Gene expression in soft-shell clam (*Mya arenaria*) transmissible cancer reveals survival mechanisms during host infection and seawater transfer

Authors: Samuel F.M. Hart [1,2,3], Fiona E. S. Garrett [1], Jesse S. Kerr[4], Michael J. Metzger [1,2]

[1] Pacific Northwest Research Institute, Seattle, WA, USA

[2] Molecular and Cellular Biology Program, University of Washington, Seattle, WA, USA

[3] Genome Sciences Department, University of Washington, Seattle, WA, USA

[4] PEI Department of Fisheries, Tourism, Sport and Culture, Canada

ABSTRACT

Transmissible cancers are unique instances in which cancer cells escape their original host and spread through a population as a clonal lineage, documented in Tasmanian Devils, dogs, and ten bivalve species. For a cancer to repeatedly transmit to new hosts, these lineages must evade strong barriers to transmission, notably the metastasis-like physical transfer to a new host body and rejection by that host's immune system. We quantified gene expression in a transmissible cancer lineage that has spread through the soft-shell clam (*Mya arenaria*) population to investigate potential drivers of its success as a transmissible cancer lineage, observing extensive differential expression of genes and gene pathways. We observed upregulation of genes involved with genotoxic stress response, ribosome biogenesis and RNA processing, and downregulation of genes involved in tumor suppression, cell adhesion, and immune response. We also observe evidence that widespread genome instability affects the cancer transcriptome via gene fusions, copy number variation, and transposable element insertions. Finally, we incubated cancer cells in seawater, the presumed host-to-host transmission vector, and observed conserved responses to halt metabolism, avoid apoptosis and survive the low-nutrient environment. Interestingly, many of these responses are also present in healthy clam cells, suggesting that bivalve hemocytes may have inherent seawater survival responses that may partially explain why transmissible cancers are so common in bivalves. Overall, this study reveals multiple mechanisms this lineage may have evolved to successfully

spread through the soft-shell clam population as a contagious cancer, utilizing pathways known to be conserved in human cancers as well as pathways unique to long-lived transmissible cancers.

INTRODUCTION

The maximum life span of a cancer is typically limited by the lifespan of its host, with cancer either regressing or dying along with its host. However, a small number of transmissible cancers in Tasmanian Devils[1,2], dogs[3,4], and bivalves[5–10] have been able to extend their life span by transmitting to a new host like an infectious parasite. In these rare cases, cancers have gained the ability to repeatedly bypass two major barriers to cancer transmission: the physical transfer between individuals and immune rejection[11]. Transmission in devils occurs during biting and engraftment of cells on the new host's facial wounds[1], in dogs the cancer is a sexually transmitted genital tumor[3], and in bivalves the cancer cells transfer through the seawater, presumably via filter feeding[11–13]. Immunologically, the vertebrate transmissible cancers are believed to evade immune detection though mechanisms such as the downregulation of MHC genes and the release of immunosuppressive cytokines[14–16]. Additionally, it is hypothesized that low genetic diversity of the devil population and of the ancestral founder pack of dogs contributed to the ability of the cancers to initially evade immune rejection before evolving additional mechanisms[3,17]. Bivalve transmissible neoplasia (BTN) has been identified in ten bivalve species[5–10], indicating that bivalves may be particularly susceptible to cancer transmission. In bivalves, as in other invertebrates, there is no adaptive immune system, and it has been assumed that this contributes to the inability to uniformly reject non-self cancer cells[11]. It is unknown if there is any host innate immune response to bivalve cancers or if there are any mechanisms in the cancer that might have evolved to escape rejection by host innate immune systems.

The first species in which BTN was identified is the soft-shell clam (*Mya arenaria*), in which a single clonal lineage has spread through the native range along the east coast of North America[5]. In a previous study we analyzed *M. arenaria* BTN (MarBTN) genome sequences and found that the cancer genome was highly mutated and unstable[18]. Though this continued mutation would be expected to mediate adaptation of the cancer to its new parasitic lifestyle, it is difficult to elucidate from mutational data alone

which genes and pathways are central to this parasitic ability. Here we turned to transcriptome-wide expression analysis of MarBTN to investigate the mechanisms by which it has been able to survive, proliferate, and spread through the soft-shell clam population.

RESULTS

Confirmation of the hemocyte origin of MarBTN

Comprehensive annotation of all genes in the soft-shell clam genome is key to identifying expression changes in MarBTN that may have played a role in its evolution as a transmissible cancer. We previously assembled a soft-shell clam genome and annotated genes using RNAseq data from six tissues from the same clam (foot, gill, hemocytes, mantle, adductor muscle, and siphon) and genome annotation pipeline MAKER[18]. In this study, we used an improved transcriptome reference, annotated using the same genome and RNAseq data used previously (NCBI eukaryotic genome annotation pipeline). This output annotation is more comprehensive, capturing a higher number of gene models, transcript isoforms, exons, characterized genes, and complete BUSCOs (**Supplementary Table 1**).

We sequenced RNA from five MarBTN isolates, six tissues each from three healthy clams (hemocytes and five solid tissues), and hemocytes from an additional five healthy clams (**Supplementary Table 2**). We then mapped RNA reads to the new genome annotation to quantify expression for each gene. Principal component analysis (PCA) of expression across all genes separated MarBTN and hemocytes from all solid tissues across the first principal component (**Supplementary Fig. 1A**). This supports previous analyses implicating hemocytes, bivalve immune cells found in the circulatory fluid, as the likely tissue of origin for MarBTN and two independent BTNs in European cockles[18,19]. Hierarchical clustering on the top 100 tissue-specific genes also supports this origin (**Supplementary Fig. 1B**). Because BTN likely arose from a normal hemocyte, we focused on the comparison of MarBTN isolates (n=5) to healthy clam hemocytes (n=8) for differential expression analysis.

Differential expression in MarBTN compared to hemocytes

An overwhelming number of genes are significantly up- (n=8,218, 19% of annotated genes, **Supplementary Table 3**) or down-regulated (n=8,660, 20% of annotated genes, **Supplementary Table 4**) in MarBTN versus healthy hemocytes (**Fig. 1A**), unsurprising given the clonal nature of MarBTN and their centuries of divergence from healthy clam cells. Orthologs of known human tumor suppressors (TUSC2 and RASF8) and oncogenes (RHEB) are among the most significant of these genes. We also see many genes involved in the cellular response to genotoxic stress, likely facilitating DNA damage repair and/or permitting cells to continue proliferating despite the ongoing genome instability observed in this lineage[18,20]. These include orthologs to *PUM3*, a gene highly expressed in some human cancers[21] that inhibits the degradation of PARP1 following genotoxic stress[22], RAD18, an E3 ubiquitin protein ligase involved in post-replication repair of DNA lesions[23], and ECT2, which is expressed during DNA synthesis and can lead to genotoxic stress-induced cell death[24], but is downregulated in MarBTN. Among genes with the largest fold difference in MarBTN versus healthy hemocytes we see genes involved in cell adhesion (TENX and CNTN5). This likely contributes to MarBTN's distinctive non-adhesive spherical phenotype and may facilitate the hyper-metastatic ability of MarBTN to engraft and release from tissues repeatedly. An innate immune signaling gene, GBP1, is also among the most downregulated and could play a role in the cancer's ability to evade host immune rejection of non-self cells, an open question across all transmissible cancers. Thousands of other genes are highly mis-regulated and likely to play important roles in MarBTN, but most of these are either uncharacterized or do not have an obvious link to cancer. This is not unexpected, since in addition to known cancer-associated genes, we would expect this set to include undiscovered cancer-associated genes, genes specific to bivalve oncogenesis, genes specific to transmissible cancer cell survival, and genes that do not provide a selective advantage but are differentially regulated either by chance or as a byproduct of selection on genes in related pathways.

