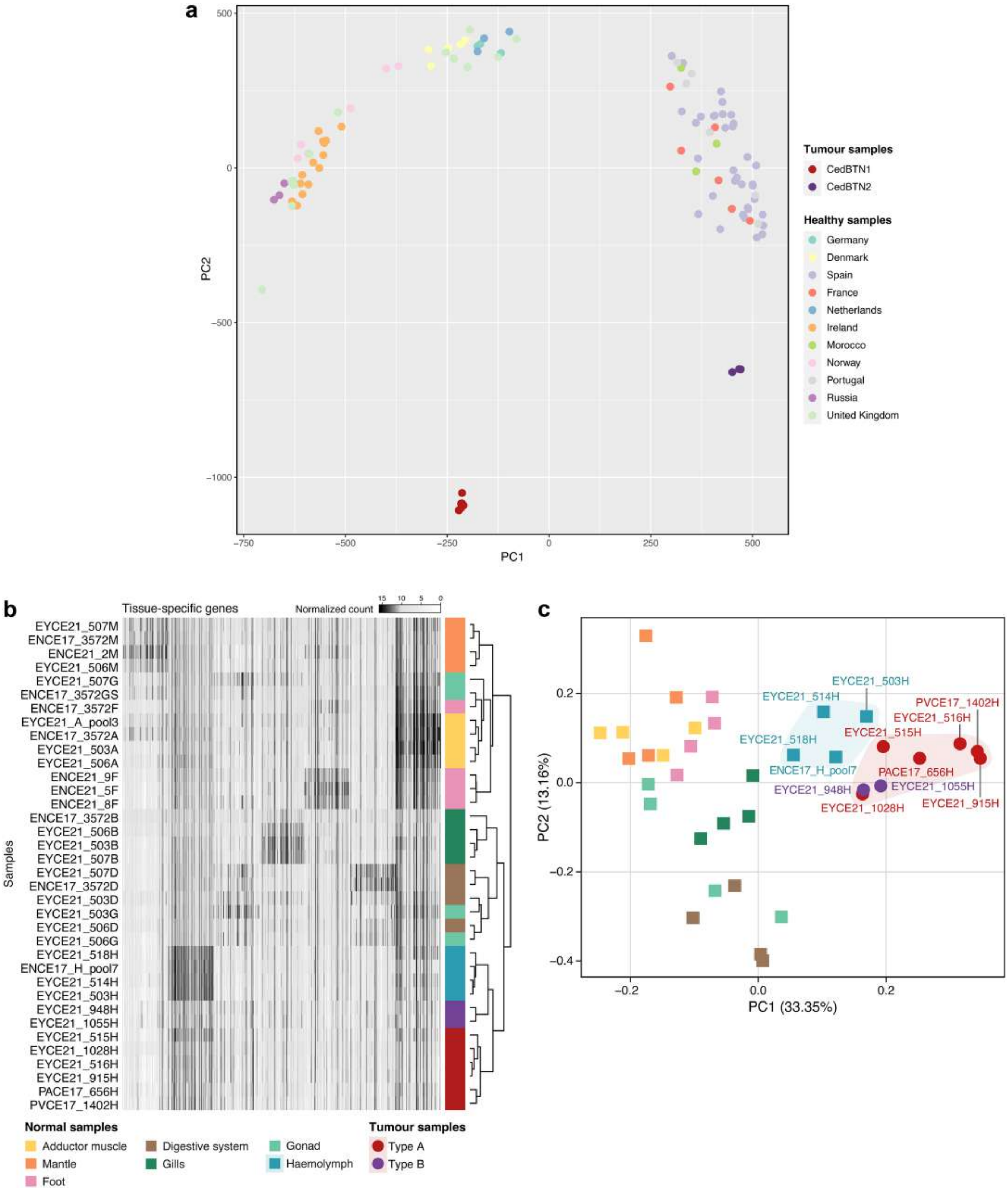


Extended Data Fig. 2 | Distribution of repetitive elements in the cockle genome. Frequency of classifiable repeats (26% of all repeats) along the reference cockle genome, displayed in terms of number of copies per 100-kb genomic segment. Repetitive element types with more than 1000 annotated copies are represented: long interspersed nuclear elements (LINE, 172,722 copies, 33.0%), transfer RNA repeats (tRNA, 81,766 copies, 15.6%), long terminal repeat elements (LTR, 78,009 copies, 14.9%), simple repeats (70,016 copies, 13.3%), short interspersed nuclear elements (SINE, 55,434 copies, 10.6%), DNA repeat elements (42,917 copies, 8.2%), low complexity repeats (12,171 copies, 2.3%), rolling circle repeats (RC, 8,843 copies, 1.7%), satellite repeats (2,100 copies, 0.4%). Genomic segments along the ideogram are classified as GC-low or GC-high based on whether their average nucleotide content is below or above the estimated average genomic G+C content (35.6%).

