Evolution of soft-shell clam transmissible cancer

Samuel F. M. Hart

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Michael Metzger, Chair

Kelley Harris

Harmit Malik

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Evolution of soft-shell clam transmissible cancer

Samuel F. M. Hart

Chair of the Supervisory Committee:
Dr. Michael Metzger
Assistant Investigator, Pacific Northwest Research Institute
Affiliate Faculty, Genome Sciences, University of Washington

Transmissible cancers are infectious parasitic clones that metastasize to new hosts, living past the death of the founder animal in which the cancer initiated. Using genomic and transcriptomic analyses, we investigated the evolutionary history of a recently identified transmissible cancer lineage that has spread though the soft-shell clam (*Mya arenaria*) population. We first assembled a chromosome-scale soft-shell clam reference genome and characterized somatic mutations in the cancer lineage, discovering a novel mutational signature that estimates the lineage to be >200 years old and observing a wide variety of mutation types indicative of an extremely unstable cancer genome. We next quantified gene expression, observing differential expression of stereotypical cancer hallmarks in addition to key pathways that may have facilitated the cancer's ability to survive seawater transfer to new hosts and evade immune rejection. Finally, we developed genetic assays to track bivalve transmissible neoplasia in the soft-shell clam and identify a novel lineage in the Baltic clam (*Macoma balthica*). Taken together, this study reveals the long-term survival of an invertebrate cancer lineage and identifies adaptive mechanisms to evade physical and immune barriers to cancer transmission, which may have been facilitated in part by the adaptive potential of its unstable genome.

Table of Contents

Acknowledgements

First, I want to thank Michael for the freedom he gave me to make this project my own, the support to grow as a scientist over these last five years, the countless hours of meetings and feedback, and for diligently making sure I got my weekly donut. I also couldn't ask for a better team of techs to work with in the lab, the field, and side by side on our computers (let's be honest, mostly that) than Rachael, Marisa, Jordana, Fiona, Finola, and Karyn. Thanks to the PNRI community, particularly Rick for being my backup mentor downstairs, Patrick for making everything work behind the scenes, and Ricky for bravely accompanying me on our adventure as the sole MCB students on Cap Hill. Thanks to my committee - Kelley, Alice, Harmit, and Steven - for their encouragement, support, and tolerance despite my inauspicious scheduling of our first two meetings in 2020. Thanks to the MCB community, especially the 2018 cohort for being the best, Maia for also being the best, and Denise for addressing my quarterly panicked emails about missing tuition payments. Thanks to the Harris, Mclaughlin, Tepolt/Tarrant and Feder labs for broadening my horizons in their lab meetings and journal clubs. Finally, thanks to John and Amanda for launching me on my science journey as a wide-eyed undergrad at UVM, and to Wenying for equipping me as a tech with the tools to succeed in graduate school.

Many lovely friends and family also made this project possible through their unwavering support. Thanks to Mom for making sure I took care of my health before everything else, to Dad for exemplifying what it means to be a lifelong learner and for passing on your uncanny focusing gene, to Annie for letting me live vicariously through her cougar research (sorry, clams, you can't compare), Kiva for looking after my parents while I'm away, and to Gram for her endless love. Thanks to my Seattle family – Sayre, Andrew, Thatcher, Owen, Oscar, and Tula – for Sunday dinners, covid escapes, and swimming in Lake Washington. Thanks to my covid lockdown roommates – Pete, Dre, Benji and Pico - for Gloomhaven, vibe-setting, and all the presents left outside my door. Thanks to Seattle Garbage and the Wednesday workout crew – Bronson, Petter, Stump, Blair, and Mike – for keeping up my endorphins. Thanks to Chelsey for letting me beat her at cribbage even after nine years. Thanks to Mike and Julian for making sure I left work behind to take an occasional adventure, even though they happen less often and are less dangerous than they once were. Thanks to Todd for talking me through it all. Last and anything but least, thank you Addy for being my rock, confidant, sounding board, editor, forced roommate, public health advisor, gym buddy, caregiver, and everything else I needed along the way.

Dedication

For the clams, and all other animal models that make the unchosen and ultimate sacrifice in the name of science.

Preface

This dissertation is organized into three results chapters, each written as a stand-alone manuscript, and a short conclusion chapter. As such, each results chapter has its own introduction and can be read individually, while the conclusion chapter takes a high-level look at the state of the field at the conclusion of my PhD. Together these chapters represent my PhD work in the Metzger lab studying the biology of soft-shell clam transmissible cancer. Though the bulk of the research and writing was done by me, shaped by an enormous amount of guidance, feedback and groundwork laid by Michael, the contributions section of each chapter lays out where others have contributed to this work. Thanks for reading!

Chapter 1: Centuries of genome instability and evolution in soft-shell clam transmissible cancer

Authors: Samuel F.M. Hart [1,2], Marisa A. Yonemitsu [1,2], Rachael M. Giersch [1], Fiona E. S. Garrett [1], Brian F. Beal [3,4], Gloria Arriagada [5,6], Brian W. Davis [7,8], Elaine A. Ostrander [9], Stephen P. Goff [10,11], Michael J. Metzger [1,2]

Affiliations:

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

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

[3] Division of Environmental and Biological Sciences, University of Maine at Machias, Machias, ME, USA

[4] Downeast Institute, Beals, ME, USA

[5] Instituto de Ciencias Biomedicas, Facultad de Medicina y Facultad de Ciencias de la Vida, Universidad Andres Bello, Santiago, Chile

[6] FONDAP Center for Genome Regulation, Santiago, Chile

[7] Department of Veterinary Integrative Biosciences, Texas A&M University School of Veterinary Medicine, College Station, TX , USA

[8] Department of Small Animal Clinical Sciences, Texas A&M University School of Veterinary Medicine, College Station, TX , USA

[9] Cancer Genetics and Comparative Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, 20892, USA

[10] Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA

[11] Department of Microbiology and Immunology, Columbia University, New York, NY, USA

1.1 Abstract

Transmissible cancers are infectious parasitic clones that metastasize to new hosts, living past the death of the founder animal in which the cancer initiated. We investigated the evolutionary history of a cancer lineage that has spread though the soft-shell clam (*Mya arenaria*) population by assembling a chromosome-scale soft-shell clam reference genome and characterizing somatic mutations in transmissible cancer. We observe high mutation density, widespread copy number gain, structural rearrangement, loss of heterozygosity, variable telomere lengths, mitochondrial genome expansion, and transposable element activity, all indicative of an unstable cancer genome. We also discover a previously unreported mutational signature associated with overexpression of an error-prone polymerase and use this to estimate the lineage to be >200 years old. Our study reveals the ability for an invertebrate cancer lineage to survive for centuries while its genome continues to structurally mutate, likely contributing to the evolution of this lineage as a parasitic cancer.

1.2 Introduction

Most cancers arise from oncogenic mutations in host cells and remain confined to the body of that host. However, a small number of transmissible cancer lineages exist in which cancer cells metastasize repeatedly to new hosts, living past the death of their original hosts as asexually reproducing unicellular organisms [1]. Observed cases of transmissible cancer in nature include Canine Transmissible Venereal Tumor (CTVT) in dogs [2,3], two unrelated lineages of Devil Facial Tumor Disease (DFTD) in Tasmanian devils [4,5], and at least eight Bivalve Transmissible Neoplasia (BTN) lineages observed in several marine bivalve species [6–10]. Although transmissible cancers and their host genomes have been well characterized in dogs [11–13] and devils [14–16], little is known about the evolutionary history of the BTN lineages, which have only recently been recognized as transmissible cancers. Here we perform the first genome-wide analysis of a BTN lineage, focusing on the single lineage found in the soft-shell clam (*Mya arenaria*), or MarBTN.

BTN is a fatal leukemia-like cancer characterized by high numbers of cancer cells in the circulatory fluid of the bivalve and dissemination into tissues in the later stages of disease. BTN cells can survive for

days to weeks in seawater [17,18] and likely spread from animal to animal by transmission through the water column. This cancer, referred to in the literature as disseminated neoplasia or hemic neoplasia, was first reported in soft-shell clams in the 1970s [19,20] and has since been found across much of the soft-shell clam native range along the east coast of North America (**Fig. 1.1A**). In the 1980s in New England and in the 2000s in Prince Edward Island, Canada, severe outbreaks were documented with prevalence as high as 90% followed by severe population losses [21,22]. The disease is still observed throughout this range, although no more recent large-scale population die-offs have been reported. All disseminated neoplasia isolates tested in a 2015 study were shown to be of clonal origin, and it was hypothesized that historical observations of the cancer dating back to the 1970s were occurrences of this same clonal lineage [6]. However, it is not known how long this lineage has propagated, or how the genome has evolved since the original cancer initiated. To address these and other questions, we assembled a high-quality soft-shell clam reference genome and characterized the genome evolution of the MarBTN lineage by comparative analysis of healthy clam and MarBTN sequences. We show a striking pattern of mutation occurrence and evolution, suggestive of an unstable genome with the potential to rapidly mutate despite its long-term survival.

1.3 Results

1.3.1 Sample sequencing and genome assembly

We assembled a soft-shell clam reference genome from a single healthy female clam collected from Larrabee Cove, Machiasport, Maine, USA (**Fig. 1.1B-C**, MELC-2E11). We assembled PacBio long reads into contigs using FALCON-Unzip [23], scaffolded contigs to the chromosome-level with Hi-C sequences using FALCON-Phase (**Supplementary Fig. 1.1**), polished the scaffolds using 10X Chromium reads, and annotated with RNAseq reads using Maker to yield a high quality reference genome. The final reference genome is 1.22 Gb organized into 17 phased scaffolds, matching the 17 chromosomes expected based on karyotype data [24]. The contig N50 is 3.4 Mb and the metazoan BUSCO (Benchmarking Universal Single Copy Orthologs [25]) score is 94.9%. Our assembly is similar in size, GC, and repeat content of a recently published *Mya arenaria* genome [26] but with drastically improved contiguity and completeness

(**Supplementary Table 1.1**), allowing for comprehensive genomic investigation into the evolutionary history of MarBTN.

We performed whole genome sequencing (WGS) on three healthy uninfected clams and eight isolates of MarBTN from the hemolymph of highly infected clams (e.g. **Fig. 1.1D, Supplementary Fig. 1.2**) sampled from five locations across the established MarBTN range [27] (**Fig. 1.1A, Supplementary Table 1.2**), and called single nucleotide variants (SNVs) against the reference genome. Contaminating host variants were removed from MarBTN sequences via variant calling thresholds, rather than using paired tissue sequences as has been done for other transmissible cancers, since MarBTN hemolymph isolates were high purity (>96% cancer DNA) while paired tissue samples from the host often contained high cancer DNA due to dissemination (**Supplementary Fig. 1.3**).

To investigate somatic evolution of the MarBTN lineage, it is important to distinguish between founder variants, those present in the genome of the founder clam from which the cancer initially arose, and somatic mutations, which occurred during the propagation and evolution of the cancer lineage. We observed that 10.7 million SNVs were shared by all MarBTN samples but not present in the reference genome. Of these, 8.1 million were found in at least one of the three healthy clams, indicating that these variants are likely from the germline of the founder.

A MarBTN phylogeny, built from pairwise SNV differences between samples, confirmed the previous analysis identifying two distinct sub-lineages of MarBTN [6], here referred to as the Prince Edward Island (PEI) and United States of America (USA) sub-lineages **(Fig. 1.1E)**. While the original founder clam is lost, we are able to leverage this deep split between the sub-lineages to identify those mutation likely to be somatic and not founder, as SNVs that occurred after the divergence of the two subgroups would be somatic. Most SNVs identified in the cancers and also found in healthy animals (and therefore highly likely to be founder variants) were present in both sub-lineages of MarBTN, but we observed some genomic regions with clusters of these founder SNVs in one sub-lineage but not the other. These are unlikely to be somatic mutations, instead they likely indicate loss of heterozygosity (LOH) events which took place after divergence of the sub-lineages. LOH was identified in 8% and 13% of the USA and PEI sub-lineage

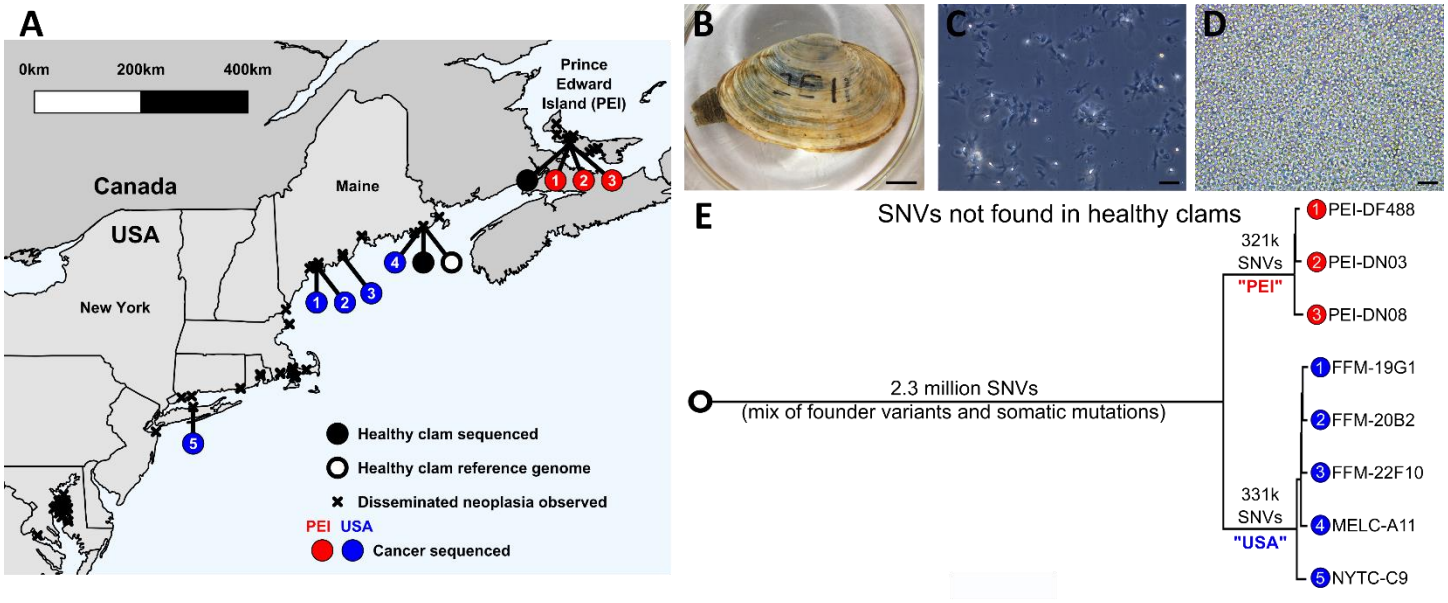


Figure 1.1. MarBTN distribution and sequencing.

(A) Locations of samples sequenced (circles) and disseminated neoplasia observations (x's) along the east coast of North America. Circles colored for healthy clams (black) and MarBTN sampled from the PEI (red) or USA (blue) coast. **(B)** Image of healthy clam used to assemble reference genome (MELC-2E11) and **(C)** hemolymph of the same clam, with hemocytes extending pseudopodia. The healthy reference clam (open black circle in A) was included in whole genome sequencing analysis. **(D)** Hemolymph from a clam infected with MarBTN (FFM-22F10), with distinct rounded morphology and lack of pseudopodia of cancer cells. Scale bars are 10 mm for the clam and 50 µm for the hemolymph. **(E)** Phylogeny of cancer samples built from pairwise differences of SNVs not found in healthy clams, excluding regions that show evidence of LOH. Numbers along branches indicate the number of SNVs unique to and shared by individuals in that clade. All nodes have 100/100 bootstrap support.

genomes, respectively **(Supplementary Fig. 1.4)**. LOH regions were excluded during identification of somatic mutations in the following SNV analysis unless otherwise noted, since we are unable to determine which mutations are founder and which are somatic in these regions. SNVs found in all cancer samples, but no healthy samples, represent a mix of both founder variants and somatic mutations (2.3 million), while SNVs found in just one or the other sub-lineage represent likely somatic mutations (700 k). The majority of these SNVs were shared by all individuals in a sub-lineage and are herein referred to as "high confidence somatic mutations" (321 k for PEI and 331 k for USA).

1.3.2 Mutational biases in MarBTN

By analyzing all identified SNVs and their trinucleotide context, we observed a distinct SNV mutational bias in somatic mutations within both the PEI and USA sub-lineages that was not found in healthy clams (**Fig. 1.2B).** These biases are nearly identical in somatic SNVs from both sub-lineages and were also present in more recent mutations, such as SNVs unique to each MarBTN sample **(Supplementary**

Fig. 1.5). *De novo* signature extraction, which deconvolutes mutational biases in their trinucleotide context between samples [28], yielded four mutational signatures (**Supplementary Fig. 1.6).** Three signatures were found in both healthy clams and MarBTN samples, and thus are likely endogenous within the germline of clam genomes. One signature closely resembles COSMIC signature 1 (termed Sig1'), showing a characteristic bias for C>T mutations at CpG sites, which is associated with the deamination of methylated CpGs in humans [29]. Sig1' represents a greater fraction of mutations in the PEI sub-lineage **(Supplementary Fig. 1.7)**, which may indicate that PEI has more methylated CpG sites than USA. Sig1' also represents a greater fraction of mutations in coding regions **(Supplementary Fig. 1.8)**, fitting prior observations that methylation is elevated in gene regions in bivalves [30]. The other two signatures are "flatter" and less distinctive, most closely resembling COSMIC signatures 5 and 40 (termed Sig5' and Sig40'), which are both associated with aging in humans [31,32].

A single signature captured the biases specific to the Somatic mutations in MarBTN, termed SigS (**Fig. 1.2A).** The closest analog in the COSMIC database of human mutational signatures is signature 9, which shares a T>G bias in A/T trinucleotide contexts [31]. Signature 9 in humans represent mutations induced by polymerase eta during somatic hypermutation and translesion synthesis in humans [31,33]. This may indicate that an error-prone polymerase with similar biases to human polymerase eta is broadly upregulated in the cancer or induced due to a high level of DNA lesions during MarBTN replication. In addition to the striking T>G bias in A/T contexts, there is also a notable bias towards C>A mutations compared to healthy clam SNVs, particularly CC>CA and TCT>TAT. Interestingly, both C>A and T>G mutations have been linked to oxidative DNA damage [34]. Clam hemolymph is strongly hypoxic in late stages of the disease [35], so this environment may also be contributing to these mutational biases.

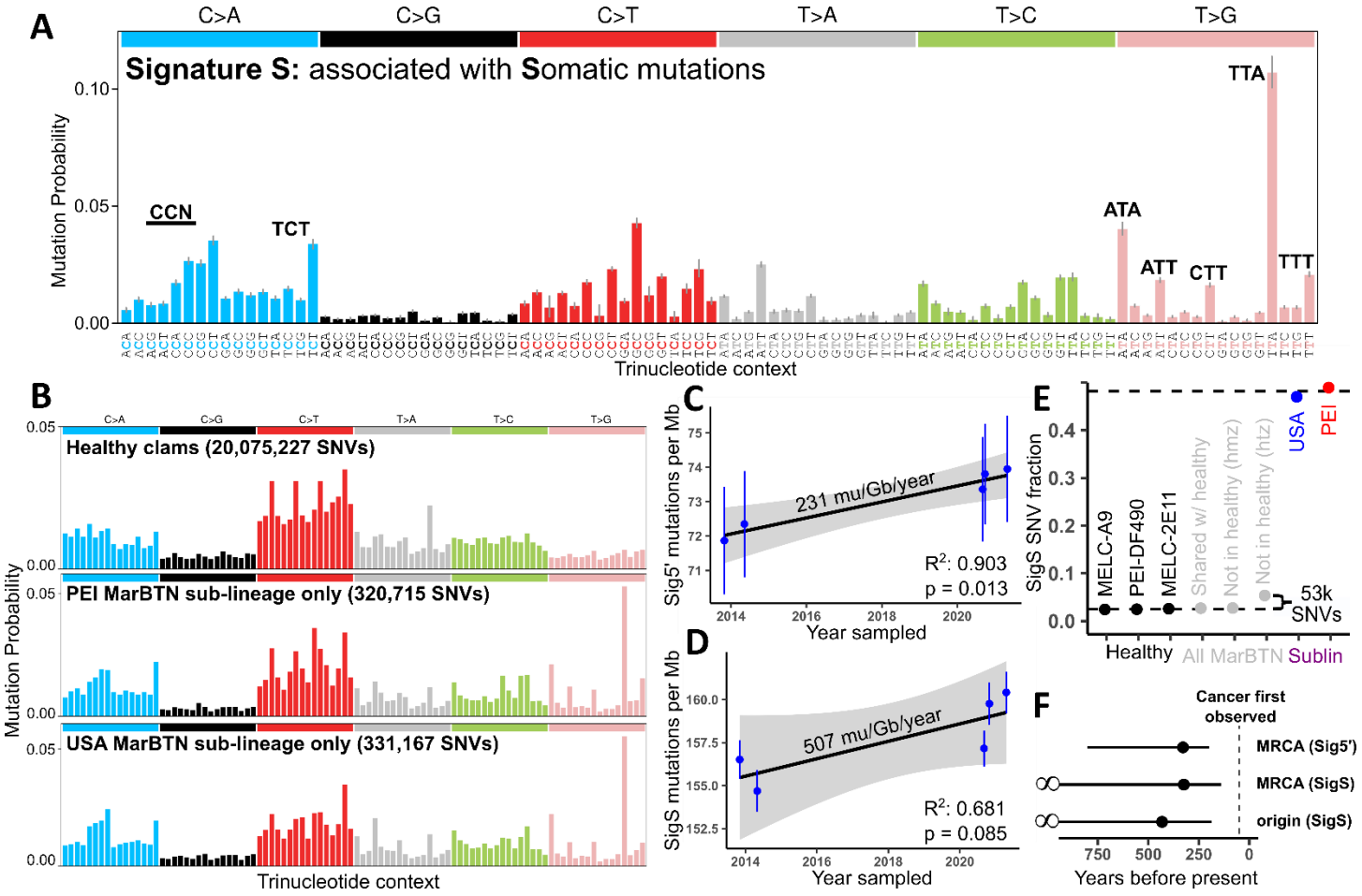


Figure 1.2. Unique mutational signature found in somatic mutations dates cancer to >200 years old.

(A) *De novo* extracted mutational biases for SigS. **(B)** Trinucleotide context of SNVs found in healthy clams (top) and high confidence somatic mutations in PEI (middle) or USA (bottom) sub-lineages, corrected for mutational opportunities in the clam genome. Trinucleotide order same as in A. **(C)** Sig5' mutations and **(D)** SigS mutations per Mb across USA MarBTN samples correlated with sampling date, with linear regression and 95% confidence interval (grey) overlaid. SNVs found in healthy clams, PEI MarBTN samples, or LOH regions are excluded. **(E)** Fraction of SNVs attributed to SigS from healthy clams (black), variants found in all MarBTN samples (grey), and high confidence somatic mutations (colored). Variants found in all MarBTN samples are divided by whether they are found in healthy clams and whether they are homozygous (hmz) or heterozygous (htz). Dashed lines display SigS fraction estimates for likely somatic mutations and likely founder variants. **(F)** Age estimate of the most recent common ancestor (MRCA) of the USA and PEI sub-lineages using Sig5' and SigS, and of the BTN origin from SigS mutations. Error bars in all plots display 95% confidence intervals.

1.3.3 MarBTN is several centuries old

Signatures 1 and 5 are considered clock-like in humans and other mammals [36,37], and signature 1 was used to date CTVT's origin to 4,000-8,500 years before present [12]. We took advantage of the temporal distribution of our USA samples to test whether any signatures were clock-like in MarBTN. We fit somatic mutations for each sample (SNVs not in other sub-lineage and outside LOH regions) to the four extracted

signatures and regressed mutations attributed to each signature against sample collection date (**Supplementary Fig. 1.9**). Sig1' did not correlate with time, perhaps due to methylation changes affecting CpG>TpG mutation rates and/or inherent differences between clams and mammals. Sig5' mutations did display a strong correlation with time within the USA samples (**Fig. 1.2C**, p=0.013). Assuming Sig5' mutation rate has remained steady since USA diverged from PEI, this corresponds to the sub-lineages diverging 319 years ago (95% CI: 199-801 years). However, PEI samples have 33% fewer Sig5' mutations than USA samples, indicating that Sig5' mutation rate differs between sub-lineages. SigS mutations also appear to increase with time, and although the correlation is not statistically significant within the USA sub-lineage (**Fig. 1.2D**, p=0.085), the number of SigS mutations in PEI samples fall within the range predicted by the linear regression of USA samples (**Supplementary Fig. 1.9**). Minimal deviation in the SigS accumulation over time across both sub-lineages, despite their deep divergence, indicates that mechanism producing SigS mutations is remarkably steady, although the lack of recent PEI samples does not allow us to independently test whether SigS continues to accumulate at the same rate in PEI. Based on the rate calculated from the USA samples, the sub-lineages diverged 315 years ago (95% CI: 139-Inf years), in close agreement with our Sig5' estimate. This estimate lacks an upper bound due to the small number of USA samples and higher deviation of SigS in comparison to Sig5'. However, we can be more confident in the stability of SigS mutation rate than Sig5' given the consistency in SigS between the sub-lineages.

Since SigS is specific to somatic mutations, we can use it to estimate how many of the mutations shared by all cancers are somatic mutations, and therefore estimate how long prior to the sub-lineage divergence the cancer first arose in the founder clam and began horizontal transmission. SigS contributed roughly half of high confidence somatic mutations in each sub-lineage but was virtually absent from SNVs in the healthy clam population (**Fig. 1.2E)**. If we assume the SigS mutation rate has remained constant since oncogenesis and that the founder clam SNVs have a similar profile of genomic SNVs to those observed in healthy clams, we estimate that 3.1% of heterozygous SNVs found in all cancer samples, but no healthy samples, are somatic mutations attributed to SigS. This corresponds to 108 years by the SigS rate estimate

above, for a total cancer age estimate of 423 years (95% CI: 187-Inf years) (**Fig. 1.2F**), long before the first recorded observations of disseminated neoplasia in soft-shell clams in the 1970s [19,20].

If we also assume the fraction of SigS somatic mutations has remained constant since oncogenesis, we estimate that, in addition to the 3.1% SigS SNVs estimated above, approximately 3.7% (95% CI: 3.4-4.0%) of heterozygous SNVs found in all cancer samples, but no healthy clams, are somatic mutations due to the other three signatures. Combining this estimate (117k mutations) with sub-lineage-specific mutations (321k and 331k) we calculate a total somatic SNV estimate of 441 and 452 mu/Mb for the PEI and USA sub-lineages, respectively. This is a much higher mutation density than that estimated for the <40-year-old DFTD lineages (DFT1: <3.1 mu/Mb, DFT2: <1.3 mu/Mb) [15], but less than the >4000-year-old CTVT (~867 mu/Mb from exome data) [12], showing that mutation density generally scales with age across the small number of characterized transmissible cancer lineages.

1.3.4 Selection on SNVs is largely neutral

We used the ratio of non-synonymous to synonymous coding changes (dN/dS) to infer selection acting on coding regions in our sample set. After correcting for mutational opportunities in coding regions, a ratio of one indicates neutral selection, >1 indicates positive selection, and <1 indicates negative/purifying selection. We used dNdScv[38] to determine that the global dN/dS for healthy clam SNVs was 0.454 (95% CI: 0.451-0.457), indicating that genes are generally under negative selection in clam genomes, as expected. On a gene-by-gene basis, 70% of intact coding genes (16,222/23,273) in healthy clams have significantly negative dN/dS, while 0.4% (88/23,273) are significantly positive. Genes under positive selection in hosts may be those at the host-pathogen interface that are under selection for continued nonsynonymous mutation. In the case of clams, some of these genes may be a response to MarBTN evolution itself, though this hypothesis cannot be tested by the current study.

High confidence somatic mutations had a global dN/dS of 0.982 (95% CI: 0.943-1.024), indicating that MarBTN is largely dominated by neutral selection, reflecting observations in human cancers [39] and CTVT [12] **(Supplementary Fig. 1.10)**. We found no genes with a dN/dS ratio significantly <1, indicating no

genes are under significant negative (or purifying) selection, but we did identify five genes with a dN/dS ratio significantly >1, indicating positive selection (**Supplementary Table 1.3**). For all five of these genes, nearly all somatic mutations were found in a single sub-lineage. Only one of these genes has a dN/dS ratio above one in healthy clams, suggesting that four of five genes are truly under positive selection in only a single sub-lineage and they are not founder or host clam SNVs. The only characterized gene among the four is a *TEN1-like* gene that is under positive selection in the USA sub-lineage. TEN1 is a component of the CTC1-STN1-TEN1 complex, which plays a crucial role in telomere replication and genome stability [40].

1.3.5 Widespread structural mutation

Polyploidy has been described in disseminated neoplasia in several bivalve species [27,41]. In *Mya arenaria*, disseminated neoplasia cells have approximately double the chromosome count and genome content of healthy clam cells [24]. Given the discovery that these cells are of clonal origin [6], we had hypothesized that a full genome duplication occurred early in the cancer's evolution and that most of the MarBTN genome should be 4N. To test this theory, we called copy number states across each non-reference sample genome based on read depth **(Fig. 1.3A)**. As expected, both healthy clams were 2N across nearly the entire genome **(Fig. 1.3B)**. Surprisingly, MarBTN samples displayed a wide variety of copy number states.

PEI samples were predominantly 4N with substantial 3N and 2N portions, while USA samples were more evenly distributed between 4N, 3N, and 2N **(Fig. 1.3B)**. Copy number calls in cancer samples displayed close agreement within sub-lineages (R^2>0.94). There was a positive correlation between copy number calls between the two sub-lineages, but large differences could be observed suggesting that copy number changes have occurred since sub-lineage divergence (R^2=0.53-0.56) **(Supplementary Fig. 1.11)**. Variant allele frequencies (VAF) for high confidence somatic mutations largely support copy number calls **(Supplementary Fig. 1.12)**, with some off-target VAF peaks, most notably in the lower copy number regions (<3N), indicating that some of these regions are higher copy number than called through this method and appeared lower likely due to reduced read mapping in polymorphic genome regions.

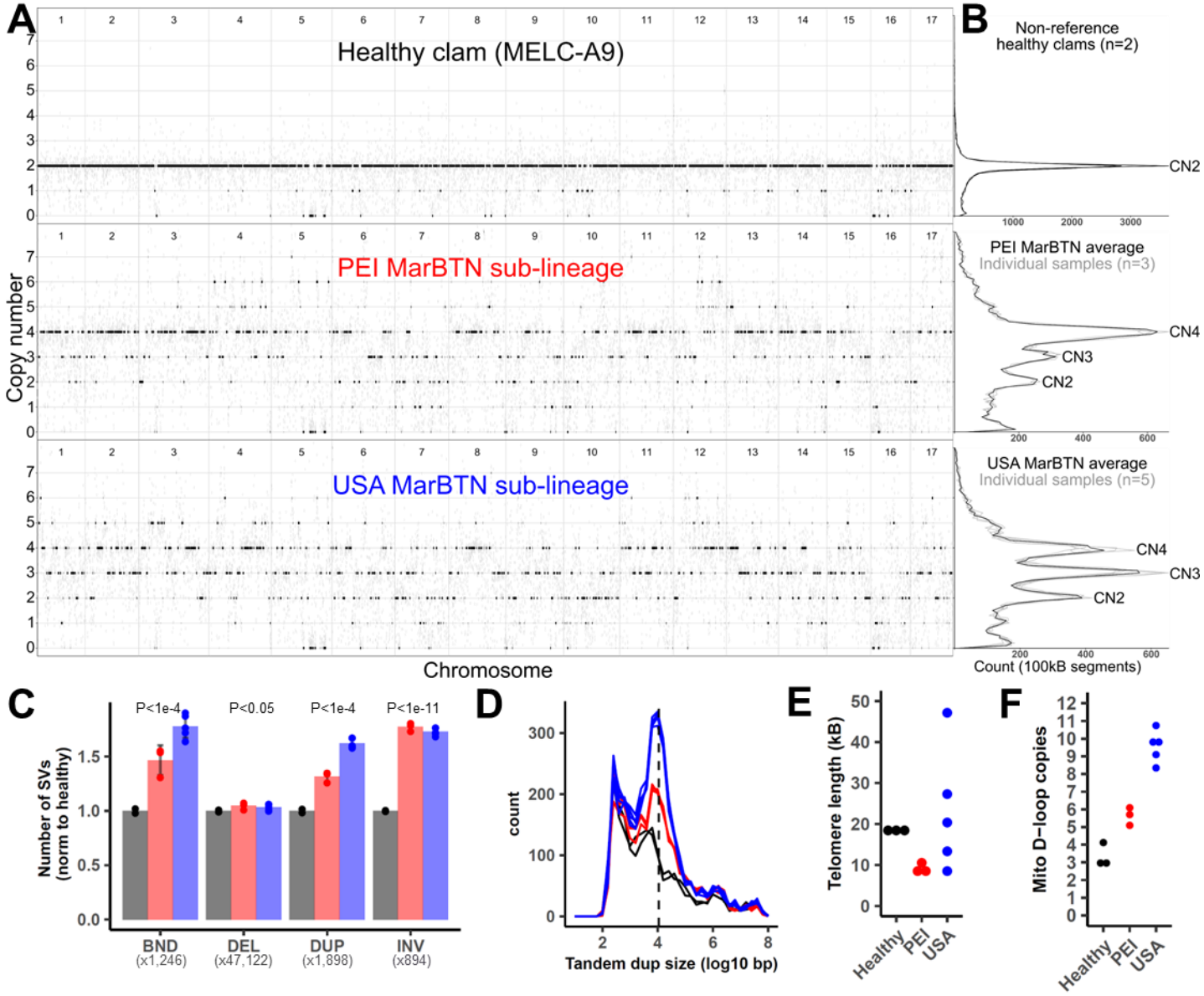


Figure 1.3. Widespread copy number gain and structural mutation

(A) Copy number calls across clam genome, rounded to the nearest integer (black) and unrounded (grey) in 100 kB segments. The healthy clam is a representative individual, and the MarBTN sub-lineages are averages of each individual sample from that sub-lineage, which were in close agreement. **(B)** Summary of copy number states across entire genomes for two non-reference healthy clams and MarBTN sub-lineages. Grey lines display copy number summaries for individual samples within each sub-lineage, which are in close agreement. **(C)** Number of SVs in each sample. Reference clam was excluded, since one haplotype from that animal was used to build the reference genome and thus does not contain SVs. Values are normalized to the average number of SVs in non-reference healthy clams for each SV type (numbers below SV type labels). P-values are from two-sided unequal variance t-test between BTN samples (n=8) and non-reference healthy clams (n=2). Labels follow delly abbreviations of SV types: BND = translocations, DEL = deletions, DUP = tandem duplications, INV = Inversions. Error bars indicate standard deviation. **(D)** Size distribution of tandem duplications in each non-reference sample. Dashed line indicates 11 kB. **(E)** Telomere length estimated by TelSeq for each sample. **(F)** Tandem duplicate copies of the mitochondrial D-loop region per sample. Healthy normal clams are black, MarBTN from PEI are red, and MarBTN samples from USA are blue.

To estimate timing of duplication events we looked at VAF in regions called CN4 across both sub-lineages (14% of the genome, **Supplementary Fig. 1.13**). While the majority of founder variants were distributed around a VAF of 0.5 (2/4 alleles) in both sub-lineages, as expected for a CN2>CN4 duplication, USA also had VAF distributions around 1/4 and 3/4 that were absent in PEI, indicative of CN2>CN3>CN4 duplication where not all haplotypes duplicated evenly. Additionally, we observe more 2/4 high confidence somatic mutations in PEI than USA, indicative of later duplication events. The fraction of 2/4 somatic mutations in the USA sub-lineage was low in nearly all CN4 segments of the genome, indicating most segments duplicated before or shortly after the USA-PEI sub-lineage split, with a low rate of duplications occurring after that time. In contrast, many segments in PEI sub-lineage have around 20% of the mutations at 2/4, suggesting a burst of duplications at some point after the USA-PEI sub-lineage split. Overall, these frequencies indicate the USA and PEI sub-lineages arrived at CN4 largely via independent duplication events, rather than the assumed single whole genome duplication, and that duplication events have occurred at multiple points throughout MarBTN evolution.

Many mid-chromosome breakpoints were apparent in the copy number calls, indicating that the MarBTN genome has likely undergone widespread structural alterations in addition to whole-chromosome and within-chromosome copy number gain. We are unable to resolve the structure of the MarBTN genome with the short sequence reads in our current data set but were able to call likely structural variants (SVs) from split reads. Relative to non-reference healthy clams, MarBTN samples had a significantly higher number of deletions, inversions, tandem duplications, and inter-chromosomal translocations, indicating substantial somatic structural alterations **(Fig. 1.3C).**

Comparing likely somatic structural variants specific to each sub-lineage, USA samples had significantly more translocations and tandem duplications than PEI **(Supplementary Fig. 1.14)**. Median somatic tandem duplication sizes displayed a distinct distribution around a mode of ~11 kB **(Fig. 1.3D, Supplementary Fig. 1.15)**. In human cancers, tandem duplication phenotypes of this same size distribution are thought to be driven by the loss of *TP53* and *BRCA1* [42], indicating that a parallel mutational process may be influencing the observed genome instability in MarBTN and more active in the USA sub-lineage.

Maintenance of telomere length is a requirement for an immortalized cell line such as MarBTN and would be necessary for long-term survival. We estimated telomere lengths for each sample and found them to be highly variable within the USA sub-lineage (8-47 kB), while short but relatively stable within the PEI sub-lineage (8-11 kB) compared to healthy clams (18-19 kB) **(Fig. 1.3E).** Variable telomere lengths in the USA sub-lineage may relate to the *TEN1-like* gene that is under positive selection in that sub-lineage, as the CTC1-STN1-TEN1 complex inhibits telomerase and is involved in telomere length homeostasis [40].

1.3.6 Mitochondrial genome evolution

A tree built from pairwise mitochondrial SNV differences between samples reflects a similar phylogeny to that built from genomic SNVs **(Supplementary Fig. 1.16)**. This indicates no evidence of mitochondrial uptake or recombination with host mitochondria, which has been observed in other transmissible cancers [8,43,44]. Transitions were highly overrepresented in both healthy and cancer samples, with C>T mutations comprising 41/50 likely somatic mutations **(Supplementary Fig. 1.17)**. Somatic mutations resulted in missense mutations in at least 10 of the 12 mitochondrial genes and appear to be under relaxed selection, with dN/dS ratios of 0.97 (95%CI: 0.45-2.1) versus 0.26 (95%CI: 0.11-0.58) for SNVs in healthy clams **(Supplementary Fig. 1.18)**.

When aligned to the published *Mya arenaria* mitochondrial genome [45], short read sequences from all MarBTN and healthy samples display increased coverage across the mitochondrial D-loop (**Supplementary Fig. 1.19)**, indicating the region is multi-copy. The D-loop is part of the non-coding control region of the mitochondrial genome and is the origin of both replication and transcription. We resolved this region with PacBio long reads from the healthy reference clam, revealing three copies in tandem. Two of the copies contain a 236 bp insertion not found in the published mitochondrial genome. The insert includes an 80 bp region with 70% guanine content, likely complicating previous PCR-based efforts to resolve it. Altogether, the observed copies extend the D-loop region of the reference clam genome from 845 bp to 2,727 bp and the full mitochondrial genome to 19,815 bp.

Read coverage of the D-loop region suggest that there have been additional somatic tandem duplications in the MarBTN mitogenome. While read coverage indicates 3-4 copies in the non-reference healthy clams, PEI MarBTN samples have 5-6 copies and USA MarBTN samples have 8-11 (**Fig. 1.3F)**. These somatic tandem duplications likely arose via replication errors and the trend towards increased copies in cancer suggests that they may be under selection. Selection can act on the level of the mitogenome itself, giving it a replicative advantage over other mitogenomes (as hypothesized for CTVT), or on the level of the cancer cell, if this duplication provides cancer cells a replicative advantage over others. Notably, the mitogenome site suspected to be under selection during repeated mitochondrial capture in CTVT is in the control region [44], the same region we see amplified in MarBTN.