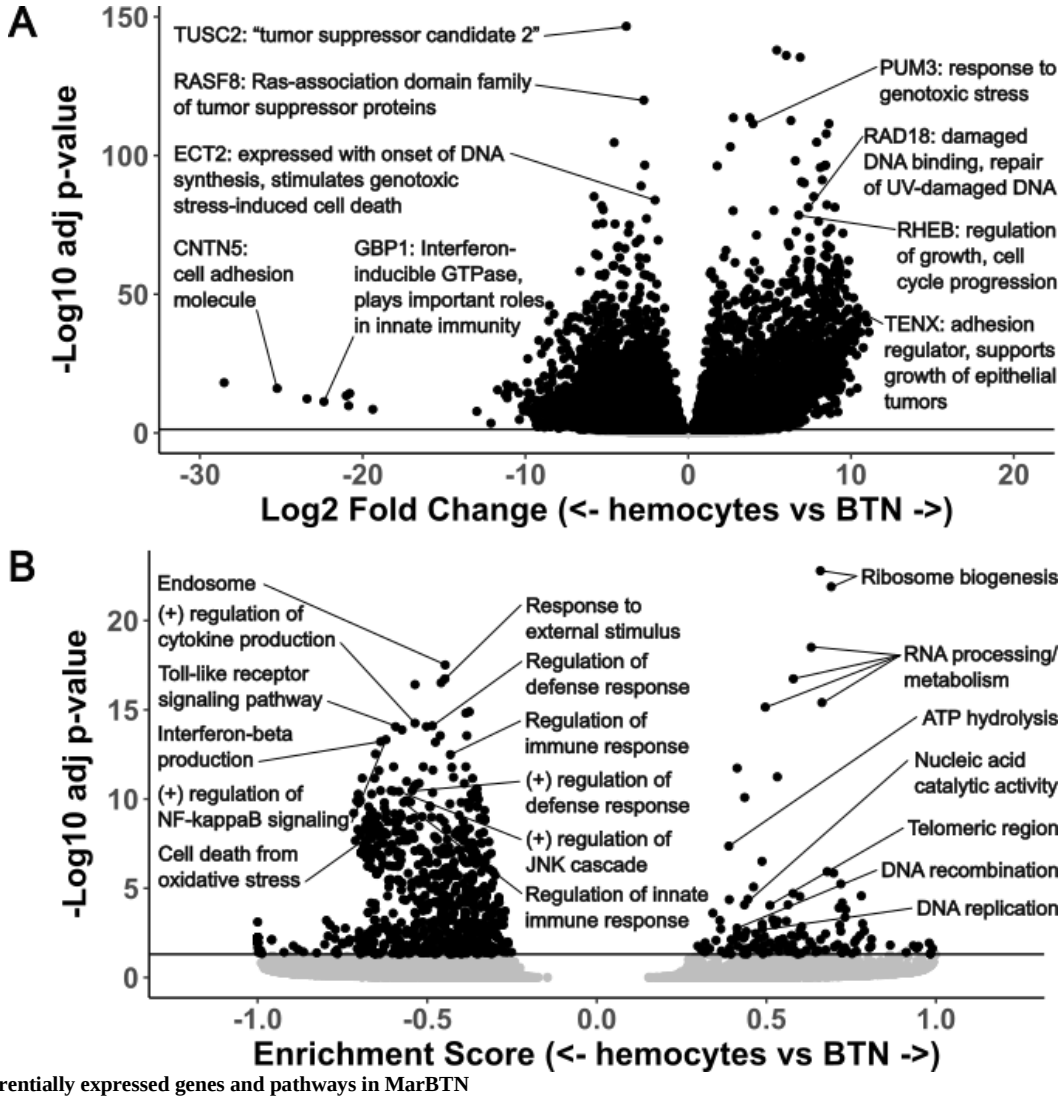


Figure 1. Top differentially expressed genes and pathways in MarBTN

Volcano plots of differentially expressed genes **(A)** and gene sets **(B)** from the comparison of MarBTN isolates (n=5) versus healthy hemocytes (n=8). Genes of note are labeled with annotations and abbreviated descriptions. Line marks false discovery rate adjusted significance threshold ($p < 0.05$), with genes below threshold colored in grey.

To investigate transcriptome-wide expression trends we turned to gene set enrichment analysis (GSEA), which order ranks genes by their differential expression and tests whether genes related to a particular process, function or localization are disproportionately up or down regulated[25]. Of 15,473 pathways tested, we observed 135 significantly upregulated pathways and 756 significantly downregulated pathways (**Fig. 1B, Supplementary Fig. 2**). The most highly upregulated pathways involved RNA processing and ribosome biogenesis (**Supplementary Table 5**), which are recognized as important for cell growth and proliferation of cancer cells[26]. ATP hydrolysis and DNA replication/recombination are also

among the top pathways and would be key for a metabolically demanding growth and division of cancer cells. We also observe upregulation of genes whose products localize to telomeric regions and DNA repair complexes, perhaps facilitating maintenance of genome integrity in response to damage and telomere shortening, which would be critical for the survival of a long-lived transmissible cancer.

Interestingly, the top downregulated pathways all relate to immune responses, such as cytokine production, NF-κB activation, toll-like receptor signaling and defense/inflammatory/innate immune responses (**Supplementary Table 6**). We suspected this may be an evolved response allowing MarBTN to better evade host immune rejection, though alternatively it could represent the downregulation of unnecessary pathways from the cancer's origin as an immune cell itself. To test the latter possibility, we looked at differential expression comparing MarBTN isolates (n=5) to solid tissues (n=15: 5 tissues each from 3 clams). Many of the same genes and pathways were similarly up- or down-regulated as they were in the hemocyte comparison (**Supplementary Fig. 3)**, with immune pathways continuing to dominate the downregulated gene pathways. This indicates the observed immune downregulation is not primarily due to the comparison to hemocytes and instead supports the hypothesis that this is a mechanism to evade host immune rejection. Additionally, we observe downregulation of stress responses such as the JNK/MAPK cascades and oxidative stress induced cell death. These pathways likely contribute to the ability of MarBTN to survive repeated exposure to the extreme environments of hypoxic late-stage cancer infections[27] while continuing to proliferate and maintain the ability to infect new hosts.

Although we observe many differentially regulated genes and pathways in MarBTN, these samples still represent a single transmissible cancer lineage and therefore an effective sample size of one. To investigate conserved trends across BTNs, we compared our results to those from a recent study of gene expression in an independent BTN lineage that originated in *Mytilus trossulus* and circulates in various Mytilus species (MtrBTN2)[28]. To identify convergent evolution between MarBTN and MtrBTN2, we identified genes that were significantly differentially regulated in both cancers and shared an exact gene annotation match (**Supplementary Table 7,** n = 1498). More genes were either upregulated in both cancers

(n = 373) or downregulated in both cancers (n = 569) than would be expected by chance (942/1498, 63%, p = 2e-23, Chi-squared test). This indicates that some of the same genes may be playing a role in both cancers, particularly genes that are downregulated in both cancers, where the greatest overlap was observed. The genes with the strongest downregulation in both the transmissible cancers from clams and mussels included genes involved in the innate immune inflammatory response (toll-like receptors, ficolin-2), cell cycle regulation (cell division control protein 42), stress response (heat shock protein beta-1) and apoptosis (caspase-3). As more data become available from other BTN lineages, further analysis more thoroughly identifying gene homology across bivalves and comparing differentially expressed genes in their respective BTNs may help us to zero on in universally conserved mechanisms that have repeatedly evolved to allow BTNs to survive, repeatedly engraft in new animals, and evade the host response.

Genome instability affects gene expression

We previously observed that MarBTN's genome is highly unstable, displaying widespread genome rearrangement, copy number gains, and transposable element activity[18]. With gene expression data, we were interested to investigate how this genome instability affected the cancer's transcriptome, as the intermediary between genotype and phenotype. We first quantified the number of fusion transcripts in each sample, as structural mutations would be expected to generate gene fusions that may play important roles in MarBTN evolution. We observed ~10-fold more gene fusions in MarBTN isolates than the baseline number observed in healthy hemocyte samples (**Fig. 2A**, avg = 416 vs 39, p = 1.5e-4, two tailed t-test with unequal variance). In addition to true fusions generated by germline polymorphisms, the small number of fusions in healthy samples may be due to genome mapping errors, transcript read-throughs, transposable elements missed in masking, or structural variants polymorphic in the clam population, while the increased number of fusions in cancer samples is likely caused by somatic genome rearrangement. Fusions found all cancer samples but no healthy samples (n=181, **Supplementary Table 8**) include fusions from early in the cancer's somatic evolution that may have contributed to the oncogenesis and/or transmission ability of the lineage.

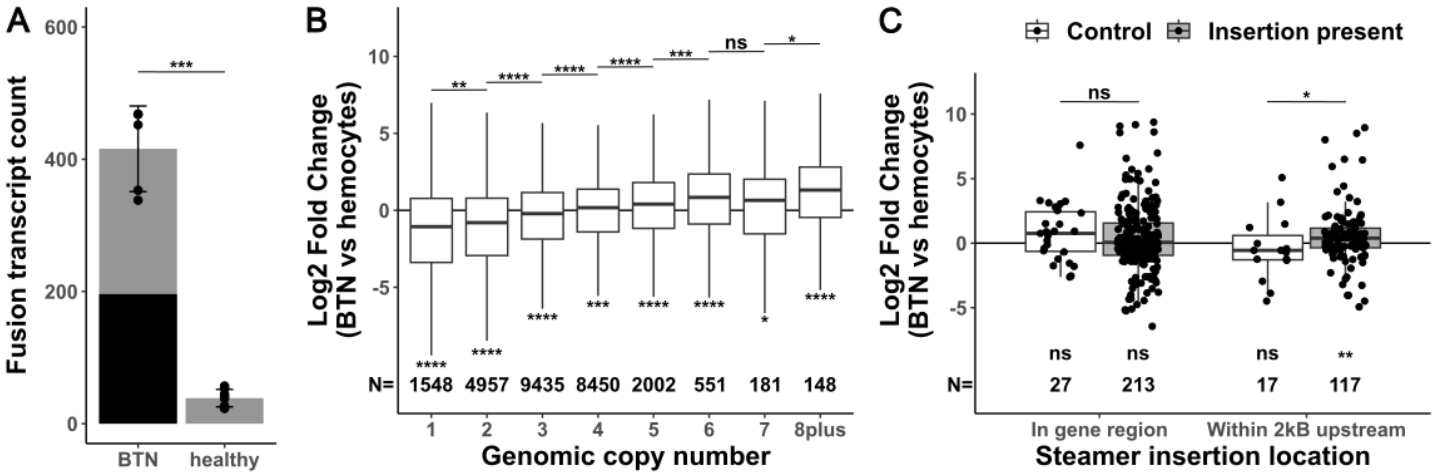


Figure 2. Fusions, copy number and transposable element insertions influence gene expression.