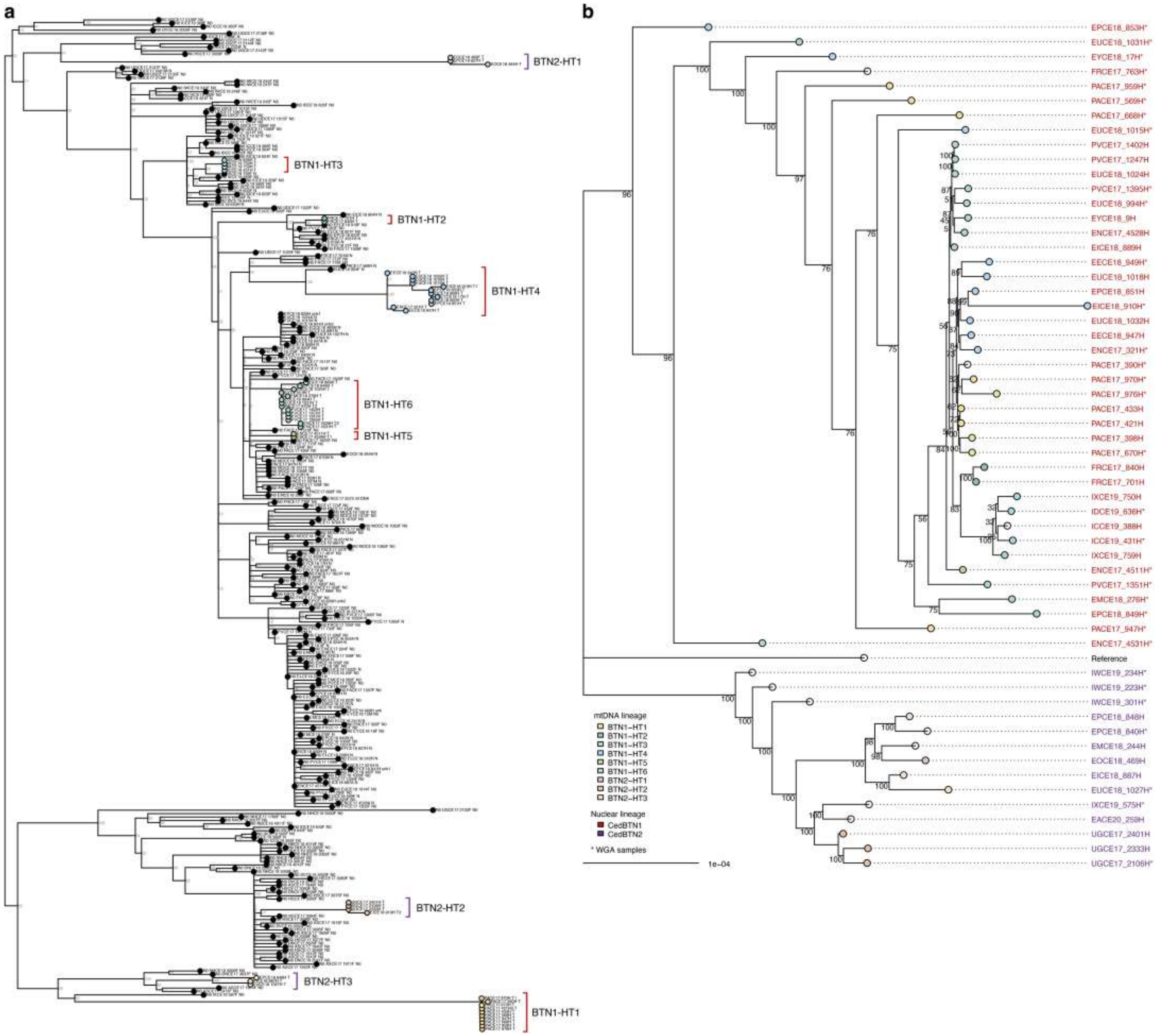


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Histology and cytology of disseminated neoplasia in *C. edule*. **a**–**j**, Micrographs of histological sections of CedBTN1 (**a**–**g**) and CedBTN2 (**h**–**j**) samples included in the 'golden set' of high-purity tumors. Micrographs show gills (**b**, **h**, **j**) and connective tissue around gonadal follicles and digestive gland (**a**, **c**–**g**, **i**), showcasing the distinctive features of the two morphological types of cockle DN: type A (**a**–**g**) and type B (**h**–**j**). **k**–**m**, Representative cell monolayers for normal hemocytes (**k**), type A DN (CedBTN1) cells (**l**), and type B DN (CedBTN2) cells (**m**). Histological sections stained with hematoxylin and eosin; cell monolayers stained with Hemacolor kit (Merck). Scale bars, 50 µm for **a**–**j**, 25 µm for **k**–**m**.