1.3.7 Transposable element mobilization

MarBTN is known to contain the LTR-retrotransposon *Steamer* at a much higher copy number than healthy clams, indicating likely somatic expansion [46]. To test whether *Steamer* activity is ongoing we identified *Steamer* insertion sites using split reads spanning *Steamer* and the reference genome. Only 5-11 sites were found in each healthy sample, versus 275-460 sites in each cancer sample. One hundred ninety-three sites are shared by all cancer samples, indicating that *Steamer* expansion likely began early in the cancer's evolution, while sub-lineage-specific *Steamer* integrations indicate that *Steamer* has continued to replicate somatically in the MarBTN genome (**Fig. 1.4A**). However, *Steamer* has generated more insertions within the USA sub-lineage (248) than the PEI sub-lineage (64), indicating the regulatory environments of the sub-lineages have not remained stable since they diverged.

We also observed strong biases for *Steamer* to insert at specific genomic sequences. *Steamer* has a palindromic bias for NATG outside the five bp target site duplication (CATNnnnnnNATG), inserting at these locations 45× more frequently than expected by chance (**Fig. 1.4B).** *Steamer* was also >3× more likely to insert within 1000 bp upstream of genes than would be expected by chance (**Fig. 1.4C**). We also observed early *Steamer* insertions (those found in all MarBTN samples) upstream of cancer-associated orthologs more often than expected by chance in the reverse but not the forward orientation **(Supplementary Fig.**

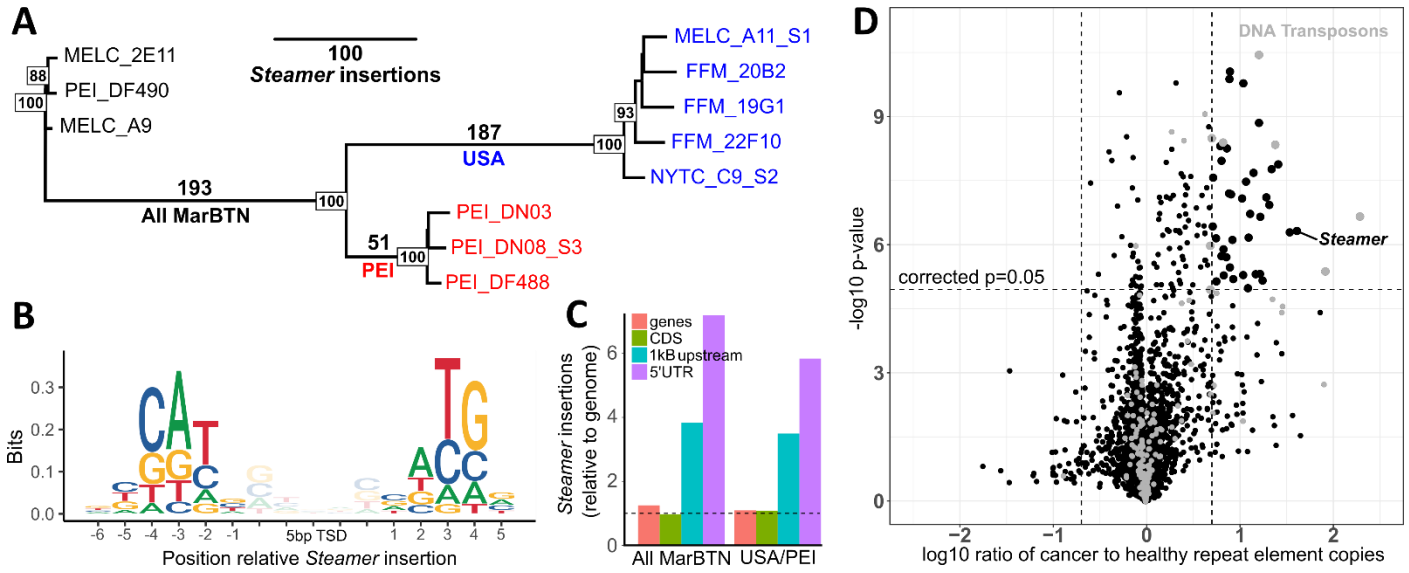


Figure 1.4. Somatic expansions of *Steamer* and other TEs

(A) Phylogeny of all samples built from pairwise differences of *Steamer* insertion sites, colored by healthy (black), USA MarBTN (blue), and PEI MarBTN (red). Numbers along branches indicate the number of insertions unique to and shared by individuals in that clade, numbers on nodes indicate bootstrap support, with bootstrap values below 75 not shown. **(B)** Logo plot of insertion bias relative to the 5 bp target site duplication (TSD) of all *Steamer* insertions, normalized by nucleotide content of the genome. **(C)** *Steamer* insertion probability in annotated genome regions, normalized by read mapping rates and relative to full genome. Displayed for insertions found in all MarBTN samples but no healthy clams, and unique to each sub-lineage but shared by all individual in that sub-lineage. Dashed line indicates expectation given random insertions. **(D)** Volcano plot comparing copy number of all repeat elements in MarBTN and healthy clam samples. Dashed lines correspond to significance threshold (p=0.05, Bonferroni corrected) and 5-fold differences. Elements annotated as DNA transposons are marked in grey.

1.20, Supplementary Table 1.4). This bias that could indicate either an insertion preference for those locations or a selective advantage to MarBTN cells associated with those insertions.

We further investigated whether other transposable elements (TEs) in addition to *Steamer* have expanded somatically by identifying a library of repeat sequences (putative TEs) found in clam genomes and counting the copy number of each TE type in each sample. Forty-five TEs were present at a significantly higher copy number in cancer samples relative to healthy clams after removing TEs with less than five-fold differences (**Fig. 1.4D)**. TEs annotated as DNA transposons were enriched in this data set (8/45: 17.8%) compared to the total TE library (171/4471: 3.8%), indicating this TE type may have been particularly successful in somatically expanding its copy number in MarBTN. LTR retrotransposons (like *Steamer*) appear to have had more success in the USA versus PEI sub-lineage. Thirty-six TEs have significantly more copies in the USA sub-lineage than PEI, and eight of those are LTR-retrotransposons, compared to zero

LTR-retrotransposons out of 20 of those more highly expanded in PEI **(Supplementary Fig. 1.21)**. Reduced copy numbers of LTR retrotransposons and other TEs in the PEI sub-lineage could be linked to the increased methylation indicated by mutational signature analysis, as methylation is thought to repress TE mobilization [30,47]. Our finding of widespread increases in TE copy numbers alongside structural mutations indicate general genome instability of the MarBTN lineage and provides further evidence of a higher rate of certain mutation types in the USA sub-lineage, which cannot be explained by the temporal distribution of the samples alone (**Supplementary Fig. 1.22**).

1.3.8 MarBTN gene expression

To investigate the role of genes implicated in MarBTN evolution we sequenced RNA from a new set of five MarBTN isolates from the USA sub-lineage, six tissues (hemocytes, foot, gill, adductor muscle, mantle, and siphon) across three healthy clams, and hemocytes from an additional two clams (**Supplementary Table 1.5**). Both principal component analysis and hierarchical clustering clearly separate MarBTN and hemocytes from all solid tissue samples (**Fig. 1.5A, Supplementary Fig. 1.23**), indicating MarBTN likely originated as a hemocyte. This origin has been hypothesized due to MarBTN being most obviously detectable in the hemolymph[6,48,49], but had not previously been tested.

MarBTN-specific SigS resembles an error-prone polymerase signature in humans, so we first compared the expression of the 28 polymerase genes identified in the clam genome. We observed widespread up-regulation across polymerases in MarBTN (**Fig. 1.5B, Supplementary Fig. 1.24**), likely facilitating increased cellular replication and/or DNA damage repair. The most highly up-regulated polymerase is homologous to polymerase Nu (*POLN*), a very low fidelity polymerase that plays a role in translesion synthesis and cross-link repair by homologous recombination[50,51]. Polymerase Nu frequently mis-incorporates dT opposite a template dG in humans[52,53], a bias which does not match SigS. However, given the distance between bivalves and humans, it is possible that this polymerase introduces different biases in clams and is in part responsible for the observed SigS biases and/or genome instability.

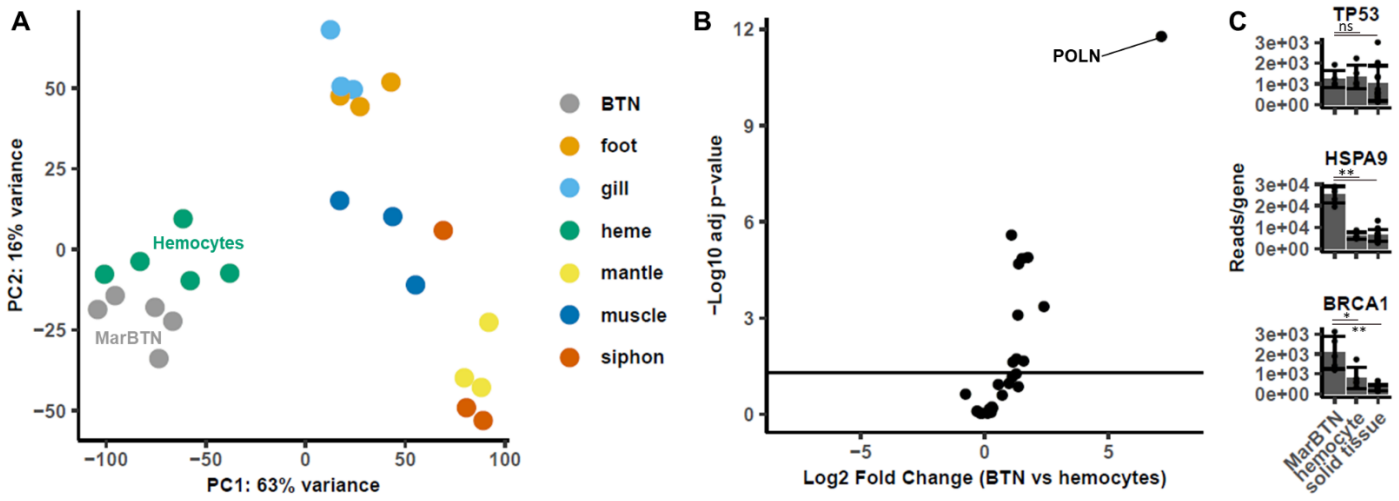


Figure 1.5. Expression indicates hemocyte origin and possible mutagenic pathways in MarBTN

(A) Principal component analysis of normalized expression across all genes, with PC1 separating MarBTN and hemocytes from all other tissues. **(B)** Volcano plot of polymerase genes expression (n=28) for MarBTN (n=5) compared with hemocytes (n=5). **(C)** Normalized expression, in reads per gene, of *TP53*, *HSPA9* (mortalin) and *BRCA1* for MarBTN (n=5), hemocytes (n=5), and non-hemocyte tissues (n=15: 5 tissues for 3 clams). Error bars display standard deviation, differential expression comparison results displayed as * = p<0.05, ** = p<1e-5, ns = not significant.

We next looked at the expression of four genes under putative positive selection as identified by dN/dS (**Supplementary Fig. 1.25**). Positive selection in cancer can indicate repeated selection for either loss-of-function or gain-of-function mutations. Two genes were not expressed in MarBTN, including the *TEN1-like* gene, indicating potential loss-of-function, while two genes were up-regulated in MarBTN versus healthy hemocytes, indicating potential gain-of-function.

Finally, we investigated genes implicated by the distinct ~11kB tandem duplication phenotype; *TP53* and *BRCA1*. Prior work identified the deactivation of p53 via cytoplasmic sequestration by overexpressed mortalin[54], so we investigated the expression of genes homologous to *TP53* and mortalin-encoding *HSPA9* (**Fig. 1.5C**). Indeed, while *TP53* had no nonsynonymous MarBTN mutations and was not differentially regulated, *HSPA9* was significantly upregulated in MarBTN samples, supporting the proposed model of inactivation by mortalin sequestration. Similarly, clam *BRCA1* homolog has no obvious loss of function mutations (three missense SNVs were observed in all MarBTN samples, which do not correspond

to known loss-of-function mutations and could be either somatic mutations or inherited founder variants). The tandem duplicator phenotype was reported to be strongly associated with loss of function of BRCA1[42] in humans, but in MarBTN *BRCA1* was upregulated (**Fig. 1.5C**). We speculate that either: A) BRCA1 is rendered non-functional by some other mechanism (similar to p53); B) it is functionally overwhelmed by genome instability over the long timescale of this cancer lineage, resulting in a similar phenotype to loss-of-function; and/or C) a different pathway in bivalves is responsible for the tandem duplication phenotype; although we are unable to test these hypotheses in this study. Overall, MarBTN gene expression illuminates possible mechanisms behind the lineage's observed genome instability, though much remains unknown about the forces generating and tolerating such widespread genomic alterations.

1.4 Discussion

Our genome analyses reveal a diverse set of somatic mutations occurring in MarBTN, with continued accumulation of SNVs and widespread structural mutations indicative of genome instability. It is unclear whether these mutations have consistently occurred over time or have been generated in multiple punctuated chromothripsis-like events, but the continued accumulation of these changes between the sub-lineages shows that this instability was not confined to a single ancestral event. Genomic studies of the dog and devil transmissible cancers have observed contrastingly stable genomes, remaining predominantly diploid despite thousands of years of evolution in the case of CTVT [11,55]. Polyploidy has been reported in other BTN lineages in other bivalves [27], indicating that genome instability may be a common driver mechanism or a tolerated by-product of conserved processes in BTN evolution. Interestingly, while there is ongoing instability in both sub-lineages, we observe differences in the number of structural mutations, duplication timing, telomere length, and TE amplification between the two sub-lineages, suggesting that genome instability or mutation tolerance may have changed over time in MarBTN after the sub-lineages diverged. These changes in fundamental mutational mechanisms observed in distinct sub-lineages post-divergence highlight the fact that oncogenesis is not a single event, but an ongoing evolutionary process.

In contrast to the above unstable and variable structural processes, we observe a pattern of consistent single nucleotide mutation biases in both sub-lineages. Most notable is the novel profile and consistent accumulation of mutational signature S. We hypothesize that this signature is due in part to an upregulated error-prone polymerase and that its consistent accumulation may be due to consistent MarBTN replication rates over time, as seen in human somatic cells with defective proofreading polymerases [56], or due to continual damage of chromosomal DNA and its repair using translesion synthesis. Both SigS and Sig5' (analogous to the clock-like signature 5 in humans) generate consistent age estimates for the most recent common ancestor of our sample set and estimate the cancer is at least 200 years old, though uncertainty in the calculated mutation rates means the actual age of the cancer could be far greater. This indicates MarBTN is likely an intermediate age compared with DFTD (<40 years [16]) and CTVT (4000-8500 years [12]).

We observe that the MarBTN genome is largely dominated by neutral selection, reflecting observations in human cancers [39] and CTVT [12], with a few notable genes under positive selection in a single sub-lineage, which may reflect selection for repeated mutations involved in critical oncogenic processes. However, we also note that selection is not simply relevant at the level of the cancer cell, but also on the level of the gene (as seen in MarBTN TE expansions), mitochondria (as seen in CTVT horizontal transfer [44] and MarBTN mitogenome expansion), and hosts (as seen in DFTD [57]). Further analysis of MarBTN and other cancers will help us to understand how these selective forces interact to influence cancer evolution, and perhaps how we can manipulate those forces to our advantage to combat conventional and transmissible cancers.

Our analysis of the MarBTN genome is presented simultaneously to an independent analysis of two lineages in the common cockle (*Cerastoderma edule*), or CedBTN, by Bruzos and colleagues [58]. CedBTN infection presents as a similar leukemia-like disseminated neoplasia phenotype to MarBTN and gene expression points toward a hemocyte origin for BTN in both species. The CedBTN genomes display signatures of ongoing instability like MarBTN, supporting the hypothesis that genome instability is a common feature of BTN evolution and confirming that long-term survival of a cancer lineage can be

maintained despite remarkably widespread and continued genome rearrangement. This level of instability might be expected to lead to an error catastrophe [59], yet these cancers have continued to replicate for centuries, changing our understanding of what is possible in cancer evolution. Tandem duplications in the mitochondrial control region were also observed in both studies and may represent convergent evolution driven the same selective mechanisms. Similar tandem duplications in the D-loop have also been observed in human cancers [60,61], though the functional consequences of these mutations remains unclear. Repeated expansion of this region in independent BTN lineages, along with their long history of co-evolution with their hosts, make BTNs unique model systems for the understanding of the functional significance of mitogenome mutations on cancer cell growth and the potential for selfish selection at the level of the mitogenome in cancer.

In contrast, our finding of a distinctive polymerase-associated mutational signature, evidence of positive selection, variable telomere length, and amplification of the *Steamer* retrotransposon and other TEs may be unique features of the BTN in clams. We find no evidence of mitochondrial genome transfer events or host co-infection by multiple clones as observed in cockles, though this may be due to the smaller sample size of our study and the low level of polymorphisms in mtDNA in soft-shell clams. Given the apparent abundance of BTNs, continuing to analyze BTN lineages in other species may reveal both common and unique pathways that have allowed these cancers to repeatedly circumvent new host immune systems and spread through host populations as contagious cancers. These cancers therefore provide unique models for the understanding of cancer evolution and exemplify what genomic changes are possible in long-lived cancers evolving together with their hosts.

1.5 Acknowledgements

We thank Carol Reinisch and James Sherry for sample collection, Charles Walker for advice and aid during the initiation of this project, Phase Genomics (Shawn Sullivan, Emily Reister, Kyle Langford, and Hayley Mandelson) for HiC scaffolding, Andrew Banman and Nikita Sakhanenko for in-house

computing support at PNRI, Sam White and Steven Roberts for consultation on the use of MAKER for gene annotation, Kelley Harris and Jed Carlson for consultation on mutational signatures, Adrian Baez-Ortega for consultation on somatypus and sigfit, and Claudia Carvalho for consultation on structural variants.

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), Intramural Program of the National Human Genome Research Institute (supporting E.A.O.), NSF EEID grant 2208081 (to M.J.M.), and ANID/ACE/210011 (to G.A.).

1.6 Author contributions

S.F.M.H, M.J.M. and S.P.G. contributed to study conceptualization and design. M.J.M., B.F.B., G.A., and S.P.G. contributed to sample collection. M.A.Y. performed high molecular weight extractions. F.E.S.G. and M.A.Y. performed tissue dissections. B.W.D, E.A.O., and M.J.M. contributed to 10X sequencing and analysis. M.J.M. assembled the reference genome. R.M.G. and S.F.M.H. contributed to disseminated neoplasia literature search. S.F.M.H. performed the data analysis. S.F.M.H. wrote the original draft of the manuscript. S.F.M.H, M.A.Y., R.M.G, B.F.B., G.A., B.W.D., E.A.O., S.P.G., and M.J.M. contributed to review and editing of the manuscript.

1.7 Methods

1.7.1 Data availability

All code is available on GitHub (https://github.com/sfhart33/MarBTNgenome), including all dependencies with version numbers. Individual commands for genome assembly are listed below with triangular bullets, and scripts corresponding to written genome analysis methods are listed in bullets at the end of each written section. 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.

Raw sequence data and the assembled genome 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.

1.7.2 *Mya arenaria* genome assembly

1.7.2a Reference animal collection and HMW DNA extraction

Due to the high rate of heterozygosity in bivalves, a single clam was chosen to be the source of all DNA used in the generation of the reference genome, and a diploid phased assembly strategy was used. The clam chosen as the reference animal (MELC-2E11, 62 mm shell length, **Fig. 1B**) was collected from Larrabee Cove, Machiasport, Maine, USA in June 2018, and shipped to the Pacific Northwest Research Institute labs. Hemolymph was drawn from the animal from the pericardial sinus using a 0.5 in 26 gauge needle on a 3 ml syringe, and it was checked for the presence of MarBTN through morphological analysis (**Fig. 1C**) and with a highly sensitive cancer-specific qPCR assay (described in [17]). There was no evidence of detectible BTN through either method. Examination of gonad region revealed the presence of eggs, showing that this individual was female.

1.7.2b High Molecular Weight DNA extraction for PacBio sequencing

High molecular weight (HMW) DNA, used for PacBio sequencing, was extracted from ~50 mg snap-frozen mantle tissue using a modified CTAB extraction protocol (adapted from [62]). CTAB isolation buffer (2% CTAB, 1.4 M NaCl, 20 mM EDTA, 100 mM Tris-HCl, pH 8.0) was preheated to 60°C in a water bath. Tissue was minced, then ground with a pestle in 500 µL 60°C CTAB isolation buffer in a 1.7 ml microcentrifuge tube. 20 µL proteinase K was added and the sample was incubated at 60°C for 10 h on a shaker (200 rpm), then held at room temperature. Sample was extracted once with the addition of 500 µL chloroform-isoamyl alcohol (24:1), mixing gently but thoroughly. This produces two phases, an upper aqueous phase which contains the DNA, and a lower chloroform phase that contains some degraded proteins, lipids, and many secondary compounds. The sample was spun at 6,000 × g for 10 min at room temperature to concentrate phases. Aqueous phase was removed with a wide bore pipet, transferred to a

new microcentrifuge tube. 2/3 volumes cold isopropanol (237 µL) was added and inverted gently to precipitate nucleic acids. HMW DNA was spooled out with a glass hook and transferred to a 2 mL microcentrifuge tube containing 1 mL wash buffer (76% ethanol, 10 mM ammonium acetate) for 20 minutes. HMW DNA was spun down (6,000 × g for 10 min) after a minimum of 20 min of washing. Supernatant was poured off carefully and allowed to air dry briefly at room temperature. HMW DNA was resuspended in 200 µL TE (10 mM Tris-HCl, 1 mM EDTA, pH 8.0). RNase A (DNase-free, 10 mg/mL, Thermo Scientific, Waltham, MA) was added to a final concentration of 25 ug/ml (0.5µL) and incubated 30 min at 37°C. Sample was diluted to 2 volumes with TE, 10 M ammonium acetate was added to a final concentration of 2.5 M, sample was mixed, 1.2 mL 100% ethanol was added, and sample was gently inverted to precipitate HMW DNA. HMW was spun down (10,000 × g for 10 min at 4°C). Sample was air dried and resuspended in 200 µL TE buffer overnight at 4°C.

1.7.2c 10X Chromium sequencing

High molecular weight genomic DNA was isolated from reference animal (MELC-2E11) tissue using the MagAttract HMW DNA Kit (Qiagen), quantified using Qubit 2.0 (Life Technologies) and fragment size determined using the Agilent 2200 TapeStation. Average fragment size exceeded 50 Kb. Approximately 1 ng of DNA was loaded on the Chromium Genome Chip (10X Genomics). Whole genome sequencing libraries were prepared using Chromium Genome Library & Gel Bead Kit v.2, Chromium i7 Multiplex Kit and Chromium Controller according to 10X Genomics instructions. The resulting library was indexed and sequenced on 0.75 lanes of a single flow cell on the Illumina HiSeq X Ten system, generating 150-bp paired-end reads.

1.7.2d Unsuccessful assembly using 10X Chromium data

An attempt to assemble a genome of the MELC-2E11 reference animal using 10X sequencing was unsuccessful at creating a highly contiguous genome, likely due to the higher amount of repeats in the bivalve genome than in the human genome. However, we report it here for transparency of our assembly

attempts. *De novo* assembly was performed using Supernova (v2.1.1). Assembly was conducted with the command:

- supernova run --id=10X_MELC-2E11-100 --fastqs=/home/mmetzger/10XChromiumSized/Data --description="MELC-2E11 sized 10X assembly " --maxreads=all --accept-extreme-coverage

Secondary attempts at assembly were conducted using down-sampled subsets of the data (75%, 666928259 reads; 50%, 444618839 reads, and 25%, 222309414 reads), using the read numbers listed above for the option "--maxreads". The pseudohap2 output was used (Mar.3.1.1 Myaare100B_pseudohap2.1.fasta).

1.7.2e Successful FALCON-Unzip diploid assembly

HMW DNA extracted using the CTAB protocol was sequenced using the PacBio core facility at the University of Washington Department of Genome Sciences. Subreads were converted to fasta using:

- bam2fasta -o Marenaria.3.2 *.subreads.bam

Due to the high heterozygosity in bivalve genomes, the FALCON-Unzip pipeline was run to generate a diploid-aware *de novo* assembly. The resulting assembly can be expressed as either as two pseudo-haploid reference genomes or as a primary assembly with alternate "haplotigs" in genomic regions where the two copies of the diploid genome in the reference individual differ. This was done using the commands:

- fc_run fc_run_marenaria.cfg &> run1.log &
- mv all.log all0.log
- fc_unzip.py fc_unzip_marenaria.cfg &> run1.std &

Several modifications were made to default configuration parameters, including changing to "pwatcher_type=blocking" and lowering the memory per job in fc_unzip_marenaria.cfg. Configuration files are found on github (fc_run_marenaria.cfg and fc_unzip_marenaria.cfg). FALCON-Unzip version was pbbioconda-0.0.5 and was used with python 3.7. This FALCON-Unzip pipeline resulted in a primary contig assembly (Mar.3.2.2_cns_p_ctg.fasta) and an alternate haplotig assembly (Mar.3.2.2_cns_h_ctg.fasta).

High heterozygosity in species such as bivalves can lead to under-calling of haplotype homology. The purge_haplotigs pipeline[63] was used to remove pairs of contigs that were called as separate primary

contigs by FALCON-Unzip but which are more likely to be alternate alleles. This tool identifies pairs of syntenic contigs and moves one from the primary assembly to the haplotig assembly generating a new curated assembly (Mar.3.2.3_curated.FALC.fasta).

1.7.2f Scaffolding with Hi-C using FALCON-Phase

Chromatin conformation capture data was generated using a Phase Genomics (Seattle, WA) Proximo Hi-C Animal Kit, which is a commercially available version of the Hi-C protocol[64]. Following the manufacturer's instructions, intact cells from two adductor muscle samples from the same reference clam (MELC-2E11) were crosslinked using a formaldehyde solution, digested using the Sau3AI restriction enzyme, and proximity ligated with biotinylated nucleotides to create chimeric molecules composed of fragments from different regions of the genome that were physically proximal *in vivo*, but not necessarily genomically proximal. Molecules were pulled down with streptavidin beads and processed into an Illumina-compatible sequencing library. Sequencing was performed on an Illumina NextSeq 500, generating a total of 313,340,002 PE150 read pairs.

The Hi-C reads, primary contigs, and alternate haplotigs (Mar.3.2.3_curated.haplotigs.FALC.fasta) were provided as input to FALCON-Phase (https://phasegenomics.github.io/2019/09/19/hic-alignment-and-qc.html) to correct likely phase switching errors. All other options were set to default, except for the options which specify restriction enzyme motifs in the library (GATC) and the number of iterations to perform (100,000,000). Phased contigs were output in pseudohap format, creating one complete set of contigs for each of the two phased assemblies from the diploid genome of the reference individual (arbitrarily named Phase 0 and Phase 1).

Reads were aligned to the resulting Phase 0 contig assembly 25etwork_mussel.phased.0.fasta following Phase Genomics' standard Hi-C alignment protocol [65]. Briefly, reads were aligned using BWA-MEM[66] with the -5SP and -t 8 options specified, and all other options default. SAMBLASTER [67] was used to flag PCR duplicates, which were later excluded from analysis. Alignments were then filtered with samtools [68] using the -F 2304 filtering flag to remove non-primary and secondary alignments. These

alignments, along with the primary contigs and alternate haplotigs were used as inputs to the scaffolding process.

Phase Genomics' Proximo Hi-C genome scaffolding platform was used to create chromosome-scale scaffolds from the Phase 0 assembly, following the same single-phase scaffolding procedure described in Bickhart et al. [69]. As in the LACHESIS method [70], this process computes a contact frequency matrix from the aligned Hi-C read pairs, normalized by the number of restriction sites (GATC) on each contig, and constructs scaffolds in such a way as to optimize expected contact frequency and other statistical patterns in Hi-C data. Approximately 120,000 separate Proximo runs were performed to optimize chromosome assignment and scaffold construction in order to make the scaffolds as concordant with the observed Hi-C data as possible. This process resulted in a set of 17 chromosome-scale scaffolds containing 1,212 Mbp of sequence (99.89% of the phase 0 assembly) with a scaffold N50 of 78.4 Mbp. Juicebox [71,72] was used to correct likely scaffolding errors, though no breaks for mis-joined contigs were introduced at this stage in order to maintain exact contig relationships with the Phase 1 assembly.

Separately, Hi-C data were aligned to a concatenated Phase 0 and Phase 1 assembly using the standard protocol cited above. Because this would cause Hi-C data for most homozygous regions to have a MAPQ of 0 (among possible other issues), this alignment emphasizes phase-specific Hi-C relationships. These alignments and the Phase 0 scaffolds were passed to FALCON-Phase's bamfilt (-f 20 -m 10), bam2 binmat (default options), and phase (-n 100000000 -s 10) steps to generate new phasing metadata intended to correct latent phasing issues not detected during the earlier contig phasing step.

Juicebox was again used to correct remaining scaffolding errors in Phase 0, including introducing a single break into each of eight suspected mis-joined contigs and two breaks into one suspected double-misjoined contig, based on the appearance of Hi-C signals consistent with chimeric joins. These scaffolding changes were replicated to Phase 1, and new scaffolds for each phase were generated using the juicebox_assembly_converter.py script (https://github.com/phasegenomics/juicebox_scripts). In these final scaffolds, both Phase 0 and Phase 1 included 17 scaffolds spanning 99.7% (1,204 Mbp in Phase 0 and 1,214 Mbp in Phase 1) of input with a scaffold N50 of 70.2 Mbp in Phase 0 and 71.4 Mbp in Phase 1

(Mar.3.3.2_p0_PGA_assembly.fasta and Mar.3.3.2_p1_PGA_assembly.fasta). The 17 scaffolds from both of these two haploid assemblies were compiled into Mar.3.3.2_p0p1_PGA_assembly_17.fasta using a custom perl script two_fasta_prefix_compile_firstX.pl

- perl two_fasta_prefix_compile_firstX.pl Mar.3.3.2_p0_PGA_assembly.fasta Mar.3.3.2_p1_PGA_assembly.fasta p0_ p1_ Mar.3.3.2_p0p1_PGA_assembly_17.fasta 17

1.7.2g Genome Gap-Filling and Polishing using long read PacBio data and 10X linked-read data

PBJelly was run to gap-fill the scaffolded assembly using pbsuite [73] (v15.8.24, slightly modified: https://github.com/esrice/PBJelly) using blasr (v5.1) and 27etwork (v2.2) with Python 2.7, with the protocol file Protocol_MELC.xml. Only captured gaps were filled (no inter-scaffold gaps) using the option "—capturedOnly" during the "support" step. PBJelly was run with the commands:

- Jelly.py setup Protocol_MELC.xml
- Jelly.py mapping Protocol_MELC.xml
- Jelly.py support Protocol_MELC.xml -x "—capturedOnly"
- Jelly.py extraction Protocol_MELC.xml
- Jelly.py assembly Protocol_MELC.xml -x "—nproc=20"
- Jelly.py output Protocol_MELC.xml

The output of PBJelly (Mar.3.3.3_jelly.out.fasta) renamed all scaffolds to Contig0-Contig33, so names were corrected manually based on PBJelly liftOverTable.json (Mar.3.3.3_jelly.out_name.fasta), and the two haploid genomes were separated (using commands listed in PBJellyRenaming.txt) to generate the haploid gap-filled assemblies (Mar.3.3.3.p0.fasta and Mar.3.3.3.p1.fasta)

Direct use of short reads to polish a highly heterozygous genome is likely to introduce more errors than it corrects, due to the mapping of reads from both haplotypes to a single haploid genome. Therefore, we used a phase-aware polishing strategy, using the 10X linked reads generate above, modified from the pipeline described in the vertebrate genome project (https://github.com/VGP/vgp-assembly/tree/master/pipeline/freebayes-polish). Both phases (p0 and p1) of the scaffolded, and gap-filled diploid genome were concatenated into a single diploid reference file with 34 scaffolds, and the linked-

read-aware mapper Longranger [74] (v2.2.2) was used to map the 10X reads to the diploid gap-filled assembly (Mar.3.3.3_jelly.out_name.fasta), and the output was indexed using samtools (v1.9). FreeBayes (v1.3.1) and Bcftools (v1.10.2) were used to call SNPs in reads that mapped uniquely to one location on one haplotype, under stringent conditions, using a Q>30 filtering of both the Longranger mapping calls and FreeBayes variant calls. For Bcftools, filters allowed only homozygous ALT alleles (GT="A"), as the reference assembly used was a concatenated diploid assembly instead of a haploid one. 1,862,877 variants were called. The resulting polished, concatenated diploid assembly (Mar.3.4.6.p0p1_Q30Q30A.fasta) was split into the two polished haploid genomes and renamed (Mar.3.4.6.p0_Q30Q30A.fasta and Mar.3.4.6.p1_Q30Q30A.fasta). Commands for running of polishing and renaming of the assembly are available (LongrangerFreeBayesBcftoolsPolishing.txt).

The Phase 0 and Phase 1 assembled, scaffolded, and polished haploid genomes represent the two genomes found in the diploid reference individual. One must be chosen to be the reference for mapping and analysis of other genomes, so Phase 1 was selected as the primary reference genome for annotation and further use, as it contained the first endogenous *Steamer* insertion site that was initially reported [46]. This site is polymorphic in *Mya arenaria* populations and was not present in Phase 0.

1.7.2h RNA extraction and transcriptome assembly

In order to create a transcriptome assembly that would include transcripts expressed in different tissue types across the clam, seven tissues from the reference animal (MELC-2E11) frozen at -80°C in RNAlater (Invitrogen, Waltham, MA) were used for RNA extraction (1, mantle; 2, foot; 3, siphon; 4, stomach; 5, adductor muscle; 6, gills; and 7, hemocytes). Solid tissues were homogenized with a disposable plastic mortar and pestle in liquid nitrogen before extraction with the Qiagen RNeasy kit (Qiagen, Hilden, Germany), eluting in 60 µL elution buffer. DNase I (2 µL, 2,000 U/ml, RNase-free, New England Biolabs, Ipswich, MA), 10× DNase buffer, and water was then added 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 (excluding the stomach due to

possible contamination with RNA from clam food) were then sequenced on a single Illumina HiSeq 4000 lane for 20-30 million reads per sample (Genewiz, Leipzig, Germany).

RNAseq reads from the six tissues were concatenated to create single files for each read direction and used to assemble a transcriptome using Trinity (v2.8.5) [75]:

- Trinity --seqType fq --max_memory 200G --CPU 16 --trimmomatic --full_cleanup --left MELC-2E11_R1_allfiles-cat.fastq.gz --right MELC-2E11_R2_allfiles-cat.fastq.gz

1.7.2i Genome annotation

Genome annotation was conducted using a strategy of first masking repeats, then annotating using a MAKER pipeline, incorporating transcriptome data (assembled above) and using homology to five previously annotated bivalve genomes, then generating putative gene identities based on homology using BLASTP. Repeat elements in the genome assembly were called using RepeatModeler (v2.0), and repeat elements were masked using RepeatMasker (v4.1.0) [76]:

- RepeatModeler -database Mar.3.4.6.p1_Q30Q30A -pa 20 -LTRStruct
- RepeatMasker -pa 20 -lib $i-families.fa Mar.3.4.6.p1_Q30Q30A.fasta

The genome was annotated using MAKER (v2.31.10), exonerate (v2.2.0), and RepeatMasker (v4.1.0) [76], with two rounds of SNAP training using custom scripts, following previously established methods [77]. The *Mya arenaria* transcriptome (as assembled above) was used as input into the MAKER annotation, along with the proteins identified from five well-annotated bivalve genomes (*Mytilus coruscus*, GCA_011752425.2; *Crassostrea virginica*, GCF_002022765.2; *Mizuhopecten yessoensis*, GCF_002113885.1; *Pecten maximus*, GCF_902652985.1; and *Crassostrea gigas*, GCF_902806645.1).

Putative gene identification was made by BLASTP search of the uniprot database (accessed 2021-03-02) and the proteins identified from the five well-annotated bivalve genomes concatenated into a single file (CgiCviMcoPmaMye_protein.fasta), using blast+ (v2.10.0). The top hit (with and e value $<1e^{-6}$) was used for gene identification. Genes were labeled based on the most similar uniprot hit (if applicable with an e value $<1e^{-6}$) and "-like" suffix or labeled as uncharacterized if only matching an uncharacterized bivalve gene or no gene at all. To account for multiple genes with the same uniprot hit, an additional numeric

suffix was added to indicate additional hits to the same uniprot gene or uncharacterized genes (e.g. (e.g. "TEN1-like_3", "uncharacterized_1199").

- wget ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz
- gunzip uniprot_sprot.fasta.gz
- makeblastdb -in uniprot_sprot.fasta -out uniprot_sprot -dbtype prot
- blastp -query /home/metzgerm/MAKER_Mya/2020-09-11-Mar.3.4.6.p1-MAKER/snap02/2020-09-11_Mar_genome_snap02.all.maker.proteins.fasta -db uniprot_sprot -evalue 1e-6 -max_hsps 1 -max_target_seqs 1 -outfmt 6 -out 2020-09-11_Mar_genome_snap02.all.maker.proteins.fasta.blastp -num_threads 20
- makeblastdb -in CgiCviMcoPmaMye_protein.fasta -out CgiCviMcoPmaMye_protein -dbtype prot
- blastp -query /home/metzgerm/MAKER_Mya/2020-09-11-Mar.3.4.6.p1-MAKER/snap02/2020-09-11_Mar_genome_snap02.all.maker.proteins.fasta -db CgiCviMcoPmaMye_protein -evalue 1e-6 -max_hsps 1 -max_target_seqs 1 -outfmt 6 -out 2020-09-11_Mar_genome_snap02.all.maker.proteins.fasta.CgiCviMcoPmaMye_blastp -num_threads 20

1.7.2j Genome assembly summary statistics

Genome assembly statistics for the current and previous *Mya arenaria* assemblies can be found in **Supplementary Table 1.1**. Genome size, GC content, scaffold N50 and contig N50 were calculated using BBTools stats.sh (v38.86) [78]. Repeat content was estimated by running RepeatMasker (v4.1.0) using the previously generated RepeatModeler repeat library. Benchmark of Universal Single Copy Orthologs (BUSCO) scores were calculated against the metazoa_odb10 database using BUSCO v3 [25].

1.7.3 MarBTN genome sequence analysis

1.7.3a Sample collection and DNA extraction

MarBTN samples were collected from highly neoplastic clams from Maine and New York, USA and Prince Edward Island, Canada (**Fig. 1.1A and Supplementary Table 1.2)**. Hemolymph DNA from several samples of MarBTN have been previously reported (those collected 2009-2014) [6,46], and remaining samples (those collected 2020-2022) were shipped live on ice from a seafood supplier in Maine. Hemolymph was drawn and screened for highly neoplastic animals (as described above), and genomic DNA was extracted using the same protocol previously used for other MarBTN samples (DNeasy Blood and Tissue Kit, Qiagen)[6,46]. Two healthy clams were collected and DNA extracted from siphon or mantle tissue as reported previously[6], in addition to the healthy reference individual described above. Previous reports of likely BTN in *M. arenaria* (**Figure 1A**, x's) were collected from reports in which "disseminated neoplasia" or "hemic neoplasia" were diagnosed in *Mya arenaria* [21,22,24,35,48,79–97].

1.7.3b Whole genome sequencing and genome mapping

All samples were sequenced by Genewiz on an Illumina HiSeq (paired end 150 bp reads). Healthy tissue and cancer hemolymph were sequenced using a full lane with a target read depth of 50× given the expected haploid genome size. Paired tissue samples for a subset of cancer samples were sequenced on a half lane with a target read depth of 30×. Illumina sequences were purged of optical duplicates using BBTools clumpify (v38.86)[78], trimmed using trimmomatic (v0.36) [98] with a read quality threshold of 20, and mapped to the reference genome using BWA-MEM [99] with default settings.