(A) Number of fusion transcripts per sample (dots), with mean and standard deviation for MarBTN (n=5) and healthy hemocyte (n=8) sample groupings. Black bar represents fusions found in all MarBTN samples (196). Statistical test is two-tailed t-test with unequal variance. **(B)** Log fold change in expression of MarBTN versus healthy hemocytes, binning genes by genomic copy number for each gene. Upper statistical tests are two-sided Wilcoxon rank-sum tests between adjacent copy numbers, lower statistical tests are one-sample two-sided Wilcoxon rank-sum tests versus no fold change. Counts for each group are listed below box plots. **(C)** Log fold change in expression of MarBTN versus healthy hemocytes for genes with a *Steamer* insertion within the gene itself or in the 2kB region upstream the gene. Each set is divided into *Steamer* insertions present in these samples (grey box) and, as a control, *Steamer* insertion sites observed in a different sub-lineage of MarBTN but not present in these samples (white box). Statical tests and counts are the same as listed in (B). ns: not significant, *: p<0.05, **: p<0.005, ***: p<0.0005, ****: p<0.00005.

Copy number alteration is a known mechanism in cancers to alter the expression of cancer-promoting genes[29]. To test whether copy number affects expression in MarBTN we binned genes by genomic copy number, observing that MarBTN expression relative to healthy hemocytes scales with copy number state (**Fig. 2B**). MarBTN has an average ploidy of ~3.5N across the genome, with >80% of genome at 2-4N[18], and we see that median relative expression of genes ≥4N is higher than average while lower than average for genes ≤3N ($p<0.05$ for all copy number states, one sample two-sided Wilcoxon rank-sum test). Given the widespread copy number changes in the MarBTN genome and ongoing instability, gain or loss of gene copies likely represents a mechanism that has helped scale expression of key genes for MarBTN to adapt as a transmissible cancer.

We were also interested in whether transposable element activity influences the expression of nearby genes, so we looked at the expression of genes near insertions of LTR-retrotransposon *Steamer* (**Fig.**

2C), one of the most active and best characterized transposable elements in MarBTN[30]. When compared to the null expectation of no fold change, expression was not biased when insertions were within gene regions ($p = 0.15$, one-sample two-sided Wilcoxon rank-sum test). However, we previously found that *Steamer* preferentially inserts in the 2kB region upstream genes[18], and here we find that expression was biased to be higher in genes with these upstream insertions ($p = 0.0034$, one-sample two-sided Wilcoxon rank-sum test), likely due to promotors or enhancers in *Steamer*'s LTR[31]. To control for the possibility this bias was due to an insertion preference for highly expressed genes due to accessible chromatin (instead of the insertion causing the expression change itself), we also compared against *Steamer* insertion sites observed in a different sub-lineage of MarBTN (found in clams in Prince Edward Island, Canada) but not observed in any samples of MarBTN in clams from the USA sub-lineage analyzed here. Insertions upstream genes also had a significant effect on expression when compared to this control set ($p= 0.038$), but insertions in gene regions did not have an effect ($p = 0.27$, two-sided Wilcoxon rank-sum test).

Overall, we see that gene fusions, copy number alterations, and *Steamer* insertions all influence expression and thus likely contributed to the adaptability of this lineage as a transmissible cancer.

Transcriptomic plasticity in response to saltwater

The late stage of MarBTN infection is one in which a highly pure sample can be obtained for sequencing, but the MarBTN infection cycle also includes transmission to engraft and proliferate in a new clam host through repeated metastasis-like jumps (**Fig. 3A**). This transfer is believed to occur through release of cells into seawater and uptake by filter-feeding, an inference supported by findings that MarBTN cells survive for weeks in saltwater and that MarBTN-specific DNA can be detected in tank water where MarBTN-infected clams are maintained[12]. This metastatic transmission stage would involve a different environment and selective pressures than those faced during infection, and it is possible that MarBTN has evolved the plasticity to respond to the two stages differently.

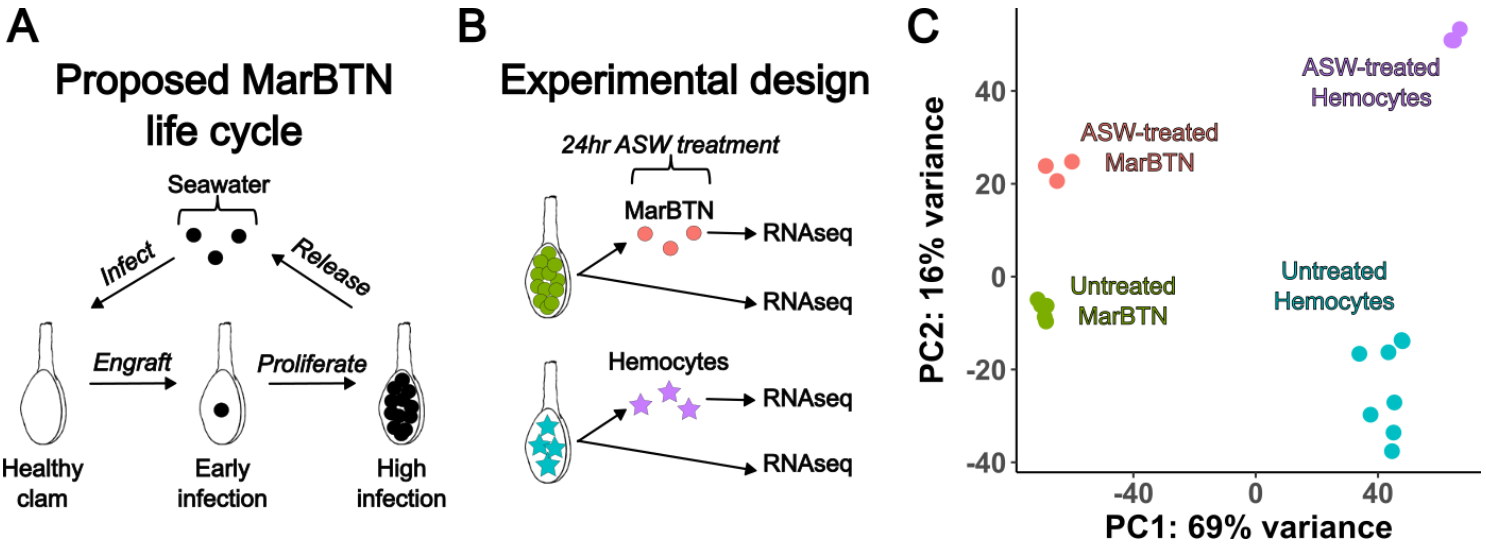


Figure 3. Transcriptomic response to seawater exposure.

(A) Proposed life cycle for MarBTN infections and **(B)** experimental design to investigate gene expression during seawater transmission. (**C**) Principal component analysis results from gene expression across all genes for ASW-treated and untreated MarBTN and hemocytes.

To test this possibility, we incubated an aliquot of three MarBTN isolates in artificial sea water (ASW) for 24 hours prior to RNA sequencing to investigate the gene expression response in this stage compared to direct RNA sequencing of another aliquot of the same isolate (**Fig. 3B**). As a control to test for the intrinsic response of clam cells to seawater, we also exposed three healthy hemocyte isolates to ASW before sequencing. We performed principal component analysis on the gene expression results of these samples, with the primary principal component separating hemocytes from MarBTN and the secondary principal component separating ASW-treated cells from untreated cells (**Fig. 3C**). Gene expression was more similar within treatment groups (ASW-treated versus pre-treatment) than source clam pairings (biological replicates before and after treatment) indicating that saltwater exposure results in a consistent transcriptomic response greater than the biological variation among our samples (**Supplementary Fig. 4**).

We compared ASW-treated vs. untreated MarBTN and ASW-treated vs. untreated hemocytes for differentially expressed genes and gene sets, dividing results into two groups: differentially regulated in both comparisons (**Fig. 4A/B**) and differentially regulated in MarBTN but not hemocytes (**Fig. 4C/D**). Genes differentially regulated in both comparisons would indicate intrinsic responses to seawater conserved

by MarBTN and healthy clam hemocytes, while genes differentially regulated in MarBTN but not hemocytes would indicate MarBTN-specific responses to seawater that may be adaptive in the cancer.

Among conserved gene responses, the outlier upregulated gene was *PCKGC*, the main control point for the regulation of gluconeogenesis, likely representing a metabolic response to the new energy-source-free environment. Similar gluconeogenesis-activating responses have been observed in glucose-deprived human cancer cells[32]. Another notable gene from the top upregulated genes (**Supplementary Table 9**) is *XIAP*, which is part of a family of apoptotic suppressor proteins and likely helps cells to avoid an apoptotic response to seawater. *XIAP* also modulates inflammatory and immune signaling via NF-kappaB and JNK activation[33], indicating these pathways, which were downregulated when comparing untreated MarBTN to healthy hemocytes, may be activated in both cell types in response to seawater. Indeed, when we use GSEA to identify differentially regulated pathways, we see pathways involved in the response to interleukin-1, tumor necrosis factor, and stress among the conserved upregulated pathways (**Supplementary Table 10**), indicating these pathways are likely to be intrinsic responses of clam cells when exposed to seawater.