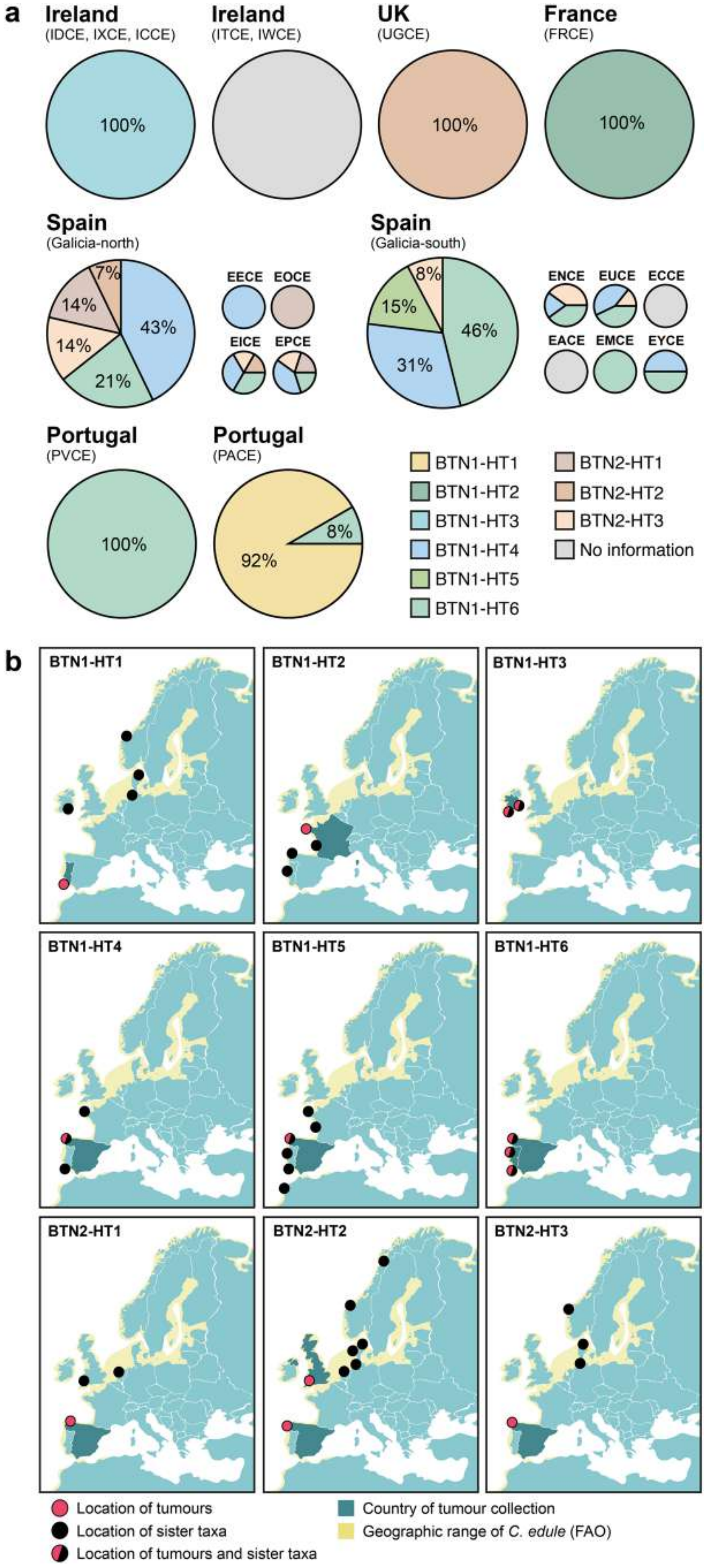


Extended Data Fig. 4 | Germline polymorphism and gene expression in cockles and CedBTN tumors. a, Principal component analysis (PCA) of germline polymorphisms in CedBTN and healthy cockle samples. Logistic PCA was performed on a randomly selected subset of 100,000 germline exonic single-nucleotide polymorphisms, genotyped across 100 non-neoplastic cockles (covering all sampling locations), seven CedBTN1 tumors, and three CedBTN2 tumors. **b**, Heatmap and unsupervised clustering of normal cockle tissue samples and CedBTN tumor samples, based on normalized gene expression values for 420 genes with tissue-specific expression (60 genes per normal tissue type). **c**, Principal component analysis of tissue-specific gene expression; normal hemolymph and CedBTN samples are labeled. Both analyses indicate a clustering of CedBTN samples with normal hemolymph.