- 02_Illumina_data_processing/02_map_to_genome.sh
- 02_Illumina_data_processing/00_sampling_map.R
- 02_Illumina_data_processing/01a_dedupe_and_trim.sh
- 02_Illumina_data_processing/01b_dedupe_and_trim_newsamples.sh

1.7.3c SNV calling

SNVs and indels were called using somatypus (v1.3), a platypus-based variant calling pipeline designed for closely related cancer data without a paired normal sample, ideal for the analysis of transmissible cancer genomes [12]. Variants were called as present in a healthy clam if they were called by somatypus and supported by more than 3 reads. For cancer samples, we used more stringent thresholds to eliminate contaminating host DNA from being called as cancer alleles. Paired host tissue samples proved to be too highly contaminated by cancer to be useful in eliminating host alleles and were only used as a downstream confirmation that we were eliminating host alleles with our read thresholds. Unlike the mammalian transmissible cancers, which form solid tumors and allow the collection of uncontaminated healthy host DNA, BTN disseminates into the tissues of the host as the cancer progresses, resulting in tissue samples that include significant BTN cells, often even more cancer cells than host cells, in late stages of the disease. However, we find DNA extracted from hemolymph to be so highly composed of BTN cells and so few host hemocytes (**Supplementary Fig. 1.3)**, that we were able to effectively remove host variants using these thresholds.

Sequencing resulted in a range of average sequencing depths of 57-90 at called SNV loci across samples. We normalized SNV calling thresholds to the average read depth for each individual sample, to avoid biasing calls in favor of more deeply sequenced samples. A variant was called as present in cancer if it was present in at least one cancer sample at a depth of greater than 1/6 the average read depth for that sample (9-15 reads). Given a variant passed that criteria for at least one cancer sample, it was called in any other cancer sample if it was present at a depth greater than 1/16 the average read depth for that sample (3-5 reads). These thresholds were chosen to minimize the calling of host alleles or mis-mapped reads as cancer alleles, while also preventing the exclusion of real cancer alleles from our variant set.

We used median allele frequency MarBTN-specific homozygous nuclear SNVs in copy number 2 regions as a proxy for cancer isolate purity and host tissue purity, as shown in **Supplementary Fig. 1.3**. For non-reference healthy clams (the reference clam has no homozygous SNVs since one of the haplotypes is the reference genome), median allele frequencies were calculated from homozygous SNVs present in all

samples. For cancer samples, median allele frequencies were slightly lower, attributed to the presence of host clam DNA, but remained >96%. Two MarBTN isolates that were excluded from this study due to high host DNA contamination were included on this analysis as contaminated sample controls. For samples for which paired tissue sequencing existed (3 of 8 cancer samples), we used the median allele frequencies of the same set of SNVs to estimate contamination of tissue by cancer DNA. Some samples contain a high amount of cancer DNA, making genome-wide differentiation between host and cancer SNVs difficult in tissue and leading us not to include paired tissue DNA in our analyses.

- 02_Illumina_data_processing/03_rename_for_somatypus.sh
- 02_Illumina_data_processing/04_run_somatypus.sh
- 03_SNV_analysis/02_initial_SNV_counts.R
- 06_Mito_analysis/05a_sample_purity_nuclear_and_mito.sh
- 06_Mito_analysis/05a_sample_purity_nuclear_and_mito.R

1.7.3d LOH region identification

To call genome regions where one of the two original founder haplotypes was lost in one sub-lineage but retained by the other sub-lineage (termed LOH for loss of heterozygosity), we focused on SNVs for which we had high confidence that they came from the founder clam germline – those found in all cancer samples and at least one healthy clam. We calculated the allele frequencies for each of these SNVs in each cancer sample and flagged SNVs that were likely homozygous (above 0.8 frequency) for all samples in one sub-lineage, while heterozygous (less than 0.8) for all samples in the other sub-lineage. We included three representative samples from the USA sub-lineage in these calls (FFM-19G1, FFM-22F10 and NYTC-C9) so that calls were not biased by there being more USA samples than PEI. A region with SNVs transitioning to homozygous from heterozygous (with the ancestral heterozygous state being captured in the other sub-lineage) would indicate regions that had lost a parental haplotype in the homozygous sub-lineage. We looked at sliding windows of 50 heterozygous founder SNVs across each scaffold (independently for the PEI and USA sub-lineages) and counted the number of SNVs that were heterozygous

-> homozygous discordant in the other sub-lineage. We found that windows with 10 or more discordant SNVs were the most effective for calling LOH regions (see below for validation of this threshold). We merged overlapping windows, for a total of 1,098 LOH windows for PEI (155 Mb, 12.8% of the genome) and 817 LOH windows for USA (98 Mb, 8.1% of the genome).

The amount of the genome that was called as in an LOH region was highly sensitive to the threshold of heterozygous -> homozygous discordant SNVs used to call LOH windows. We used two metrics to determine the best threshold, signature S mutation fraction and dN/dS. Both metrics are proxies for somatic mutations, with lower values corresponding to more founder variants and higher values corresponding to more somatic mutations. A higher threshold results in a high confidence in the LOH regions, but with missed true regions of LOH, resulting in more founder variants in regions called as non-LOH, while a lower threshold results in over-calling of LOH regions but with less founder variants in the regions called as non-LOH. We tested the calling of LOH regions as described above for all possible thresholds between 0 and 50 SNVs in the 50 heterozygous SNV window. We then divided high confidence somatic mutations into the regions called in these test sets as LOH and non-LOH. We then calculated signature S mutation fraction and dN/dS ratio for each and plotted the values against the threshold used for the test calling (**Supplementary Fig. 1.26**). Over-calling of LOH drops dramatically before flattening out around 10/50 SNVs, while missed LOH appears to rise consistently as the threshold is increased. Overall, a threshold of 10/50 SNVs maximizes the difference between somatic mutations in non-LOH vs LOH regions and was used in all other analyses to call LOH regions, so that they could be excluded from somatic mutation analysis.

To validate that our LOH calling method was successfully removing LOH regions we filtered for a different set of SNVs than those used to call LOH: sub-lineage-specific founder variants (variants found in a healthy clam and all individuals of one sub-lineage but none in the other sub-lineage). The density of USA-specific founder variants SNVs was 36x higher in PEI LOH regions versus non-LOH regions, and PEI-specific founder variants SNVs was 20x higher in USA LOH regions versus non-LOH regions (**Supplementary Fig. 1.4**), confirming these regions were likely lost from the other sub-lineage.

- 03_SNV_analysis/04a_LOH_calling_upstream.R
- 03_SNV_analysis/04b_LOH_merge_and_helmsman.sh
- 03_SNV_analysis/04c_exclude_LOH_SNVs.sh
- 03_SNV_analysis/04d_LOH_founder_SNV_density.R
- 03_SNV_analysis/06b_run_dNdS.sh
- 03_SNV_analysis/07_LOH_threshold_validation.R

1.7.3e MarBTN phylogeny

To build the phylogeny in **Fig. 1.1E** we concatenated all variant loci into an alignment for all 8 cancer samples with the reference genome sequence at those loci as the tree root. SNVs found in any healthy clam samples were excluded prior to this analysis, since nearly all those SNVs were likely present in the founder clam. SNVs in LOH regions were also excluded to remove founder variants from the sub-lineage branches. We then used R package "ape" (v5.5) [100] to calculate the pairwise distance between sequences using the dist.dna(model="raw") function, build a neighbor-joining tree using the and nj() function and calculated bootstrap support using the boot.phylo() function, revealing high confidence (100/100) at all nodes.

- 03_SNV_analysis/01_pairwise_phylogeny.sh
- 03_SNV_analysis/02_initial_SNV_counts.R

1.7.3f Mutational signature extraction and fitting

We categorized SNVs into 25 bins based on which samples they were found in and the MarBTN phylogeny (see **Supplementary Fig. 1.27** or code reference below). We further divided each SNV bin by annotated genome regions into additional nested bins (full genome, genes, exons, CDS, 5'UTR, 3'UTR), with the thought that some mutational processes may have different exposures across the genome. We used Helmsman (v1.5.2) [101] to count SNVs for each bin in their trinucleotide context, and R package "Biostrings" (v2.54.0) to count trinucleotide opportunities in each genome region. We performed *de novo* signature extraction on this data set using R package "sigfit" (v2.0.0) [102], correcting for opportunities in each genome

region. The unbiased estimate for the best number of signatures to fit our data was 3, though extracting 4 signatures revealed a signature of unmistakable resemblance to COSMIC signature 1 (CpG>TpG), so we proceeded with 4 signatures. SNV bins were then reanalyzed with these 4 signatures, again correcting for mutational opportunities, to reveal the fraction of SNVs in each category that could be attributed to each signature.

- 03_SNV_analysis/03a_sig_extraction_upstream.R
- 03_SNV_analysis/03b_sig_extraction_count_trinuc.sh
- 03_SNV_analysis/03c_sig_extraction_fitting_sigfit.R
- 01_Genome_assembly/03a_trinucleotide_counting.sh
- 01_Genome_assembly/03b_trinucleotide_counting.R

1.7.3g Cancer dating and estimation of total somatic SNVs

To estimate the age of the MarBTN lineage we only wanted to consider likely somatic mutations, so we excluded regions that were called as LOH in either sub-lineage from these analyses (as true founder SNVs in a region lost in one sub-lineage would appear to be unique to the other sub-lineage and could be falsely considered to have occurred after the divergence of the sub-lineages if those regions were not removed). We only included genomic SNVs for this analysis, as there were a limited number of MarBTN-specific mitochondrial SNVs and they displayed a different mutational profile than genomic mutations (**Supplementary Fig. 1.17**). We then filtered remaining SNVs from each MarBTN sample to remove any SNVs that were found in a healthy clam or the other sub-lineage using the same thresholds as described above (SNV calling). We then have high confidence that the remaining SNVs for each sample should be somatic mutations which occurred since the time the two sub-lineages diverged (since the most recent common ancestor, or MRCA). We counted the number of mutations in their trinucleotide contexts using Helmsman [101] for each MarBTN sample and fit this to our *de novo* extracted mutational signatures to estimate contributions of each of the 4 signatures. We then performed a linear regression of the mutation count attributed to each signature for each sample against the date the sample was collected (**Fig. 1.2C/D**

and **Supplementary Fig. 9**). We performed regression across USA samples only, with the thought that this set would be less susceptible to small changes in mutation rates between the sub-lineages and would not be confounded by the timing or number of copy number differences between the sub-lineages. Within the USA sub-lineage, Sig5' was the best fit with time. When considering PEI samples, SigS appeared more clock-like in that PEI samples fall within the 95% confidence interval of the USA regression. Additionally, to test whether structural mutations types that were higher in USA than PEI was due to sampling date, we performed the same analysis on somatic tandem duplications, somatic translocations, total steamer insertion sites, and total mitochondrial D-loop copies (**Supplementary Fig. 1.22**).

The x-intercept of the regressions calculated above indicates the age of the most recent common ancestor of the two sub-lineages (i.e., when mutation count separating them equals zero). To estimate the total age of the cancer, we first estimated the number of somatic SigS mutations in the trunk of the MarBTN lineage: SNVs shared by all MarBTN samples. We continued to exclude LOH regions, to work with the same region for which the mutation rate was calculated, and to exclude SNVs shared with any healthy clams, which are presumably founder variants. Somatic mutations in non-LOH regions with copy number >1 would have been heterozygous when they occurred, so we filtered for SNVs with an average allele frequency under 0.8 across the 8 MarBTN samples. For comparison, we also analyzed the following SNV bins: likely homozygous SNVs (those with an average allele frequency over 0.8); SNVs in healthy clams (from each of the three healthy clams individually as well as all SNVs found in any healthy clam); SNVs found in all cancer samples and shared with a healthy sample; and SNVs found in all samples in one sub-lineage but not the other sub-lineage or healthy clams (high confidence somatic mutations). We counted mutations in their trinucleotide contexts and fit the 4 *de novo* extracted signatures as described previously.

The fraction of SNVs found in healthy clams attributable to signature S was taken to be the baseline SigS fraction (0.025). The SigS mutation fraction was near this baseline for individual healthy clam SNVs and for likely founder variants – those found in all MarBTN samples and either shared with a healthy clam or not shared with one of the sequenced healthy clams but homozygous. Heterozygous SNVs found in all MarBTN samples but no healthy samples had noticeably higher SigS fraction (0.056). The difference

between this fraction and the baseline was taken to be from SigS mutations in the early somatic evolution of the MarBTN lineage (0.056-0.025=0.031). This fraction is equivalent to 53,350 mutations, or 108 years (95% CI: 48-Inf) by the previous SigS mutation rate calculation. This confidence interval was determined solely from uncertainty in the mutation rate since error estimates from signature fitting with sigfit were negligible in comparison.

Using the SigS mutation estimations above, we then estimated the total number of somatic mutations in each of the MarBTN sub-lineages, which would be a combination of mutations occurring post-MRCA (high confidence somatic mutations) and mutations in the lineage trunk. To first estimate somatic mutations in the trunk, we assumed that somatic SigS mutation fraction had remained steady since oncogenesis at 0.48 (based on high confidence somatic mutations in the two sub-lineages). Given the following equations describing the total number of mutations being comprised of a fraction of founder SNVs and a fraction of somatic SNVs (a) and that each fraction has a known percentage of mutations due to SigS (b), we then solve for $Fraction_{somatic}$ (c-e):

a) $Fraction_{founder} = 1 - Fraction_{somatic}$

b) $SigS_{observed}$(heterozygous all BTN, no healthy)=$Fraction_{somatic}*SigS_{somatic}+Fraction_{founder}*SigS_{founder}$

c) $SigS_{observed}=Fraction_{somatic}*SigS_{somatic}+ (1 - Fraction_{somatic})*SigS_{founder}$

d) $Fraction_{somatic} = (SigS_{observed} - SigS_{founder}) / (SigS_{somatic} - SigS_{founder})$

e) $Fraction_{somatic} = (0.056-0.025) / (0.48-0.025) = 0.068$

This is equivalent to 116,765 somatic mutations. We then added high confidence somatic SNVs unique to each sub-lineage (those present in all samples from that sub-lineage but none in the other sub-lineage or healthy clams) and corrected for genome size of the non-LOH portion of the clam genome to get mutation density estimates for each sub-lineage (441 and 452 mu/Mb for the PEI and USA sub-lineages, respectively). Note that although we can estimate total mutation count in the lineage trunk, we cannot differentiate individual SNVs as somatic mutations or founder variants.

- 03_SNV_analysis/05a_cancer_dating_prelim.R
- 03_SNV_analysis/05a2_cancer_dating_prelim_withLOH.R

- 03_SNV_analysis/05b_cancer_dating_helmsman.sh
- 03_SNV_analysis/05c_cancer_dating_regression.R
- 03_SNV_analysis/05d_cancer_dating_trunk.R
- 04_CNV_and_SV_analysis/05b_SVs_by_time.R

1.7.3h Rate of nonsynonymous to the rate of synonymous mutations (dN/dS)

We calculated global dN/dS, the overall ratio across all genes in the genome, after identifying the following SNV subsets:

- SNVs found in all healthy clams
- SNVs found in any healthy clam
- SNVs unique to each of the three healthy clams
- SNVs in all MarBTN samples and shared with a healthy clam
- SNVs in all MarBTN samples and not found in any healthy clams in our data set
- SNVs found in all samples for each sub-lineage, but not found in the other sub-lineage or healthy clams. This resulted in three subsets: USA, PEI and SNVs from each sub-lineage combined. These were further filtered to include only SNVs outside called LOH regions.

We ran R package "dNdScv" (v0.0.1.0)[38] to calculate the dN/dS ratio for each SNV subset. dNdScv is designed to quantify selection during somatic evolution and corrects for trinucleotide context-dependent biases to estimate a dN/dS ratio normalized to the expected ratio for each gene or the entire genome. dNdScv is designed to be run on datasets of many samples, but in our case we ran it individually on the above SNV subsets. We ran dNdScv with default settings except for setting max_coding_muts_per_sample and max_muts_per_gene_per_sample to 1 billion each, effectively removing these maximum settings, which were designed for conventional cancers. We calculated global dN/dS across the whole genome, including all annotated genes**.** Likely somatic SNVs show largely neutral global dN/dS (0.98, 95%CI: 0.94-1.02), indicating that there is minimal contamination from founder variants, which are assumed to have been predominantly under negative selection, as seen in healthy clams SNVs.

We also calculated dN/dS for individual genes to search for signals of positive or negative selection. We filtered for genes under significantly positive or negative selection (corrected p-value < 0.05). For the five hits generated when dN/dS was run for somatic mutations, we performed an NCBI blastp query for each of these genes, getting hits for three of the five genes. We checked each gene visually/manually using IGV, noting that in each case nearly all SNVs appear to be on a single haplotype. We calculated the dN/dS for SNVs found in any healthy clam for each of these 5 genes, and none were under significantly positive selection in the observed healthy clam genomes (this would have been expected if these hits were due to missed founder variants in a gene under positive selection in the healthy clam population). Results and notes for each gene are summarized in **Supplementary Table 1.3**.

- 03_SNV_analysis/06a_bin_for_dNdS.R
- 03_SNV_analysis/06b_run_dNdS.sh
- 03_SNV_analysis/06c_dNdS_outputs.R

1.7.3i Copy number calling

Most cancer copy number calling tools rely on having paired tissue samples, so we developed a custom copy number calling script based on read depth relative to the clam used to build the reference genome (MELC-2E11), with the assumption that this reference clam is diploid. First we used R package "cn.mops" (v1.32.0) [103] to divide the genome into 1 kB windows and count the number of reads mapping to each window for each of the samples: healthy (3) and MarBTN (8). Any window with low mapping in the reference clam (less than ¼ the average read depth) was excluded from calling as a low-mapping region. Read depth for each window for each non-reference sample was then divided by the reference clam read depth for that window to normalize. Each window was then divided by the average read depth for that sample to yield a log2 read depth score. We then calculated the median log2 read depth for every 100 1 kB windows to form larger windows of 100 kB.

We then wanted to convert log2 read depth to copy number without prior knowledge of the average ploidy for each sample. We observed distinct peaks corresponding to copy number integers when we plotted

histograms of log2 read depth scores genome-wide for each sample. We chose the best fitting average ploidy value for each sample (the value which lined up copy number calls with integer values when multiplied by 2^{log2_score}). This was 3.6 for PEI MarBTN samples, 3.3 for USA MarBTN samples, and 1.9 for non-reference healthy samples. Note that since healthy samples should be diploid, an average ploidy just under 2 is expected, given read mapping will be slightly less efficient for non-reference clams relative to the reference clam (whose reads are mapped to a reference genome built from itself). We multiplied log2 read depth scores by this value to get copy number estimate for each 100 kB window for each sample. Observing close agreement between the samples within each sub-lineage (**Supplementary Fig. 1.11**), we calculated the average copy number calls for each sub-lineage. Finally, we smoothed copy number calls in 1 Mb windows to minimize noise in final calls. For each 100 kB window we calculated the standard deviation for the preceding ten 100 kB windows, the following ten 100 kB windows and the surrounding ten 100 kB windows (five 100 kB windows on either side). We replaced the copy number call with the median of the 1 Mb window with the smallest standard deviation, provided the standard deviation was small, defined as less than 1 on the ploidy scale. If the standard deviation was larger than 1 for all windows, we left the original unsmoothed copy number. Finally, we rounded all calls to the closest integer value for the final copy number call for each 100 kB window. However, we kept the unrounded calls for the purpose of visualizing error in our figures (**Fig. 1.3B** and grey bars in **Fig. 1.3A**).

To validate our copy number calls, which were based solely on read depth, we used variant allele frequency of somatic mutations. If calls are correct, genome regions that are of a particular copy number should exclusively have certain allele frequencies, such as 0.33/0.67 for CN3 regions or 0.25/0.5/0.75 for CN4 regions. We calculated variant allele frequencies for high confidence somatic mutations, some of which likely occurred after copy number alteration events and therefore should have a frequency distribution peak around the low frequency value (e.g 0.33 for CN3 or 0.25 for CN4). We separated SNVs specific to each sub-lineage based on the copy number calls at their locations using bedtools (v2.29.1) [104] and calculated average variant allele frequency across each ploidy level in all of the samples. A plot of the variant allele frequency distribution shows that the major peak corresponds to the expected frequency for

each copy number bin. There is evidence of some off-target peaks indicating some degree of error in these copy number calls (for example, 0.5 peak in CN3, 0.33 peak in CN2, 0.5 or less in CN1). Some of these peaks indicate regions that are called as lower copy number than the true value (e.g. 0.33 peak in CN2, 0.5 or less in CN1), which is likely due to sequence polymorphism leading to lower mapping than expected. Other off-target peaks, particularly those indicating copy number is called too high, may be due to other causes, such as the confounding effects of repetitive elements in the genome. Overall, copy-number-specific the variant allele frequencies support the conclusion that this copy number calling strategy is accurate and that much of the MarBTN genome has increased in copy number from its diploid founder ancestor.

To estimate duplication timing, we filtered for 100kB segments that were called CN4 in both USA and PEI sub-lineages. We calculated VAFs for founder germline variants (found in all cancers and at least one healthy samples) and for high confidence somatic mutations in each sub-lineage by taking the mean VAF for each SNV across the five USA samples and the three PEI samples. For each 100kB segment, we calculated the fraction of 2/4 somatic mutations by taking mutations with VAF 0.375-0.625 and dividing by total mutations.

- 04_CNV_and_SV_analysis/01_CNV_calling.R
- 04_CNV_and_SV_analysis/02_SNVs_by_CNV.sh
- 04_CNV_and_SV_analysis/03_SNV_freq_by_CNV.R
- 04_CNV_and_SV_analysis/03b_CN4_doubling_time.R

1.7.3j Structural variant and telomere calling

We used Delly (v0.8.5) [105] to call deletions, small (<100 bp) insertions, tandem duplications, inversions, and translocations in each sample individually from split read mapping. Delly is sensitive to read depth, so we subsampled all sample sequences to only include 600,000,000 reads (which is a lower count than the lowest sequenced sample) prior to running Delly using “samtools view -s”. We only considered SVs supported by reads mapping to precise breakpoints in the genome. We used default settings, except for setting a minimum paired end read mapping quality threshold to 30 to minimize false positives.

We merged all called SVs into a single file based on shared breakpoints. We removed SVs called in the reference clam from all samples and compared the number of each SV type and size of each intra-chromosomal SV type. To narrow in on high confidence somatic SVs we then filtered out SVs found in any healthy clam or the opposite sub-lineage from each sample (similar to our approach for identifying somatic SNVs) and compared number and size of SVs. To compare SV counts between healthy/MarBTN and USA/PEI, we used a two-sided t-test (unequal variance) and to compare sizes we used a two-sided Wilcoxon signed-rank test.

We used telseq (v0.0.2) [106] using default settings to estimate telomere lengths. Telseq takes raw bam alignments for all samples (generated above) as an input and uses TTAGGG-repeat content to estimate mean telomere length for each sample as an output, as plotted in **Fig. 1.3E.**

- 04_CNV_and_SV_analysis/04_SV_calling_delly.sh
- 04_CNV_and_SV_analysis/05_SV_analysis.R
- 04_CNV_and_SV_analysis/06a_telomeres.sh
- 04_CNV_and_SV_analysis/06b_telomeres.R

1.7.3k Identifying Steamer insertion sites

We called *Steamer* insertion sites in all samples via a custom pipeline which uses split reads that map to both the reference genome and *Steamer* itself. First, we used BWA-MEM [99] to map reads for each sample to the 177 bp *Steamer* long terminal repeat (LTR), a sequence that flanks either side of the internal coding sequence of LTR-retrotransposons [46]. For all reads that mapped, we extracted just the externally flanking portion of each read, discarding reads that extended into the internal sequence of *Steamer* and discarding the portion of each read that mapped to the *Steamer* LTR. We then mapped these flanking portions to the reference genome, keeping only reads that mapped to a single location in the genome with high confidence (MAPQ score ≥ 30). For reads mapping to the genome with lower confidence (MAPQ score < 30), we rematched each flanking read with its pair and re-mapped to the genome using bwa sampe[66], and if the mapping of the flanking fragment together with its mate generated a MAPQ score ≥ 30 then it

was included as a specific mapped read. Finally, we took all flanking reads that did not map to the genome with high confidence in either step and mapped them to the RepeatModeler2-generated repeat library, since many flanking reads that do not map to a specific site in the genome are likely to be in repetitive regions. We then generated a BED file format for each flanking read mapped to the genome or repeat library, calling each *Steamer* site by its 5 bp target site duplication, which is generated upon insertion and means that upstream and downstream flanking reads will overlap by 5 bp, and whether it was forward or reverse-face relative to the mapped chromosome.

We merged reads by their mapped locations to get the total number of upstream and downstream flanking reads supporting each insertion site, keeping all sites supported by at least five total flanking reads or at least one each of upstream and downstream flanking reads. We corrected for the six *Steamer* insertions that exist in the reference genome, which would otherwise result in upstream and downstream reads mapping 4.7 kB apart if that insertion is present in a sample (before and after the *Steamer* copy in the reference genome), so that upstream and downstream reads were still counted as in support of the same insertion. We estimated total read depth at *Steamer* insertion sites by averaging the read depth 10 bp before and10 bp after the 5 bp target site duplication, only considering reads with MAPQ ≥ 30. We then estimated insertion allele frequency at each site by dividing the number of *Steamer* insertion supporting reads by the total read depth. We then merged these insertion calls for each sample into a single table, converting presence/absence values to 0/1 and creating a distance matrix using dist.gene() and building a neighbor joining tree and bootstrapping using R package "ape" as described for the nuclear phylogeny [100].
We also counted shared insertion sites between samples (e.g. all MarBTN, PEI only, USA only) as shown on tree branches and reported in the text.

We noticed a bias for ATG in positions 7-9 in both our upstream and downstream *Steamer* flanking reads. To investigate this bias, we extracted the 35 bp surrounding each *Steamer* insertion sites from the reference genome (15 bp upstream, 5 bp target site duplication, and 15 bp downstream) using bedtools getfasta[104]. We then counted the number of occurrences of each nucleotide at each position, normalized by the GC content of the genome (35%), and created logo plots using ggseqlogo [107]. This bias held whether

we looked at *Steamer* sites across all samples, just cancer samples, sites shared by all cancer samples, sites unique to the USA sub-lineage, and sites unique to the PEI sub-lineage. For sites found in any cancer sample, we also counted the number of sites that had an ATG in positions 7-9 upstream, downstream (note ATG in read in reverse is CAT), and both upstream and downstream. Compared to the frequency expected based on the frequency of ATG in the genome (2.2% of trinucleotides), these sites were 8.5, 7.4, and 44.6 times more frequent than expected by chance, respectively.

To investigate where *Steamer* inserted relative to genes, we found the closest gene to each insertion site using bedtools closest [104], excluding insertion sites within genes. There was a noticeable bias in the 1-2 kB upstream genes (**Supplementary Fig. 1.28A**). To ensure this was not due to read mapping bias, we generated a similar plot based on whole genome sequence read mapping by mapping 0.1% of MELC-2E11 reads to the genome and treating the first 5 bp as a *Steamer* insertion site. This test set did not display this bias (**Supplementary Fig. 1.28B**). We then counted the number of *Steamer* insertions in annotated regions in the genome (genes, coding sequences, 5'UTR and 3'UTR) in addition to the 1 kB regions upstream of annotated gene regions. We then normalized for both the size of those portions of the genome and how likely reads were to map to these regions (to correct for biases that might skew insertions toward more mappable portions of the genome), yielding the plot found in Figure 4C.

To see whether these genes might be more likely to be cancer-associated, we conducted a blastp search of predicted intact *M. arenaria* gene models for the 729 cancer-associated genes from the COSMIC database, which generated hits of e-value>1e-6 in 14% of *Mya* genes. We then compared the number of *Steamer* insertions that intersect with these genes. In the absence of selection for insertion near these genes we would expect 14% of *Steamer* insertions to intersect with these genes. Observed versus expected insertions were compared with a Chi-squared test. See **Supplementary Fig. 1.20** for results.

- 05_TE_analysis/01_identify_steamer_in_ref_genome.sh
- 05_TE_analysis/02_steamer_calling_pipeline.sh
- 05_TE_analysis/03_steamer_downstream_analysis.R

- 05_TE_analysis/04a_steamer_ATG_bias.sh
- 05_TE_analysis/04b_steamer_ATG_bias.R
- 05_TE_analysis/05a_steamer_upstream_bias.sh
- 05_TE_analysis/05b_steamer_upstream_bias.R
- 05_TE_analysis/05c_steamer_upstream_cosmic_bias.sh

1.7.3l TE copy number analysis

We did not observe *Steamer* in our RepeatModeler run on the reference genome, likely due to it being present at low copy number and thus not clearing the threshold to be called as a repeat element. In order to capture other repeat elements like *Steamer* that might be high copy number in MarBTN but low in the reference genome, we also ran REPdenovo [108], a repeat element identifier that can be run on raw WGS data, as opposed to the assembled genome required for RepeatModeler. We ran REPdenovo on the healthy reference clam (MELC-2E11), a USA MarBTN sample (MELC-A11) and a PEI MarBTN sample (PEI-DN08) to capture repeat elements at high copy number in either sub-lineage, as well as a healthy clam to control for biasing repeat element identification towards MarBTN. We then ran RepeatClassifier, a component of RepeatModeler used for classifying repeats based on sequences, on the output repeat elements.

To generate a consensus repeat library, we used CD-HIT (v4.8.1) [109] to merge the libraries generated from the RepeatModeler and REPdenovo runs, using the same CD-HIT settings as those used by RepeatModeler itself to merge repeats with greater than 80% identity (-aS 0.8 -c 0.8 -g 1 -G 0 -A 80 -M 10000). We then used BWA-MEM to map reads from each sample to the repeat library and calculated the average read depth across each repeat element. We then normalized by read depth across the genome, calculated previously, to control for variation in sequencing depth between sequencing runs, to yield an estimate of the number of copies of each repeat element in each sample. Note that this copy number is relative to the haploid genome for all samples, so ploidy differences between our samples should not affect our copy number comparisons.

For each repeat element, we calculated the average copy number among our three healthy clams, eight MarBTN samples, and each MarBTN sub-lineage individually (five USA samples and three PEI samples). We calculated the ratio of copies in healthy clams versus MarBTN samples and PEI sub-lineage versus the USA sub-lineage, followed by a t-test to calculate the significance of each difference. We removed repeats with less than 1 copy in any sample, as these likely represent TEs that are only present in a subset of the clam population and would yield a highly significant difference simply due to the absence in some samples and presence in others. The remaining elements are plotted in the volcano plot in **Fig. 1.4D** and **Supplementary Fig. 1.21**. We calculated a significance threshold using the Bonferroni correction for multiple tests for a corrected $p<0.05$. We additionally divided and plotted the data set by repeat type classified by RepeatClassifier (DNA transposon, LTR, LINE, rolling circle, rRNA, simple repeat, SINE, snRNA, or tRNA). We performed Chi-squared tests to determine whether certain elements were higher copy number in one group versus another. We also note that although we can conclude differences in copy number, many differences may be due to variation between the founder clam and the three healthy clams sequenced in this study, as opposed to being due to somatic expansions. The magnitude of repeat expansions may be overestimated, since we are comparing an average from three difference clams to an average from eight samples of a clonal lineage. However, the strong skew towards more copies in MarBTN compared to healthy clams indicates that either A) the founder clam had more copies of many TEs than the healthy animals sequenced here or B) many TEs have increased their copy number through somatic expansion.

- 05_TE_analysis/06a_REPdenovo.sh
- 05_TE_analysis/06b_merge_repeats_and_maps_reads.sh
- 05_TE_analysis/07_TE_coverage_analysis.R

1.7.3m Mitochondrial analysis

We mapped each whole genome sequenced sample to the previously published mitochondrial genome [45] using BWA-MEM [99]. We then ran somatypus [12] using default settings to call SNVs and indels. We excluded SNVs around the multi-copy region in positions 12,060-12,971. We did not see evidence of

heteroplasmy outside this region, so an SNV was counted as present if it was present in a sample at >0.5 VAF. To infer relatedness of mitochondrial genotypes we built a neighbor-joining tree, as done for genome SNVs, from an alignment of sequences built by concatenating all variant allele positions versus the reference mitochondrial genome (170 loci).

We used median allele frequency of cancer-specific mitochondrial SNVs as a proxy for cancer isolate purity and host tissue purity to complement the similar estimation from genomic SNVs in **Supplementary Fig. 1.3**. For healthy clams, median allele frequencies were slightly below 100% likely due to sequencing, mapping, or contamination errors, and yield a maximal value for "pure" target DNA. For cancer samples, median allele frequencies were slightly lower, attributed to the presence of host clam DNA, but remained >97%. Two MarBTN isolates that were excluded from this study due to high host DNA contamination were included on this analysis as contaminated sample controls. For samples for which paired tissue sequencing existed (3 of 8 cancer samples), we used the median allele frequency of cancer-specific mitochondrial SNVs to estimate contamination of tissue by cancer DNA. Genomic and mitochondrial SNVs largely agreed in their estimation of percent cancer in hemolymph and tissue, with mitochondria SNVs giving a more precise median estimate due to higher coverage, while nuclear SNVs corrected for potential copy number differences by only focusing on CN2 regions.

To look at mutational biases, we included 12 possible single-nucleotide substitution types rather than the traditional 6, since the heavy/light strand differences of mtDNA result in unequal C/G and A/T in the forward or reverse direction (in fwd direction: A=0.29%, T=0.37%, C=0.12%, G=0.23%). We counted SNVs of each substitution type for SNVs found in healthy clams (39), shared among all MarBTN samples but not found in healthy clams (13), those found in all samples of the USA (21) or PEI (26) sub-lineages, and all high confidence somatic mutations (50: those found in only a subset of MarBTN samples). We also calculated the expected number of substitutions of each type based on the nucleotide content of the mitochondrial genome assuming no mutational biases for comparison.

We used dndscv [38] as described previously to calculate global dN/dS in the mitochondrial genome. We calculated dN/dS for SNVs found in healthy clams, SNVs shared among all cancer samples but not

found in healthy clams, and high confidence somatic mutations (i.e. those found in just the USA or PEI sub-lineages). 95% confidence intervals from dndscv are quite large due to the small number of coding mitochondrial mutations in our samples used for this calculation.

We calculated read depth at each position using samtools depth [68]. To estimate the number of copies of the D-loop region, we calculated the average read depth in positions 12,300-12,500 relative the average read depth across the full mitochondrial genome excluding that region. This region was chosen because it is within the multi-copy D-loop region but should not have reads that border the duplication breakpoint or the insertion that is only present in some copies and may cause errors in amplification due to its G-rich sequence. Copy numbers were compared between the groups using a t-test (two-sided, unequal variance).

The new reference mitochondrial genome was assembled by taking the previously published mitogenome reference with only a single, collapsed copy of the repeated D-loop region (NC_024738.1), replacing the 696 bp putative repetitive region (12,163-12,857) with a gap of 3×696 Ns, and running PBJelly to fill the gap using PacBio long reads. PBJelly was run to gap-fill the scaffolded assembly using pbsuite [73] (https://github.com/esrice/PBJelly) using using blasr v5.1, networkx 2.2, and Python 2.7 as above, with the protocol file Protocol_MELC.xml. Only captured gaps were filled (no inter-scaffold gaps) using the option "--capturedOnly" during the "support" step. PBJelly was run with the commands:

- Jelly.py setup Protocol_ MELCmtmultifastq.xml
- Jelly.py mapping Protocol_ MELCmtmultifastq.xml
- Jelly.py support Protocol_ MELCmtmultifastq.xml -x "--capturedOnly"
- Jelly.py extraction Protocol_ MELCmtmultifastq.xml
- Jelly.py assembly Protocol_ MELCmtmultifastq.xml -x "--nproc=20"
- Jelly.py output Protocol_ MELCmtmultifastq.xml

Polishing of the mitochondrial genome assembly was done with Arrow (using pbsuite as described above and pbbioconda-0.0.5 with python 3.7). First, the PBJelly output was renamed to PBJelly_mt_genome.fasta, and polishing was run using the commands:

- module load pbsuite/esrice

- pbalign --verbose --nproc 40 /home/metzgerm/MELC-2E11/Marenaria.3.2_bam.fofn PBJelly_mt_genome.fasta MELC-mtalignedall.bam 2>&1 | tee pbalign_stderrout.txt
- module load conda/4.7.10_py3.d/genomicconsensus/2.3.3
- samtools faidx PBJelly_mt_genome.fasta
- arrow --verbose --annotateGFF --reportEffectiveCoverage -j 40 MELC-mtalignedall.bam -r PBJelly_mt_genome.fasta -o MELC-2E11mtvariants.gff -o MELC-2E11mtconsensus.fasta -o MELC-2E11mtconsensus.fastq 2>&1 | tee arrow_stderrout.txt

The polished mitogenome alignment with the completely assembled repeat region (MELC-2E11mtconsensus.fasta) was renamed to mtGenome_PBJelly_polished.fasta. We confirmed the presence of a D-loop tandem duplication in a healthy clam using inverse PCR (**Supplementary Fig. 1.29)**, with outward facing primers that would only amplify if the copies or the region are in tandem (**Supplementary Table 1.4**). Amplifcation of the products of these inverse primers confirms tandem duplication of the region. However, amplicon sizes from primers spanning the D-loop support a single copy of the D-loop. Additionally, the inverse primers spanning the G-rich insertion has a dim band at expected size, but two brighter bands at smaller sizes. Given the highly G-rich region, it is likely that when primers spanning the D-loop are used that the PCR products are recombining to lose the extra copies, with selection in the PCR reaction favoring removal of the G-rich stretch that interferes with amplification. Given all samples in this study support the presence of tandem D-loop repeats, it is possible that the clam used for the previously published mitochondrial genome that contains a single D-loop copy [45] may have also been multi-copy and missed due to short-read sequences and recombination during cloning to resolve gaps in the mitochondrial genome.

- 06_Mito_analysis/00_create_coding_dndscv_input.sh
- 06_Mito_analysis/01_mapping_and_SNV_calling.sh
- 06_Mito_analysis/02_host contamination.R
- 06_Mito_analysis/03_dloop_coverage.R

- 06_Mito_analysis/04_SNV_analysis.R
- 06_Mito_analysis/05a_sample_purity_nuclear_and_mito.sh

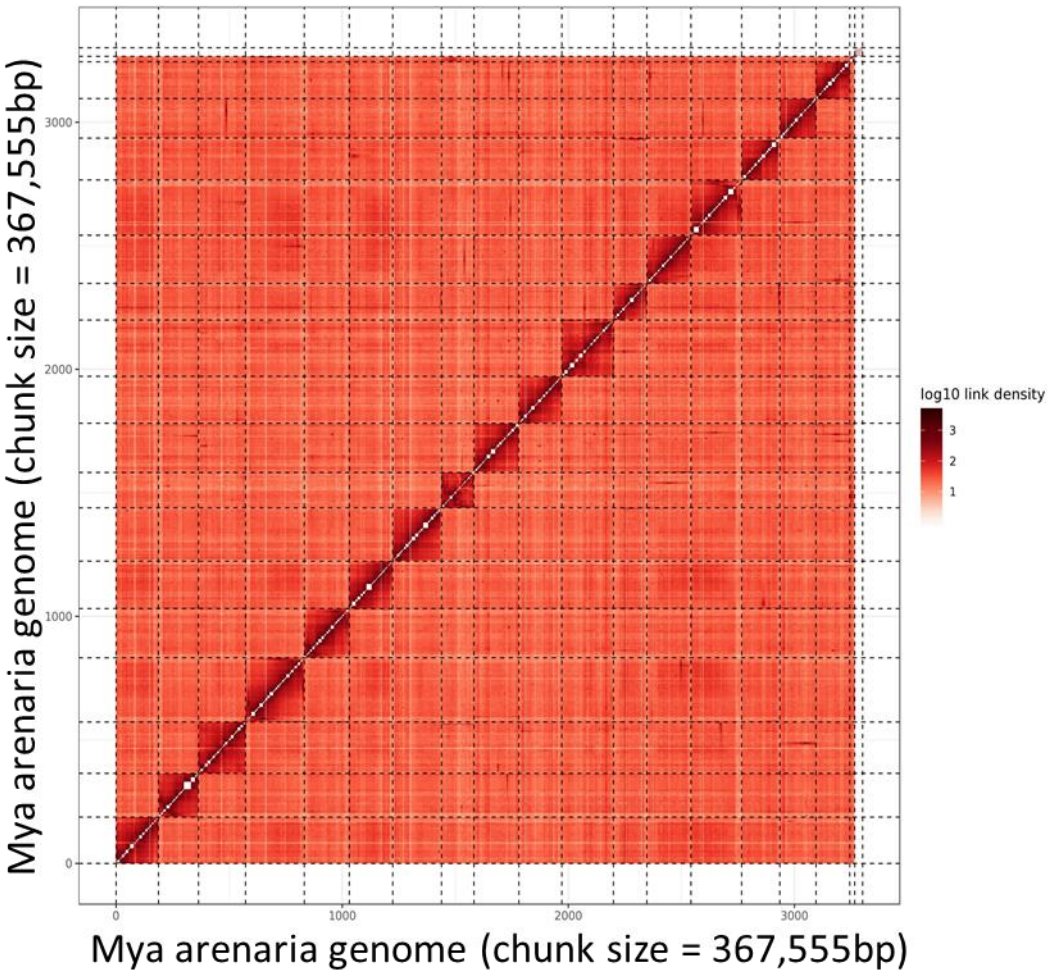
1.7.3n RNA sequence analysis

Samples from multiple tissues were collected ("Sample collection and DNA extraction") and RNA was extracted/sequenced ("RNA extraction and transcriptome assembly") as described above for two healthy clams (to add to the previously RNA sequenced reference clam, MELC-2E11), and for hemocytes only for two additional healthy clams. Hemolymph was drawn from five heavily diseased clams, and MarBTN isolates were further purified by allowing to settle for 1 hour in a 24-well plate at 4C. Remaining host hemocytes adhered to the plate and purified MarBTN cells were gently collected by pipetting. RNA was extracted using the sample protocol as for hemocytes and sequenced as described above at the sample depth (6 samples per Illumina HiSeq 4000 lane for 20-30 million reads per sample).