The top conserved downregulated gene is *ZNFX*, which encodes an RNA-binding protein involved in antiviral response[34], though it is unclear why this gene might be lower expressed in seawater. Among the other top conserved downregulated genes (**Supplementary Table 11**) are *CCNG1*, a member of the cell cycle controlling cyclin family[35], and *CTDSL*, which is involved regulating the G1/S transition[36]. Many of the top downregulated pathways are also involved with cell cycle progression (**Supplementary Table 12**), likely representing mechanisms to halt proliferation in the absence of host nutrients and may help all cells to survive the seawater environment by entering a quiescent state. These conserved responses to seawater could represent a starting point that could be built upon during MarBTN evolution and selection for transmission ability.

Although the most significant differentially regulated gene and gene sets were conserved, many genes (**Fig. 4C**) and gene sets (**Fig. 4D**) were differentially regulated in MarBTN but not hemocytes, indicating MarBTN may have evolved additional mechanisms to survive seawater transfer. Among the top

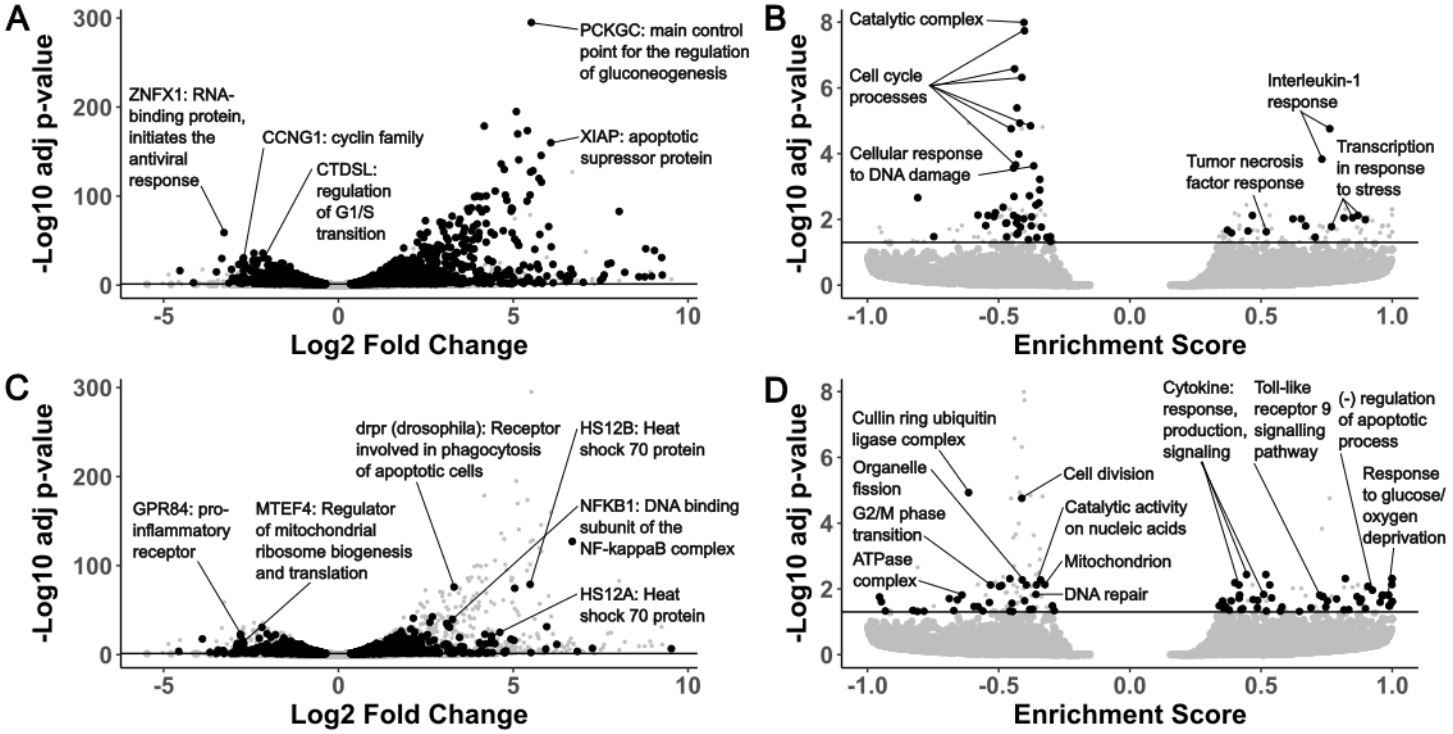


Figure 4. Top differentially expressed genes and pathways in MarBTN after exposure to seawater

Volcano plots of fold change and significance of differentially expressed genes (**A/C**) and gene sets (**B/D**) in ASW-treated MarBTN (n=3) versus untreated MarBTN (n=3). Genes of note are labeled with annotations and abbreviated descriptions. Volcano plots are filtered for genes/sets that are (**A/B**) or are not (**C/D**) differentially expressed in ASW-treated hemocytes (n=3) versus untreated hemocytes (n=3). For comparison, the reciprocal comparison is included in small grey points on each plot. Line marks false discovery rate adjusted significance threshold ($p < 0.05$), with genes below threshold colored in grey.

upregulated genes (**Supplementary Table 13**) are two heat shock protein family A paralogs (HS12A, HS12B), which can help protect cells from heat, cold, hypoxia or low glucose[37] and may be helping MarBTN cells survive one of these seawater extremes. The response to glucose/oxygen deprivation and negative regulation of apoptotic processes are among the top upregulated gene sets (**Supplementary Table 14**), indicating that these survival responses may be more broadly MarBTN-specific across additional genes.

Interestingly, several MarBTN-specific upregulated gene sets are immune response pathways, including cytokine production/response/signaling and toll-like receptor 9 signaling. This is an interesting reversal of the downregulation observed in untreated MarBTN cells, although most of these pathways are still lower expressed in ASW-treated MarBTN than healthy hemocytes (**Supplementary Fig. 5**). This could indicate that the downregulation of these immune pathways may not be as important outside the context of

a host immune system, that some innate immune system processes are reactivated in response to pathogens outside of host, or that these processes are required for some aspect of MarBTN-specific survival/engraftment .

The outlier MarBTN-specific downregulated gene sets (**Supplementary Table 15, Supplementary Table 16**) are cell division and the cullin-RING ubiquitin ligase complex, which controls cell cycle progression and other cellular processes[38]. Other metabolic pathways are downregulated in MarBTN but not hemocytes, such as catalytic activity on nucleic acids, organelle fission, ATPase complex localization, and mitochondrial localization. These responses may reflect metabolic processes that are active in proliferative MarBTN, but not hemocytes, that are shut down in response to seawater exposure. Alternatively, some may represent additional mechanisms that MarBTN has evolved to survive transfer in nutrient-poor seawater. This experiment reveals consistent MarBTN-specific seawater responses that likely facilitate transfer to new hosts and contribute to its success as a transmissible cancer.

DISCUSSION

All cancers must evolve to evade intrinsic and extrinsic barriers to successfully develop as a cancer[39]. In addition to overcoming these barriers, transmissible cancers also evolve to repeatedly transfer to new hosts and proliferate despite anti-tumor and non-self rejection mechanisms[11]. This all occurs while having no evolutionary history as a transmitting parasite prior to oncogenesis[40]. By analyzing the MarBTN transcriptome during infection and transfer, we identify possible mechanisms by which this transmissible cancer has adapted to overcome these barriers, most notably the widespread downregulation of immune signaling pathways when in hosts and survival responses to seawater exposure.

We observe mis-regulation of many gene types in MarBTN that would be expected in any cancer, such as genes involved in metabolism, cell cycle progression, adhesion, tumor suppression, genome instability and immune evasion[41]. The downregulated biological processes overwhelmingly relate to immune signaling functions (**Fig. 1B**) and likely represent an adaptive mechanism to repeatedly evade host

detection/rejection as MarBTN spread through the soft-shell clam population. Innate immune-related biological processes were also significantly downregulated in a mussel transmissible cancer[28], while the mammalian transmissible cancers display MHC downregulation[14–16]. Together this indicates that downregulation of immune processes is a conserved mechanism among transmissible cancers, though which processes likely depend on the host context and whether an adaptive immune system is present. As more BTNs are identified and characterized, a systematic comparison of differentially expressed genes and pathways would likely identify additional examples of convergent evolution and reveal underlying mechanisms of transmissible cancer evolution. Such mechanisms may also highlight more generally how conventional cancers are able to evade innate immune responses or identify metastasis-promoting mechanisms, since all BTNs have strong selective pressure for repeated metastasis.

Overcoming barriers to repeated transmission events and challenge by new host immune systems would suggest a highly adaptable cellular lineage. Indeed, widespread mutation and genome instability were observed in our prior MarBTN genomics study[18], and here we observe cases in which that genome instability directly affects the cancer transcriptome. Copy number alterations, which are highly variable across BTNs[18,19,42], may represent a particularly malleable mutation type for fine-tuning gene expression up or down to maximize cancer fitness in the face of changing selective pressures. Examples of the expressed genes influencing genome instability are also apparent, such as upregulation of genotoxic stress response gene *PUM3*. Previous work also identified the upregulation of an error-prone polymerase (*POLN*) and upregulation of *HSP9* (mortalin), which has been shown to sequester DNA damage response molecule p53[18,43]. This tolerance of genome instability, in combination with the generation of innovative mutations that affect gene expression, creates prime conditions for MarBTN to adapt and spread as a transmissible cancer. This cancer has successfully spread for at least 200 years[18], but it remains to be seen whether this lineage can continue to survive with widespread genome instability and mutation, or whether adaptability is solely a short-term benefit with the long-term cost of deleterious mutation accumulation in an asexual lineage, the process known as Mueller's ratchet[44].