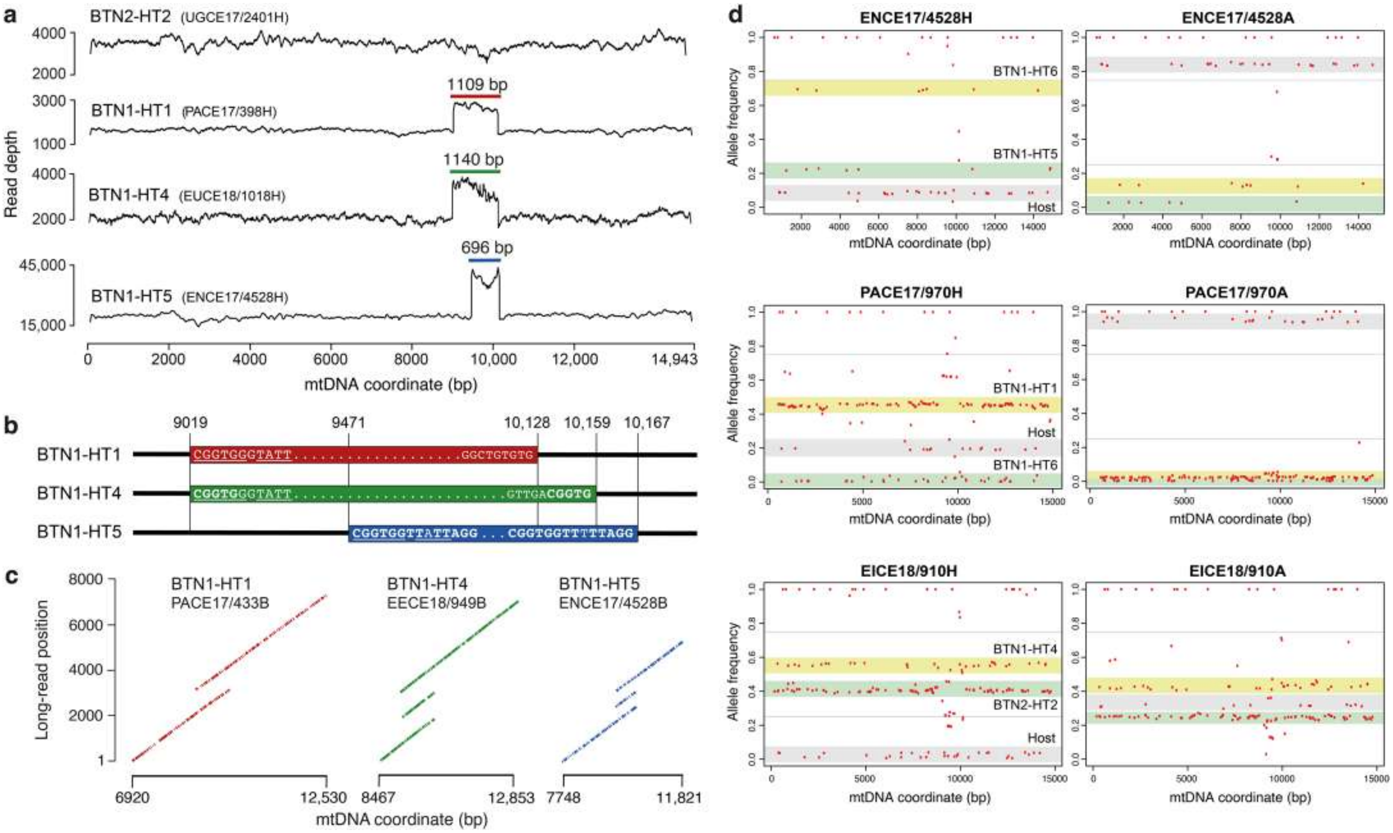
Extended Data Fig. 5 | Maximum likelihood phylogenies of cockle mtDNA and CedBTN genomes. a, Maximum likelihood cockle mtDNA phylogeny. Midpoint-rooted tree of deconvoluted mtDNA haplotypes, including sample codes for normal ('N0') and tumor samples ('T'; colored by mtDNA lineage). The nine identified mtDNA lineages are labeled. Bootstrap support values (*n* = 1000 replicates) are shown for all nodes. **b**, Maximum likelihood CedBTN nuclear phylogeny from genotyped SNVs. Phylogenetic tree inferred from a subset of 833,007 BTN-specific SNVs, including 30,000 randomly selected SNVs from each of the ancestral variant sets ('A0', 'A1', 'A2') and all the non-ancestral (postdivergence) SNVs in each nuclear lineage, which were genotyped across 61 tumor samples. Tips are colored according to mtDNA lineage (where information is available); sample labels are colored according to nuclear CedBTN lineage. Bootstrap support values (*n* = 1000 replicates) are shown for all nodes. Samples subjected to whole-genome amplification (WGA) are indicated by asterisks. Sample EICE18_910H is a case of co-infection by cells from mtDNA lineages BTN1-HT4 and BTN2-HT2 (Fig. 2e,f).

https://doi.org/10.1038/s43018-023-00641-9



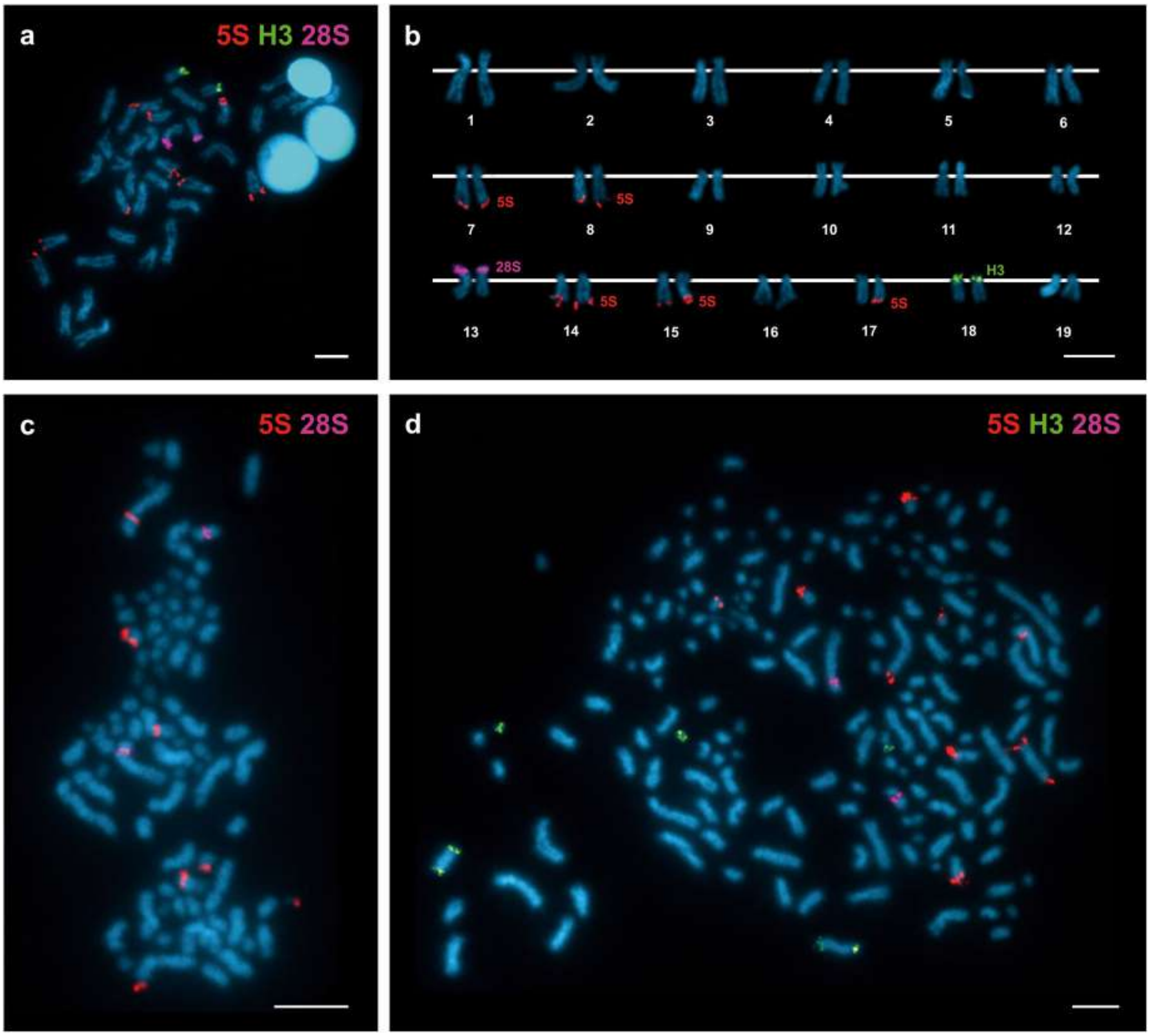
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Distributions of tumors from each CedBTN mtDNA lineage. a, Percentages of tumor samples from each CedBTN mtDNA lineage in each cockle population. Sampled cockle populations (corresponding to sampling locations; Supplementary Table 1) are grouped by country, except for Spain. Populations from Spain are divided into two groups (northern and southern Galicia), and are also presented individually to demonstrate the variability in mtDNA lineage composition across populations. **b**, Maps displaying the locations of tumor samples and their sister taxa for each identified CedBTN mtDNA lineage.