We aligned reads for all samples to the indexed annotated genome using STAR [110] and quantified 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, 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 [111], using tissue (or BTN) as condition on which to test differential expression. We performed principal component analysis by applying variance stabilizing transformation using vst() and plotPCA() from the DESeq2 package. We determined the top tissue-specific genes for each tissue by comparing each to the five others using DESeq2, 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 600 tissue-specific genes using the pheatmap package with clustering_distance_cols = "canberra". For individual gene comparisons of MarBTN versus healthy, we compared MarBTN separately to hemocytes and to

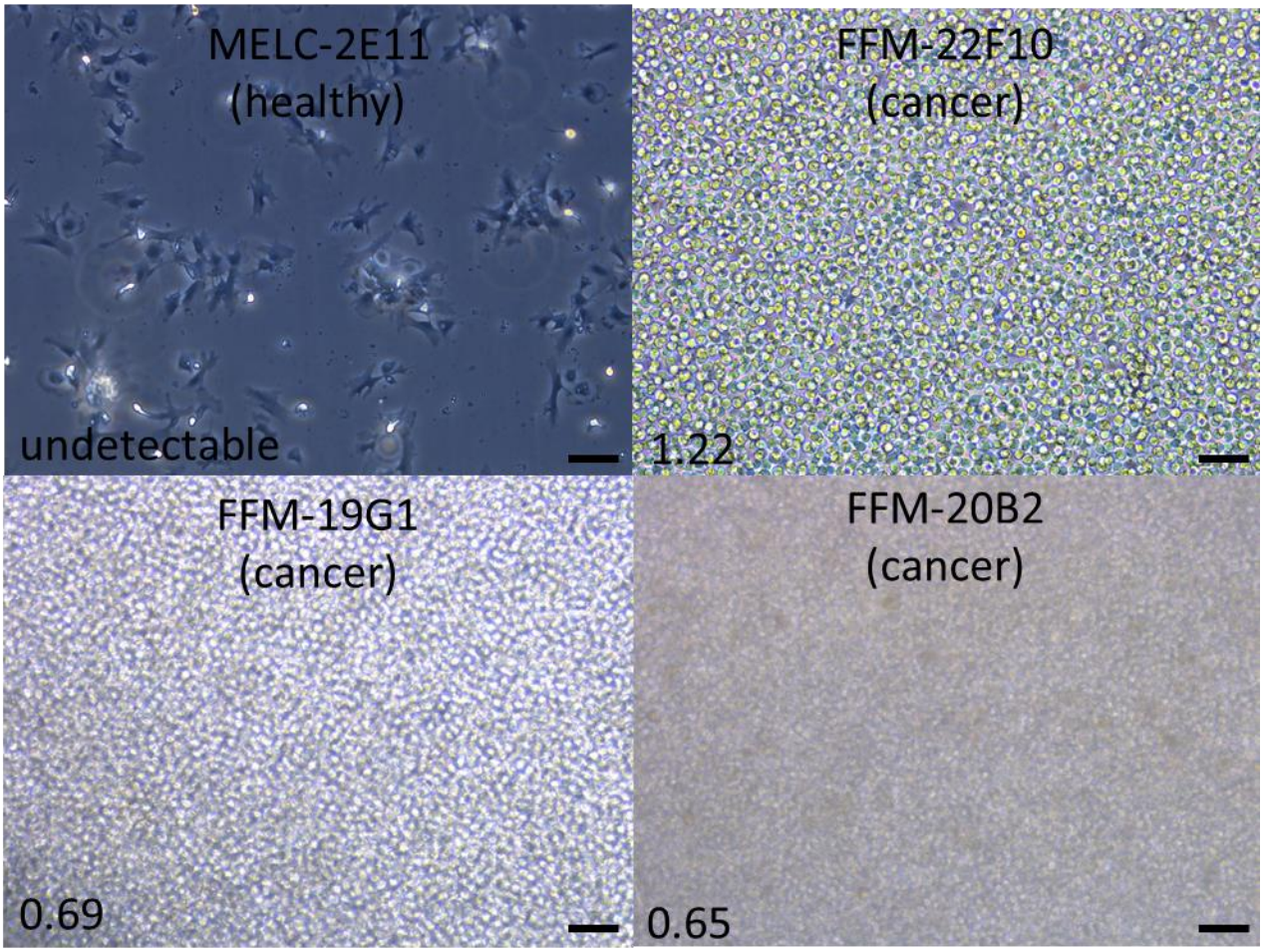
non-hemocyte solid tissues. Bar plots are comparisons of normalized read counts per gene, while statistical support for differential expression are adjusted p-values from DESeq2.

1.8 Supplemental figures



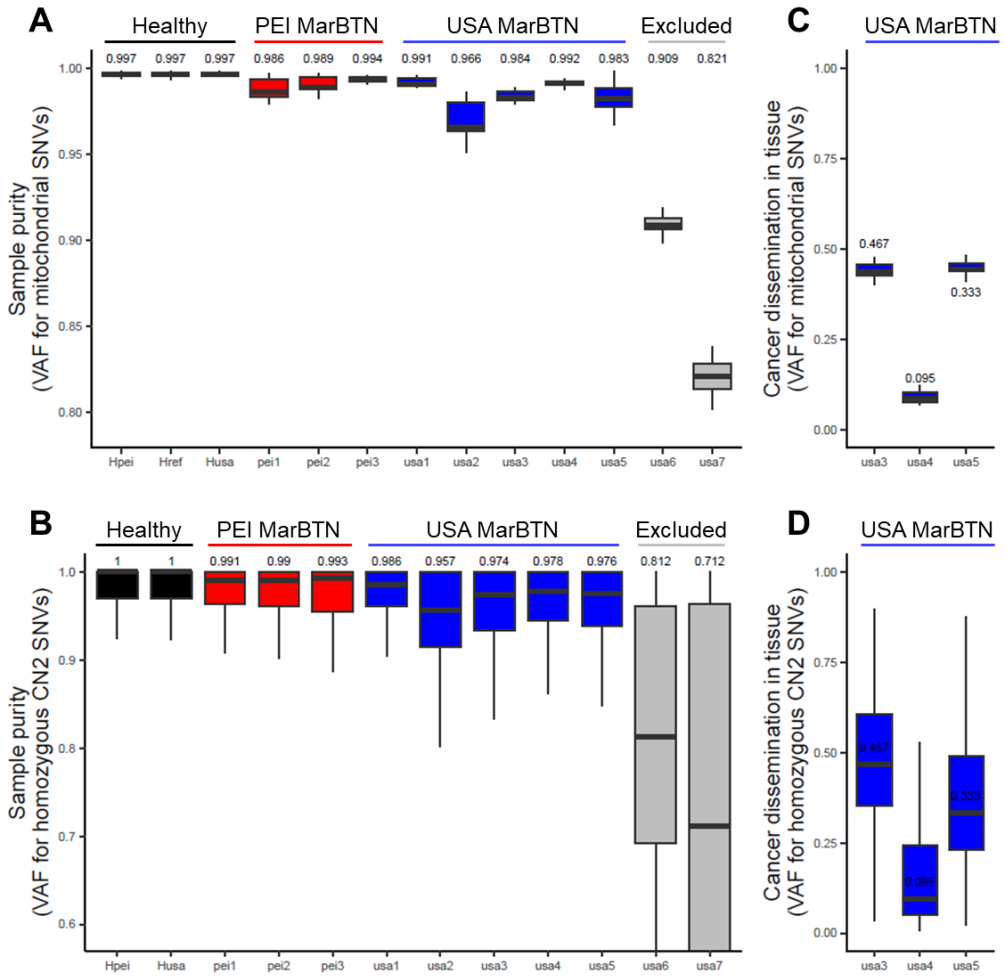
Supplementary Figure 1.1: Hi-C scaffolding yields 17 presumptive chromosomes

Heatmap of link density from Hi-C scaffolding, showing proximity of DNA segments in physical space across sequenced cells and clustering by chromosome. Results clearly yielded 17 scaffolds, matching the expected number of chromosomes in *M. arenaria*.



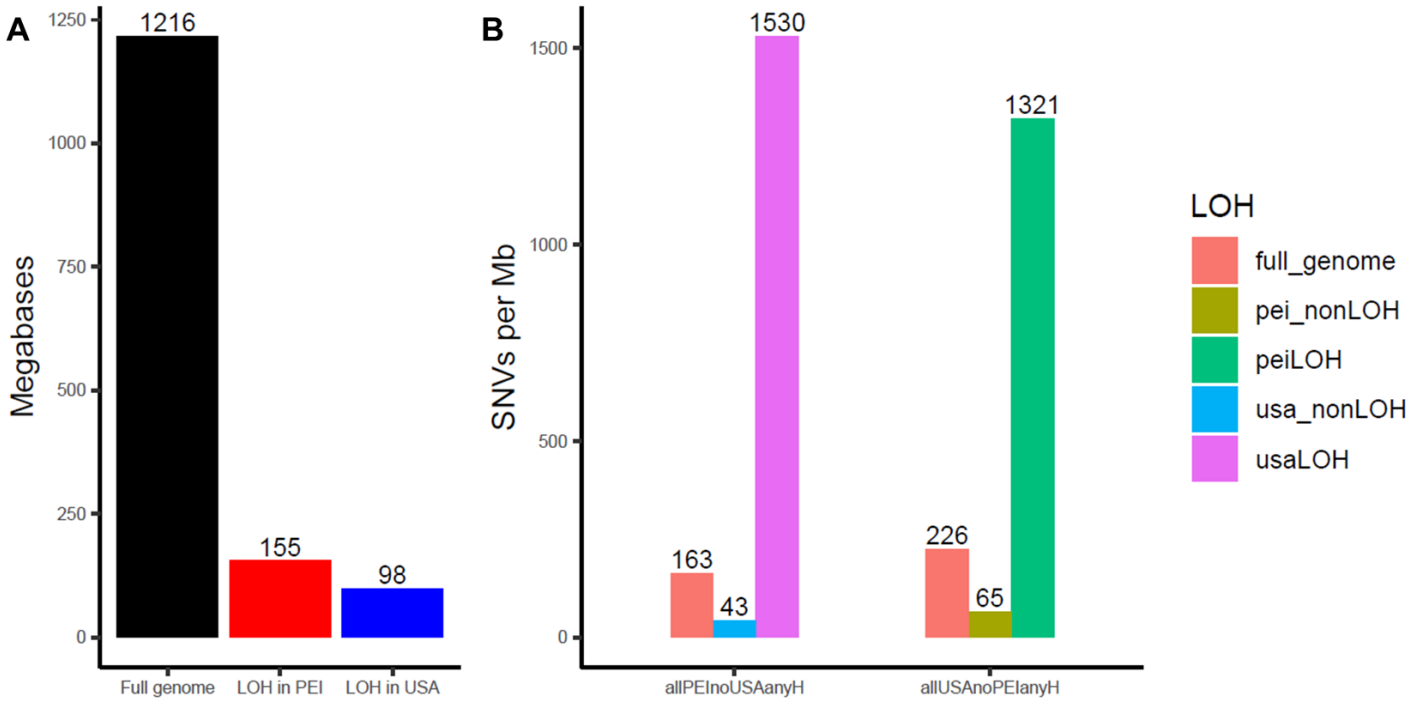
Supplementary Figure 1.2: Images of clam hemolymph for newly sampled clams

Hemolymph images for the four clams in this study sampled 2018-21. The other seven clams sampled 2010-14 were reported in past studies by Arriagada & Metzger et al. (2014) and Metzger et al (2015)[6,46]. Scale bars are 50 µm. Fraction of cancer cells detected by MarBTN-specific qPCR, as reported by Giersch et al.[17], are included in the lower left of each image. Note that while this assay is highly sensitive for the detection of low levels of MarBTN infection in animals, the fraction is a ratio of two qPCR values and minor variation in qPCR values can lead to large variation in the fraction when it is close to 100% cancer. All three cancer samples here were confirmed to be highly pure using the more sensitive genomic analysis (**Supplementary Figure 1.3**).



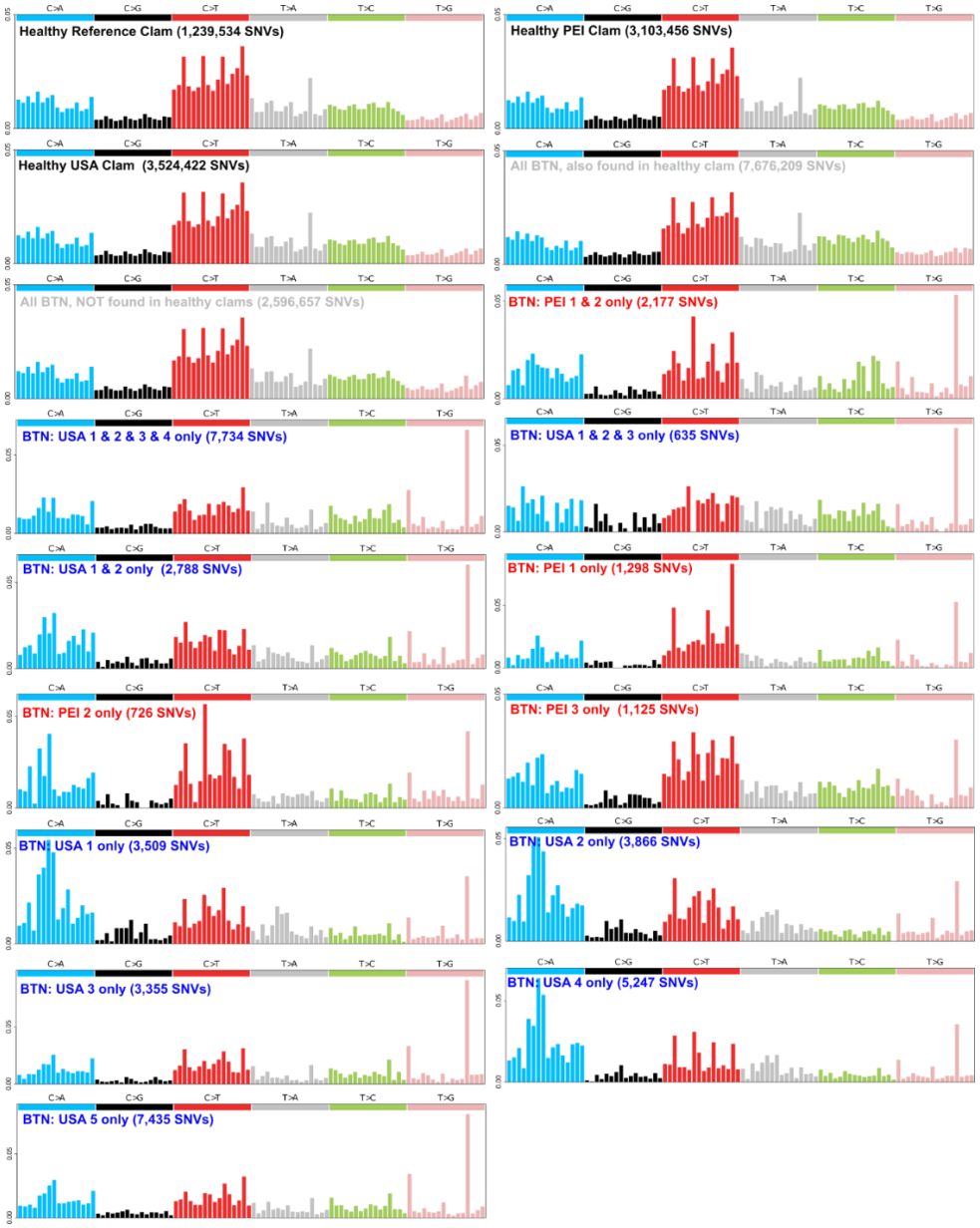
Supplementary Figure 1.3: Minimal host DNA is found in cancer hemolymph samples, while high cancer DNA is found in some tissues

(A) We identified SNVs in mitochondrial DNA in each individual sample and used the median VAF of those SNVs to estimate the purity of the sample (number of loci: 13-21 for healthy clams, 46-53 likely somatic for MarBTN samples). **(B)** Since mitochondrial genome copy numbers may differ between host and MarBTN cells, we also identified homozygous nuclear SNVs in regions called as copy number 2 in both sub-lineages and used the median VAF of those SNVs to estimate the purity of the sample (number of loci: ~250k for non-reference healthy clams, ~15k MarBTN-specific loci for MarBTN samples). Values for pure samples would be expected to be slightly below one due to mapping/sequencing errors, as evidenced by the healthy clams, which serve as pure sample controls (black, all DNA is from one individual). In cancer samples, deviation below this near-one value is attributed to the presence of contaminating host DNA (DNA is a mixture of two individuals – the cancer and the host). Two MarBTN isolates that were excluded from this study due to high host DNA contamination are included on this plot as contaminated sample controls (grey). Both nuclear and mitochondrial markers calculations yield similar estimates of cancer cell purity 96% or greater. MtDNA has the advantage of all loci being "homozygous" and much greater depth than nuclear, giving more resolution as to the exact cancer cell percentage. However, mtDNA copies per cell may vary from sample to sample and between host and cancer. We also extracted DNA from tissue samples for a subset of the USA cancers and estimated the fraction of cancer DNA disseminated into tissue using the same methodology for mitochondrial **(C)** and nuclear **(D)** loci. Tissue samples contain variable, and in some cases quite high, fractions of cancer DNA. This made genome-wide differentiation between host and cancer SNVs difficult in tissue and lead us to not include paired tissue DNA in our analyses, instead relying on variant calling thresholds to eliminate host variants from our cancer variant calling pipelines.



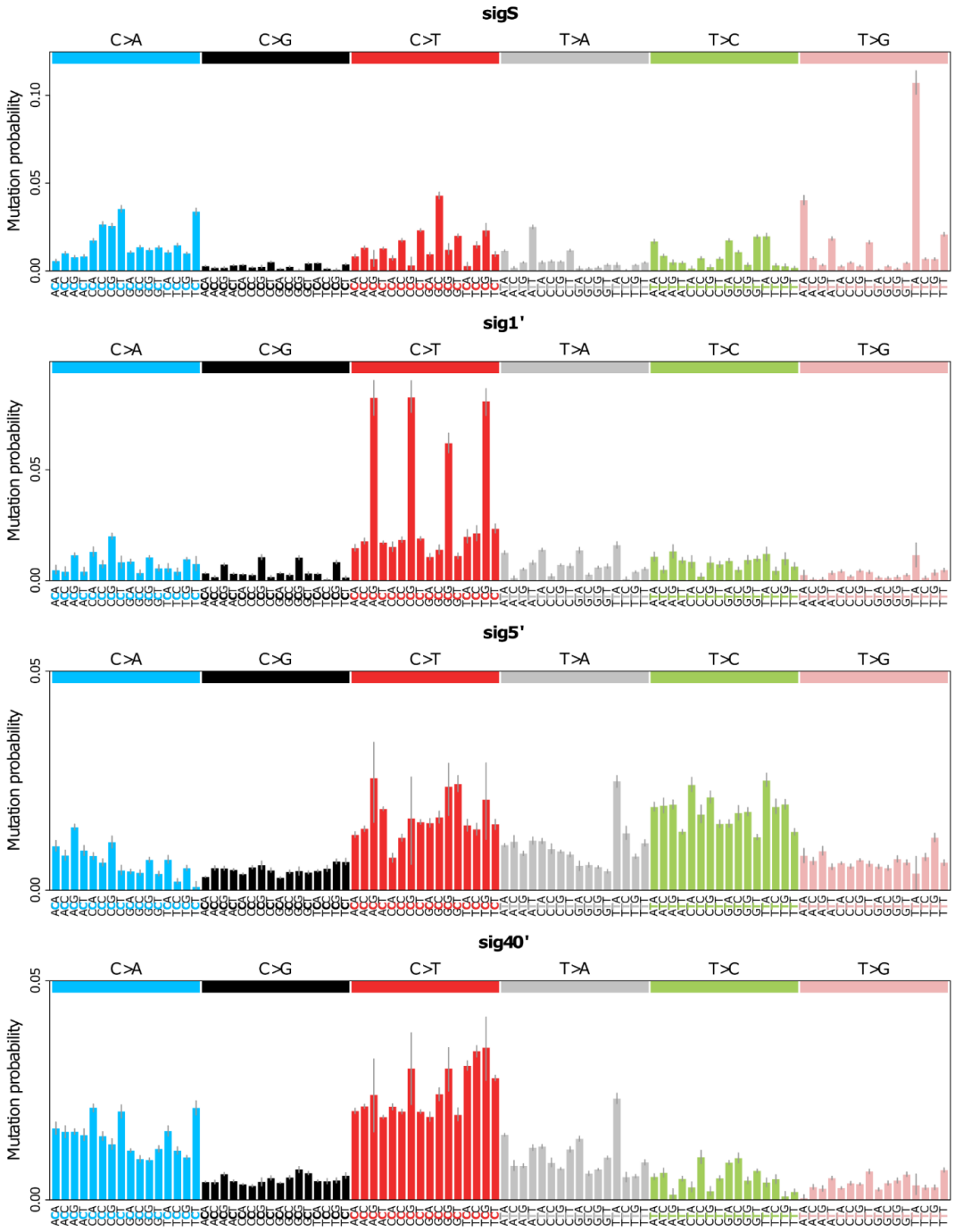
Supplementary Figure 1.4: Called LOH regions have high concentration of sub-lineage-specific founder variants

(A) Comparative sizes of the assembled genome and the fractions called as LOH in the PEI (red) and USA (blue) sub-lineages. **(B)** SNV density of sub-lineage-specific founder variants (variants found in a healthy clam and all individuals of one sub-lineage but none in the other sub-lineage) across the genome and LOH regions called in the other sub-lineage. Density is 36× greater for PEI mutations in USA LOH regions versus non-LOH regions and 20x greater for USA mutations in PEI LOH regions versus non-LOH regions. LOH regions were ignored for somatic mutation analysis to reduce the influence of remaining founder variants in sub-lineage specific SNVs, which should otherwise consist of somatic mutations.



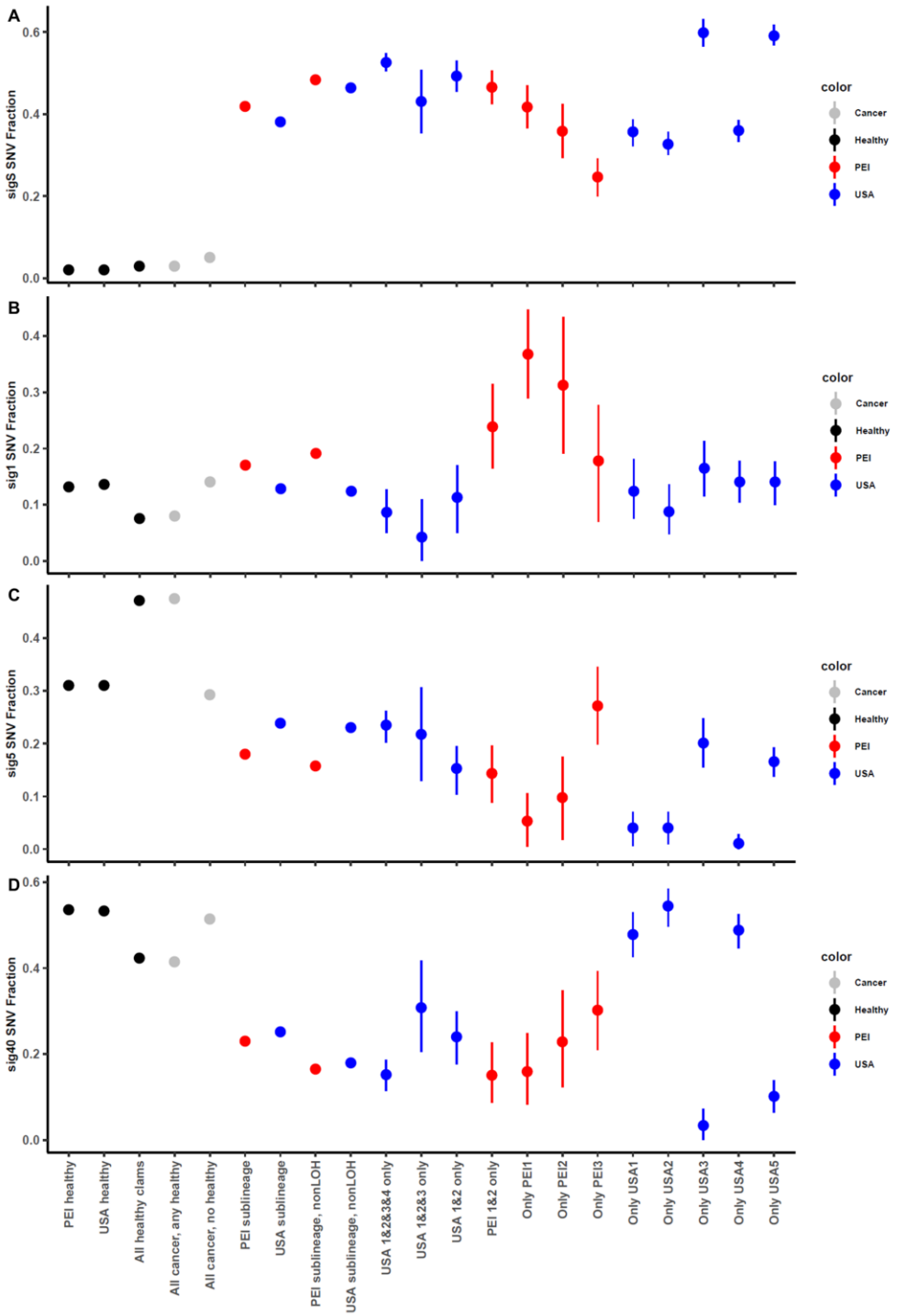
Supplementary Figure 1.5: Mutational spectra of SNV from healthy and BTN samples

Plots show the mutational probability of SNVs in all trinucleotide contexts that were identified in various samples after filtering. Trinucleotide order is the same as shown in Figure 2. Healthy clam SNVs (black labels - top) refer to SNVs that were unique to that clam and not found in other clams, resulting in no overlap of SNVs but still very similar spectra. SNVs found in all BTN samples (grey labels – upper middle) are divided into those found in a healthy clam (likely all from the founder clam genome) and those not found in any of the three healthy clams (includes a mixture of founder and early somatic mutations). Likely somatic SNVs found within the USA (blue labels) and PEI (red labels) sub-lineages show those SNVs that are either shared between all samples (Figure 2b - not shown here), multiple samples (lower middle), or unique to individual samples (bottom). SNVs found in All mutational probabilities are corrected for mutational opportunities in the clam genome, and total mutation counts in each image are shown in the label.



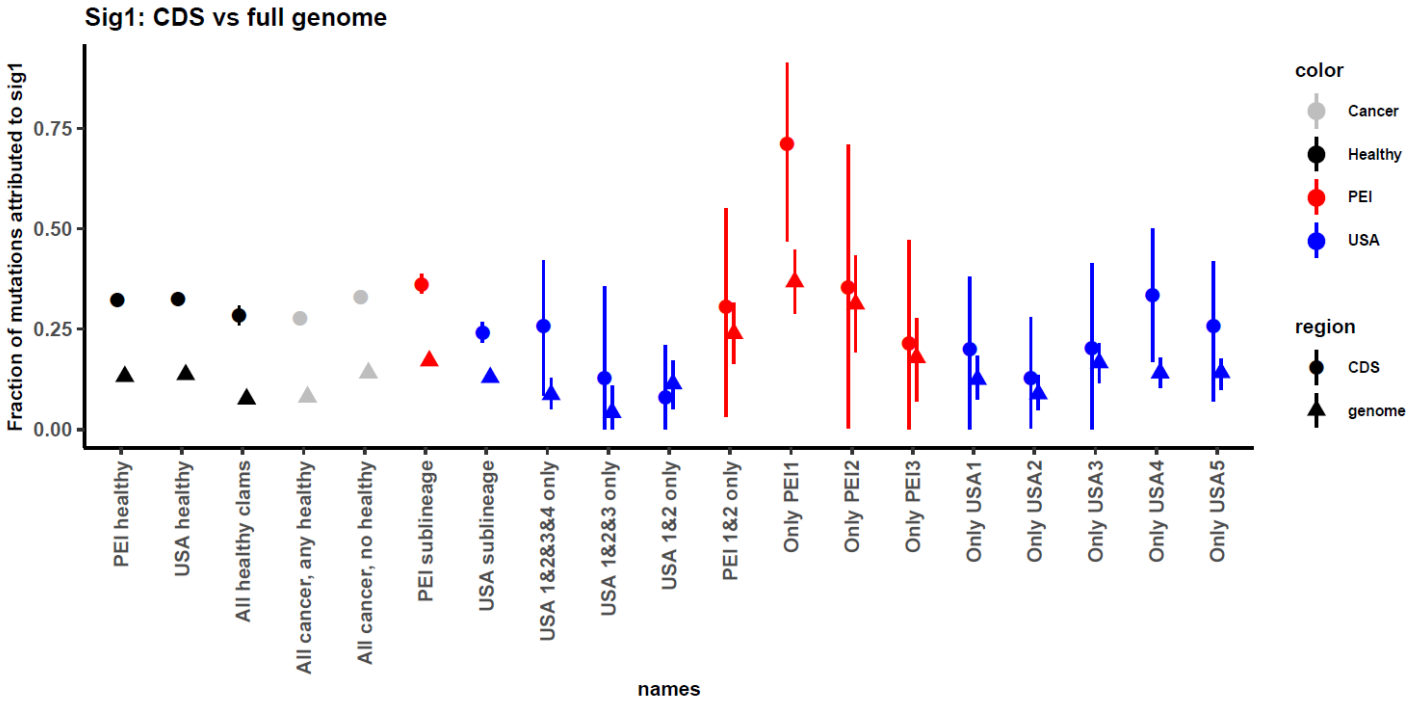
Supplementary Figure 1.6: Four mutational signatures identified by *de novo* signature extraction

We performed *de novo* mutational signature extraction to identify trinucleotide SNV differences between the various samples in this study, yielding four mutational signatures with mutational probabilities corrected for mutational opportunities in the clam genome. Error bars display 95% confidence intervals as determined by the extraction software, sigfit. Signatures sig1', sig5' and sig40' are named after the closest signature in the COSMIC database, as determined by cosine similarity. SigS was named to reflect that it was specific to **S**omatic mutations in cancer samples.



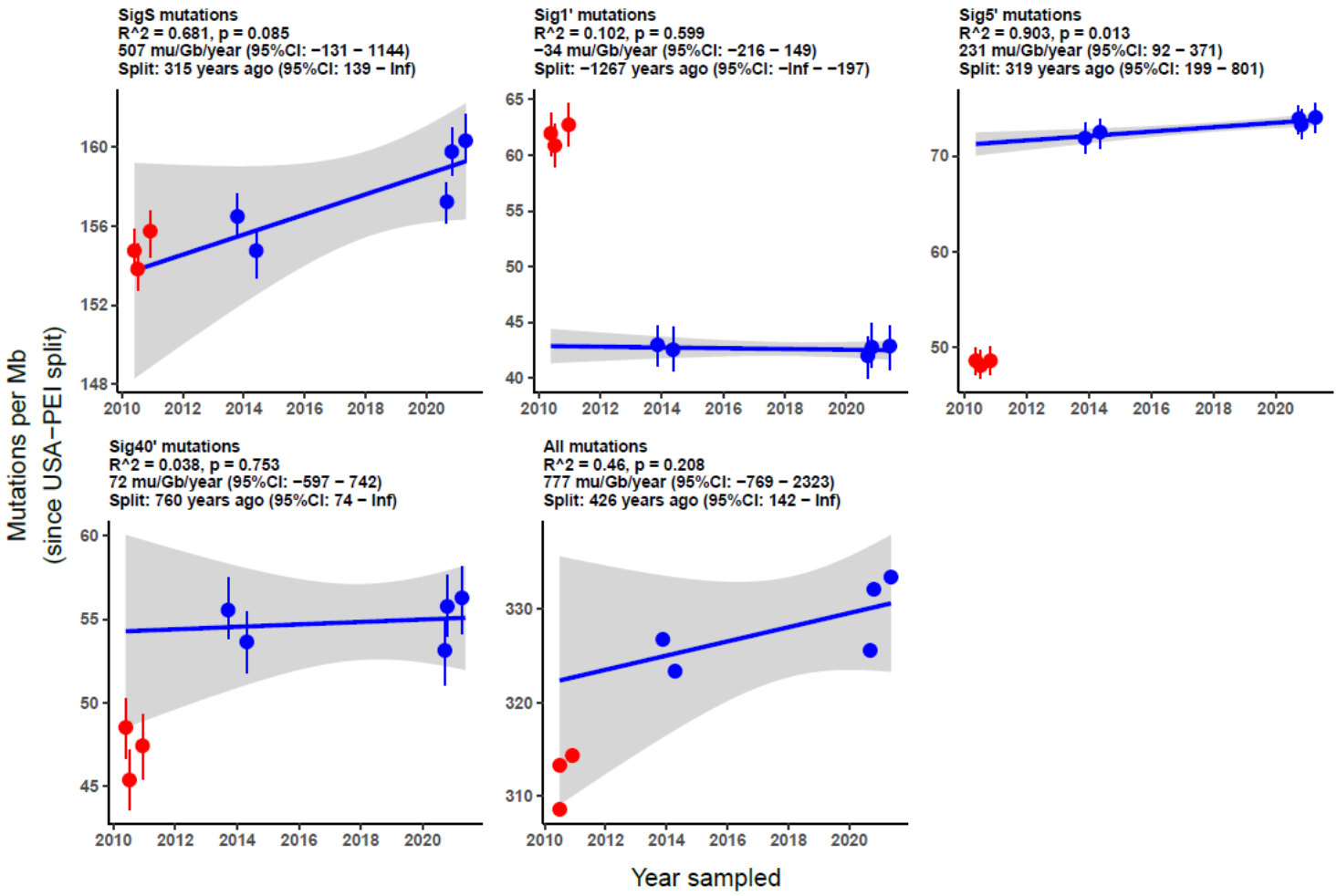
Supplementary Figure 1.7: SigS is a large fraction of both USA and PEI, but Sig1 is more highly represented in PEI

Plots shows the fraction of SNVs attributed to **(A)** signature S, **(B)** signature 1', **(C)** signature 5', and **(D)** signature 40' across healthy and cancer samples, divided and filtered as described in **Supplementary Figure 1.4** and methods (mutational signature extraction and fitting) and diagramed in **Supplementary Figure 1.20**. "All healthy clams" refers to SNVs found in all 3 healthy clams in our data set, but not in the reference genome. Error bars display 95% confidence intervals of mutation fractions from fitting error of SNVs to the four mutational signatures.



Supplementary Figure 1.8: Sig1 is more highly represented in coding regions

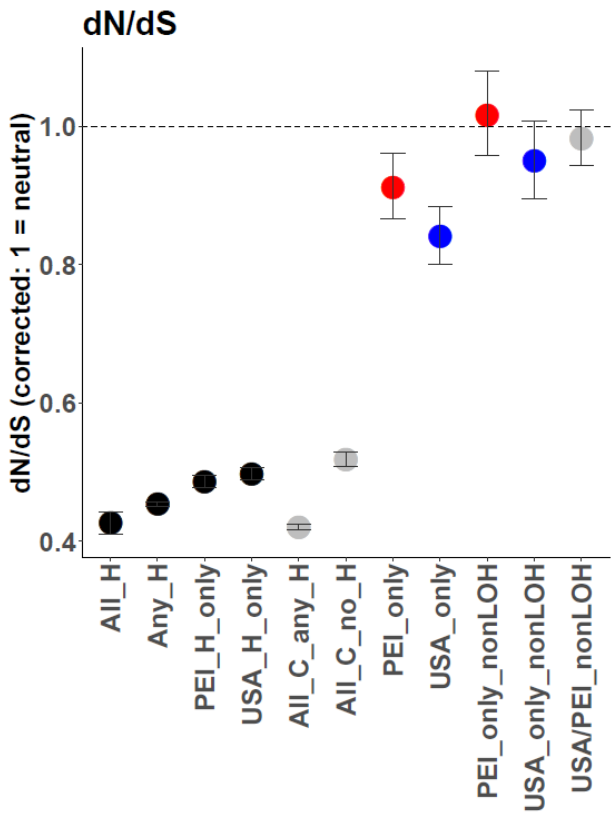
Fraction of mutations attributed to signature 1 across the whole genome (triangles, same data as in **Supplementary Figure 1.6A**) is shown compared to the fraction of signature 1 in coding regions alone (CDS, circles). Note that trinucleotide contexts of mutational opportunities are different in coding regions versus the full genome, which was factored into in the signature fitting process. Error bars display 95% confidence intervals of mutation fractions from fitting error of SNVs to the four mutational signatures.