In our seawater exposure experiment, we were surprised that the strongest responses, involved in metabolism, stress response, and cell cycle arrest, were conserved in both cancer and healthy clam hemocytes. A recent paper observed that mussel (*Mytilus edulis desolationis*) hemocytes are regularly released into seawater and transfer live into new mussels, postulating that this may facilitate the transfer of pathogen infection via hemocytes themselves[45]. This raises the intriguing possibility that bivalve hemocytes, for some unknown reason, are already adapted to survive for extended periods of time outside the bivalve body. Since BTNs originated as hemocytes[18,19], this would mean they already have the inherent ability to survive in seawater and enter new hosts and may in part explain why transmissible cancers are so common in bivalves. On top of these conserved responses, we observe cancer-specific responses to seawater that are absent in hemocytes, indicating that MarBTN may have evolved additional mechanisms to survive even better in seawater and increase its transmission ability.

In this study we investigated gene expression at two key stages of the hypothesized MarBTN life cycle: late-stage cancer infection and saltwater transfer. To gain a comprehensive understanding of MarBTN infection and progression, future work should also investigate gene expression at the early stages of cancer engraftment and proliferation, which would require sorting MarBTN from host cells. Host cell gene expression would also be informative about the clam defense response to MarBTN infection, and what defense regimens succeed at keeping the cancer contained versus succumbing to the infection. BTNs appear to be a common occurrence in bivalve populations and are likely to impose a strong selective pressure for resistance[40,46]. Identification of innate immune system cancer resistance mechanisms of hosts and countering evasion mechanisms in transmissible cancers, selected for by repeated infection, may each have broader implications in our understanding of the host-pathogen relationship of conventional cancers.

ACKNOWLEDGEMENTS

We thank Metzger lab members Jordana Sevigny, Karyn Tindbaek, and Finola Schmahl-Waggoner for feedback, Marisa Yonemitsu for dissecting the original reference clam, and Sophie Kogut for

investigating fusion genes. This work was supported by NIH training grants T32-HG000035 and T32-GM007270 (to S.F.M.H.), career transition award K22-CA226047 and R01-CA255712 (to M.J.M).

AUTHOR CONTRIBUTIONS

S.F.M.H. and M.J.M. contributed to study conceptualization. F.E.S.G. performed clam dissections. J.S.K. collected clams. S.F.M.H. extracted RNA, analyzed the data, and wrote the original manuscript. All authors contributed to the review and editing of the manuscript.

METHODS

Data availability

All code is available on GitHub (https://github.com/sfhart33/MarBTNtranscriptome), including all dependencies with version numbers. Raw sequence data are available via NCBI BioProject PRJNA874712 (https://www.ncbi.nlm.nih.gov/bioproject/874712). Data outputs can be obtained by running the supplied code on the raw data or on request. Note that code was written for our institute's working environment and thus some scripts may need to be altered manually to reproduce this analysis. Analysis was performed with an on-premises Linux server running Ubuntu 16.04. The Linux server was equipped with four Intel Xeon Gold 6148 CPUs and 250 GiB system memory.

Genome annotation

To utilize the NCBI Eukaryotic genome pipeline we supplied NCBI with the previously assembled *M. arenaria* genome[18] and RNAseq data for six tissues (foot, gill, hemocytes, mantle, adductor muscle, and siphon) from the clam that was used to assemble the reference genome. The output genome and annotation can be found at https://www.ncbi.nlm.nih.gov/assembly/GCF_026914265.1. We compared the completeness of the NCBI genome annotation to the original MAKER-annotated genome with Benchmark of Universal Single Copy Orthologs (BUSCO v3[47]) using the command: busco -m prot -l metazoa_odb10 and calculated other stats in **Supplementary Table 1** using custom scripts.

Sample collection

Clams were collected in Prince Edward Island, Canada (PSH samples, healthy clams) or by a commercial shellfish supplier in Maine (MELC and FFM samples, healthy and cancerous clams) and shipped live on ice to the Pacific Northwest Research Institute in Seattle, WA (**Supplementary Table 2**). Upon arrival, hemolymph was drawn from the pericardial sinus and checked for the presence of MarBTN with a highly sensitive cancer-specific qPCR assay (as described in [12]). The selected healthy clams were undetectable for the cancer-specific qPCR marker, while the selected MarBTN-infected clams had only cancerous cells (no host hemocytes) visible in hemolymph under a microscope. From healthy clams, 1 mL of hemolymph was spun at 500 × g for 10 min at 4 °C and hemolymph was pipetted off to leave a hemocyte cell pellet. For three of the healthy clams, dissections were performed to isolate foot, gill, mantle, adductor muscle, and siphon tissues. From MarBTN-infected clams, 1 mL of hemolymph was left for 1 hour in a 24-well plate at 4 °C to allow host hemocytes to adhere to the plate and the non-adherent MarBTN cells were collected by pipette. These isolates were spun at 500 × g for 10 min at 4 °C and hemolymph was removed to leave a MarBTN cell pellet. For three MarBTN isolates, half of the cells were resuspended in artificial sea water (ASW, 36 g/L Instant Ocean, Blacksburg, VA, USA) with antibiotics (1× concentration of penicillin/streptomycin , GenClone: Genesee Scientific, and 1 mM voriconazole, Acros Organics: Thermo Fisher Scientific) as described in [12], incubated at 4 °C for 24 hours to simulate seawater transfer, spun at 500 × g for 10 min at 4 °C, and hemolymph was pipetted off to leave ASW-treated MarBTN cell pellets. For three healthy hemocyte isolates, which are adherent, half of the cells (0.5mL) for each sample were left to adhere to a 24-well plate for 1 hour before pipetting off hemolymph, adding ASW plus antibiotics as above, incubating at 4 °C for 24 hours, pipetting off ASW, proceeding directly with RNA extraction with the digestion step directly on the plate that cells were adhered to. All other samples (healthy hemocytes, tissues, MarBTN isolates, and ASW-treated MarBTN isolates) were covered in RNAlater and stored at -80 °C until RNA extraction.

RNA extraction

RNA was extracted from each sample using the Qiagen RNeasy kit (Qiagen, Hilden, Germany), eluting in 60 µL elution buffer. Solid tissues were homogenized with a disposable plastic mortar and pestle in liquid nitrogen prior to extraction. DNase I (2 µL, 2,000 U/ml, RNase-free, New England Biolabs, Ipswich, MA), 10× DNase buffer, and water were then added to the eluted RNA to a total of 100 µL, and the reaction was incubated for 1 h at room temperature. Then 250 µL ethanol was added and mixed by pipette, and it was added to a second Qiagen RNeasy column. The RNeasy protocol was followed, skipping the RW1 step, adding 500 µL RPE 2×, and eluting in 40 µL elution buffer. RNA samples were then sequenced on a single Illumina HiSeq 4000 lane for 20-30 million reads per sample (Genewiz, Leipzig, Germany).

Differential expression analysis

We indexed the annotated genome and aligned reads for all samples using STAR [48], quantifying reads mapped per gene using --quantMode GeneCounts. We confirmed MarBTN isolates were all part of the USA sub-lineage at 48/48 mitochondrial loci differentiating USA vs PEI (see [18]), and the VAFs of USA-specific mitochondrial SNVs were 96-99+% in all samples, confirming high BTN purity.

We merged counts per gene for all samples and ran DESeq2 [49], using sample groupings (healthy tissues, hemocytes, or MarBTN) as conditions on which to test differential expression. We performed principal component analysis by applying variance stabilizing transformation using vst() and then plotPCA() from the DESeq2 package. We determined the top tissue-specific genes for each tissue by comparing each to the five others (e.g. gills versus all five non-gill tissues) using DESeq2 on read counts per gene, sorting by the “stat” output and taking the top 100 overexpressed genes for each tissue. We normalized read counts for each sample by calculating total mapped reads and multiplying so that each sample totaled the same number of reads as the maximum sample. We then performed hierarchical clustering on expression of the combined 6 sets of 100 top overexpressed genes for each tissue using the pheatmap package with clustering_distance_cols = "canberra". ASW-treated samples were excluded from

the original clustering analysis (**Supplementary Fig. 1**), then included alongside untreated hemocytes and MarBTN for principal component analysis and hierarchical clustering using expression of all genes and the same packages/functions as described above (**Supplementary Fig. 4**).

For the comparison of MarBTN to solid tissues, we combined all five solid tissue types and ran DESeq2 versus MarBTN. We ran similar comparisons for ASW-treated versus untreated MarBTN and hemocytes. For the comparison of differential expression results from multiple DESeq2 runs (e.g. **Supplementary Fig. 3**) we calculated a "+/- directional" p-value by taking the -log10 of the adjusted p-value when the log2 fold change was positive and log10 of the adjusted p-value when the log2 fold change was negative.

Gene set enrichment analysis

For gene set enrichment analysis, we first had to determine gene sets for *M. arenaria* genes. We used blastp to determine the closest uniprot hit for each gene, taking the gene with the highest e-value and leaving excluding genes that did not have a hit <1e-6. We then merged this list of genes with the msigdbr[50] *Homo sapiens* ontology gene set ("C5") to get putative *M. arenaria* gene sets. Separately, genes were rank-ordered using "stat" DESeq2 parameter using ties.method = "random" for each comparison (MarBTN vs hemocytes, MarBTN vs solid tissues, ASW-treated MarBTN vs untreated MarBTN, etc.). We then ran GSEA (clusterProfiler package) on each ranked gene lists with additional parameters: eps = 1e-1000, pvalueCutoff = 1, seed = 12345.