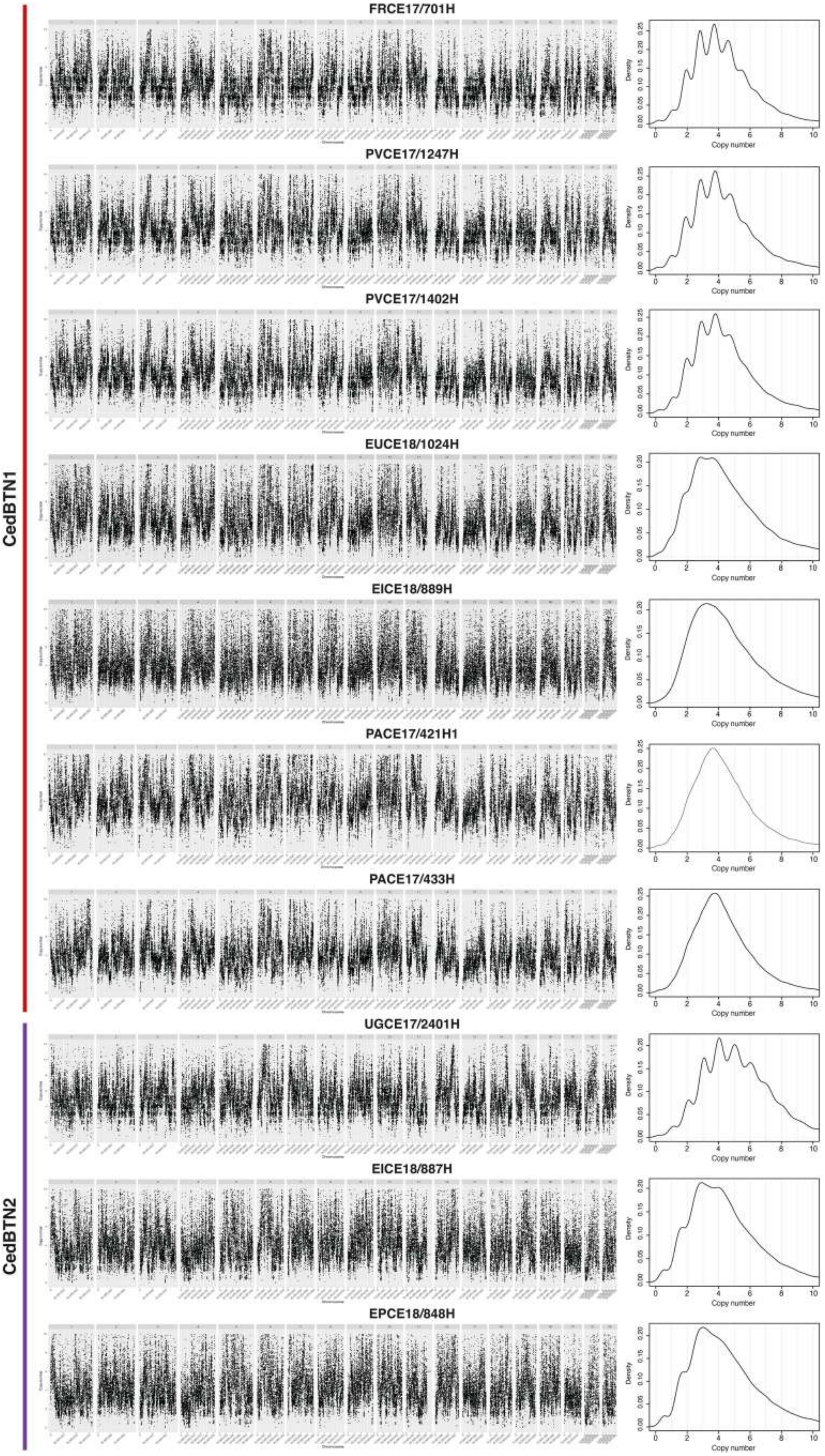
Extended Data Fig. 7 | Recurrent mtDNA D-loop amplification and host co-infection in CedBTN. a, Sequence read depth along the mitochondrial genome in representative samples from four CedBTN mtDNA lineages, showing the independent mtDNA amplifications identified in three mtDNA lineages within CedBTN1. A sample from BTN2-HT2 (top) is shown as representative of the read depth distribution in CedBTN2 samples. Amplification lengths are indicated. **b**, Schematic representation of the three mtDNA amplification events, two of which share the same start coordinate. Identity among the start sequences is marked by underlining, while overlapping microhomology at the boundaries of two of the amplified regions is highlighted in bold. **c**, Diagonal plots of position along long sequence reads (Oxford Nanopore) against mtDNA coordinate, showing the number of copies gained in each mtDNA lineage (duplication in BTN1-HT1, triplication in BTN1-HT4 and BTN1-HT5). **d**, mtDNA allele frequency plots evidencing the presence of two tumor mtDNA haplotypes (green/yellow) and one host haplotype (gray) in hemolymph (left) and adductor muscle (right) samples from three cockles presenting evidence of co-occurrence of multiple CedBTN lineages (top to bottom: ENCE17/4528, PACE17/970, EICE18/910; Supplementary Table 10). Each dot represents a mitochondrial SNV. Identified tumor mtDNA haplotypes are labeled as in Fig. 2a. As expected, tumor and host mtDNA haplotypes present lower and higher allele frequencies, respectively, in adductor muscle compared to hemolymph.