Supplementary Figure 1.9: Mutations attributed to each signature versus sampling date

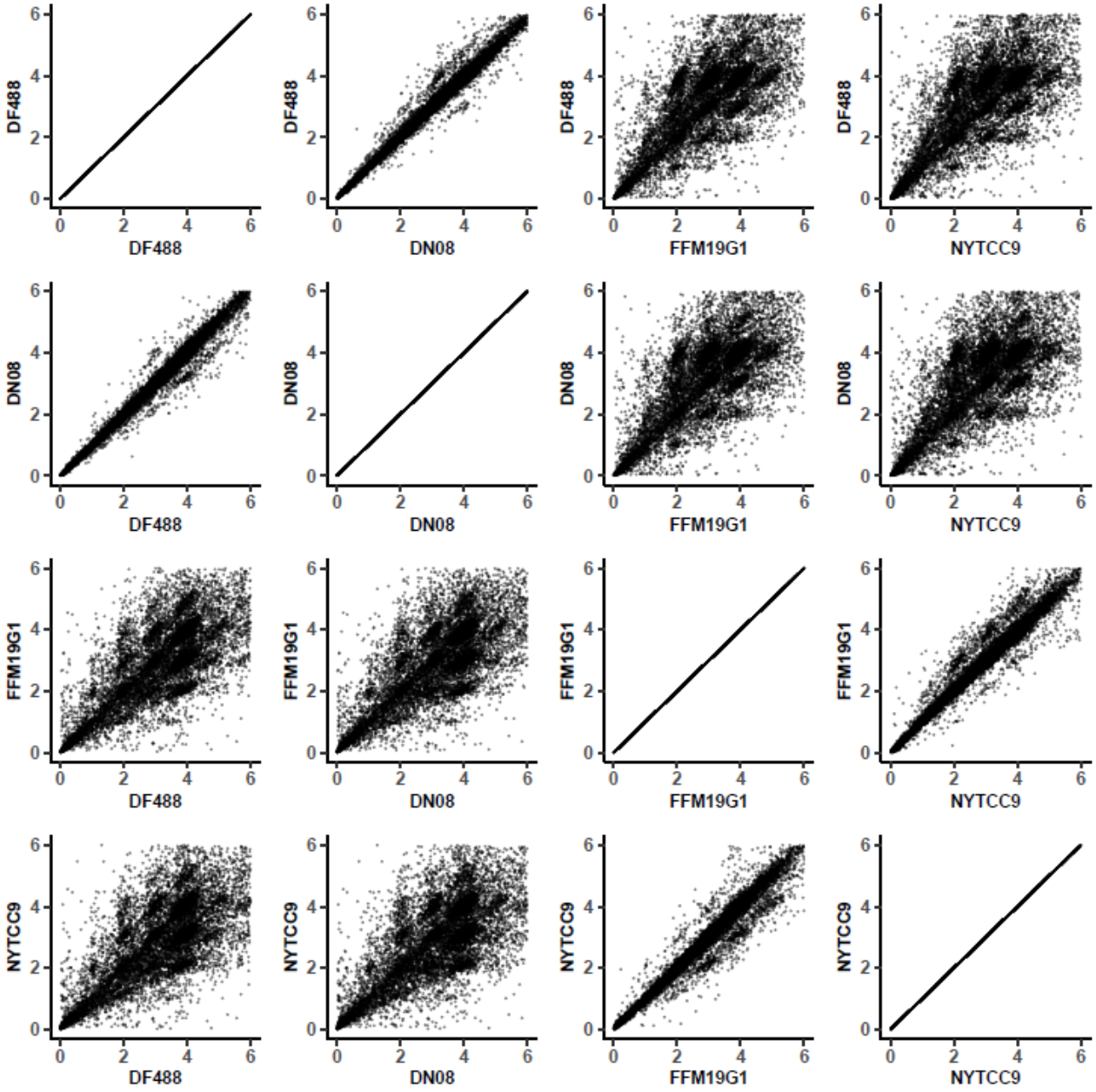
Mutations attributed to each mutational signature versus sampling date for MarBTN samples. SNVs found in healthy clams, all BTN samples, or LOH regions are excluded prior to analysis to remove founder variants. Results from linear regression of USA samples are shown above each plot, including R squared, p-value, mutation rate estimate and the corresponding x-intercept (indicating date the two sub-lineages diverged from one another). Error bars indicate 95% confidence intervals of mutation counts estimated from mutational signature fitting. PEI samples are included on plots to compare relative mutation counts attributed to each signature but are not included in the linear regression. It is apparent that sig1'mutation counts are higher in PEI, while sig5' and sig40 mutations are higher in USA. SigS mutations in PEI line up well with the USA sample regression, indicating that sigS mutation rate has stayed stable since the sub-lineages diverged.

.



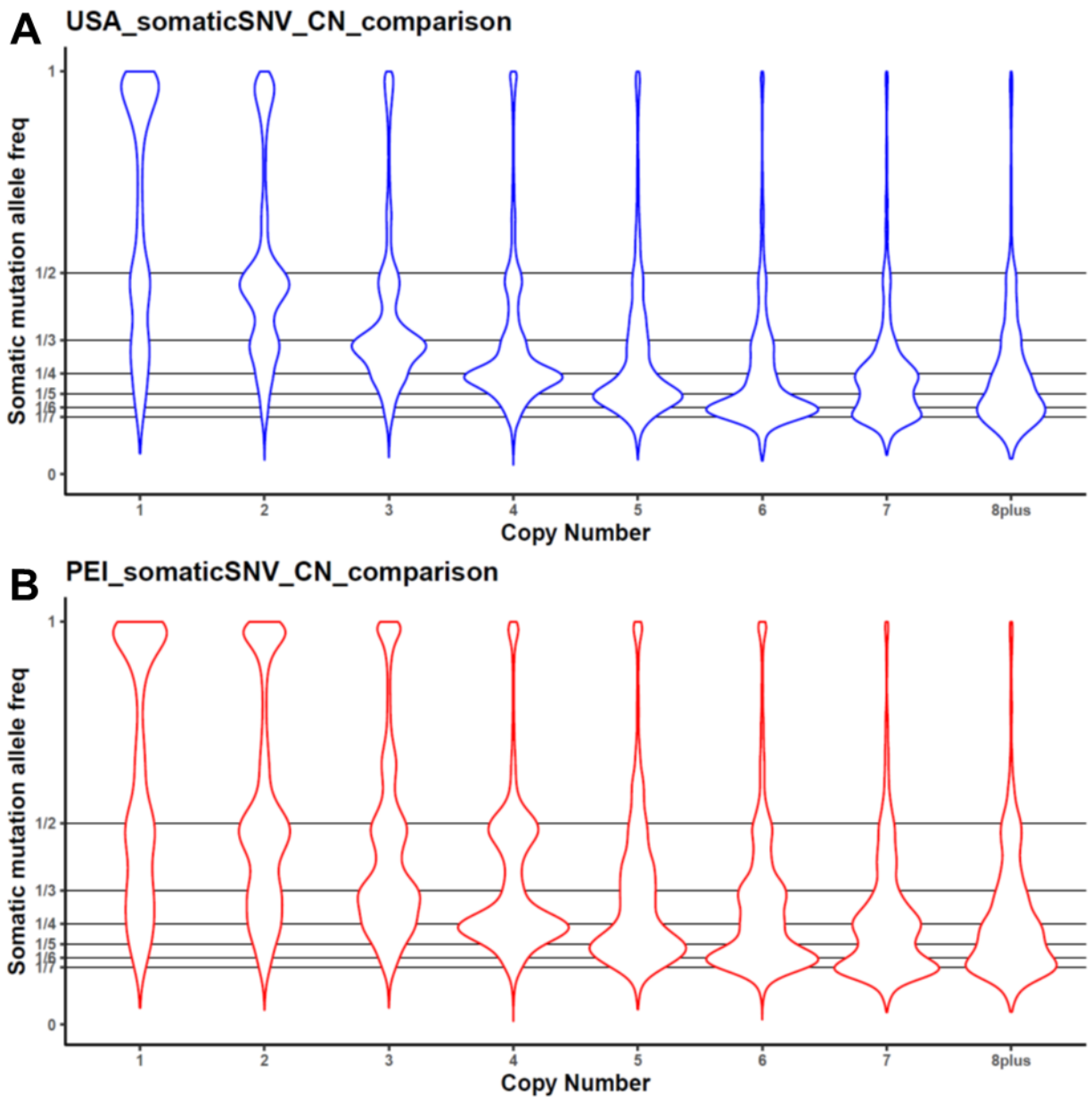
Supplementary Figure 1.10: Global dN/dS across germline and somatic SNVs

dN/dS ratios, corrected so that a ratio of 1 indicates neutrality, across sample bins. Error bars indicate 95% confidence intervals as estimated by dndscv. Sample labels along x-axis are as follows: All_H – SNVs found in all three healthy clams; Any_H – SNVs found in any healthy clam; PEI and USA_H_only – SNVs found in only that healthy clam; All_C_any_H – SNVs found in all cancer samples and at least one healthy clam; All_C_no_H – SNVs found in all cancer samples and no healthy clams; PEI/USA_only – Somatic mutations before excluding LOH regions; PEI and USA_only_nonLOH – High confidence somatic mutations outside putative LOH regions; and USA/PEI_nonLOH – Combined high confidence somatic mutations from both sub-lineages and outside putative LOH regions. After removing LOH regions, dN/dS for high confidence somatic mutations is not significantly different than 1.



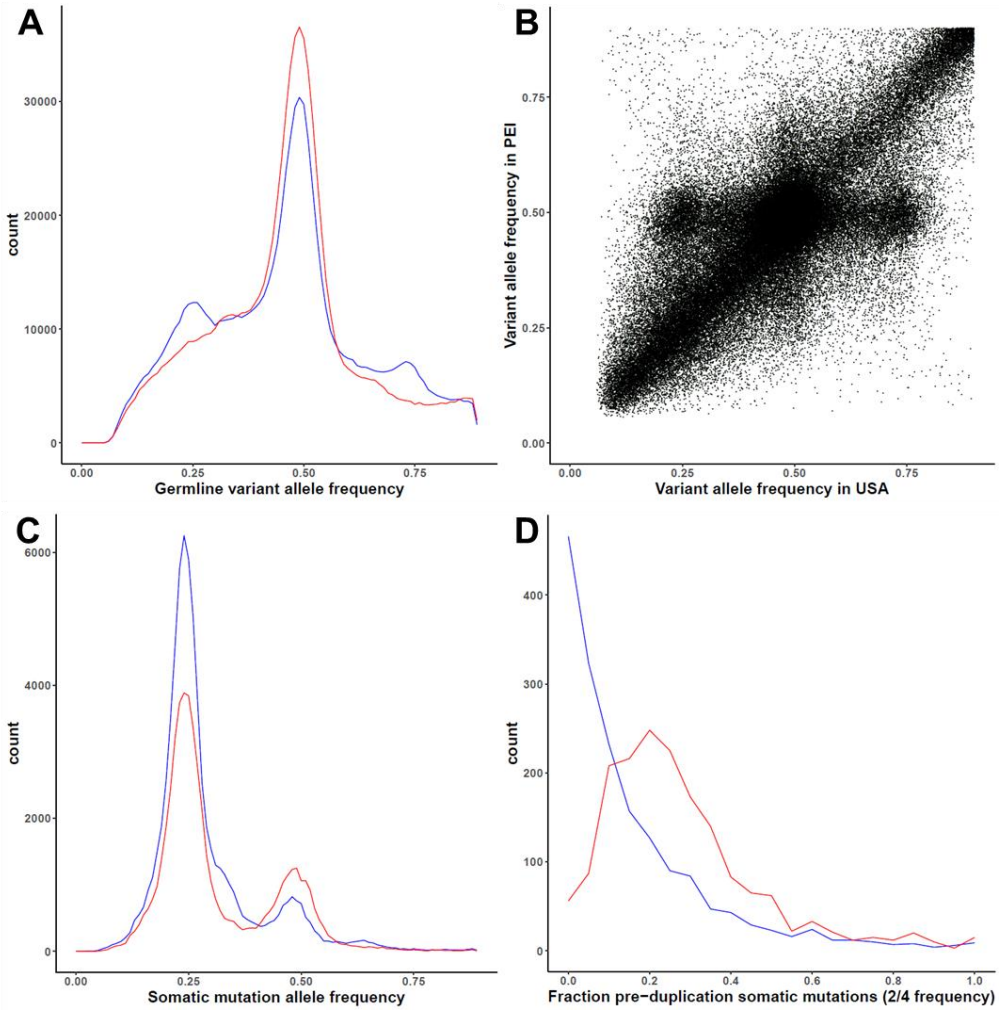
Supplementary Figure 1.11: Copy number calls agree closely within sub-lineages, but differ between sub-lineages

We called copy number across the genome in 100 kB chunks for each sample individually. Here we plot pairwise comparisons of the copy number call for each 100 kB chunk between two representative PEI BTN samples (DN08 and DF488) and two representative USA BTN samples (FFM19G1 and NYTCC9: notably, the two most distantly related USA samples). There is a close correlation (R^2>0.94) within sub-lineages (DN08 vs DF488, FFM19G1 vs NYTCC9) and a weaker correlation (R^2=0.53-0.56) when comparing between sub-lineages (DN08 or DF488 vs FFM19G1 or NYTCC9). Copy number differences between samples can be seen here as denser groupings of points around integer values that deviate from equal values along the diagonal.



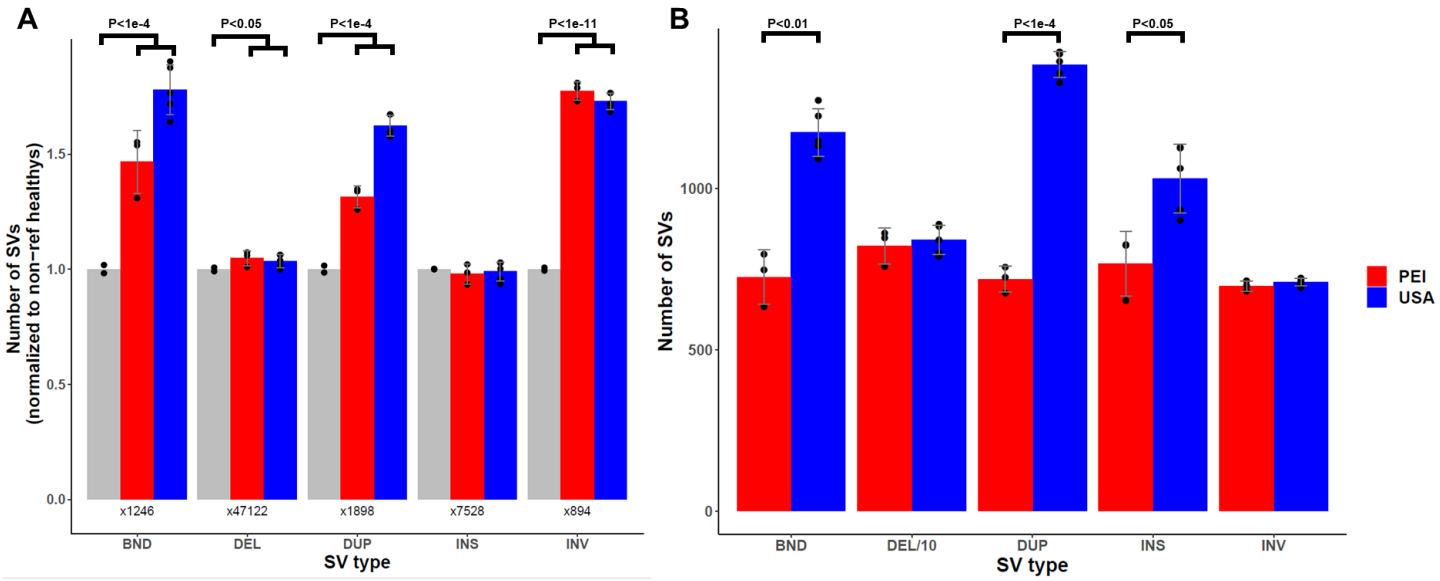
Supplementary Figure 1.12: Somatic mutation allele frequencies support copy number calls

Copy number was called across the genome and the variant allele frequencies of all high confidence somatic mutations were calculated separately for BTN from **(A)** USA) and **(B)** PEI. Violin plots show probability densities of allele frequencies of high confidence somatic mutations, divided into portions of the genome called at each copy number. The peak allele frequency in each case is distributed around the expected value of 1/copy number. In addition to the main, expected peaks for each copy number, in some cases, additional peaks can be seen that indicate somatic mutations prior to copy number gain (e.g. VAF of 0.5 in regions with CN4 that could be due to mutation followed by duplication of the region). Some minor peaks also indicate possible errors in copy number calling or allele frequency counting (e,g, VAF of 0.5 in CN3 regions). These errors could be due to lower read mapping due in polymorphic region, errors caused by repeat regions, regions spanning a CN breakpoint, among other possibilities.



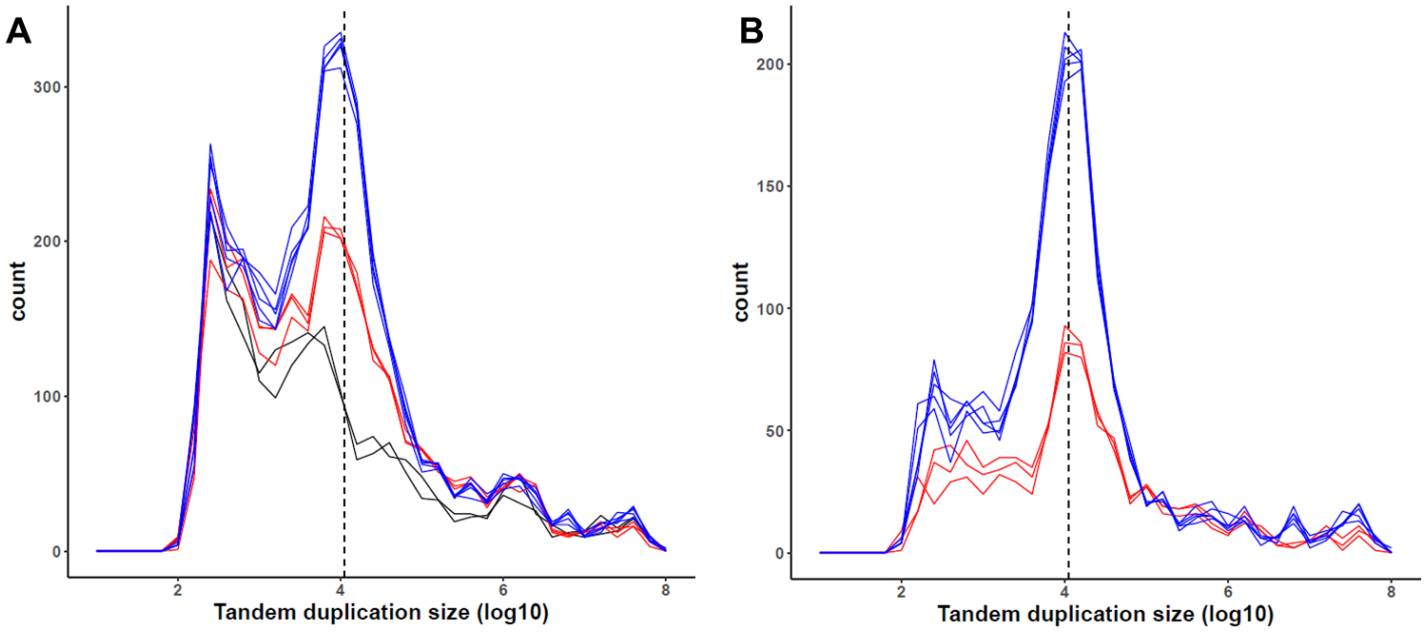
Supplementary Figure 1.13: USA and PEI sub-lineages arrived at CN4 via independent duplication events

(A) Distribution of variant allele frequencies for founder germline variants (found in all cancers and at least one healthy sample) in USA (blue) and PEI (red) sub-lineage, restricted to regions that are CN4 in both sub-lineages. **(B)** A random subset of 100,000 germline variants plotted as a scatter plot. Alleles at 1/4 and 3/4 in the USA sub-lineage are incongruent with a simple CN2>CN4 duplication. **(C)** Distribution of variant allele frequencies for high confidence somatic mutations, restricted to regions that are CN4 in both sub-lineages, showing a higher proportion of 2/4 mutations (pre-duplication SNVs) in PEI than USA. **(D)** The genome was subdivided into 100kb segments (as done for copy number analysis), and for all shared CN4 segments the plot shows the fraction of mutations in each 100kB segment that were at 2/4 frequency compared to the total amount of 2/4 and 1/4 SNVs, corresponding to mutations occurring before or after duplication of the allele, respectively. While the USA distribution peaks at 0, indicating most 100kB segments duplicated before or shortly after the USA-PEI sub-lineage split, with a low rate of duplications occurring after that time, the distribution for PEI centers around 0.2, indicating that one-fifth of mutations occurred between the USA-PEI sub-lineage split and duplication of the corresponding regions, suggesting a burst of duplications at some point in the PEI sub-lineage.



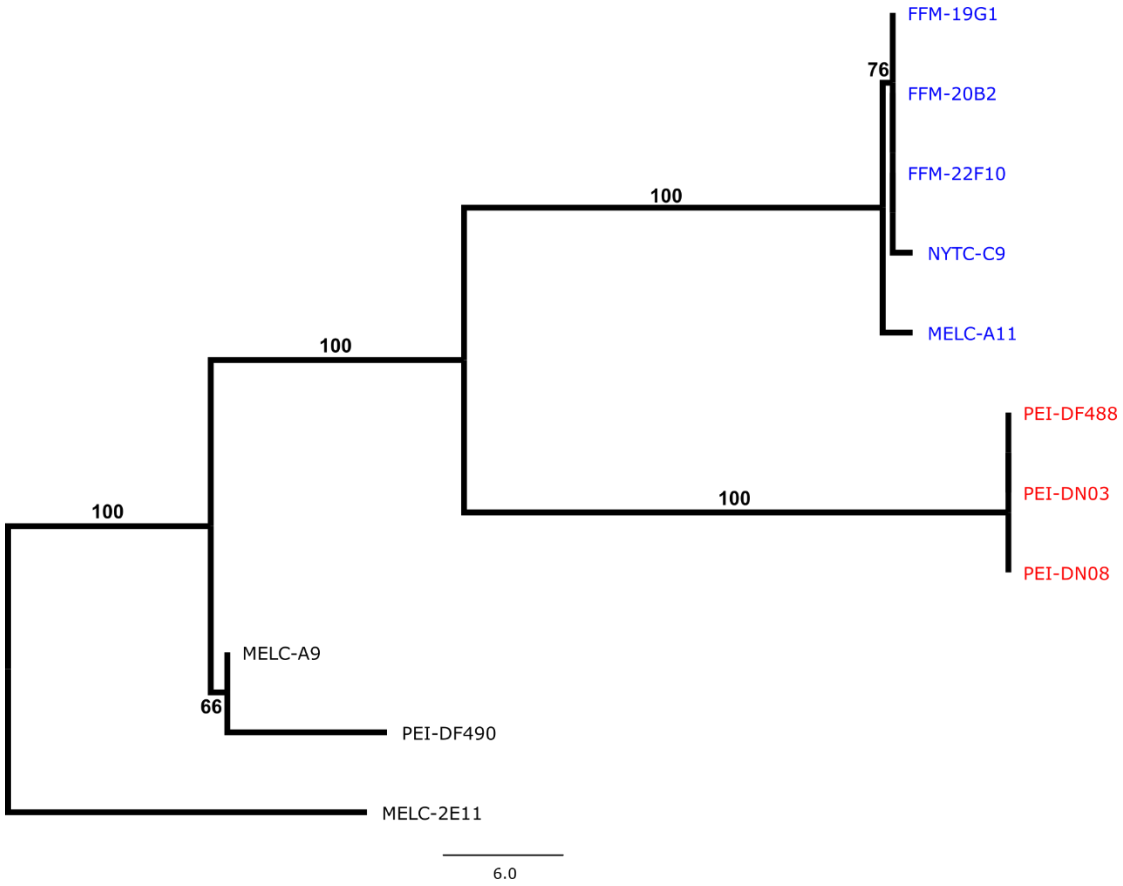
Supplementary Figure 1.14: Evidence for elevated somatic structural variants of several types

Structural variants were called in healthy and BTN samples using Delly, only including only those with precise breakpoints and excluding SVs found in the reference clam. **(A)** The number of called SVs of each type are normalized to the average number of SVs in non-reference healthy clams for each SV type (value above x-axis). Dots represent individual samples, while bars summarize averages for each group: healthy clams, PEI BTN, and USA BTN. Error bars indicate standard deviation. P-values are from two-sided unequal variance t-test between BTN samples (n=8) and non-reference healthy clams (n=2). **(B)** Number of called SVs of each type that are unique to each sub-lineage were calculated by removing SVs found in any healthy clams or in any BTN samples from the other sub-lineage. P-values are from two-sided unpaired unequal variance t-test between PEI BTN samples (n=3) and USA BTN samples (n=5). Labels follow delly abbreviations of SV types: BND = translocations, DEL = deletions, DUP = tandem duplications, INS = small insertions, INV = Inversions. Deletion counts were much higher than other SV types, so were divided by 10 in (B) for visualization ("DEL/10").



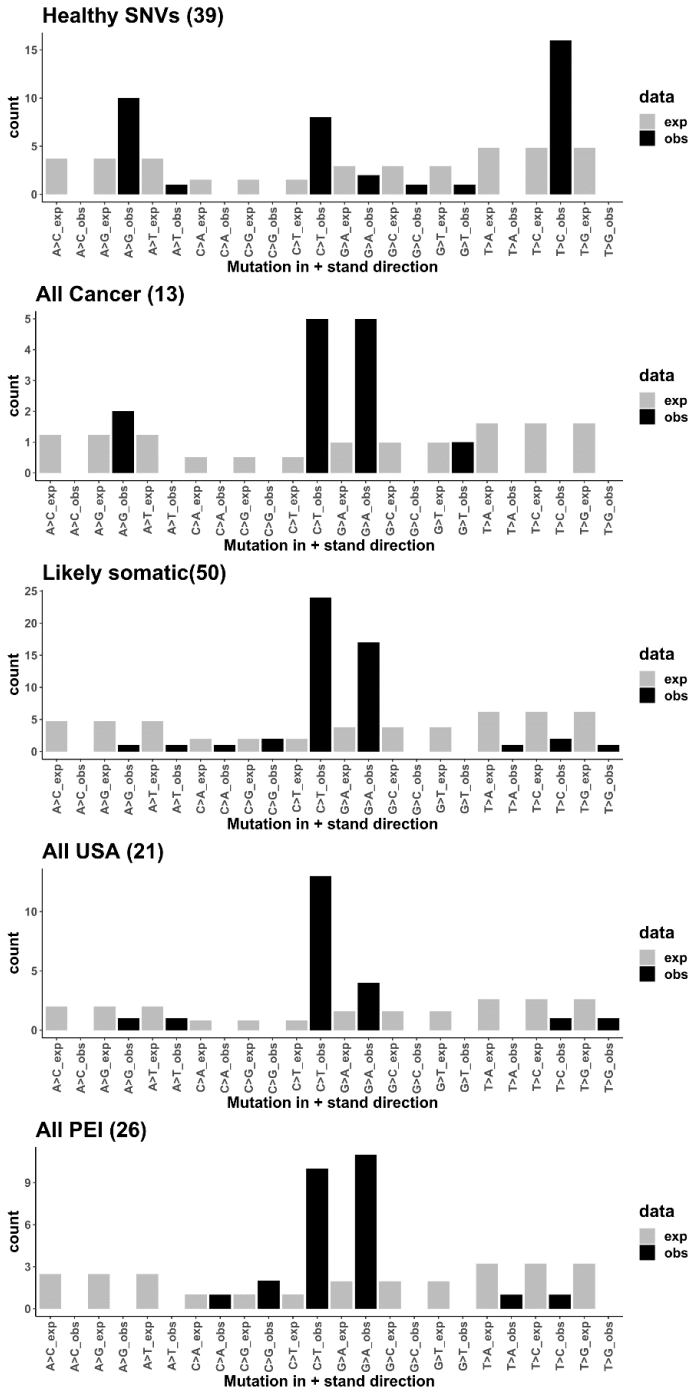
Supplementary Figure 1.15: Somatic tandem duplications are distributed around 11 kB

Plot shows the size distribution of tandem duplications in each sample, after removing SVs found in the reference clam **(A)**, and after removing SVs found in any healthy clams or in any BTN samples from the other sub-lineage **(B)**. Black lines indicate non-reference healthy clams, red lines indicate PEI BTN samples, and blue lines indicate USA BTN samples. Dashed line indicates 11 kB, the median tandem duplication size reported in a tandem duplication phenotype observed in human and mouse cancers with mutant p53 and BRCA1. We observe a bias towards an increase in similarly sized somatic tandem duplications in both sub-lineages of BTN.



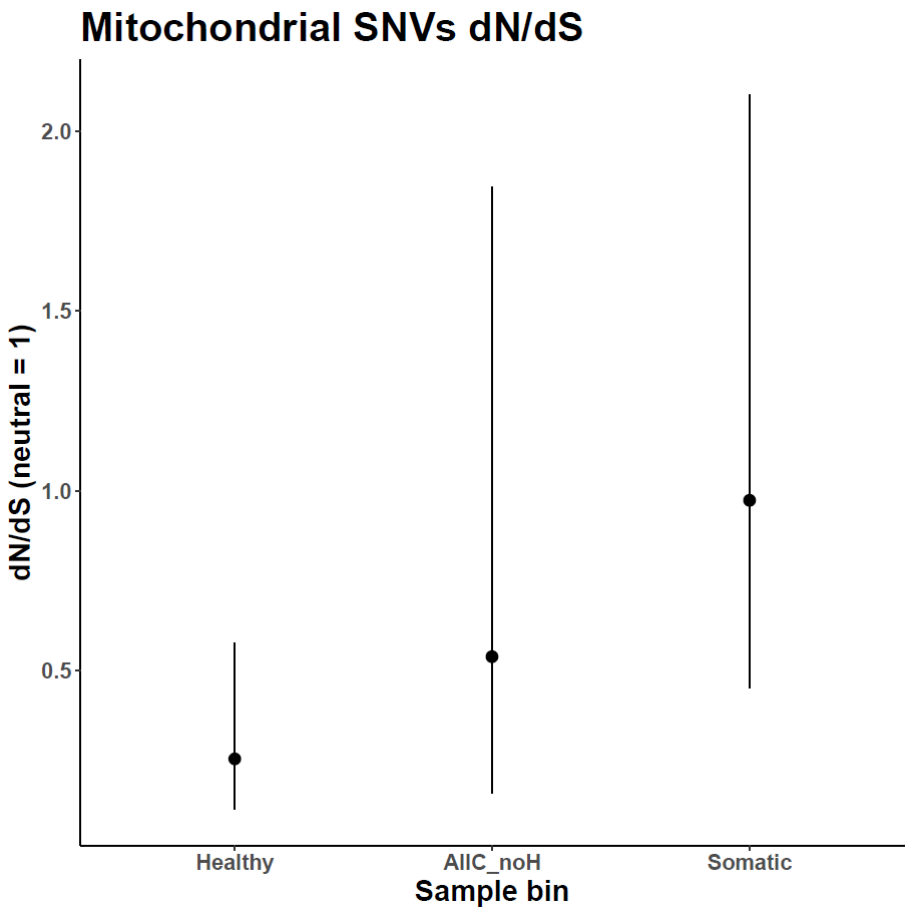
Supplementary Figure 1.16: No evidence for mitochondrial transfer

Neighbor joining tree built from variants called in all samples (170 SNVs) against the previously published *M. arenaria* reference mitogenome (excluding the repeated region) for USA MarBTN samples (blue), PEI MarBTN samples (red) and healthy clams (black). Bootstrap values in support of each clade are included on the preceding branch (bootstraps under 50 are not shown). The phylogenetic relationship generally reflects that built from genomic SNVs (i.e., monophyletic MarBTN group with separate USA and PEI sub-lineages). The phylogeny within the USA sub-lineage deviates from that built from the nuclear genome, but only three SNVs are variable within the USA sub-lineage: one SNV unique to NYTC-C9 and two SNVs unique to MELC-A11. This causes the other samples to cluster more often with NYTC-C9 due to only one difference (versus two versus MELC-A11), but this relationship is still compatible with the USA branch structure from the nuclear phylogeny



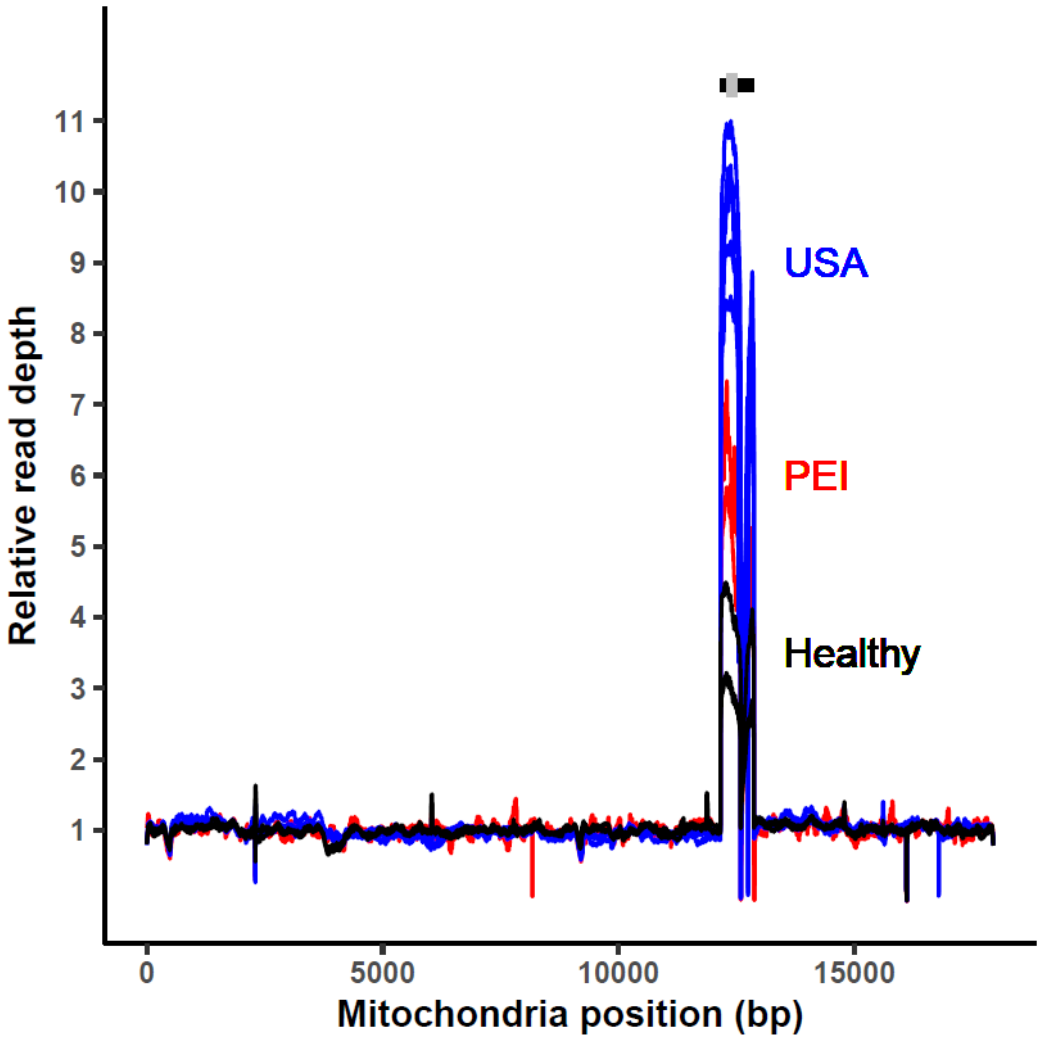
Supplementary Figure 1.17: Mitochondria are enriched for transitions in healthy clams, and C>T specifically in somatic mutations

Observed SNVs (black) compared with expected counts estimated from nucleotide frequencies of the *M. arenaria* mitogenome and assuming equal mutation probability. This calculation was not collapsed to the usual 6 mutation types due to the imbalance of nucleotides in mitochondrial genomes (unequal frequencies of G/C and A/T). Likely somatic refers to SNVs found in a subset of BTN samples, while All USA and All PEI refer to SNVs found in all individuals from that sub-lineage, but not the other sub-lineage.



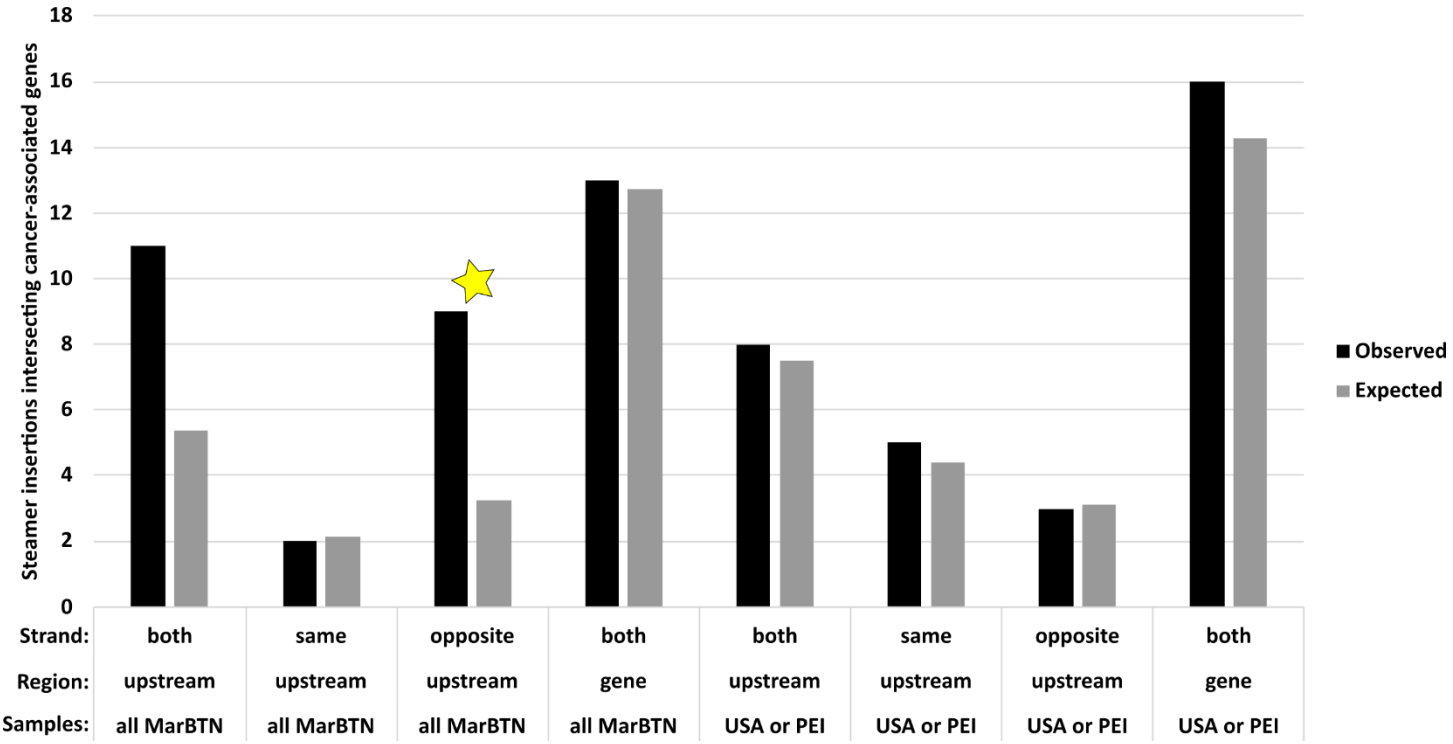
Supplementary Figure 1.18: Mitochondria are under relaxed selection in BTN

dN/dS ratios, where a ratio of 1 indicates neutrality, were calculated for mitochondrial SNVs found in healthy clams, all BTN samples but not healthy clams, and likely somatic mutations (those found in a subset of BTN samples). Error bars indicate 95% confidence intervals as estimated by dndscv and are quite large, due to the low number of mitochondrial SNVs.