Identification of fusion genes

We identified fusion transcripts using STAR-Fusion (v1.11.0[51,52]). We first generated a custom genome index using prep_genome_lib.pl on the annotated genome with "--pfam_db current --dfam_db human" as default run parameters. We then ran STAR-Fusion each sample individually with default setting plus additional parameters: --FusionInspector validate, --examine_coding_effect, --denovo_reconstruct. We determined fusions shared by multiple samples by identical left and right breakpoints, excluding fusions

that were found in all samples (n=16) as likely genome assembly or annotation artifacts from our results. To compare the number of fusions in MarBTN samples versus hemocytes, we used a two-tailed t-test with unequal variance.

Copy number effects

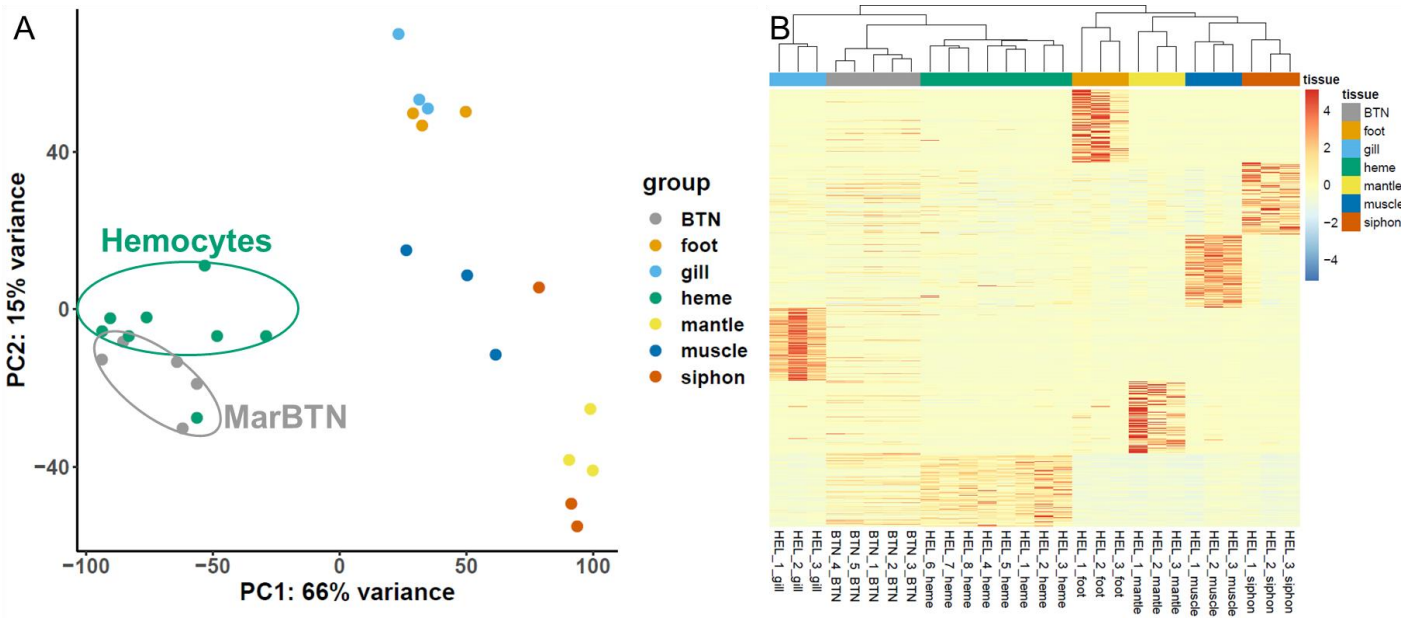
Genomic copy number calls were determined in 100 kB segments for USA sub-lineage MarBTN samples in previous work [18]. The copy number regions were observed to be nearly identical between the samples of the MarBTN from the USA sub-lineage, so while there are likely minor differences in these samples, these copy number calls are likely to be similar for the samples of this current study. We used bedtools intersect to link each gene to its genomic copy number state, excluding genes that were not at >90% at a single copy number state (e.g. gene spans a breakpoint in copy number). We dropped CN0 due to issues with reliably calling CN0 vs off-target mapping due to repetitive regions. We then created a boxplot for each copy number state of the log2 fold change of MarBTN versus healthy hemocytes, observing that higher copy number genes tend to have increased expression versus their diploid healthy references. We applied two tests for significance: two-sided Wilcoxon rank-sum tests between adjacent copy numbers and one-sample two-sided Wilcoxon rank-sum tests versus no fold change (log2 fold change = 0).

Steamer insertion effects

We had also determined *Steamer* insertion sites in previous work [18], and assumed that insertions previously found in all USA sub-lineage MarBTN samples would also be present in the samples of this current study, which were all confirmed to be from the USA sub-lineage (see above). We determined where *Steamer* had inserted within genes or within 2 kB upstream genes using bedtools intersect. As a control, we took genes intersecting insertions that were found in the PEI sub-lineage but not USA sub-lineage as sites that are unlikely to be present in the samples of this current study but that were accessible for *Steamer* insertion. We applied two tests for significance: two-sided Wilcoxon rank-sum tests between genes with

steamer insertions and controls, and one-sample two-sided Wilcoxon rank-sum tests versus no fold change (log2 fold change = 0) for genes with steamer insertions.

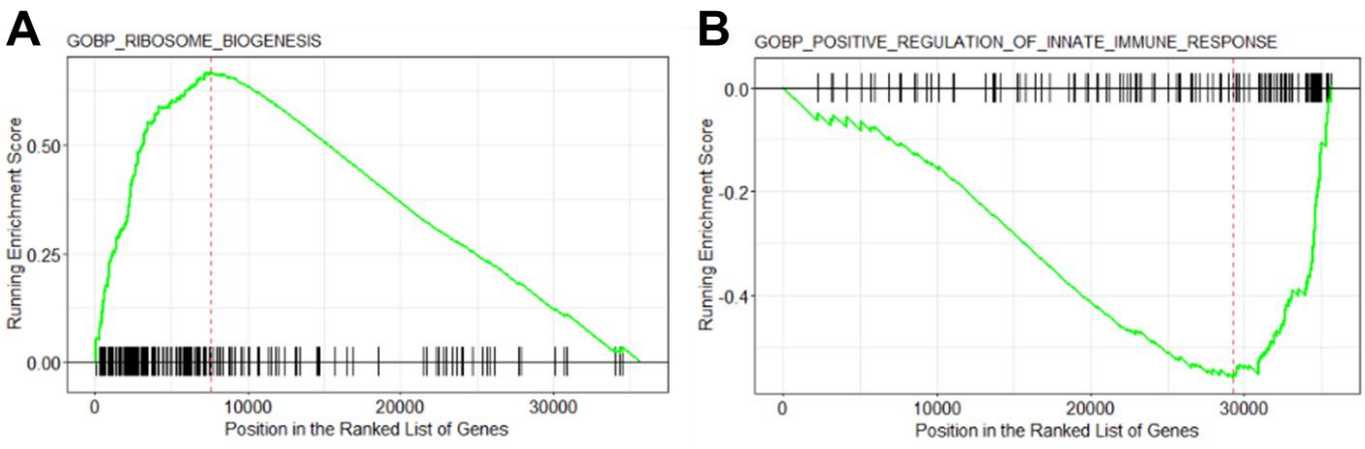
SUPPLEMENTAL FIGURES



Supplementary Figure 1. Hemocyte origin of MarBTN supported by PCA and clustering with new gene annotations

(A) Principal component analysis of normalized expression across all genes, with PC1 separating MarBTN and hemocytes from all other tissues. **(B)** Hierarchical clustering of all RNA sequenced samples (excluding ASW-treated samples) by the expression of the top 100 most significant genes expressed in each specific healthy tissue relative to all other tissues, with heatmap of normalized relative gene expression for each gene. MarBTN ("BTN") clusters most closely with hemocytes ("heme"), supporting principal component analysis results. Results for both panels closely match similar analyses with previous genome annotation and a smaller sample set [18].