https://doi.org/10.1038/s43018-023-00641-9

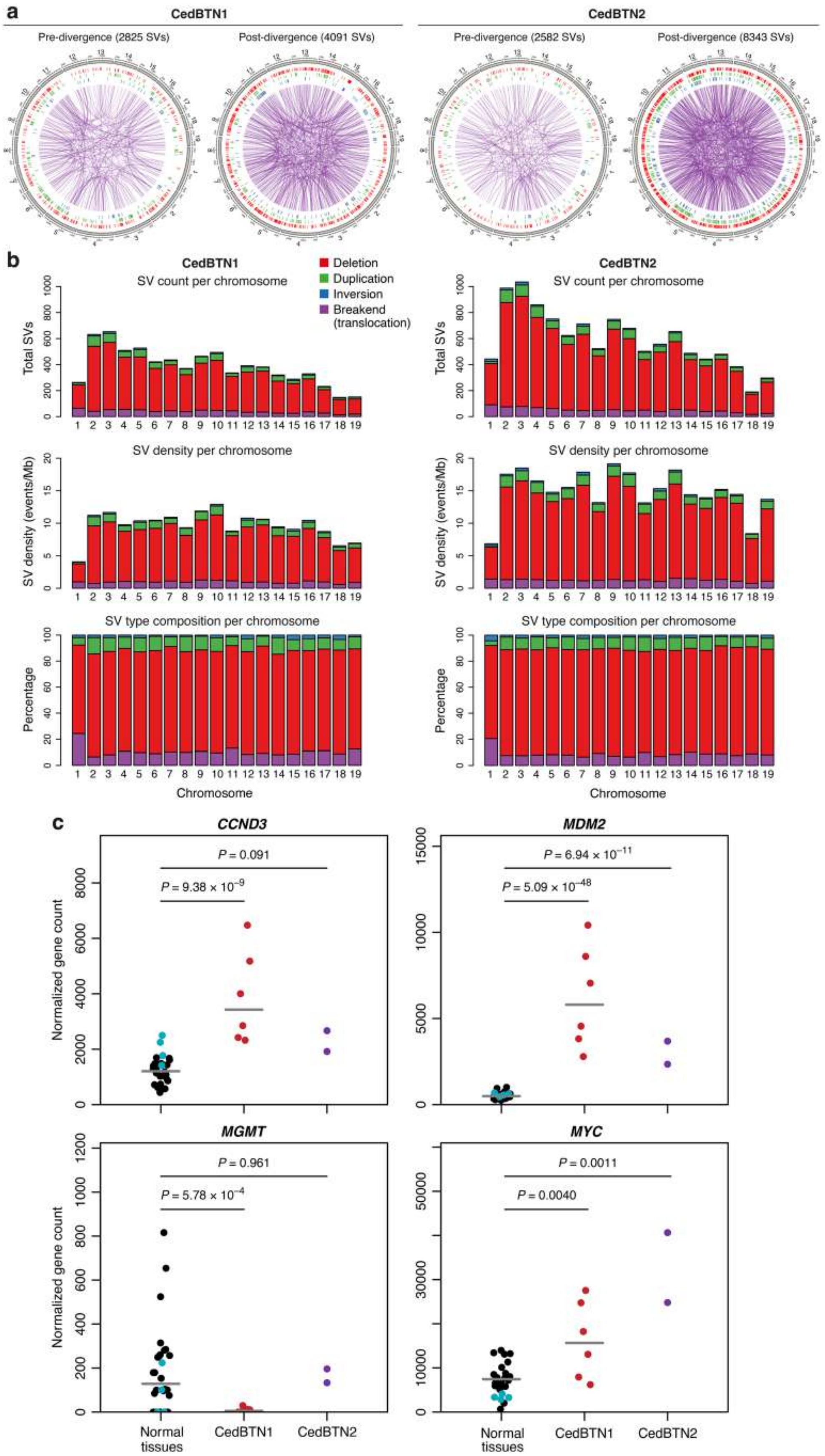


Extended Data Fig. 8 | Molecular cytogenetic results from metaphases of healthy and neoplastic specimens. a–b, FISH of 28S ribosomal DNA (rDNA; violet), 5S rDNA (red) and H3 histone gene (green) probes mapped onto a metaphase plate of a healthy specimen of *C. edule* and its corresponding karyotype with $2n = 38$ chromosomes. As previously described[80], up to five chromosome pairs hold subtelomeric clusters of 5S rDNA on their long arms, while 28S rDNA and histone H3 probes hybridize to the short arm of subtelocentric chromosomes. **c–d**, FISH mapping of the probes above onto example neoplastic metaphases, revealing abnormal location and number of these clusters. Scale bars, 10 µm.

https://doi.org/10.1038/s43018-023-00641-9



Extended Data Fig. 9 | Copy number profiles of CedBTN samples. Plots of unrounded copy number along the reference genome (left) and copy number density (right) for each sample in the 'golden set' of high-purity tumors, grouped by CedBTN lineage. Each dot represents a genomic bin containing 10 kb of mappable sequence.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Structural variant distribution and candidate driver gene expression in CedBTN. a, Circos plots representing the distribution of BTN-specific structural variants within the predivergence (ancestral) and postdivergence phylogenetic variant sets in CedBTN1 and CedBTN2. Deletions and duplications of size <10 kb are omitted for interpretability. **b**, Distributions of structural variant frequency, density and type composition (top to bottom) per reference chromosome, for variants identified in CedBTN1 (left) and CedBTN2 (right). **c**, Expression of genes with potential early driver CN alterations in CedBTN. For each of the four genes with potential early driver CN alterations, normalized gene expression counts are shown for normal tissue samples ($n = 28$), CedBTN1 samples ($n = 6$) and CedBTN2 samples ($n = 2$). Each dot represents one sample, and gray lines denote the median expression for each group. Normal hemolymph samples ($n = 4$) are marked in light blue. Adjusted p-values are shown for comparisons between normal tissues and each CedBTN lineage, obtained via differential expression analysis (two-sided Wald tests with Benjamini–Hochberg correction). Seven normal tissue samples presented null *MGMT* expression: ENCE17/3572B (gill), EYCE21/503H (hemolymph), EYCE21/507B (gill), EYCE21/507G (gonad), EYCE21/514H (hemolymph), ENCE21/2M (mantle), ENCE21/5F (foot). Normalized gene count values are comparable across samples for the same gene, but are not comparable across genes.