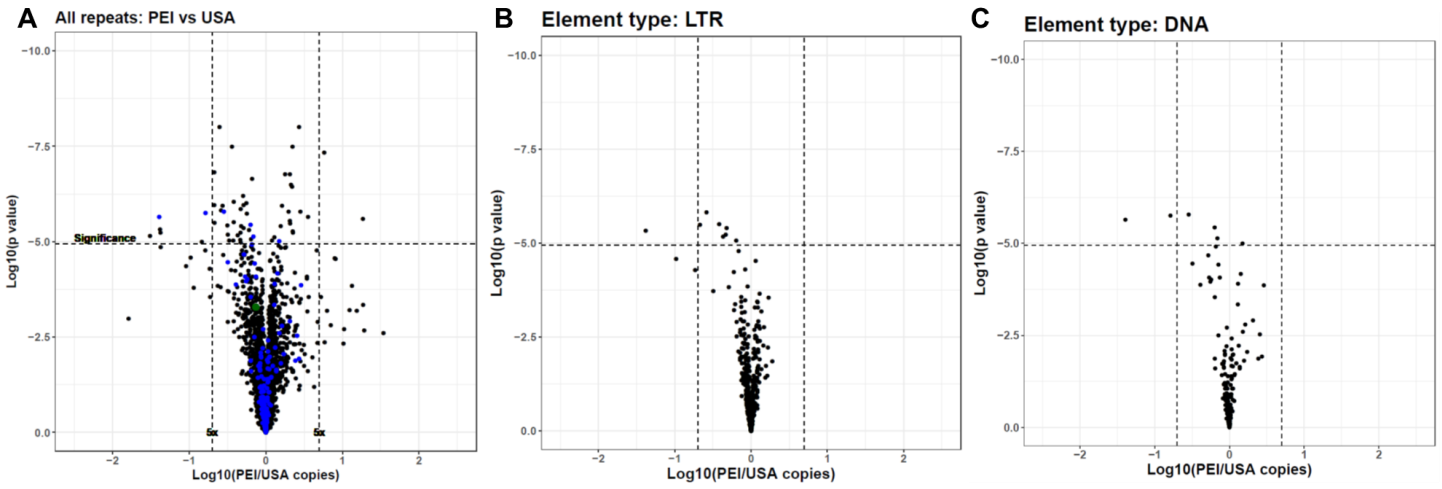
Supplementary Figure 1.19. Somatic tandem duplications in mitochondrial D-loop

Read depth across the mitochondrial genome for healthy clams (black), PEI MarBTN (red) and USA MarBTN (blue), normalized to mean depth outside D-loop. Bars above indicate the D-loop region (12,164-12,870 bp, black) and the region used to estimate duplicated region copy number (12,300-12,500 bp, grey), as shown in Fig 3F.



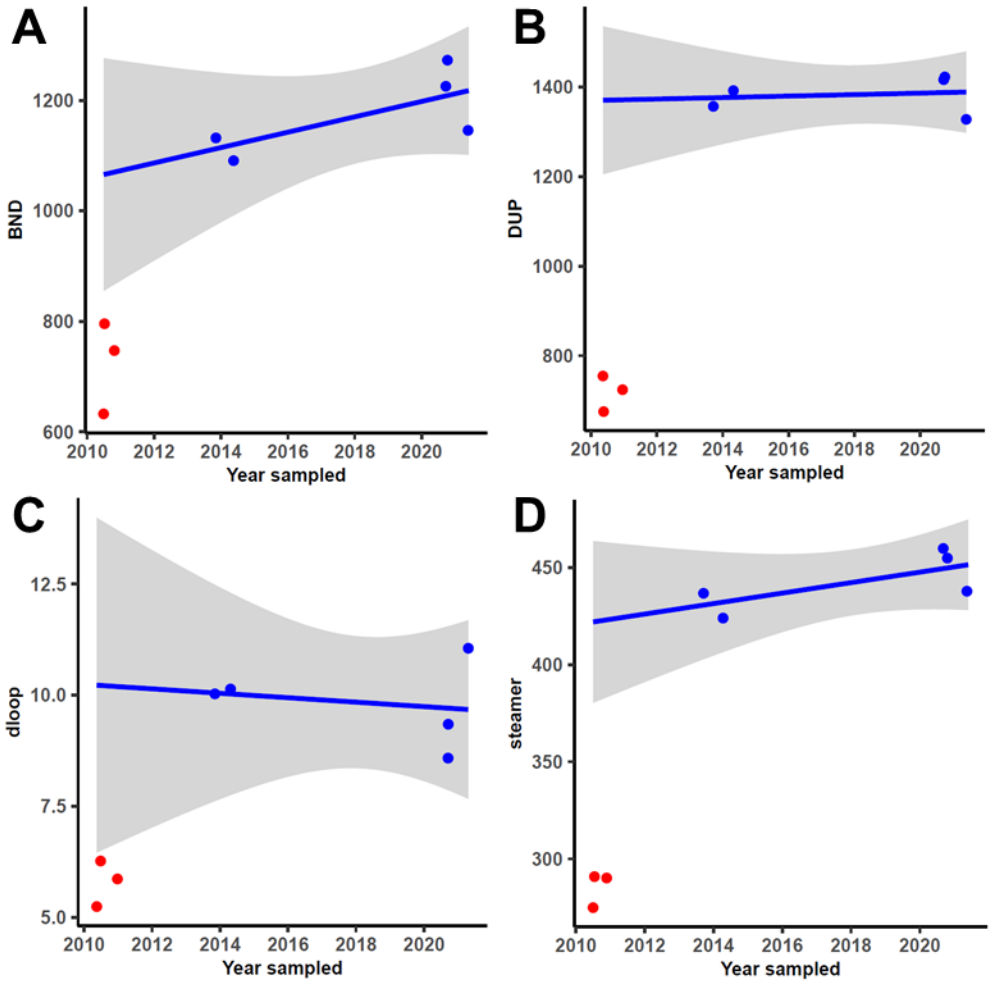
Supplementary Figure 1.20: The *Steamer* retrotransposon inserts upstream of cancer-associated orthologs in the opposite direction more often than expected

We conducted a BLASTP search for the 729 cancer-associated genes in the COSMIC database and found hits in 5,430 of the 38,609 predicted *M. arenaria* genes (14%). If there is not selection for insertion near these genes, we would expect 14% of *Steamer* insertions with a *M. arenaria* gene to intersect with these genes. We counted the number of steamer insertions in genes ("gene") and in the 2 kB upstream genes ("upstream") for early steamer insertions in the lineage trunk ("all MarBTN") and after the divergence of the sub-lineages ("USA or PEI"). We plotted these counts (black) against that expected by chance (grey). Counts match expected closely for late insertions (in only the USA or PEI sub-lineage – right side of plot), either upstream genes or within them, but were higher than expected for early insertions. We further divided upstream insertions by whether the steamer insertion was in the same strand/direction as the gene or opposite, to compare with counts regardless of directionality ("both"). The early insertion bias to insert upstream cosmic genes can be fully explained by a bias to insert in the opposite strand (yellow star), here with 9/23 (39%) of the genes being cancer associated (would expect 3/23: Chi-squared Bonferroni-corrected p-value = 0.004)



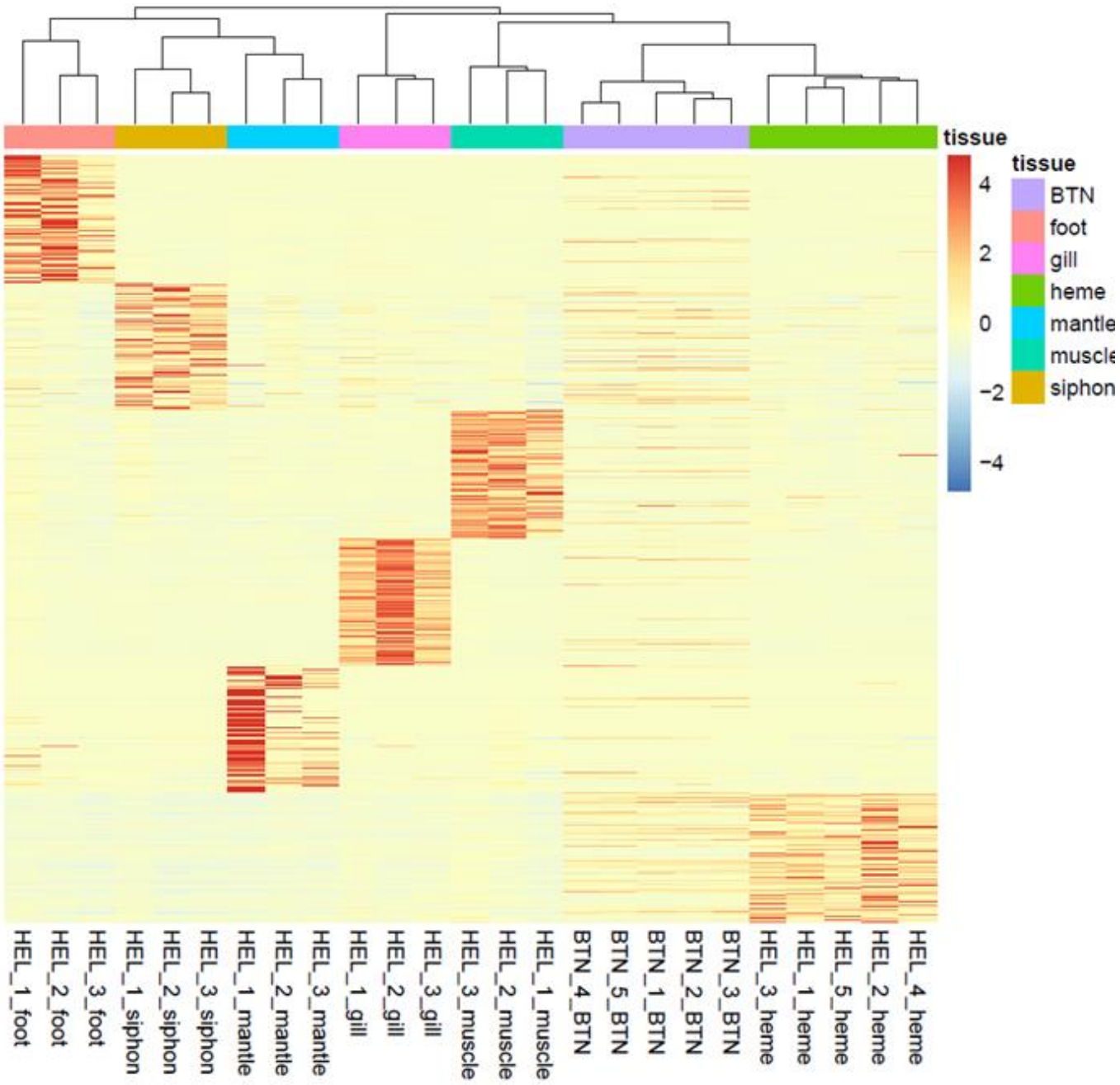
Supplementary Figure 1.21: More TEs are expanded in USA vs PEI, particularly LTR elements

Volcano plot shows estimated copy number of each TE, comparing copy number in MarBTN from PEI with USA for all TE types **(A),** LTR elements **(B**), and DNA transposons **(C)**. TEs more highly amplified in PEI MarBTN are to the right and TEs amplified more highly in USA MarBTN are to the left. Dashed lines correspond to significance threshold (p=0.05, Bonferroni corrected) and 5-fold differences. **(A)** DNA transposons are labeled in blue and *Steamer* is labeled in green. Eight LTR retrotransposons and five DNA transposons are significantly amplified in the USA sub-lineage compared to the PEI sub-lineage, while no identified LTR retrotransposons and a single DNA transposon TEs are significantly amplified in the PEI sub-lineage compared to the USA sub-lineage.



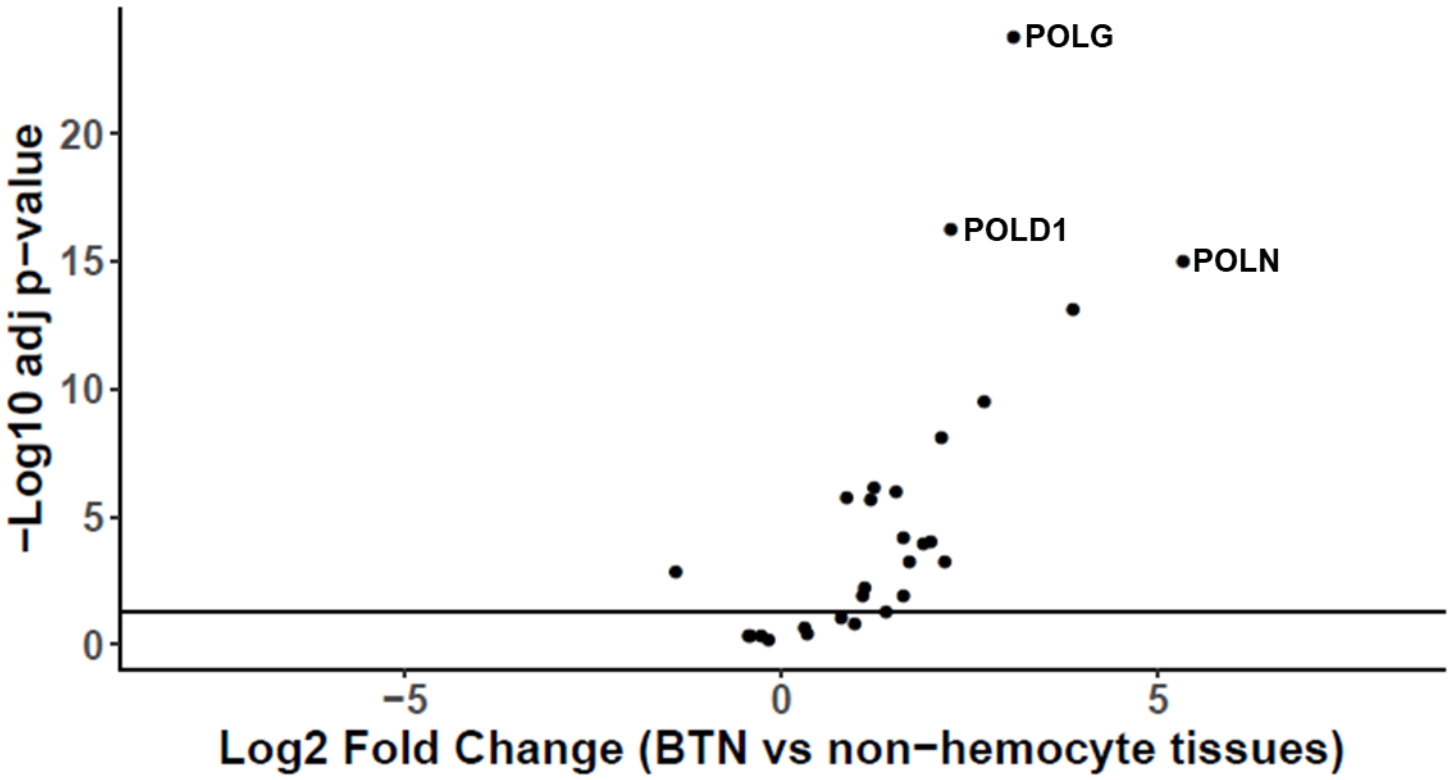
Supplementary Figure 1.22: The USA sub-lineage has higher amounts of structural mutation

Number of **(A)** translocations (BND) and **(B)** tandem duplications (DUP) since the divergence of the sub-lineages, **(C)** copies of the mitochondrial D-loop, and **(D)** total *Steamer* insertions per sample, each plotted against sampling date. Linear regression (blue line) and 95% confidence interval (grey) were calculated for the USA samples. No regression was statistically significant. No PEI samples fell within 95% confidence intervals of regression lines, indicating the higher mutation counts in USA samples cannot be explained by the later sampling of USA samples.



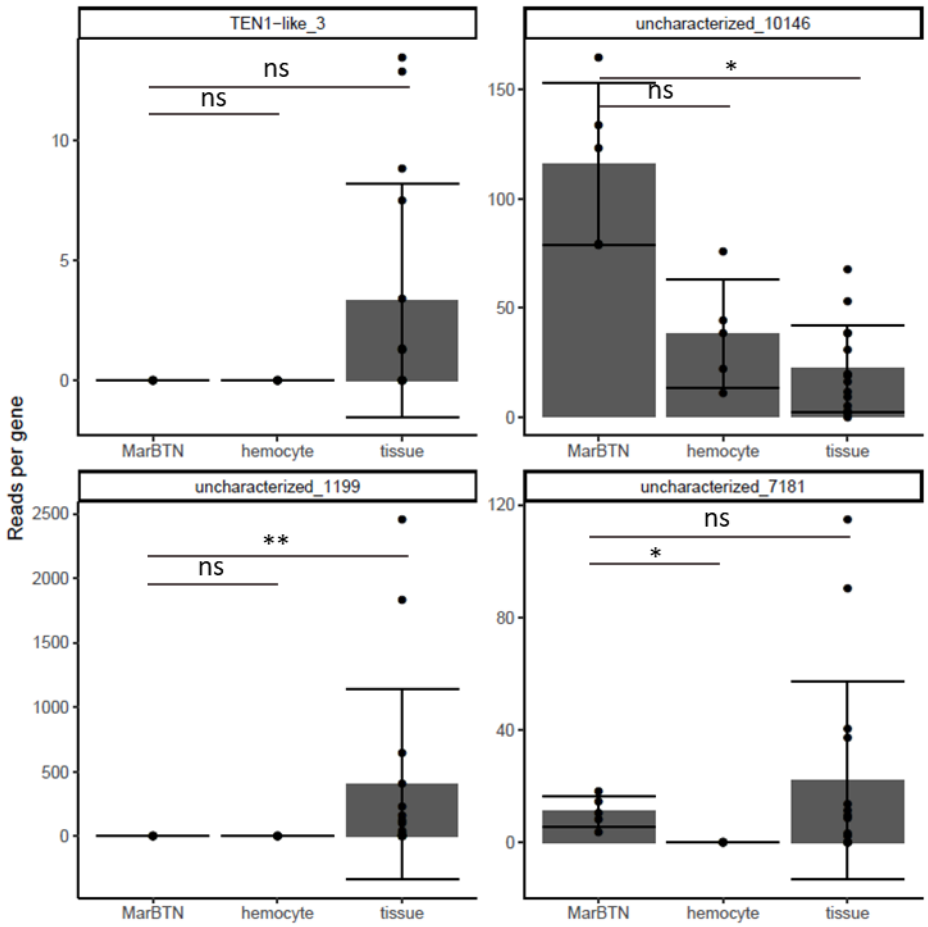
Supplementary Figure 1.23: Hierarchical clustering supports hemocyte origin of MarBTN

Hierarchical clustering of all RNA sequenced 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.



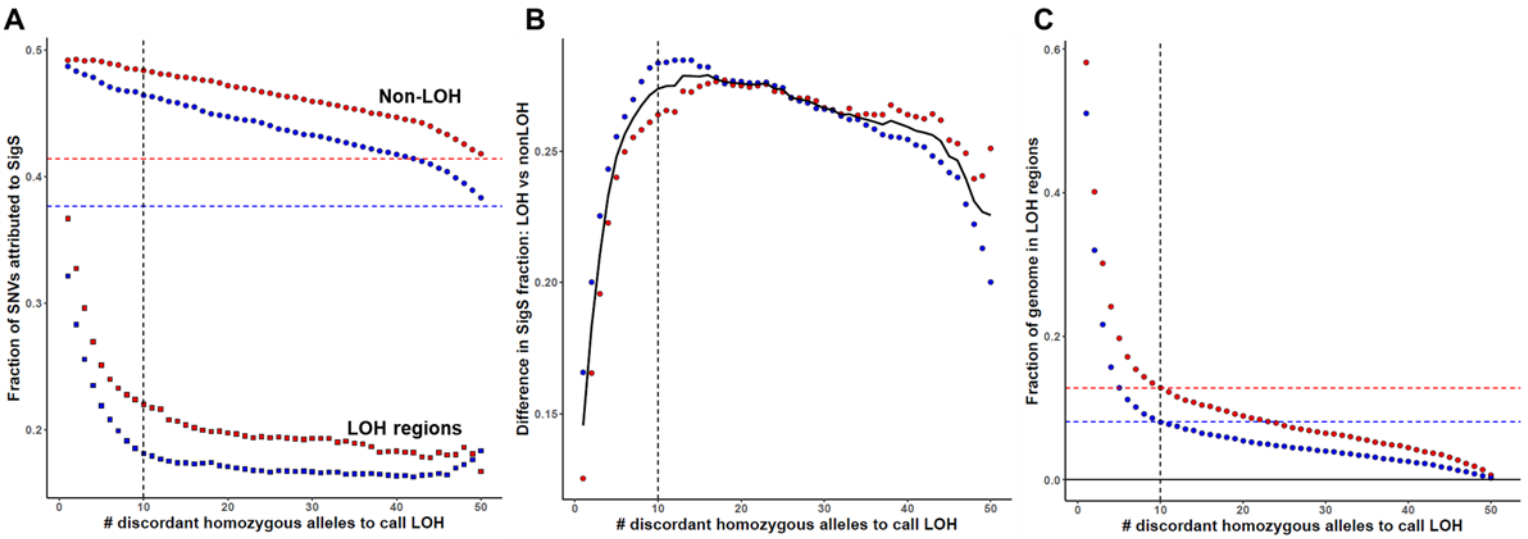
Supplementary Figure 1.24: Differential expression of polymerase genes for MarBTN vs tissue

Volcano plot of polymerase genes expression (n=28) for MarBTN (n=5) compared with non-hemocyte tissues (n=15: 5 tissues for 3 clams).



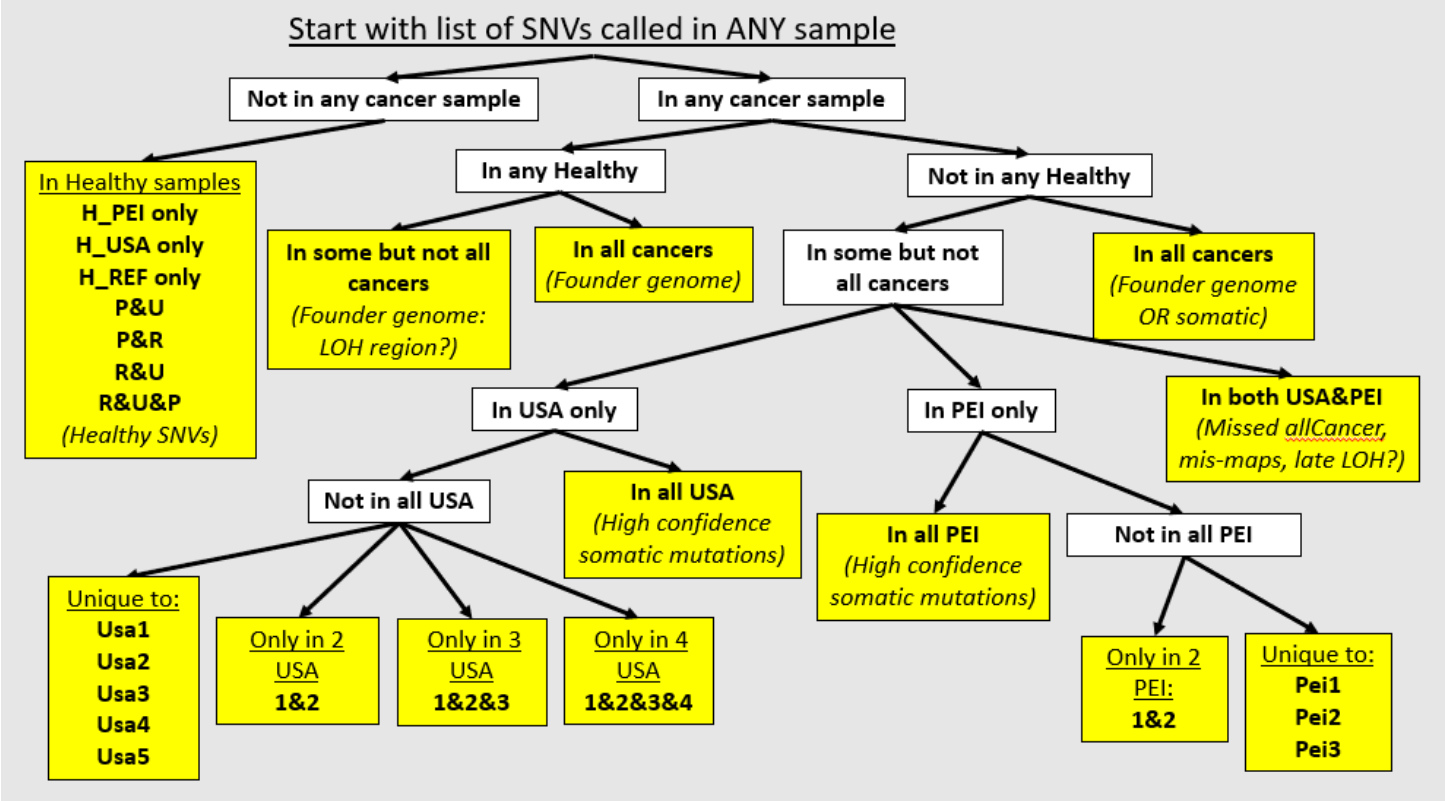
Supplementary Figure 1.25: Expression of genes with positive dN/dS

Normalized expression, in reads per gene, of four genes with detectable positive dN/dS for MarBTN (n=5), hemocytes (n=5), and non-hemocyte tissues (n=15: 5 tissues for 3 clams). Error bars display standard deviation, differential expression comparison results displayed as * = p<0.01, ** = p<1e-7, ns = not significant.



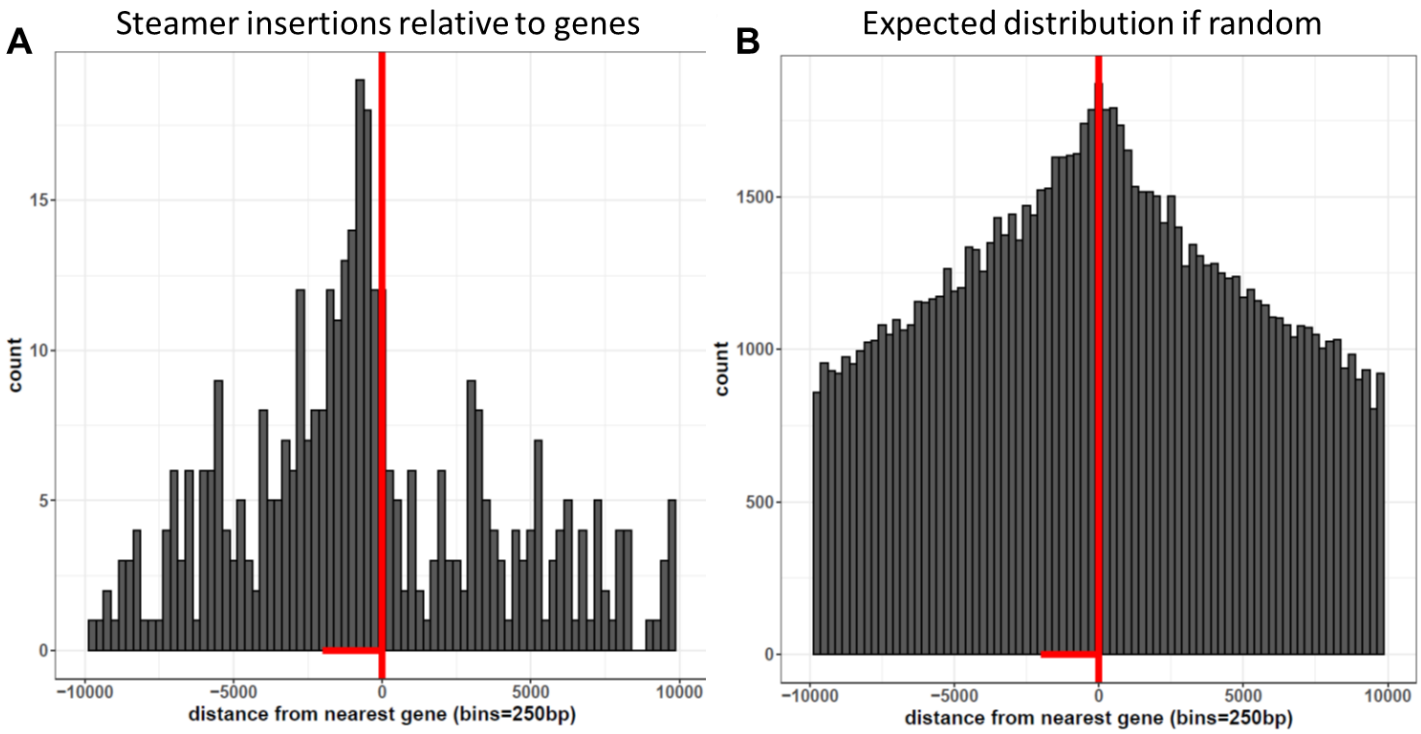
Supplementary Figure 1.26: Calibrating LOH calling thresholds

(A) We used various thresholds of stringency to call LOH across the genomes of each sub-lineage based on the number of shared SNVs that were homozygous in one sub-lineage but heterozygous in the other across a window of 50 SNVs (x-axis). After calling LOH, we calculated the fraction of likely somatic mutations attributed to signature S in LOH (squares) and non-LOH (circles) (y-axis). Values are shown separately for the BTN subgroups from USA (blue) and PEI (red). Vertical dashed line indicates the threshold used for LOH-calling. Horizontal dashed lines indicated baseline signature S fractions without LOH region removal. **(B)** Plot of the difference between non-LOH and LOH regions as shown in (A) (calculated by subtracting the square from the circle). Black line shows the average difference, which peaks around the threshold used (10). **(C)** Proportion of the genome that is called LOH for each sub-lineage based on calling threshold. Dashed lines indicate the fraction of the genome called as LOH for each sub-lineage for the final threshold used.



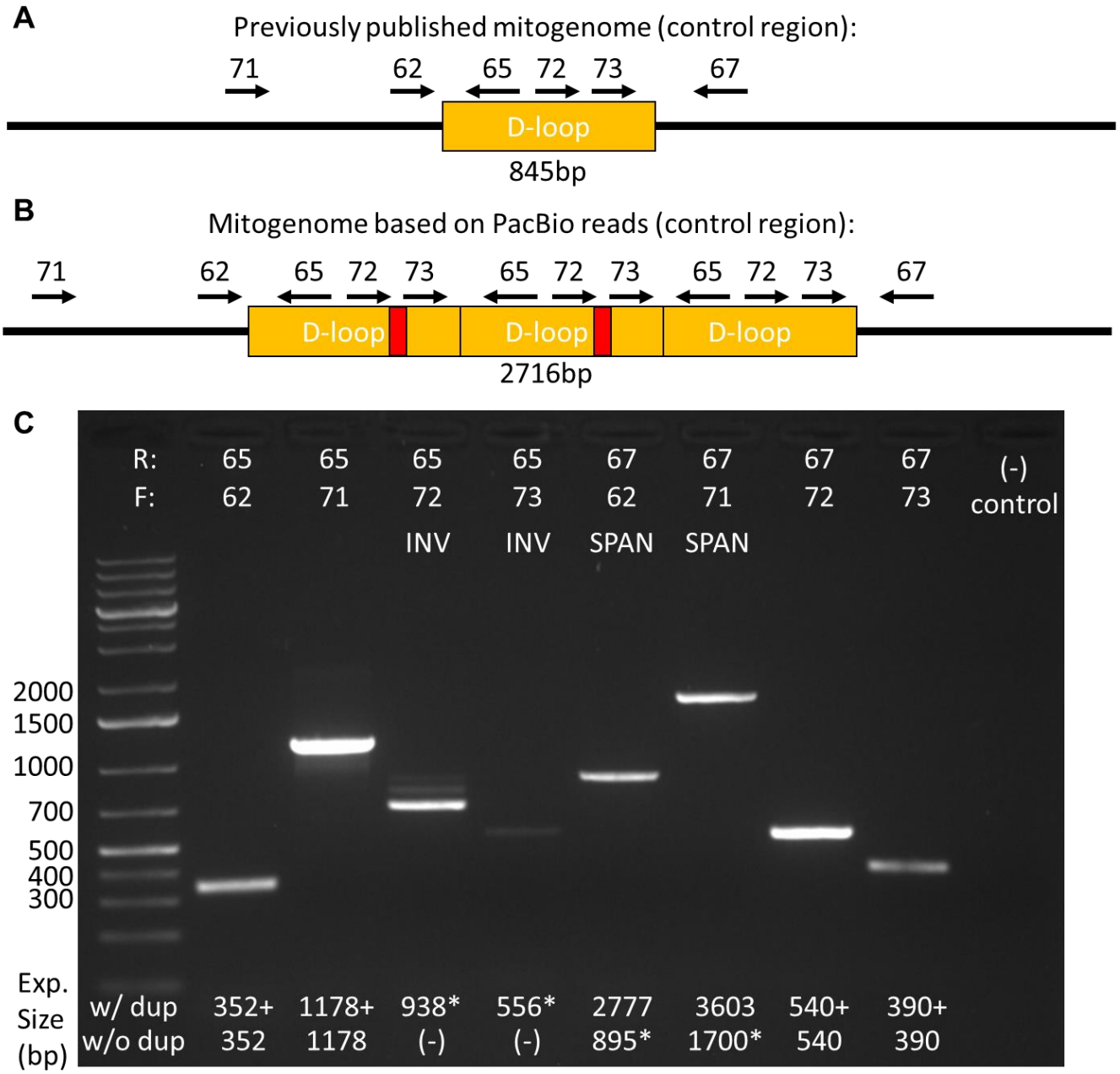
Supplementary Figure 1.27: SNV binning strategy for *de novo* signature extraction

Flowchart of our strategy to separate SNVs into bins for *de novo* signature extraction, based on which sample(s) each SNV was called in. Many of these bins were also used in other analyses, as indicated in the manuscript. The starting point refers to a vcf file of every SNV that was called in at least one of the eleven sample (three healthy, eight cancer) sequenced in this study. Bins highlighted in yellow indicate non-overlapping SNV bins used to for signature extraction.



Supplementary Figure 1.28: *Steamer* preferentially inserts upstream genes

(A) The histogram shows the distance to nearest gene for *Steamer* insertions found in any cancer sample (n=550). If an insertion was within an annotated gene, the distance to the next nearest insertion was used. 0 (vertical red line) corresponds to the first or last nucleotide of the annotated gene for when the insertion is upstream (negative) or downstream (positive) relative to the gene, respectively. Horizontal red segment highlights 2 kB upstream genes with elevated *Steamer* insertions. **(B)** The histogram shows a distribution of randomly generated insertion sites (n=224,134) based off the observed read mapping in the genome assuming insertions are random.



Supplementary Figure 1.29: PCR validation of mitochondrial repeat in health clam

(A) Schematic of the control region of the *M. arenaria* control region in the previously published mitogenome with a single d-loop copy and placement of primers (not to scale). **(B)** Schematic of the proposed mitochondrial genome with three d-loop copies and G-rich insertions and placement of primers. **(C)** PCR results. Primer pair combinations are listed on top and expected sizes are listed on bottom. Amplicon sizes from primers spanning the D-loop (67 with 62/71) support a single copy of the D-loop. However, we suspect this is a result of recombination and selection for the smaller product and loss of the G-rich insertions. Inverse PCR with outward-facing primers (65 with 72/72) indicates a tandem duplication allowing outward-facing primers to amplify. The inverse primers spanning the G-rich insertion (65 with 72) has a dim band at expected size, but two brighter bands at smaller sizes.

1.9 Supplemental tables

Supplementary Table 1.1: Improved genome contiguity and completeness

Genome	Sequencing	Source	Length (Gb)	GC (%)	Repeat (%)	Scaffolds	Contigs	Scaffold N50 (kB)	Contig N50 (kB)	BUSCO score (metazoa)
Mya.genome.v1.01	Illumina, low cov PacBio	2 clams, Plachetzki et al.	1.32	34.98	35.54	152,330	226,958	14.7	10.6	C:71.4%[S:55.6%,D:15.8%],F:16.7%,M:11.9%,n:954
Mar.3.1.1	10x	MELC-2E11, This paper	1.29	35.27	39.64	1,029,422	1,100,210	22.1	11.0	C:73.9%[S:62.8%,D:11.1%],F:13.1%,M:13.0%,n:954
Mar.3.4.6.p1	**PacBio, HiC, 10x**	**MELC-2E11, This paper**	**1.22**	**35.32**	**41.72**	**17**	**539**	**58,023**	**3,381**	**C:94.9%[S:92.5%,D:2.4%],F:1.2%,M:3.9%,n:954**

Supplementary Table 1.2: List of whole genome sequenced samples

Name	Map code	Alternate aliases	Healthy/BTN	Date sampled	Location
MELC-2E11	Href		Healthy	6/1/2018	Larrabe Cove, Machiasport, ME, USA
MELC-A9*	Husa		Healthy	9/18/2013	Larrabe Cove, Machiasport, ME, USA
PEI-DF490*	Hpei	Dfar490	Healthy	5/1/2010	Dunk Estuary, PEI, Canada
PEI-DF488*	pei1	DF-488, Dfar-488	BTN	10/28/2010	Dunk Estuary, PEI, Canada
PEI-DN03*	pei2	DN-HL03, Dnear-HL03	BTN	5/1/2010	Dunk Estuary, PEI, Canada
PEI-DN08*	pei3	Dnear-08	BTN	5/1/2010	Dunk Estuary, PEI, Canada
FFM-19G1	usa1		BTN	8/31/2020	Brunswick, ME, USA
FFM-20B2	usa2		BTN	8/31/2020	Brunswick, ME, USA
FFM-22F10	usa3		BTN	3/31/2021	Waldoboro, ME, USA
MELC-A11*	usa4		BTN	9/18/2013	Larrabe Cove, Machiasport, ME, USA
NYTC-C9*	usa5		BTN	4/12/2014	Long Island Northshore, NY, USA

*previously reported in Metzger *et al.* 2015

Supplementary Table 1.3: dN/dS positive selection hits table

Gene annotation information			*Somatic mutations*				*Healthy clam SNVs*			
annotated_name	**AA #**	**top bivalve blastp hit**	**n_syn**	**n_mis**	**wmis_cv**	**qmis_cv**	**n_syn**	**n_mis**	**wmis_cv**	**qmis_cv**
uncharacterized_1199	312	no hits	2	30	35.96	2.42E-08	0	1	0.06	9.22E-07
TEN1-like_3	112	receptor-type tyrosine-protein phosphatase mu-like (Mizuhopecten yessoensis)	2	12	34.87	1.64E-04	13	14	0.43	4.53E-02
uncharacterized_7181	80	no hits	1	9	105.56	1.64E-04	3	3	0.57	5.11E-01
uncharacterized_10146	211	uncharacterized protein (Crassostrea virginica)	4	15	11.32	9.25E-03	4	0	0.00	1.06E-05

AA #: gene length in amino acids

n_syn: number of synonomous SNVs

n_mis: number of misense SNVs

wmis_cv: dN/dS ratio after corrections

qmis_cv: significance after corrections

Supplementary Table 1.4: Primers for inverse PCR

ID code	Oligo name	Sequence	Direction
SHO_062	SHO_062_MyaMT_dloop_F1	TACGAGCAAAAGCCGTTCCT	F
SHO_065	SHO_065_MyaMT_dloop_R1	CCCATAACGCCCGATTTTGC	R
SHO_067	SHO_067_MyaMT_dloop_R3	AACCGAGCTGACCTCATTCA	R
SHO_071	SHO_071_MyaMT_dloop_F7	TCCTGTGTGCCGAAAGAGTC	F
SHO_072	SHO_072_MyaMT_dloop_F8	CGTGGCGGGAGTATACAGTG	F
SHO_073	SHO_073_MyaMT_dloop_F9	GGAGAGGGGAGAGGGGATTT	F

Supplementary Table 1.5. Cancer-related genes with steamer antisense upstream insertion bias

gene name	cancer hit	Cancer hit description from genecards.org
TADBP-like_1	MSI2	RNA-binding protein that is a member of the Musashi protein family
ANR10-like_1	ANK1	Ankyrins are a family of proteins that link the integral membrane proteins to the underlying spectrin-actin cytoskeleton
DHSDB-like_1	SDHD	This gene encodes a member of complex II of the respiratory chain, which is responsible for the oxidation of succinate
ZN865-like_1	PRDM16	Binds DNA and functions as a transcriptional regulator.
ZN878-like_1	ZNF429	Predicted to enable DNA-binding transcription factor activity, RNA polymerase II-specific and RNA polymerase II cis-regulatory region sequence-specific DNA binding activity
KPSH1-like_1	PRKACA	This gene encodes one of the catalytic subunits of protein kinase A, which exists as a tetrameric holoenzyme with two regulatory subunits and two catalytic subunits, in its inactive form
BTAF1-like_1	SMARCA4	The protein encoded by this gene is a member of the SWI/SNF family of proteins and is similar to the brahma protein of Drosophila
DAPP1-like_1	PLCG1	The protein encoded by this gene catalyzes the formation of inositol 1,4,5-trisphosphate and diacylglycerol from phosphatidylinositol 4,5-bisphosphate
PTPRA-like_9	PTPRT	The protein encoded by this gene is a member of the protein tyrosine phosphatase (PTP) family

Supplementary Table 1.6: List of RNA sequenced samples

Name	Alternate aliases	Healthy/BTN	Tissues	Date sampled	Location
MELC-2E11	HEL_1	Healthy	hemocytes, foot, gill, adductor muscle, mantle, and siphon	6/1/2018	Larrabe Cove, Machiasport, ME, USA
FFM-27C11	HEL_2	Healthy	hemocytes, foot, gill, adductor muscle, mantle, and siphon	10/17/2022	Friendship, ME, USA
FFM-27H7	HEL_3	Healthy	hemocytes, foot, gill, adductor muscle, mantle, and siphon	10/17/2022	Friendship, ME, USA
FFM-15G1	HEL_4	Healthy	hemocytes only	6/30/2020	Brunswick, ME, USA
FFM-16B11	HEL_5	Healthy	hemocytes only	6/30/2020	Brunswick, ME, USA
FFM-28E5	BTN_1	BTN	BTN isolate	10/13/2022	Perry, ME, USA
FFM-28E6	BTN_2	BTN	BTN isolate	10/13/2022	Perry, ME, USA
FFM-28E7	BTN_3	BTN	BTN isolate	10/13/2022	Perry, ME, USA
FFM-15G11	BTN_4	BTN	BTN isolate	6/30/2020	Brunswick, ME, USA
FFM-16G1	BTN_5	BTN	BTN isolate	6/30/2020	Brunswick, ME, USA

Chapter 2. Soft-shell clam transmissible cancer transcriptome reveals downregulation of immune processes

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

Affiliations:

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

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

2.1 Abstract

Transmissible cancers are unique instances in which cancer cells escape their original host and spread through a population as a clonal lineage, evading expected barriers such as non-self immune rejection. 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 parasitic cancer lineage, observing extensive differential expression of genes and gene pathways. We see upregulation of genes involved with genotoxic stress response, ribosome biogenesis and RNA processing, while downregulation of genes involved in tumor suppression and immune response, including the top downregulated gene in an independent transmissible cancer in mussels. We also observe evidence that catastrophic genome instability affects the cancer transcriptome via gene fusions, copy number variation, and transposable element insertions. Finally, we observe a consistent transcriptomic response and resumed expression of a subset of immune pathways after extended exposure to seawater, the presumed host-to-host transmission vector. Overall, this study reveals multiple mechanisms this lineage may have evolved to successfully spread through the soft-shell clam population as a contagious cancer.