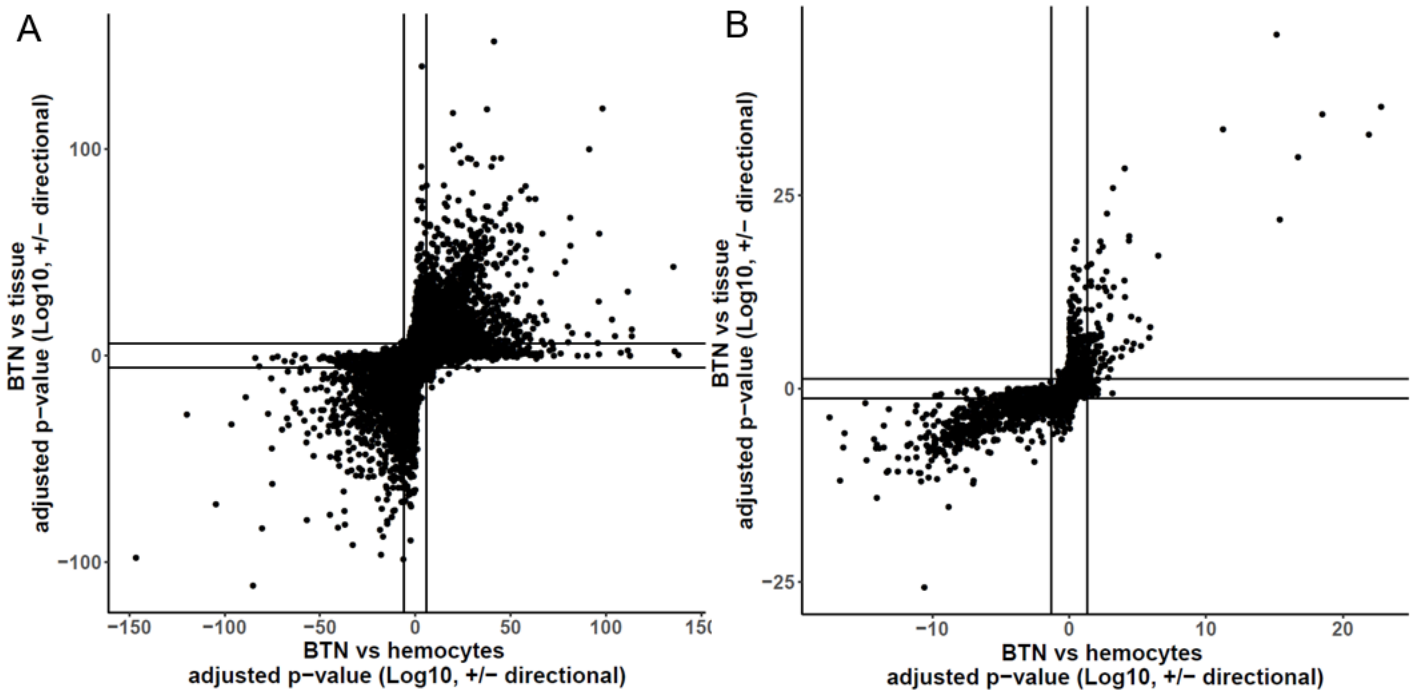
1



Supplementary Figure 2. Example GSEA results for one each of the top up- and downregulated pathways.

Running enrichment score (green), which increases each time it hits a gene in the gene set (black bars along x-axis) for the ribosome biogenesis biological process **(A**), one of the top upregulated pathways, and positive regulation of innate immune response biological process **(B)**, one of the top downregulated pathways.

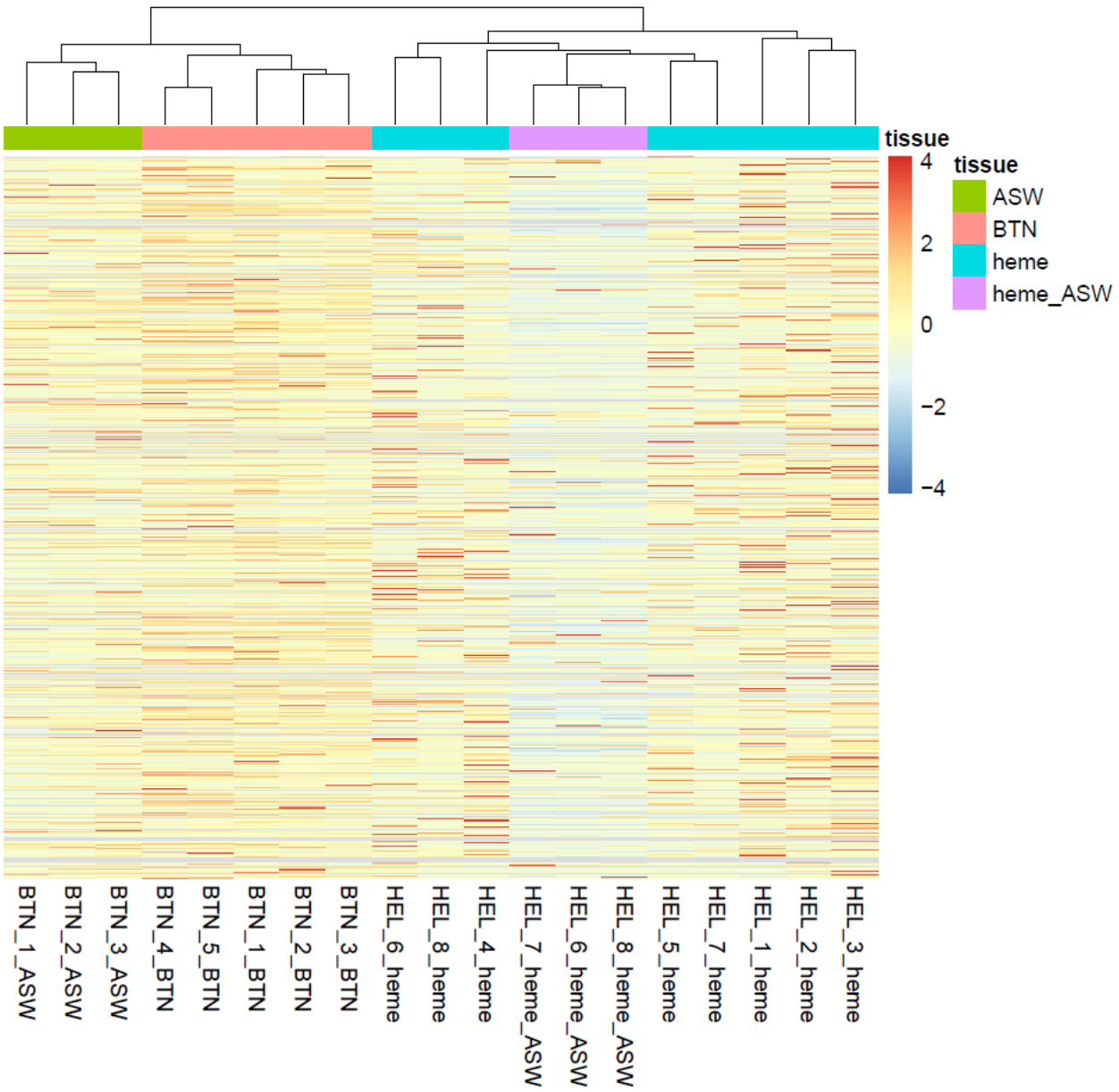
2



Supplementary Figure 3. Differential expression is similar whether comparing MarBTN to hemocytes or solid tissue.

Adjusted p-values, further adjusted to be positive for upregulation and negative for downregulation, from MarBTN versus healthy hemocytes differential expression results (x-axes) and MarBTN versus solid tissues differential expression results (y-axes) for individual genes **(A)** and gene pathways **(B)**. In general, genes and pathways that are upregulated versus hemocytes are also upregulated versus solid tissues, indicating that major differential expression results and conclusions are not artifacts of the comparison with hemocytes. Lines represent false discovery-corrected $p < 0.05$ significance thresholds.

1

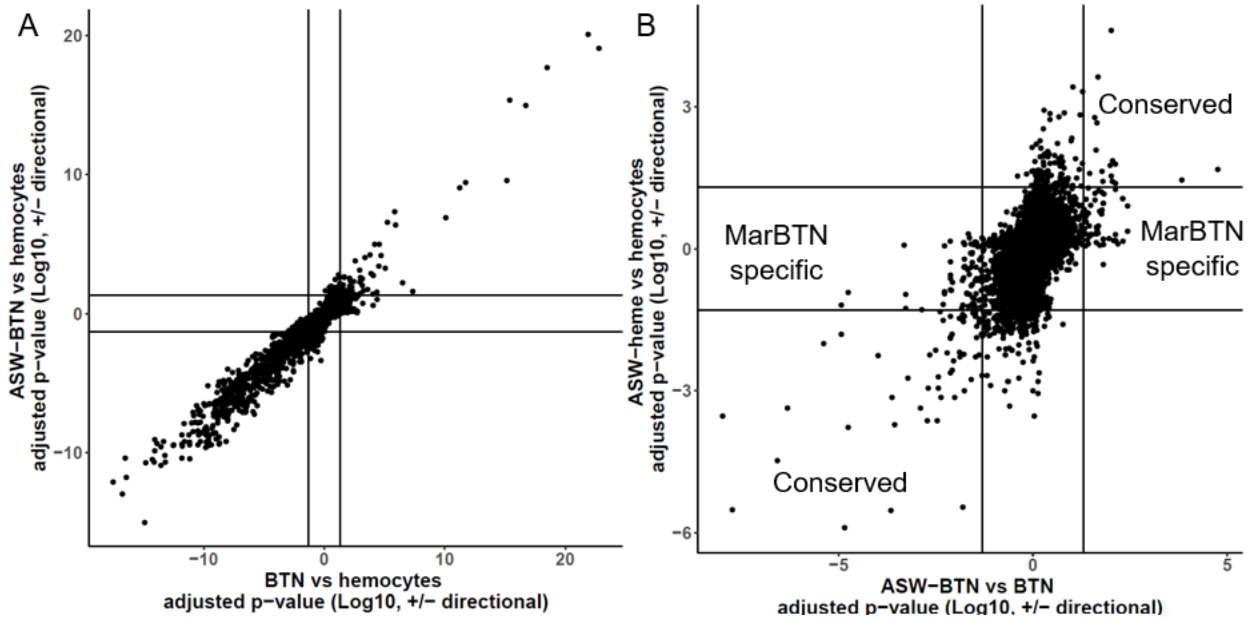


Supplementary Figure 4. Hierarchical clustering separates samples by seawater treatment.