nature portfolio

Corresponding author(s): Adrian Baez-Ortega
Jose M. C. Tubio

Last updated by author(s): Jun 28, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	□	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
□	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Guppy (v2.3.1), MinKNOW (v18.12.09).
Data analysis	3D-DNA pipeline (v180922), Adobe Photoshop CS6 (v13.1.3), BEAST (v2.6.2), biobambam2 (v2.0.87), Blast2GO (v1.4.5), BUSCO (v3.0.2), BWA-MEM (v0.7.17), circlize (v0.4.14), ComplexHeatmap (v2.10.0), DELLY (v0.7.9), DESeq2 (v1.34.0), dicey (v0.1.8), dNdScv (v0.0.1.0), EggNOG (v4.5.1), GATK (v4.1.6.0), Genious Prime (v.11.03), GenomeScope (v1.0.0), GraphTyper (v2.0), HaploMerger2 (v3.4), KAT (v2.3.2), HISAT2 (v2.1.0), InterProScan2 (v2), IQ-TREE2 (v2.1.1), Jellyfish (v2.2.10), KAAS (v2.1), logisticPCA (v0.2), LUMPY (v0.2.13), Maker2 (v3.01.03), Manta (v1.6.0), MaSuRCA (v3.2.4), ModelTest-NG (v0.1.6), Nikon NIS-Elements (v5.42.01), PAUP* (v4.0a168), phangorn (v2.8.1), Platypus (v0.8.1), PretextMap (v0.0.2), PretextView (v0.0.2), purge_haplotigs (v1.1.0), Qualimap2 (v2.2.1), R (v4.1.3), RAxML (v8.2.12), RAxML-NG (v0.8.1), RepeatExplorer (v2.3.8.1), RepeatMasker (v4.1.0), RepeatModeler (v1.0.11), RSEM (v1.3.1), samtools (v1.9), sigfit (v2.2.0), SNAP (v0.15), STAR (v2.7.3a), StringTie (v2.1.1), svimmer (v0.1), Tracer (v1.7.1), TreeAnnotator (v2.6.2), Variant Effect Predictor (v104.3), varibase (v1.0), WindowMasker (v1.0.0), wtdgb2 (v2.5). Custom code deposited on GitLab (gitlab.com/mobilegenomesgroup/scuba_cancers).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The reference genome sequence for Cerastoderma edule and sequencing data supporting the findings of this study have been deposited in the European Nucleotide Archive (www.ebi.ac.uk/ena) under overarching accession code PRJEB58149. Human mutational signatures were retrieved from the COSMIC v3.2 database (cancer.sanger.ac.uk). Source data for Figures 1–4 and Extended Data Figures 1, 2, 4, 6, 7, 9 and 10 have been provided as Source Data files. All other data supporting the findings of this study are available from the corresponding authors on reasonable request.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences □ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The study analyzed genomic data from samples of transmissible neoplasia in cockles of the species Cerastoderma edule.
Research sample	To investigate the prevalence of disseminated neoplasia in the common cockle, Cerastoderma edule, adult specimens of C. edule were collected along the species' geographic range, spanning from the northern Barents Sea to the Atlantic coast of Morocco.
Sampling strategy	A total of 6854 specimens of C. edule were collected from natural seabeds at 36 locations in 11 countries (Portugal, Ireland, Spain, United Kingdom, France, Germany, Denmark, Morocco, the Netherlands, Norway, Russia). This sample size was determined by the available resources, and was considered sufficient for the study.
Data collection	Data recorded for all specimens include: sampling country, location, coordinates, year, disseminated neoplasia diagnosis. Data for specimens diagnosed with disseminated neoplasia include: neoplasia stage, neoplasia type, percentage of circulating neoplastic cells. (see Supplementary Tables 1 and 2).
Timing and spatial scale	Specimens of C. edule were collected during several sampling periods between March 2016 and March 2021 (see Supplementary Table 1). These periods were determined by the ecological features of the species and the available resources.
Data exclusions	Specimens were excluded from the study if they were found to belong to a species other than Cerastoderma edule, as determined through morphological and molecular markers.
Reproducibility	Reproducibility of experimental findings was assessed using standard approaches, including replication of experiments, phylogenetic bootstrap analysis, and variant calling via independent software tools, among others; see Methods for details.
Randomization	The experiments were not randomized.
Blinding	The investigators were not blinded to allocation or diagnosis during experiments and outcome assessment.

Did the study involve field work? ☒ Yes □ No

Field work, collection and transport

Field conditions	Field conditions were variable, as field work was carried out at multiple locations and times (described above). Animals were collected from inter-tidal areas with traditional methods used by locals.
Location	Location information, including latitude and longitude, is provided for all sampling points in Supplementary Table 1.
Access & import/export	Collection of animals from natural sand beds was carried out after obtaining the permits required by local and/or national authorities. Animals were transported in isotermal boxes monitored according to European Commission Decision 2003/623/EU, 599/2004/EU and 1251/2008/EU. The Intra Trade Animal Health Certificate (TRACES) was obtained when required by EU Regulation 511/2014 on compliance measures for users from the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization in the Union.
Disturbance	To avoid disturbance, we contacted local organizations to ensure that sampling in each location was conducted in proportion to cockle abundance, such that less animals were collected at locations presenting lower numbers of cockles. Total numbers of samples per sampling point are provided in Supplementary Table 1. Sampling was performed twice at some locations only when this seemed to allow collection of cockles affected by late-stage neoplasia.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
□	☒	Animals and other organisms
☒	□	Clinical data
☒	□	Dual use research of concern

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	The study did not involve laboratory animals.
Wild animals	Live cockles (Cerastoderma edule) from a range of ages (2-8 years) were sampled from sandbeds at different depths; coordinates of sampling locations are provided in Supplementary Data 1. Cockles were collected manually at low depths and with a catcher from a boat in places with greater depth. All specimens arrived at the laboratory alive and were kept in a Level 2 biosecurity laboratory between 2 and 4 days, in 50-litre tanks filled with filtered seawater. Sample processing consisted in extraction of haemolymph samples and subsequent euthanasia and dissection of animals with a scalpel, with the aim of obtaining histological, cytological and DNA samples for diagnosis and sequencing.
Reporting on sex	The study investigates a transmissible cancer which was assumed to affect host animals of both sexes to an equal extent. Therefore, sex was not considered in the study design, and was not recorded. The conclusions of the study do not apply only to one sex.
Field-collected samples	Maintenance in seawater tanks was carried out in two facilities: Toralla Marine Science Station, Universidade de Vigo (ECIMAT, Illa de Toralla, Vigo, Spain; REGA: ES360570181401) and in the Aquatic Facilities of the Faculty of Biology, Universidade de Santiago de Compostela (Rúa Constantino Cadeira, Campus Vida, Santiago de Compostela, Spain; REGA: ES150780263301). As we were aware of the potential ecological threat of work with animals affected by contagious cancers, in terms of environmental protection, international specimens were carefully processed in a biosecurity facility (ISO 9001:2015) to minimise the potential biological risks. Animals from different sampling locations were never mixed in the same tank and bleach cleaning of tanks was performed between sample arrivals. Water temperature was set at 12 °C and photoperiod was set to 12 h light, 12 h darkness.
Ethics oversight	Animal samples were obtained with the approval of the Standing Committee on Conflict of Interest, Scientific Misconduct and Ethical Issues (CoIME) of the European Research Council, and under regional licenses for mollusc extractions and trading authorizations.

Note that full information on the approval of the study protocol must also be provided in the manuscript.