2.2 Introduction

The maximum life span of a cancer is typically limited by the lifespan of its host, with cancer developing during that host's life and surviving until its death. However, a small number of transmissible cancers in Tasmanian devils[4,112], dogs[2], and bivalves[6–10] 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[113]. Transmission in devils occurs during biting and engraftment of cells on the new host's facial wounds[4], in dogs the cancer is a sexually transmitted genital tumor[2], and in bivalves cancer cells are presumed to transfer in seawater via filter feeding[17,18,113]. 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[114–116]. 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[2,117]. In bivalves, it has been assumed that the lack of an adaptive immune system contributes to the inability to reject non-self cancer cells[113], though the specific mechanisms bivalve cancers might have evolved to escape rejection by host innate immune systems remain unknown.

Bivalve transmissible neoplasia (BTN) has been identified in nine bivalve species[6–10], indicating that bivalves may be particularly susceptible to cancer transmission. One species that is affected 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[6]. In a previous study we looked at *M. arenaria* BTN (MarBTN) genome sequences and were able to determine that the cancer was highly mutated with an unstable genome[118]. 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 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.

2.3 Results

2.3.1 MarBTN transcriptome

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. In previous work[118], we 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. To improve this annotation, we submitted the same input data (genome and RNAseq) to the 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 (**Table 2.1**), so we proceeded to analyze MarBTN gene expression using the NCBI annotation.

We sequenced RNA from five MarBTN isolates, six tissues across three healthy clams (the reference clam mentioned above and two others), and hemocytes from an additional two clams (**Supplementary Table 2.1**). We mapped RNA to the new genome annotation to quantify expression for

Table 2.1: Improved *Mya arenaria* genome annotation

	MAKER annotation	NCBI annotation
Number of gene models	38609	44373
Number of characterized genes	21125	26005
Number of transcripts	38609	71569
Number of exons	278756	739854
Number of gene models <200bp	28	2392
Average gene length (bp)	12036.1	14382.8
Complete BUSCOs (C)	89.0%	98.7%
Complete and single-copy BUSCOs (S)	73.3%	64.2%
Complete and duplicated BUSCOs (D)	15.7%	34.6%
Fragmented BUSCOs (F)	3.0%	0.1%
Missing BUSCOs (M)	8.0%	1.2%
Total BUSCO groups searched	954	954
% genome in gene regions	35.8%	51.0%
% genome in coding sequence	4.0%	5.1%

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. 2.1A**). 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. Hierarchical clustering on just the top 100 tissue-specific genes also points toward this origin (**Supplementary Fig. 2.1B**). Because of this relationship, we focused on the comparison of MarBTN isolates (n=5) to healthy clam hemocytes (n=5) for differential expression analysis.

2.3.2 Differential expression

Unsurprisingly, many genes are significantly up- (n=7,905, **Supplementary Table 2.2**) or down-regulated (n=8,299, **Supplementary Table 2.3**) in MarBTN versus healthy hemocytes (**Fig. 2.1**). The most significant of these is an ortholog to *PUM3*, whose product inhibits the degradation of PARP1 by CASP3 following genotoxic stress[119] and highly expressed in some human cancers[120]. *PUM3* is upregulated in MarBTN and likely plays a role in continued proliferation despite DNA damage and/or DNA repair in response to catastrophic genome instability observed in this lineage[24,118]. Another of the top upregulated genes is an ortholog to *RHEB*, which encodes a small GTPase in the TOR signaling pathway that regulates cell cycle progression and is a known oncogene in humans[121]. Thousands of other genes are highly upregulated 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 oncogenes we would expect this set to include genes specific to clam 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.

Among the top eight most significantly downregulated genes are four orthologs to tumor suppressors that have been implicated in humans: *TUSC2*[122], *NISCH*[123], *LLR1*[124] and *RASF8*[125]. The second most significantly downregulated gene is an ortholog to *FGL1* (fibrinogen-like protein 1), which encodes an immune suppressive molecule that inhibits antigen-specific T-cell activation in vertabrates[126]. FGL1 is

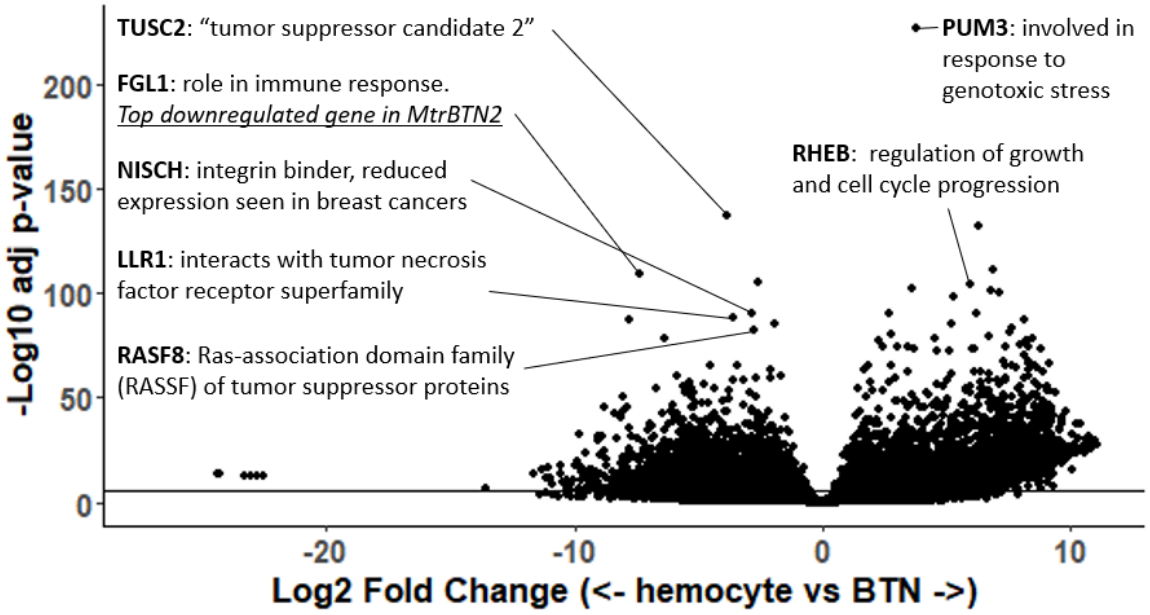


Figure 2.1. Top differentially expressed gene sets: MarBTN vs Hemocytes.

Volcano plot of fold change and significance of differentially expressed genes in MarBTN cells versus healthy hemocytes. Top genes are labeled with annotations and abbreviated descriptions. Line marks conservative significance threshold $p < (0.05 / \text{total gene count})$.

differentially regulated across many human cancers[126,127], though it has both pro-tumor and anti-tumor functions that appear to be context dependent across tumor types[128,129]. Bivalves do not have a canonical adaptive immune system and thus lack T-cells, but hemocytes, the likely cell of origin of MarBTN, are bivalve's leukocyte-like cells that mediate the innate immune defense response[130,131]. Interestingly, an *FGL1*-like gene was also the most significantly downregulated gene in a recent mussel transmissible cancer transcriptomic study[132], indicating it may be more universally important across BTNs. Fibrinogen-related proteins (FREPs) are believed to function as pathogen recognition receptors in bivalves[130], so the importance of *FGL1* across these two BTNs likely relates to immune signaling, whether to stimulate growth of the hemocyte-derived cancer or to evade rejection by the host. Though it is unclear exactly what role *FGL1* downregulation might play in BTNs, its observation as a top hit in two independent lineages from different taxonomic orders strongly indicates it is a key gene in each lineage.

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[133]. We observed

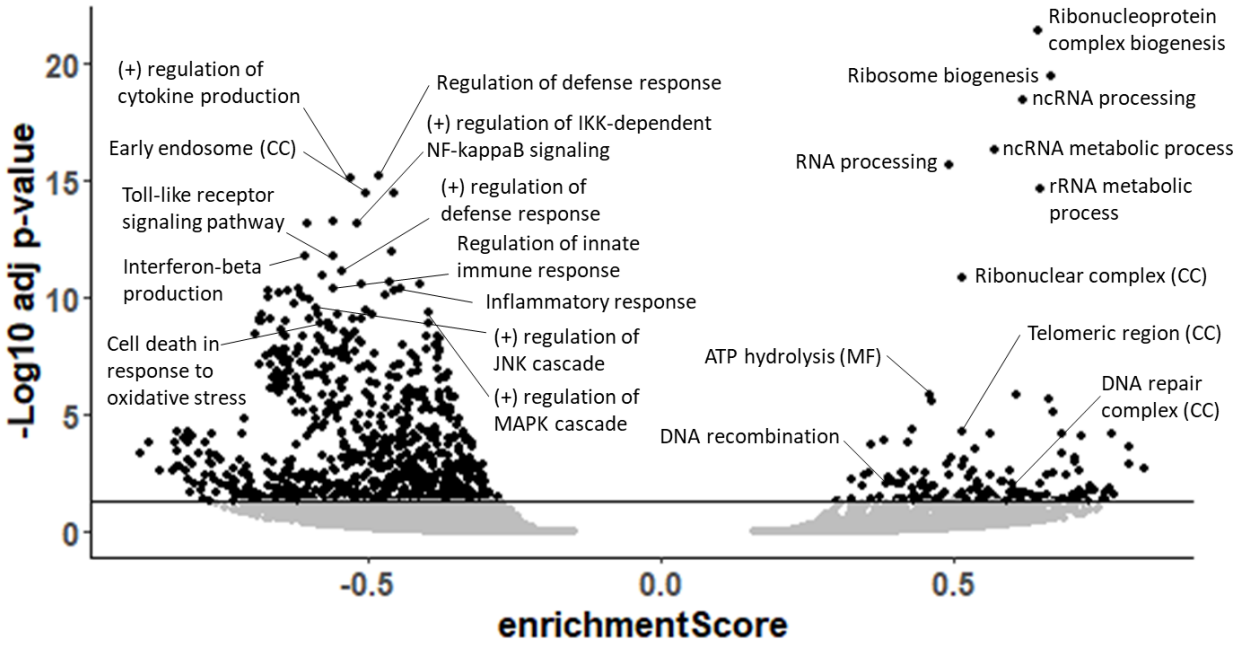


Figure 2.2. Top differentially expressed pathways: MarBTN vs Hemocytes.

Volcano plot of enrichment score and significance of differentially expressed pathways in MarBTN cells versus healthy hemocytes. Top pathways are labeled with annotations. Pathways are biological processes (BP) unless marked as molecular function (MF) or cellular component (CC). Line marks significance threshold $p < 0.05$, with pathways below line colored in grey.

163 upregulated pathways and 828 downregulated pathways (**Fig. 2.2, Supplementary Fig. 2.2**). The most upregulated pathways involved RNA processing and ribosome biogenesis (**Supplementary Table 2.4**), which are recognized as important for cell growth and proliferation of cancer cells[134]. 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 key 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 2.5**). We suspected this could be either an evolutionary response for MarBTN to evade host immune rejection or downregulation of unnecessary pathways from the cancers hemocyte origins. To test the latter possibility, we looked at differential expression between MarBTN isolates (n=5) and solid tissues (n=15: 5 tissues across 3 clams). Many of the genes and pathways were

similarly mis-regulated as they were in the hemocyte comparison (**Supplementary Fig. 2.3)**, with immune pathways continuing to dominate the downregulated gene pathways (**Supplementary Table 2.6**). This indicates the observed immune downregulation is not primarily due to the cancer's hemocyte origin and instead likely represents cellular mechanisms to evade immune recognition and 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 while continuing to proliferate and maintain the ability to infect new hosts.

2.3.3 Genome instability affects expression

We previously observed that MarBTN's genome is highly unstable, displaying widespread genome rearrangement, copy number gains, and transposable element activity[118]. 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. 2.3A**). Fusions in healthy samples may be due to genome assembly 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=183, **Supplementary Table 2.7**) include fusions from early in the cancer's somatic evolution that may have contributed to the oncogenesis and/or transmission ability of the lineage.

Copy number alteration as a known mechanism in cancers to alter the expression of cancer-promoting genes[135]. 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. 2.3B**). MarBTN has an average ploidy of ~3.5N across the genome, with >80% at

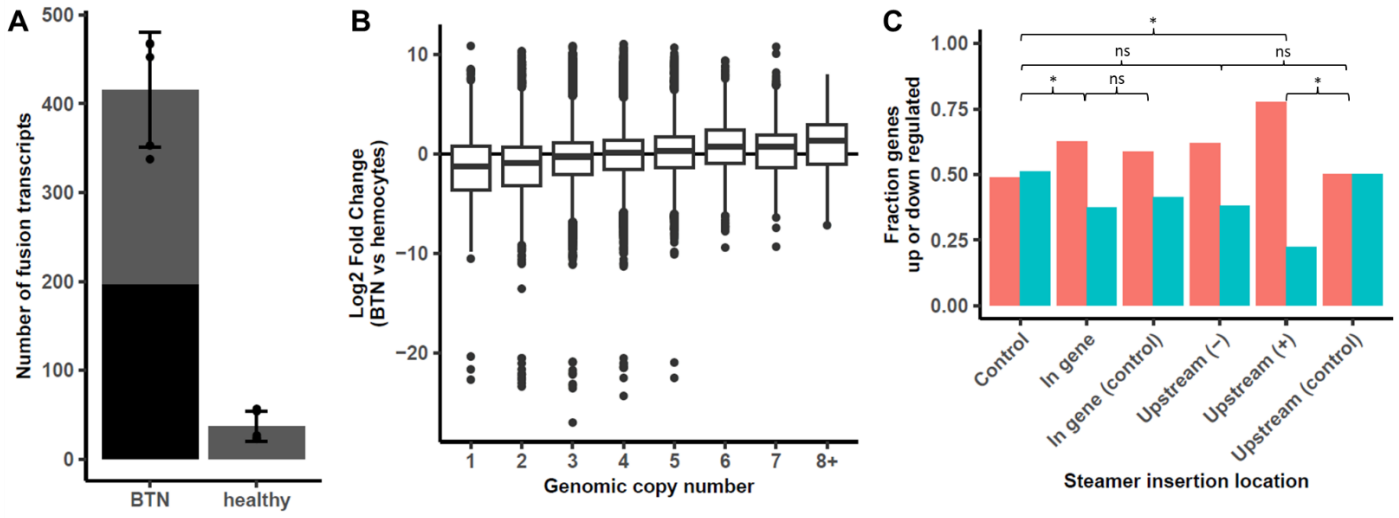


Figure 2.3. Genome instability influences transcriptome.

(A) Number of fusion transcripts per sample (dots), with mean and standard deviation for MarBTN and healthy hemocyte sample groupings. Black bar represents fusions found in all MarBTN samples (196). **(B)** Boxplots of expression change of MarBTN versus healthy hemocytes separated out by genomic copy number for each gene. **(C)** Relative fraction of genes that were significantly up or down regulated. Genes that were not significantly differentially regulated were excluded. Control = all genes genome-wide (n=14,391), In gene = genes with *Steamer* insertions in them (n=110), In gene (control) = genes without *Steamer* insertions in this sample set but observed insertions in other MarBTN samples (n=17), Upstream (-) = genes with *Steamer* insertions within 2 kB upstream in the antisense direction (n=21), Upstream (+) = genes with *Steamer* insertions within 2 kB upstream in the same direction (n=27), Upstream (control) = genes without *Steamer* insertions within 2 kB upstream in this sample set but observed insertions within 2 kB upstream in other MarBTN samples (n=10). Significance of differences are from Chi squared tests, using the control proportions as expected.

2-4N[118], and we see that median relative expression of genes ≥4N is higher than average while lower than average for genes ≤3N. 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 testing whether transposable element activity was influencing gene expression of nearby genes. We looked at the expression of genes near insertions of LTR-retrotransposon *Steamer* (**Fig. 2.3C**), one of the most active and best characterized transposable elements in MarBTN[46,118]. Compared to all MarBTN genes, which have a roughly equal chance of being up or down regulated, *Steamer* was more likely to insert in upregulated genes. However, we suspect that is because these genes are more accessible to insertions, rather than mid-gene-insertions causing upregulation. Indeed, when we look at

genes that lack *Steamer* insertions in this sample set but that have insertions in other MarBTN samples, these genes were also disproportionately upregulated. *Steamer* has a strong bias to insert upstream of genes[118], so we also looked at whether the expression of genes with insertions within 2 kB upstream were affected. When *Steamer* inserted in the same direction as the gene it was much more likely to be upregulated than expected, but when the insertion was in the opposite orientation it was not significantly different than predicted by either control (genome-wide or insertion sites in from a different sample set). This indicates that upstream insertions in the same direction of the gene are causing the upregulation of the downstream gene in some cases, which is not surprising given the 3' LTR contains a promoter and this effect has been reported for other LTR-retrotransposons[136]. Overall, though we are unable to test which individual genes affected by structural mutations may be phenotypically important in this study, we see that all three of these mutation types influence expression and thus likely contributed to the adaptability of this lineage as a transmissible cancer.

2.3.4 Transcriptomic response to saltwater

We generally focus on late-stage MarBTN infections for genomic/transcriptomic analysis since we can purify highly pure cancer samples from hemolymph with minimal host contamination at this stage. However, this is just one stage in a MarBTN infection cycle, which is likely to also include engrafting in a naïve clam, proliferating until the late stages of disease and transferring to infect a new clam (**Fig. 2.4A**). This transfer is believed to occur through release into seawater and uptake by filter-feeding, an inference supported by MarBTN cells surviving for days to weeks in saltwater and being detected in tank water where MarBTN-infected clams are housed[17]. This transfer 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. To test this possibility, we exposed an aliquot of three MarBTN isolates to artificial sea water (ASW) for 24 hours prior to RNA sequencing to investigate the transcriptomic response in this stage.

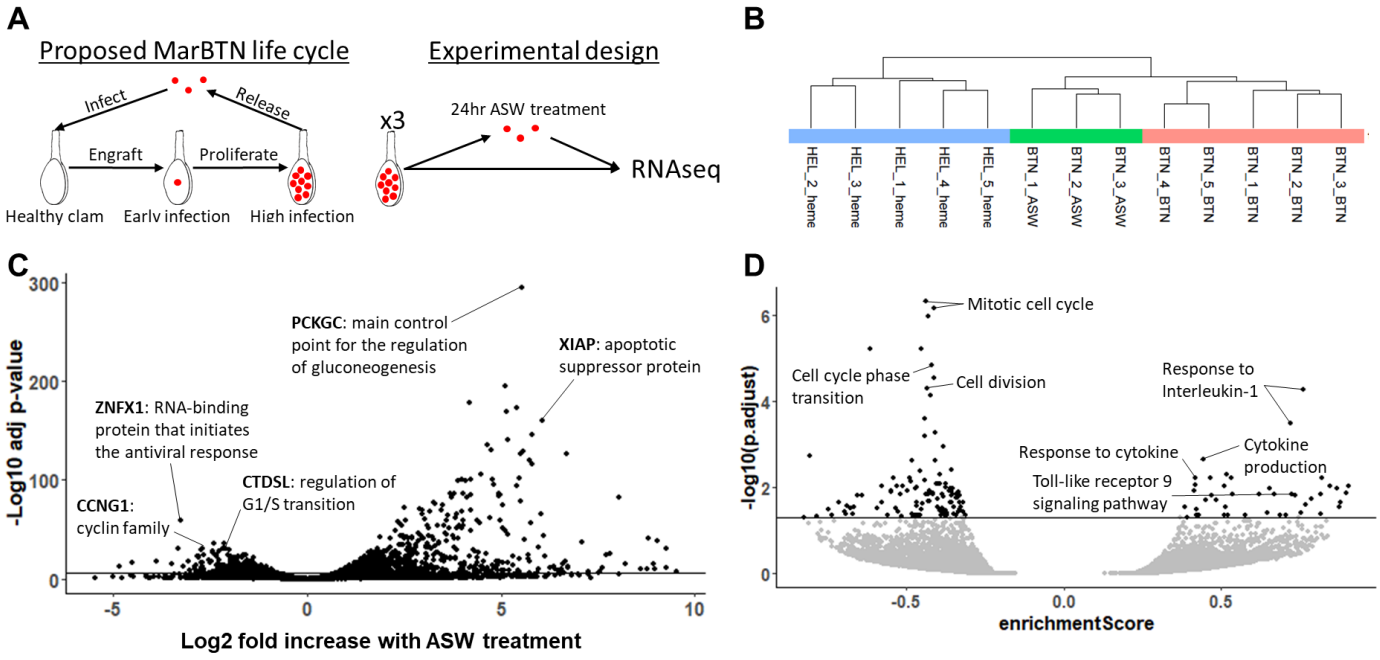


Figure 2.4. Transcriptomic response to seawater exposure.

(A) Proposed life cycle for MarBTN infections and experimental design to investigate gene expression during seawater transmission. **(B)** Hierarchical clustering of healthy hemocytes ("heme" - blue), untreated MarBTN ("BTN" - red) and ASW-treated MarBTN ("ASW" - green). Samples are labeled by their source clam and treatment. **(C)** Volcano plot of fold change and significance of differentially expressed genes in ASW-treated MarBTN versus untreated MarBTN. Genes discussed in text are labeled with annotations and abbreviated descriptions. Line marks conservative significance threshold $p < (0.05 / \text{total gene count})$. **(D)** Volcano plot of enrichment score and significance of differentially expressed pathways in ASW-treated MarBTN versus untreated MarBTN. Select pathways discussed in text are labeled with annotations. Line marks significance threshold $p < 0.05$, with pathways below line colored in grey.

Gene expression was more similar within treatment groups (ASW compared to no treatment) than source infection pairings (BTN isolate 1, 2, or 3) using both hierarchical clustering (**Fig. 2.4B**) and principal component analysis (**Supplementary Fig. 2.4**), indicating that saltwater exposure results in a consistent transcriptomic response. Marine animal cells are likely to have some inherent response to this environment, which may be the starting point that could be built upon during MarBTN evolution and selection for transmission ability. We suspect some of these reproducible gene expression changes are MarBTN-specific and the result of selection for survival and transmission ability.

We compared ASW-treated to untreated MarBTN isolates for differentially expressed genes (**Fig. 2.4C**) and pathways (**Fig. 2.4D**). 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 has been observed in glucose-deprived human

cancer cells[137]. Another notable gene from the top upregulated genes (**Supplementary Table 2.8**) is *XIAP*, which is part of a family of apoptotic suppressor proteins and likely helps MarBTN cells to avoid an apoptotic response to seawater. *XIAP* also modulates inflammatory and immune signaling via NF-κB and JNK activation, indicating these pathways, which were downregulated when comparing untreated MarBTN to healthy hemocytes, may be reactivated in absence of a host. Indeed, when we use GSEA to identify differentially regulated pathways, we see immune response pathways such as cytokine production/response, toll-like receptor 9 signaling, and T-cell activation among the 45 significantly upregulated pathways (**Supplementary Table 2.9**). 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. 2.5**). Overall, this finding indicates that the downregulation of these immune pathways may not be as important outside the context of a host immune system, supporting the hypothesis that immune pathway downregulation is an adaptive response to evade host immune rejection.

The top downregulated gene is *ZNFX*, which encodes an RNA-binding protein involved in antiviral response, though it is unclear why this gene might be lower expressed in seawater. Among the other top downregulated genes (**Supplementary Table 2.10**) are *CCNG1*, a member of the cell cycle controlling cyclin family, and *CTDSL*, which is involved regulating the G1/S transition. The top downregulated pathways are also involved with division and cell cycle progression (**Supplementary Table 2.11**), likely representing mechanisms to halt proliferation in the absence of host nutrients to survive the seawater transfer environment by entering a quiescent state.

2.4 Discussion

All cancers must evolve to evade intrinsic and extrinsic barriers to successfully develop as a cancer[138]. 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[113]. This all occurs while having no evolutionary history as a transmitting parasite prior to oncogenesis[139]. 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 many key immune signaling pathways.

We observe misregulation of many gene types in MarBTN that would be expected in any cancer, such as genes involved in metabolism, cell cycle progression, tumor suppression, genome instability and immune evasion[140]. The downregulated biological processes overwhelmingly relate to immune signaling functions (**Fig. 2.2**, **Supplementary Table 2.5, Supplementary Table 2.6**) 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[132], while the mammalian transmissible cancers display MHC downregulation[114–116]. 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. The down-regulated immune gene *FGL1* from both mussel and clam transmissible cancers is a striking example of convergent evolution of independent cancers. As more BTNs are identified and characterized, a systematic comparison of differentially expressed genes and gene sets would likely identify additional examples of convergent evolution and allow identification of underlying mechanisms of transmissible cancer evolution. Such mechanisms may also highlight more generally how cancers are able to evade innate immune responses.

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[118], and here we observe cases in which that genome instability directly affects the cancer transcriptome. Copy number alterations in this highly aneuploid cancer 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 transcriptome 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*[118], which has been shown to sequester DNA damage response molecule p53[54]. 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[118], 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[1].

Here we investigate 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[57,139]. Identification of innate immune system cancer resistance mechanisms of hosts and countering evasion mechanisms in cancers, selected for by repeated infection, may each have broader implications in our understanding of the host-pathogen relationship of conventional cancers.

2.5 Acknowledgements

We thank Sophie Kogut for investigating fusion genes and Metzger lab members Jordana Sevigny, Karyn Tindbaek, and Finola Schmahl-Waggoner for feedback. This work was supported by NIH training grants T32-HG000035 and T32-GM007270 (to S.F.M.H.), career transition award K22-CA226047, R01-CA255712 and NSF award number 2208081 (to M.J.M).

2.6 Author contributions

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

2.7 Methods

2.7.1 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.

2.7.2 Genome annotation

To utilize the NCBI Eukaryotic genome pipeline we supplied NCBI with the previously assembled *M. arenaria* genome[118] 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[25]) using the command: busco -m prot -l metazoa_odb10 and calculated other stats in Table 1 using custom scripts.

2.7.3 Sample collection

Clams were collected by a commercial shellfish supplier in Maine (**Supplementary Table 2.1**) and shipped live on ice to the Pacific Northwest Research Institute in Seattle, WA. 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 [17]). The five selected healthy clams were undetectable for the cancer-specific qPCR marker, while the five 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 (36 g/L Instant Ocean, Blacksburg, VA, USA) with antibiotics (penicillin, streptomycin and voriconazole) as described in [17], 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. All samples (healthy hemocytes, tissues, MarBTN isolates, and ASW-treated MarBTN isolates) were covered in RNAlater and stored at -80 °C until RNA extraction.

2.7.4 RNA extraction

RNA was extracted from each sample using the Qiagen RNeasy kit (Qiagen, Hilden, Germany) and 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 was then added 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).

2.7.5 Differential expression analysis

We indexed the annotated genome and aligned reads for all samples using STAR [110], 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 [118]), 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 [111], 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 600 tissue-specific genes using the pheatmap package with clustering_distance_cols = "canberra". ASW-treated MarBTN samples were excluded from the original clustering analysis (**Supplementary Fig. 2.1**), then included alongside only hemocytes and untreated MarBTN for principal component analysis and hierarchical clustering using expression of all genes and the packages described above (**Supplementary Fig. 2.4**).

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 MarBTN versus untreated MarBTN and versus hemocytes. For the comparison of differential expression results from multiple DESeq2 runs (e.g. **Supplementary Fig. 2.3**) 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.

2.7.6 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[141] *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.

2.7.7 Fusions gene identification

We identified fusion transcripts using STAR-Fusion (v1.11.0[142,143]). We first generated a custom genome index using prep_genome_lib.pl on the annotated genome with the "current" Pfam database and "human" Dfam database 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.

2.7.8 Copy number effects

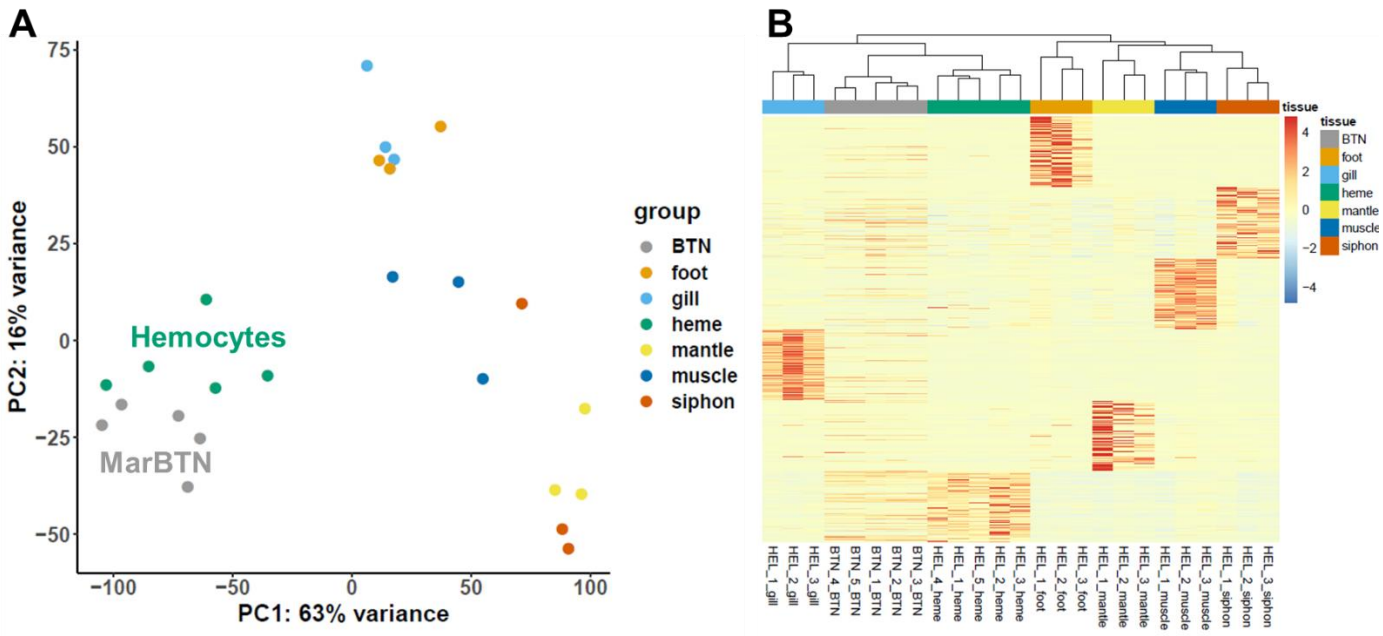
Genomic copy number calls were determined in 100 kB segments for USA sub-lineage MarBTN samples in previous work [118]. The copy number regions were observed to be nearly identical between the samples of the MarBTN from the USA sub-lineage, so while there are likely minor differences in these samples, 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) or in a CN0 region. 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.

2.7.9 *Steamer* insertion effects

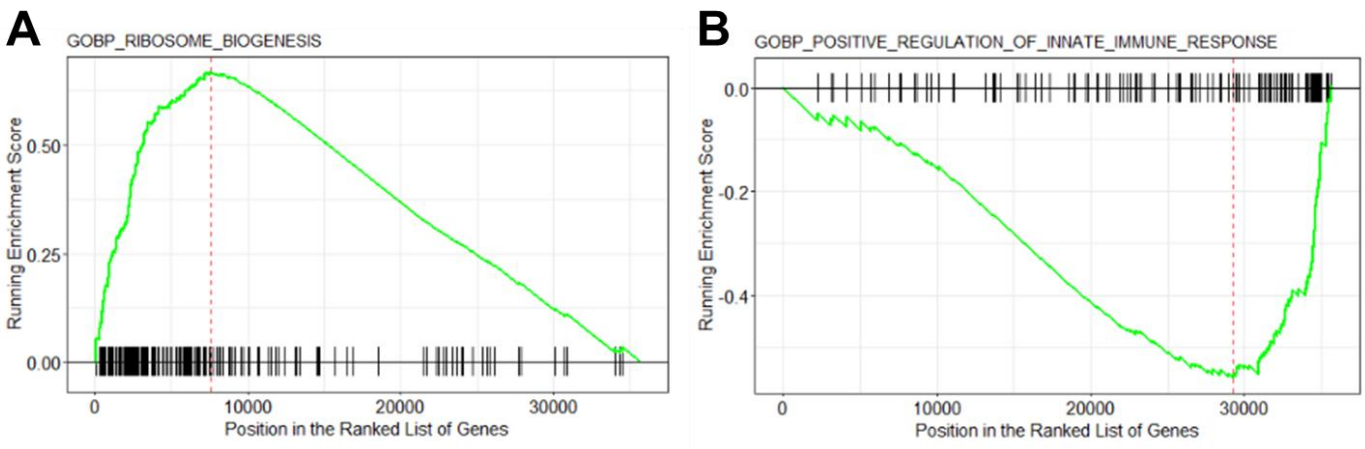
We had also determined *Steamer* insertion sites in previous work [118], and assumed that insertions previously found in all USA sub-lineage MarBTN samples would also be present in the samples of this current study. We determined where *Steamer* had inserted within genes or within 2 kB upstream genes 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 then tallied for each set of genes how many of them were significantly (adjusted p-value < 0.05) up- or downregulated in the MarBTN versus healthy hemocytes DESeq2 comparison. As another control, we also considered the overall proportion of upregulated to downregulated genes for all genes (which was close to 50-50). Finally, for pairwise statistical comparisons we used Chi squared tests to compare each gene set to what would be expected from the control set.

2.8 Supplemental figures



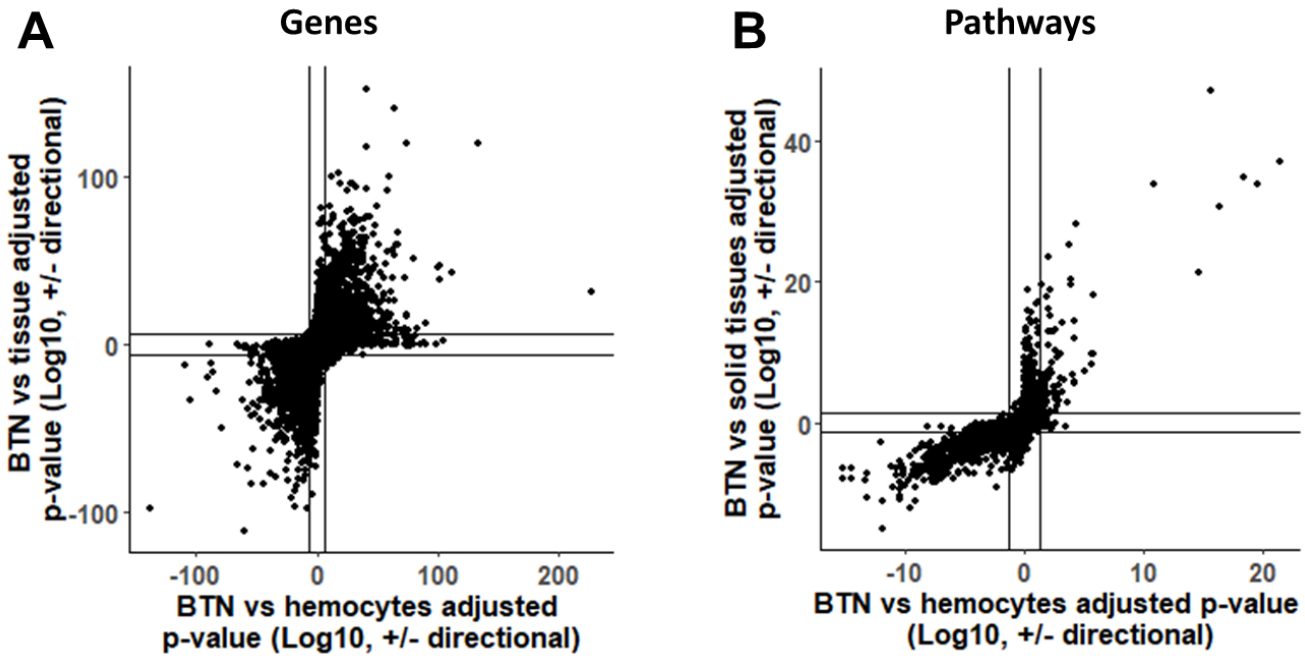
Supplementary Figure 2.1. Hemocyte origin of MarBTN supported by PCA and clustering with new gene annotations

(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 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[118].



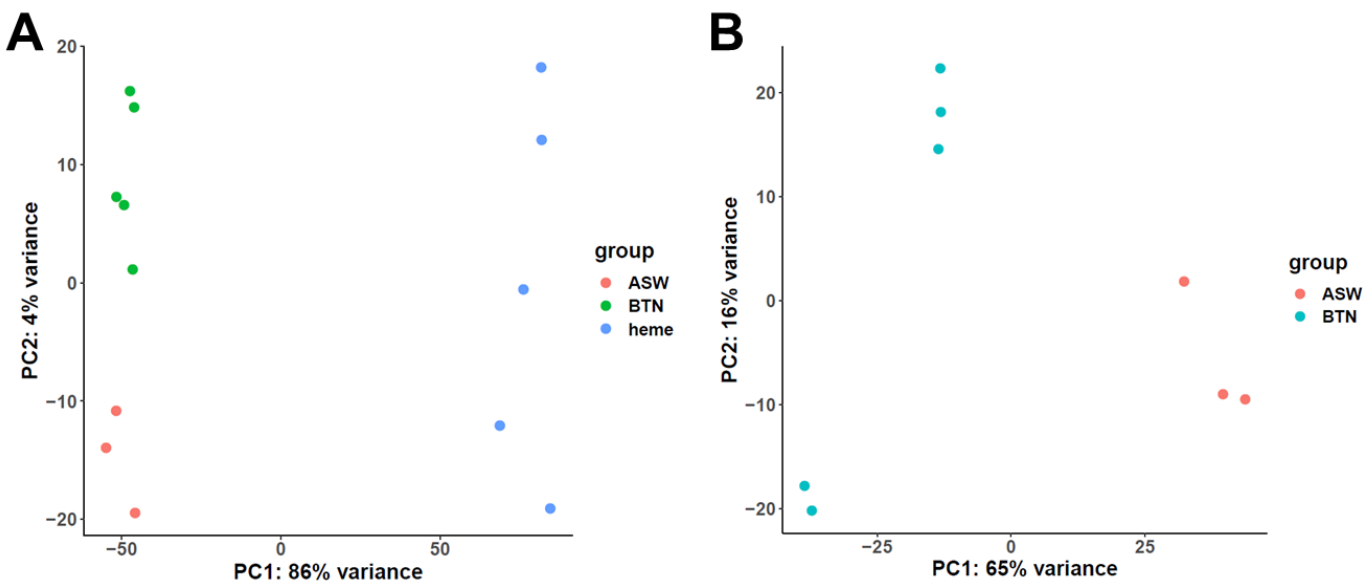
Supplementary Figure 2.2. Example GSEA results for one each of the top up- and downregulated pathways.