Hierarchical clustering using all genes for untreated healthy hemocytes ("heme" - blue), ASW-treated healthy hemocytes ("heme_ASW" - pink), untreated MarBTN ("BTN" - red) and ASW-treated MarBTN ("ASW" - green). Samples are labeled by their source clam and treatment.

2

1



Supplementary Figure 5. Differential expressed pathways show similarities across comparisons.

Adjusted p-values, further adjusted to be positive for upregulation and negative for downregulation, to compare gene set enrichment analysis results from **(A)** ASW-treated MarBTN and untreated MarBTN each versus hemocytes and **(B)** ASW-treated versus untreated for each of MarBTN and hemocytes. In **(A)** results are highly correlated, with the same top up- and downregulated pathways regardless of which treatment is compared to healthy hemocytes. In **(B)**, gene set groupings corresponding to conserved and MarBTN-specific seawater responses from Figure 4 are labeled. Lines represent false discovery-corrected $p < 0.05$ significance thresholds.

2

REFERENCES

1. Pearse, A.-M. & Swift, K. Transmission of devil facial-tumour disease. *Nature* **439**, 549 (2006).
2. Pye, R. J. *et al.* A second transmissible cancer in Tasmanian devils. *PNAS* **113**, 374–379 (2016).
3. 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).
4. Rebbeck, C. A., Thomas, R., Breen, M., Leroi, A. M. & Burt, A. Origins and Evolution of a Transmissible Cancer. *Evolution* **63**, 2340–2349 (2009).
5. 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).
6. Metzger, M. J. *et al.* Widespread transmission of independent cancer lineages within multiple bivalve species. *Nature* **534**, 705–709 (2016).
7. 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**, (2019).
8. 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).
9. 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. *Molecular Ecology* **31**, 3128–3136 (2022).
10. Yonemitsu, M. A. *et al.* Multiple lineages of transmissible neoplasia in the basket cockle (Clinocardium nuttallii) with repeated horizontal transfer of mitochondrial DNA. 2023.10.11.561945 Preprint at https://doi.org/10.1101/2023.10.11.561945 (2023).
11. Metzger, M. J. & Goff, S. P. A sixth modality of infectious disease: contagious cancer from devils to clams and beyond. *PLOS Pathogens* **12**, e1005904 (2016).

12. Giersch, R. M. *et al.* Survival and Detection of Bivalve Transmissible Neoplasia from the Soft-Shell Clam Mya arenaria (MarBTN) in Seawater. *Pathogens* **11**, 283 (2022).

13. Burioli, E. A. V. *et al.* Traits of a mussel transmissible cancer are reminiscent of a parasitic life style. *Sci Rep* **11**, 24110 (2021).

14. Siddle, H. V. *et al.* Reversible epigenetic down-regulation of MHC molecules by devil facial tumour disease illustrates immune escape by a contagious cancer. *PNAS* **110**, 5103–5108 (2013).

15. Siddle, H. V. & Kaufman, J. Immunology of naturally transmissible tumours. *Immunology* **144**, 11–20 (2015).

16. Yang, T. J., Chandler, J. P. & Dunne-Anway, S. Growth stage dependent expression of MHC antigens on the canine transmissible venereal sarcoma. *Br J Cancer* **55**, 131–134 (1987).

17. Miller, W. *et al.* Genetic diversity and population structure of the endangered marsupial Sarcophilus harrisii (Tasmanian devil). *Proceedings of the National Academy of Sciences* **108**, 12348–12353 (2011).

18. Hart, S. F. M. *et al.* Centuries of genome instability and evolution in soft-shell clam, Mya arenaria, bivalve transmissible neoplasia. *Nat Cancer* 1–14 (2023) doi:10.1038/s43018-023-00643-7.

19. Bruzos, A. L. *et al.* Somatic evolution of marine transmissible leukemias in the common cockle, Cerastoderma edule. *Nat Cancer* **4**, 1575–1591 (2023).

20. Reno, P. W., House, M. & Illingworth, A. Flow cytometric and chromosome analysis of softshell clams, Mya arenaria, with disseminated neoplasia. *Journal of Invertebrate Pathology* **64**, 163–172 (1994).

21. Cho, H.-C. *et al.* Puf-A promotes cancer progression by interacting with nucleophosmin in nucleolus. *Oncogene* **41**, 1155–1165 (2022).

22. Chang, H.-Y. *et al.* hPuf-A/KIAA0020 Modulates PARP-1 Cleavage upon Genotoxic Stress. *Cancer Research* **71**, 1126–1134 (2011).

23. Huang, J. *et al.* RAD18 transmits DNA damage signalling to elicit homologous recombination repair. *Nat Cell Biol* **11**, 592–603 (2009).

24. Srougi, M. C. & Burridge, K. The Nuclear Guanine Nucleotide Exchange Factors Ect2 and Net1 Regulate RhoB-Mediated Cell Death after DNA Damage. *PLOS ONE* **6**, e17108 (2011).

25. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550 (2005).

26. Elhamamsy, A. R., Metge, B. J., Alsheikh, H. A., Shevde, L. A. & Samant, R. S. Ribosome Biogenesis: A Central Player in Cancer Metastasis and Therapeutic Resistance. *Cancer Res* **82**, 2344–2353 (2022).

27. Sunila, I. Respiration of sarcoma cells from the soft-shell clam Mya arenaria L. under various conditions. *Journal of Experimental Marine Biology and Ecology* **150**, 19–29 (1991).

28. Burioli, E. a. V. *et al.* Transcriptomics of mussel transmissible cancer MtrBTN2 suggests accumulation of multiple cancer traits and oncogenic pathways shared among bilaterians. *Open Biology* **13**, 230259 (2023).

29. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).

30. Arriagada, G. *et al.* Activation of transcription and retrotransposition of a novel retroelement, Steamer, in neoplastic hemocytes of the mollusk Mya arenaria. *PNAS* **111**, 14175–14180 (2014).

31. Romanish, M. T., Lock, W. M., van de Lagemaat, L. N., Dunn, C. A. & Mager, D. L. Repeated Recruitment of LTR Retrotransposons as Promoters by the Anti-Apoptotic Locus NAIP during Mammalian Evolution. *PLoS Genet* **3**, e10 (2007).

32. Grasmann, G., Smolle, E., Olschewski, H. & Leithner, K. Gluconeogenesis in cancer cells – repurposing of a starvation-induced metabolic pathway? *Biochim Biophys Acta Rev Cancer* **1872**, 24–36 (2019).

33. Damgaard, R. B. *et al.* The ubiquitin ligase XIAP recruits LUBAC for NOD2 signaling in inflammation and innate immunity. *Mol Cell* **46**, 746–758 (2012).

34. Vavassori, S. *et al.* Multisystem inflammation and susceptibility to viral infections in human ZNFX1 deficiency. *J Allergy Clin Immunol* **148**, 381–393 (2021).

35. Xu, Y. *et al.* CCNG1 (Cyclin G1) regulation by mutant-P53 via induction of Notch3 expression promotes high-grade serous ovarian cancer (HGSOC) tumorigenesis and progression. *Cancer Medicine* **8**, 351–362 (2019).

36. Zhang, L. *et al.* The miR-181 family promotes cell cycle by targeting CTDSPL, a phosphatase-like tumor suppressor in uveal melanoma. *J Exp Clin Cancer Res* **37**, 15 (2018).

37. Sørensen, J. G., Kristensen, T. N. & Loeschcke, V. The evolutionary and ecological role of heat shock proteins. *Ecology Letters* **6**, 1025–1037 (2003).

38. Jang, S.-M., Redon, C. E., Thakur, B. L., Bahta, M. K. & Aladjem, M. I. Regulation of cell cycle drivers by Cullin-RING ubiquitin ligases. *Exp Mol Med* **52**, 1637–1651 (2020).

39. Zitvogel, L., Tesniere, A. & Kroemer, G. Cancer despite immunosurveillance: immunoselection and immunosubversion. *Nat Rev Immunol* **6**, 715–727 (2006).

40. Ujvari, B., Gatenby, R. A. & Thomas, F. The evolutionary ecology of transmissible cancers. *Infection, Genetics and Evolution* **39**, 293–303 (2016).

41. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).

42. Burioli, E. A. V. *et al.* Implementation of various approaches to study the prevalence, incidence and progression of disseminated neoplasia in mussel stocks. *Journal of Invertebrate Pathology* **168**, 107271 (2019).

43. Walker, C., Böttger, S. & Low, B. Mortalin-Based Cytoplasmic Sequestration of p53 in a Nonmammalian Cancer Model. *Am J Pathol* **168**, 1526–1530 (2006).

44. Ní Leathlobhair, M. & Lenski, R. E. Population genetics of clonally transmissible cancers. *Nat Ecol Evol* **6**, 1077–1089 (2022).

45. Caza, F., Bernet, E., Veyrier, F. J., Betoulle, S. & St-Pierre, Y. Hemocytes released in seawater act as Trojan horses for spreading of bacterial infections in mussels. *Sci Rep* **10**, 19696 (2020).

46. Epstein, B. *et al.* Rapid evolutionary response to a transmissible cancer in Tasmanian devils. *Nat Commun* **7**, (2016).

47. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

48. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

49. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).

50. Dolgalev, I. msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format. (2022).

51. Haas, B. J. *et al.* STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. 120295 Preprint at https://doi.org/10.1101/120295 (2017).

52. Haas, B. J. *et al.* Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biology* **20**, 213 (2019).