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.



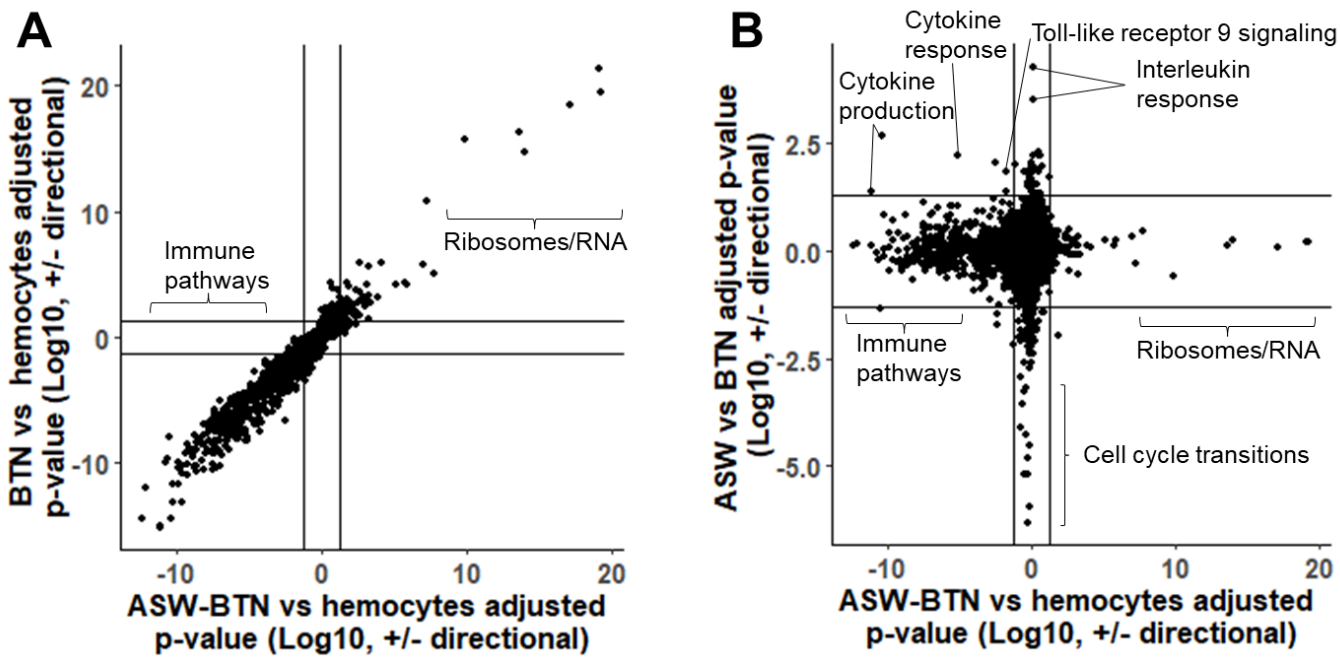
Supplementary Figure 2.3. Differential expression is similar whether comparing to hemocytes or solid tissue.

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 upregulated versus solid tissues, indicating that major differential expression results and conclusions are not artifacts of the comparison with hemocytes. Lines represent significance thresholds.



Supplementary Figure 2.4. PCA separates samples by seawater treatment.

Principal component analysis of gene expression across all genes for hemocytes, untreated MarBTN, and ASW-treated MarBTN **(A)** and for ASW-treated and untreated MarBTN alone **(B)**. In **(A)**, PC1 captures most of the variance to differentiate hemocytes from MarBTN, while PC2 separates ASW-treated from untreated MarBTN. In **(B)** PC1 captures most of the variance to differentiate ASW-treated from untreated MarBTN, while PC2 separates untreated MarBTN into two groupings that reflect the two rounds of RNAseq that the samples were sequenced in, likely due to either genetic differences, environmental differences in during sample processing, or differences in sequencing runs. These differences are minimal compared to the differences between the other groupings.



Supplementary Figure 2.5. Differential expressed pathways are similar whether comparing ASW-treated or untreated MarBTN to hemocytes.

Adjusted p-values, further adjusted to be positive for upregulation and negative for downregulation, for ASW-treated MarBTN versus healthy hemocytes GSEA results (x-axes) plotted against GSEA results for **(A)** untreated MarBTN versus healthy hemocytes and **(B)** ASW-treated versus untreated MarBTN (y-axes). 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)** we observe that some of the immune pathways (cytokine pathways) are still downregulated in ASW-treated MarBTN versus hemocytes, though upregulated versus untreated MarBTN, while some are only upregulated in ASW-treated versus untreated MarBTN (interleukin pathways). Important gene sets (lines) and gene sets categories (brackets) are labeled. Lines represent significance thresholds.

2.9 Supplemental tables

Supplementary Table 2.1: List of RNA sequenced samples

Name	Alternate aliases	Healthy/BTN	Tissues	Date sampled	Location
MELC-2E11	HEL_1	Healthy	hemocytes, foot, gill, adductor muscle, mantle, and siphon	6/1/2018	Larrabe Cove, Machiasport, ME, USA
FFM-27C11	HEL_2	Healthy	hemocytes, foot, gill, adductor muscle, mantle, and siphon	10/17/2022	Friendship, ME, USA
FFM-27H7	HEL_3	Healthy	hemocytes, foot, gill, adductor muscle, mantle, and siphon	10/17/2022	Friendship, ME, USA
FFM-15G1	HEL_4	Healthy	hemocytes only	6/30/2020	Brunswick, ME, USA
FFM-16B11	HEL_5	Healthy	hemocytes only	6/30/2020	Brunswick, ME, USA
FFM-28E5	BTN_1	BTN	BTN isolate, ASW-treatment	10/13/2022	Perry, ME, USA
FFM-28E6	BTN_2	BTN	BTN isolate, ASW-treatment	10/13/2022	Perry, ME, USA
FFM-28E7	BTN_3	BTN	BTN isolate, ASW-treatment	10/13/2022	Perry, ME, USA
FFM-15G11	BTN_4	BTN	BTN isolate	6/30/2020	Brunswick, ME, USA
FFM-16G1	BTN_5	BTN	BTN isolate	6/30/2020	Brunswick, ME, USA

Supplementary Table 2.2: Fifty most significant upregulated genes, MarBTN vs hemocytes

gene	closest uniprot hit	padj	baseMean	stat	log2FoldChange	NCBI pipeline annotation
LOC128213480	PUM3	7.63E-228	20404	-32.54	-3.80	**pumilio homolog 3-like**
LOC128203379	NA	1.95E-133	2140	-24.95	-6.35	uncharacterized LOC128203379
LOC128229303	NA	9.16E-112	539	-22.86	-6.91	uncharacterized LOC128229303
LOC128229834	LRIQ1	4.82E-105	946	-22.15	-5.98	leucine-rich repeat and IQ domain-containing protein 1-like
LOC128206567	SO4C1	9.07E-103	11033	-21.90	-3.63	solute carrier organic anion transporter family member 4C1-like
LOC128211986	NA	8.42E-102	1565	-21.80	-6.82	uncharacterized LOC128211986
LOC128244936	RHEB	1.02E-100	1747	-21.68	-7.16	**GTP-binding protein Rheb-like**
LOC128232783	CCD42	1.65E-99	1004	-21.54	-5.33	coiled-coil domain-containing protein 42 homolog
LOC128209790	CC170	2.93E-91	1470	-20.64	-6.26	coiled-coil domain-containing protein 170-like
LOC128212193	RDH11	8.34E-91	1940	-20.58	-2.67	retinol dehydrogenase 11-like
LOC128228995	NA	1.25E-88	517	-20.33	-8.13	uncharacterized LOC128228995
LOC128206740	SORCN	1.45E-86	725	-20.09	-5.19	sorcin-like
LOC128209778	SC5AC	6.02E-84	378	-19.79	-7.67	sodium-coupled monocarboxylate transporter 2-like
LOC128227628	DNAS1	5.19E-82	564	-19.56	-7.61	deoxyribonuclease-1-like
LOC128242538	CREST	3.79E-81	1938	-19.45	-2.82	uncharacterized LOC128242538
LOC128205214	NA	1.35E-80	303	-19.38	-6.74	uncharacterized LOC128205214
LOC128204997	NPHN	2.90E-79	948	-19.22	-8.52	nephrin-like
LOC128211207	CFA46	3.64E-79	1636	-19.21	-4.55	cilia- and flagella-associated protein 46-like
LOC128215984	CHSTF	7.44E-78	676	-19.05	-8.22	carbohydrate sulfotransferase 15-like
LOC128230152	TSN9	9.21E-78	1074	-19.04	-2.31	tetraspanin-9-like
LOC128241728	CO8A2	1.46E-76	722	-18.89	-8.03	complement C1q-like protein 3
LOC128240418	NA	1.28E-75	1121	-18.77	-8.34	uncharacterized LOC128240418
LOC128229905	NA	1.43E-75	775	-18.76	-7.43	uncharacterized LOC128229905
LOC128222938	NA	1.78E-75	306	-18.75	-3.48	uncharacterized LOC128222938
LOC128228782	NA	2.21E-75	459	-18.74	-2.49	uncharacterized LOC128228782
LOC128232418	BRSK2	9.95E-75	1256	-18.66	-3.01	serine/threonine-protein kinase BRSK2-like
LOC128231175	TRI33	1.21E-74	4111	-18.64	-6.16	uncharacterized LOC128231175
LOC128232121	NA	3.22E-74	1401	-18.59	-6.06	collagen alpha-1(I) chain-like
LOC128242222	TRIM1	4.10E-74	2545	-18.58	-3.70	probable E3 ubiquitin-protein ligase MID2
LOC128228609	NA	4.62E-74	860	-18.57	-8.85	uncharacterized LOC128228609
LOC128210573	PLCE1	1.11E-73	269	-18.52	-5.12	1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase epsilon-1-like
LOC128224721	ATAT	9.09E-73	693	-18.40	-4.63	alpha-tubulin N-acetyltransferase 1-like
LOC128205094	DHSO	3.03E-71	764	-18.21	-8.13	sorbitol dehydrogenase-like
LOC128224923	NA	1.97E-68	408	-17.85	-7.28	uncharacterized LOC128224923
LOC128204083	NA	3.33E-68	2095	-17.82	-8.30	uncharacterized LOC128204083
LOC128238491	GRIK2	5.35E-67	1959	-17.66	-8.32	glutamate receptor ionotropic, kainate 2-like
LOC128232007	S28A3	6.83E-67	1258	-17.65	-9.17	solute carrier family 28 member 3-like
LOC128224541	SC5A9	3.03E-66	1371	-17.56	-7.64	sodium/glucose cotransporter 5-like
LOC128231280	PI4KA	3.16E-66	6924	-17.56	-2.80	phosphatidylinositol 4-kinase alpha-like
LOC128232580	TYW1	3.21E-66	5068	-17.56	-1.90	S-adenosyl-L-methionine-dependent tRNA 4-demethylwyosine synthase TYW1-like
LOC128233851	SE1L3	3.57E-66	1100	-17.55	-8.57	protein sel-1 homolog 3-like
LOC128211406	RAD18	3.57E-66	258	-17.55	-7.40	uncharacterized LOC128211406
LOC128206512	GLSK	1.16E-65	472	-17.48	-7.83	glutaminase liver isoform, mitochondrial-like
LOC128235646	NA	2.45E-65	634	-17.44	-8.69	uncharacterized LOC128235646
LOC128245377	NA	5.71E-65	1000	-17.39	-8.65	uncharacterized LOC128245377
LOC128239808	METL2	1.24E-64	1320	-17.34	-1.71	tRNA N(3)-methylcytidine methyltransferase METTL2-like
LOC128228889	NA	1.72E-64	443	-17.32	-8.30	uncharacterized LOC128228889
LOC128208050	NA	1.51E-63	218	-17.19	-6.38	uncharacterized LOC128208050
LOC128211401	VWDE	1.81E-63	790	-17.18	-6.13	uncharacterized LOC128211401
LOC128222844	C163B	2.89E-63	471	-17.15	-8.91	deleted in malignant brain tumors 1 protein-like

Supplementary Table 2.3: Fifty most significant downregulated genes, MarBTN vs hemocytes

gene	closest uniprot hit	padj	baseMean	stat	log2FoldChange	NCBI pipeline annotation
LOC128217773	TUSC2	5.39E-138	783	25.39	3.81	**tumor suppressor candidate 2-like**
LOC128210338	TENN	9.74E-110	44264	22.64	7.37	**fibrinogen-like protein 1**
LOC128212534	PKHA1	9.60E-106	3790	22.23	2.57	pleckstrin homology domain-containing family A member 1-like
LOC128217284	NISCH	2.89E-91	1133	20.64	2.80	**nischarin-like**
LOC128208805	LLR1	9.49E-89	2797	20.35	3.54	**leucine-rich repeat protein 1-like**
LOC128224232	NA	3.60E-88	6703	20.28	7.74	keratin-associated protein 10-10-like
LOC128227340	UB2Q1	3.19E-86	5376	20.05	1.92	ubiquitin-conjugating enzyme E2 Q2-like
LOC128232857	**RASF8**	2.69E-83	747	19.71	2.72	uncharacterized LOC128232857
LOC128219963	NA	8.47E-79	1663	19.16	6.32	uncharacterized LOC128219963
LOC128243288	PDES1	3.03E-66	2167	17.56	4.48	plasmanylethanolamine desaturase-like
LOC128226789	NA	3.22E-66	863	17.56	3.40	uncharacterized LOC128226789
LOC128219778	ECT2	2.11E-64	2269	17.31	2.03	protein ECT2-like
LOC128207719	SNP25	2.34E-61	3888	16.89	1.62	synaptosomal-associated protein 25-like
LOC128235334	CARME	8.07E-61	476	16.82	5.84	carnosine N-methyltransferase-like
LOC128206985	RAC1	1.25E-60	25050	16.79	2.62	ras-related C3 botulinum toxin substrate 1-like
LOC128231872	F117B	6.76E-60	1417	16.69	2.10	protein FAM117B-like
LOC128227724	FHDC1	1.90E-59	14216	16.62	3.56	inverted formin-2-like
LOC128245641	EFHD2	3.91E-59	10053	16.58	3.13	EF-hand domain-containing protein D2-like
LOC128238733	NATT4	4.11E-58	2379	16.43	5.36	natterin-4-like
LOC128240915	CRA1B	2.70E-57	14744	16.32	4.72	collagen alpha-1(I) chain-like
LOC128240775	MRP1	8.21E-57	300	16.25	5.31	multidrug resistance-associated protein 1-like
LOC128241733	NA	1.52E-56	1759	16.21	5.58	uncharacterized LOC128241733
LOC128217702	BAP31	1.57E-55	2124	16.06	4.33	B-cell receptor-associated protein 31-like
LOC128213034	NA	4.23E-55	318	16.00	6.66	uncharacterized LOC128213034
LOC128238636	NA	4.31E-55	955	15.99	5.28	uncharacterized LOC128238636
LOC128227984	UBIQ1	1.46E-54	47318	15.91	5.35	polyubiquitin-H
LOC128218899	MRP4	1.88E-54	1812	15.90	3.75	ATP-binding cassette sub-family C member 4-like
LOC128205650	RDM1	2.66E-53	319	15.73	5.47	RAD52 motif-containing protein 1-like
LOC128220306	NA	7.00E-53	8560	15.66	5.14	uncharacterized LOC128220306
LOC128242807	ARD17	7.67E-53	5380	15.66	2.91	arrestin domain-containing protein 17-like
LOC128242823	DCA11	1.71E-51	2116	15.45	2.62	DDB1- and CUL4-associated factor 11-like
LOC128239784	NA	9.45E-51	15277	15.34	5.29	uncharacterized LOC128239784
LOC128206074	NA	9.45E-51	1045	15.34	8.06	receptor-type tyrosine-protein phosphatase kappa-like
LOC128246323	S26A2	2.57E-50	702	15.27	3.75	sulfate transporter-like
LOC128220286	NUP98	4.47E-49	159	15.08	4.02	uncharacterized LOC128220286
LOC128246874	ARHGQ	1.35E-48	11306	15.01	4.88	rho guanine nucleotide exchange factor 5-like
LOC128238720	TMTC3	7.66E-48	335	14.89	4.57	protein O-mannosyl-transferase TMTC3-like
LOC128214410	NA	7.67E-48	191	14.89	5.11	uncharacterized LOC128214410
LOC128241502	PARP8	3.46E-47	432	14.78	3.20	protein mono-ADP-ribosyltransferase PARP6-like
LOC128223123	CFAB	7.33E-47	12333	14.73	6.05	complement factor B-like
LOC128237009	NA	1.24E-46	2905	14.70	5.28	uncharacterized LOC128237009
LOC128238594	NA	1.24E-46	156320	14.70	7.84	uncharacterized LOC128238594
LOC128203576	YH2M	1.69E-46	837	14.67	5.19	probable cation-transporting ATPase W08D2.5
LOC128237455	CATL1	5.99E-46	17213	14.58	8.82	procathepsin L-like
LOC128233517	STA5B	5.79E-45	13175	14.42	2.86	signal transducer and activator of transcription 5B-like
LOC128227979	CALX	1.04E-44	3831	14.38	3.85	calnexin-like
LOC128237500	RRAS2	1.31E-44	4621	14.36	2.83	ras-related protein R-Ras2-like
LOC128212868	AVT1D	1.84E-44	977	14.34	8.01	uncharacterized LOC128212868
LOC128208898	TM260	1.88E-44	347	14.34	2.19	transmembrane protein 260-like
LOC128211600	MET	3.24E-44	1952	14.30	6.13	hepatocyte growth factor receptor-like

Supplementary Table 2.4: Fifty most significant upregulated gene sets, MarBTN vs hemocytes

gs_name	gs_exact_source	setSize	enrichmentScore	NES	p.adjust
GOBP_RIBONUCLEOPROTEIN_COMPLEX_BIOGENESIS	GO:0022613	234	0.64	2.63	4.16E-22
GOBP_RIBOSOME_BIOGENESIS	GO:0042254	182	0.67	2.65	3.00E-20
GOBP_NCRNA_PROCESSING	GO:0034470	223	0.62	2.51	3.69E-19
GOBP_NCRNA_METABOLIC_PROCESS	GO:0034660	282	0.57	2.36	4.66E-17
GOBP_RNA_PROCESSING	GO:0006396	472	0.49	2.13	2.06E-16
GOBP_RRNA_METABOLIC_PROCESS	GO:0016072	151	0.65	2.53	2.06E-15
GOCC_RIBONUCLEOPROTEIN_COMPLEX	GO:1990904	281	0.52	2.13	1.37E-11
GOBP_RIBONUCLEOPROTEIN_COMPLEX_SUBUNIT_ORGANIZATION	GO:0071826	78	0.61	2.15	1.42E-06
GOMF_ATP_HYDROLYSIS_ACTIVITY	GO:0016887	227	0.46	1.87	1.46E-06
GOCC_PRERIBOSOME	GO:0030684	58	0.66	2.22	2.09E-06
GOMF_CATALYTIC_ACTIVITY_ACTING_ON_RNA	GO:0140098	220	0.46	1.87	2.48E-06
GOBP_RIBOSOMAL_SMALL_SUBUNIT_BIOGENESIS	GO:0042274	50	0.67	2.17	8.14E-06
GOCC_CHROMOSOMAL_REGION	GO:0098687	229	0.43	1.75	4.46E-05
GOCC_CHROMOSOME_TELOMERIC_REGION	GO:0000781	122	0.52	1.95	5.55E-05
GOCC_SMALL_SUBUNIT_PROCESSOME	GO:0032040	25	0.77	2.17	5.61E-05
GOBP_RNA_MODIFICATION	GO:0009451	82	0.57	2.02	6.35E-05
GOBP_MATURATION_OF_SSU_RRNA	GO:0030490	40	0.68	2.12	6.64E-05
GOBP_RIBOSOME_ASSEMBLY	GO:0042255	32	0.72	2.11	8.20E-05
GOMF_CATALYTIC_ACTIVITY_ACTING_ON_A_NUCLEIC_ACID	GO:0140640	394	0.38	1.62	1.08E-04
GOBP_MRNA_PROCESSING	GO:0006397	232	0.42	1.72	1.36E-04
GOCC_NUCLEAR_BODY	GO:0016604	455	0.36	1.55	0.000182561
GOMF_RACEMASE_AND_EPIMERASE_ACTIVITY	GO:0016854	18	0.80	2.07	0.00021613
GOBP_CILIUM_MOVEMENT	GO:0003341	84	0.54	1.92	0.000287465
GOBP_RIBOSOMAL_LARGE_SUBUNIT_BIOGENESIS	GO:0042273	32	0.69	2.02	0.000442848
GOBP_MATURATION_OF_SSU_RRNA_FROM_TRICISTRONIC_RRNA_TRANSCRIPT_SSU_RRNA_5_8S_RRNA_LSU_RRNA	GO:0000462	28	0.71	2.03	0.000687836
GOCC_SPLICEOSOMAL_COMPLEX	GO:0005681	102	0.50	1.83	0.000711784
GOBP_TRNA_METABOLIC_PROCESS	GO:0006399	89	0.52	1.86	0.000797959
GOMF_RIBONUCLEOPROTEIN_COMPLEX_BINDING	GO:0043021	89	0.52	1.86	0.000832275
GOMF_SNORNA_BINDING	GO:0030515	26	0.71	2.02	0.001034021
GOBP_RRNA_MODIFICATION	GO:0000154	15	0.80	2.00	0.001133517
HP_SMALL_NAIL	HP:0001792	99	0.49	1.80	0.001275211
GOBP_TRNA_PROCESSING	GO:0008033	82	0.51	1.83	0.001670484
GOMF_RACEMASE_AND_EPIMERASE_ACTIVITY_ACTING_ON_CARBOHYDRATES_AND_DERIVATIVES	GO:0016857	12	0.83	1.94	0.001683384
GOBP_CLEAVAGE_INVOLVED_IN_RRNA_PROCESSING	GO:0000469	23	0.71	1.95	0.002192974
GOMF_ATP_DEPENDENT_ACTIVITY_ACTING_ON_RNA	GO:0008186	70	0.53	1.86	0.002265359
GOBP_MONOSACCHARIDE_CATABOLIC_PROCESS	GO:0046365	47	0.60	1.92	0.002379239
HP_ROD_CONE_DYSTROPHY	HP:0000510	135	0.43	1.66	0.002546364
GOMF_ISOMERASE_ACTIVITY	GO:0016853	125	0.45	1.71	0.002622688
GOBP_METHYLATION	GO:0032259	185	0.41	1.63	0.002706743
HP_ABNORMAL_PREPUTIUM_MORPHOLOGY	HP:0100587	29	0.67	1.94	0.002753203
GOMF_ATP_DEPENDENT_ACTIVITY	GO:0140657	356	0.35	1.50	0.002762549
GOBP_LIPOPROTEIN_METABOLIC_PROCESS	GO:0042157	71	0.53	1.85	0.003132369
GOBP_MATURATION_OF_5_8S_RRNA	GO:0000460	29	0.66	1.91	0.003696771
GOCC_90S_PRERIBOSOME	GO:0030686	25	0.69	1.96	0.003776012
HP_ABNORMAL_VENOUS_MORPHOLOGY	HP:0002624	78	0.49	1.75	0.003801232
GOCC_CILIUM	GO:0005929	360	0.35	1.48	0.003865499
GOBP_CILIUM_OR_FLAGELLUM_DEPENDENT_CELL_MOTILITY	GO:0001539	52	0.56	1.85	0.003905056
GOBP_NON_MEMBRANE_BOUNDED_ORGANELLE_ASSEMBLY	GO:0140694	208	0.39	1.57	0.004138843
GOCC_U2_TYPE_SPLICEOSOMAL_COMPLEX	GO:0005684	62	0.52	1.78	0.004357297
GOMF_HELICASE_ACTIVITY	GO:0004386	154	0.42	1.63	0.004359028

Supplementary Table 2.5: Fifty most significant downregulated gene sets, MarBTN vs hemocytes

gs_name	gs_exact_source	setSize	enrichmentScore	NES	p.adjust
GOBP_REGULATION_OF_DEFENSE_RESPONSE	GO:0031347	369	-0.48	-2.19	6.43E-16
GOBP_POSITIVE_REGULATION_OF_CYTOKINE_PRODUCTION	GO:0001819	269	-0.53	-2.34	7.23E-16
GOCC_EARLY_ENDOSOME	GO:0005769	307	-0.51	-2.23	3.56E-15
GOBP_CYTOKINE_PRODUCTION	GO:0001816	429	-0.46	-2.10	3.56E-15
GOBP_I_KAPPAB_KINASE_NF_KAPPAB_SIGNALING	GO:0007249	207	-0.56	-2.41	5.86E-14
GOBP_POSITIVE_REGULATION_OF_I_KAPPAB_KINASE_NF_KAPPAB_SIGNALING	GO:0043123	150	-0.60	-2.48	6.18E-14
GOBP_POSITIVE_REGULATION_OF_RESPONSE_TO_EXTERNAL_STIMULUS	GO:0032103	265	-0.52	-2.28	6.18E-14
GOCC_ENDOSOME_MEMBRANE	GO:0010008	370	-0.46	-2.08	1.01E-12
GOBP_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	GO:0002224	182	-0.56	-2.37	1.50E-12
GOBP_INTERFERON_BETA_PRODUCTION	GO:0032608	133	-0.61	-2.48	1.50E-12
GOBP_POSITIVE_REGULATION_OF_DEFENSE_RESPONSE	GO:0031349	190	-0.54	-2.31	7.76E-12
GOBP_TYPE_I_INTERFERON_PRODUCTION	GO:0032606	154	-0.58	-2.38	1.21E-11
GOCC_CELL_LEADING_EDGE	GO:0031252	329	-0.47	-2.08	2.03E-11
GOBP_REGULATION_OF_RESPONSE_TO_BIOTIC_STIMULUS	GO:0002831	228	-0.51	-2.21	2.29E-11
GOBP_CELL_ACTIVATION	GO:0001775	491	-0.41	-1.90	2.82E-11
GOBP_INTERFERON_ALPHA_PRODUCTION	GO:0032607	119	-0.62	-2.46	3.89E-11
GOBP_REGULATION_OF_INNATE_IMMUNE_RESPONSE	GO:0045088	165	-0.56	-2.33	3.89E-11
GOBP_INFLAMMATORY_RESPONSE	GO:0006954	382	-0.44	-2.02	3.89E-11
GOBP_REGULATION_OF_NIK_NF_KAPPAB_SIGNALING	GO:1901222	104	-0.64	-2.47	4.34E-11
GOMF_NADPLUS_NUCLEOSIDASE_ACTIVITY	GO:0003953	85	-0.67	-2.50	4.58E-11
GOBP_IMMUNE_RESPONSE_REGULATING_SIGNALING_PATHWAY	GO:0002764	320	-0.46	-2.04	4.69E-11
GOBP_POSITIVE_REGULATION_OF_CHEMOKINE_PRODUCTION	GO:0032722	91	-0.65	-2.48	5.38E-11
GOBP_POSITIVE_REGULATION_OF_JNK_CASCADE	GO:0046330	112	-0.62	-2.43	6.98E-11
GOBP_POSITIVE_REGULATION_OF_MAPK_CASCADE	GO:0043410	286	-0.47	-2.08	6.98E-11
GOBP_POSITIVE_REGULATION_OF_NIK_NF_KAPPAB_SIGNALING	GO:1901224	84	-0.67	-2.50	8.25E-11
GOBP_INTERLEUKIN_8_PRODUCTION	GO:0032637	117	-0.61	-2.44	8.25E-11
GOBP_NIK_NF_KAPPAB_SIGNALING	GO:0038061	121	-0.60	-2.41	1.14E-10
GOBP_POSITIVE_REGULATION_OF_INTERLEUKIN_8_PRODUCTION	GO:0032757	102	-0.63	-2.42	1.86E-10
GOBP_POSITIVE_REGULATION_OF_STRESS_ACTIVATED_PROTEIN_KINASE_SIGNALING_CASCADE	GO:0070304	129	-0.59	-2.41	2.59E-10
GOBP_PATTERN_RECOGNITION_RECEPTOR_SIGNALING_PATHWAY	GO:0002221	214	-0.50	-2.17	3.41E-10
GOBP_POSITIVE_REGULATION_OF_PHOSPHORUS_METABOLIC_PROCESS	GO:0010562	488	-0.40	-1.83	4.41E-10
GOBP_I_KAPPAB_PHOSPHORYLATION	GO:0007252	73	-0.68	-2.46	4.79E-10
GOBP_MYELOID_LEUKOCYTE_ACTIVATION	GO:0002274	151	-0.55	-2.27	5.22E-10
GOBP_POSITIVE_REGULATION_OF_INFLAMMATORY_RESPONSE	GO:0050729	123	-0.59	-2.36	5.59E-10
GOBP_REGULATION_OF_INFLAMMATORY_RESPONSE	GO:0050727	234	-0.49	-2.15	5.59E-10
GOBP_POSITIVE_REGULATION_OF_PATTERN_RECOGNITION_RECEPTOR_SIGNALING_PATHWAY	GO:0062208	111	-0.60	-2.36	6.95E-10
GOBP_PHAGOCYTOSIS	GO:0006909	201	-0.51	-2.19	7.50E-10
GOBP_REGULATION_OF_PATTERN_RECOGNITION_RECEPTOR_SIGNALING_PATHWAY	GO:0062207	180	-0.53	-2.22	7.64E-10
GOBP_MYD88_INDEPENDENT_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	GO:0002756	71	-0.69	-2.47	8.00E-10
GOBP_INTERLEUKIN_12_PRODUCTION	GO:0032615	76	-0.67	-2.44	8.36E-10
GOBP_DENDRITIC_CELL_CYTOKINE_PRODUCTION	GO:0002371	69	-0.69	-2.47	9.42E-10
GOBP_POSITIVE_REGULATION_OF_INTERLEUKIN_12_PRODUCTION	GO:0032735	72	-0.68	-2.46	9.77E-10
GOBP_POSITIVE_REGULATION_OF_INTERFERON_BETA_PRODUCTION	GO:0032728	89	-0.64	-2.42	9.95E-10
HP_UNUSUAL_INFECTION_BY_ANATOMICAL_SITE	HP:0032158	123	-0.58	-2.34	1.07E-09
GOBP_INTERLEUKIN_6_PRODUCTION	GO:0032635	139	-0.57	-2.33	1.10E-09
GOBP_RESPONSE_TO_GROWTH_FACTOR	GO:0070848	460	-0.40	-1.84	1.17E-09
GOBP_REGULATION_OF_OXIDATIVE_STRESS_INDUCED_CELL_DEATH	GO:1903201	130	-0.57	-2.33	1.26E-09
GOCC_RUFFLE	GO:0001726	130	-0.57	-2.31	1.90E-09
GOBP_NEGATIVE_REGULATION_OF_INTERLEUKIN_6_PRODUCTION	GO:0032715	80	-0.65	-2.38	2.38E-09
GOBP_CELL_DEATH_IN_RESPONSE_TO_OXIDATIVE_STRESS	GO:0036473	137	-0.56	-2.30	2.38E-09

Supplementary Table 2.6: Fifty most significant downregulated gene sets, MarBTN vs solid tissues

gs_name	gs_exact_source	setSize	enrichmentScore	NES	p.adjust
GOBP_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	GO:0002224	193	-0.58	-2.27	9.11E-16
GOBP_PATTERN_RECOGNITION_RECEPTOR_SIGNALING_PATHWAY	GO:0002221	226	-0.54	-2.12	6.67E-13
GOBP_INTERFERON_BETA_PRODUCTION	GO:0032608	136	-0.59	-2.24	6.30E-12
GOBP_REGULATION_OF_PATTERN_RECOGNITION_RECEPTOR_SIGNALING_PATHWAY	GO:0062207	191	-0.55	-2.14	8.23E-12
GOBP_IMMUNE_RESPONSE_REGULATING_SIGNALING_PATHWAY	GO:0002764	340	-0.48	-1.94	1.00E-11
GOBP_POSITIVE_REGULATION_OF_I_KAPPAB_KINASE_NF_KAPPAB_SIGNALING	GO:0043123	157	-0.57	-2.18	1.89E-11
GOBP_INTERFERON_ALPHA_PRODUCTION	GO:0032607	123	-0.61	-2.26	3.64E-11
GOMF_NADPLUS_NUCLEOSIDASE_ACTIVITY	GO:0003953	88	-0.64	-2.29	3.91E-10
GOBP_TYPE_I_INTERFERON_PRODUCTION	GO:0032606	158	-0.55	-2.11	6.91E-10
GOBP_POSITIVE_REGULATION_OF_PATTERN_RECOGNITION_RECEPTOR_SIGNALING_PATHWAY	GO:0062208	119	-0.59	-2.18	7.03E-10
GOCC_SYNAPTIC_MEMBRANE	GO:0097060	322	-0.47	-1.90	7.03E-10
GOBP_NIK_NF_KAPPAB_SIGNALING	GO:0038061	125	-0.58	-2.16	1.20E-09
GOBP_REGULATION_OF_NIK_NF_KAPPAB_SIGNALING	GO:1901222	105	-0.60	-2.19	3.04E-09
GOCC_RECEPTOR_COMPLEX	GO:0043235	320	-0.46	-1.86	3.04E-09
GOBP_CELLULAR_COMPONENT_MORPHOGENESIS	GO:0032989	494	-0.42	-1.74	3.04E-09
GOBP_POSITIVE_REGULATION_OF_INTERLEUKIN_8_PRODUCTION	GO:0032757	105	-0.59	-2.17	5.43E-09
GOBP_NEGATIVE_REGULATION_OF_HEMOPOIESIS	GO:1903707	111	-0.59	-2.16	6.76E-09
GOBP_I_KAPPAB_KINASE_NF_KAPPAB_SIGNALING	GO:0007249	214	-0.50	-1.97	7.84E-09
GOBP_DEFENSE_RESPONSE_TO_BACTERIUM	GO:0042742	241	-0.49	-1.94	9.09E-09
GOBP_REGULATION_OF_DEFENSE_RESPONSE	GO:0031347	392	-0.43	-1.77	9.71E-09
GOBP_CELL_PART_MORPHOGENESIS	GO:0032990	452	-0.42	-1.72	1.11E-08
GOBP_PEPTIDYL_TYROSINE_DEPHOSPHORYLATION	GO:0035335	260	-0.48	-1.89	1.13E-08
GOBP_INTERLEUKIN_10_PRODUCTION	GO:0032613	73	-0.65	-2.26	1.15E-08
GOBP_MYD88_DEPENDENT_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	GO:0002755	79	-0.63	-2.26	1.24E-08
GOBP_CYTOKINE_PRODUCTION	GO:0001816	460	-0.42	-1.73	1.24E-08
GOBP_POSITIVE_REGULATION_OF_INTERLEUKIN_12_PRODUCTION	GO:0032735	74	-0.65	-2.28	1.58E-08
GOMF_PROTEIN_TYROSINE_PHOSPHATASE_ACTIVITY	GO:0004725	261	-0.48	-1.90	1.89E-08
GOBP_POSITIVE_REGULATION_OF_ERK1_AND_ERK2_CASCADE	GO:0070374	138	-0.55	-2.08	2.21E-08
GOBP_POSITIVE_REGULATION_OF_INFLAMMATORY_RESPONSE	GO:0050729	127	-0.56	-2.09	2.93E-08
GOBP_NEGATIVE_REGULATION_OF_CELL_DIFFERENTIATION	GO:0045596	491	-0.41	-1.69	3.19E-08
GOBP_POSITIVE_REGULATION_OF_MAPK_CASCADE	GO:0043410	303	-0.45	-1.83	3.27E-08
GOCC_CELL_LEADING_EDGE	GO:0031252	350	-0.44	-1.78	3.33E-08
GOBP_I_KAPPAB_PHOSPHORYLATION	GO:0007252	74	-0.64	-2.25	4.32E-08
GOBP_DENDRITIC_CELL_CYTOKINE_PRODUCTION	GO:0002371	71	-0.64	-2.22	4.32E-08
GOBP_POSITIVE_REGULATION_OF_OXIDATIVE_STRESS_INDUCED_NEURON_DEATH	GO:1903223	72	-0.64	-2.22	4.72E-08
GOBP_NEGATIVE_REGULATION_OF_MYELOID_CELL_DIFFERENTIATION	GO:0045638	99	-0.59	-2.15	5.06E-08
GOBP_NEGATIVE_REGULATION_OF_INTERLEUKIN_6_PRODUCTION	GO:0032715	89	-0.60	-2.16	5.18E-08
GOBP_MYD88_INDEPENDENT_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	GO:0002756	74	-0.64	-2.24	5.55E-08
GOBP_INTERLEUKIN_1_PRODUCTION	GO:0032612	145	-0.54	-2.04	5.58E-08
GOBP_POSITIVE_REGULATION_OF_RESPONSE_TO_EXTERNAL_STIMULUS	GO:0032103	280	-0.46	-1.84	5.65E-08
GOBP_REGULATION_OF_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	GO:0034121	120	-0.55	-2.06	6.48E-08
GOBP_INTERLEUKIN_12_PRODUCTION	GO:0032615	78	-0.62	-2.22	6.66E-08
GOBP_CELL_MORPHOGENESIS_INVOLVED_IN_NEURON_DIFFERENTIATION	GO:0048667	417	-0.42	-1.72	6.66E-08
GOBP_REGULATION_OF_JNK_CASCADE	GO:0046328	143	-0.54	-2.03	6.75E-08
GOBP_NEGATIVE_REGULATION_OF_OSTEOCLAST_DIFFERENTIATION	GO:0045671	82	-0.61	-2.18	7.33E-08
GOBP_NEGATIVE_REGULATION_OF_MYELOID_LEUKOCYTE_DIFFERENTIATION	GO:0002762	86	-0.61	-2.19	7.36E-08
GOBP_INTERLEUKIN_6_PRODUCTION	GO:0032635	150	-0.53	-2.04	7.36E-08
GOBP_SYNAPTIC_SIGNALING	GO:0099536	355	-0.43	-1.74	7.49E-08
GOBP_NEGATIVE_REGULATION_OF_CYTOKINE_PRODUCTION	GO:0001818	228	-0.48	-1.89	8.83E-08
GOBP_POSITIVE_REGULATION_OF_NIK_NF_KAPPAB_SIGNALING	GO:1901224	85	-0.61	-2.17	1.01E-07

Supplementary Table 2.7: Fusion genes

https://github.com/sfhart33/MarBTNtranscriptome/outputs/fusion_gene_list.txt (too large to imbed)

Supplementary Table 2.8: Fifty most significant upregulated genes, ASW-treated vs untreated MarBTN

gene	closest uniprot hit	padj	baseMean	stat	log2FoldChange	NCBI pipeline annotation
LOC128242286	PCKGC	1.00E-295	190970	-37.02577054	-5.519020858	**phosphoenolpyruvate carboxykinase-2C cytosolic [GTP]-like**
LOC128219956	SPT13	1.18E-195	23328	-30.15130679	-5.089364883	uncharacterized LOC128219956
LOC128234666	NA	2.11E-179	43776	-28.87086699	-4.172422527	uncharacterized LOC128234666
LOC128232818	SPDEF	2.70E-174	33144	-28.45098391	-5.406427955	SAM pointed domain-containing Ets transcription factor-like
LOC128229777	RORB	1.51E-170	27069	-28.13855319	-5.128813004	nuclear receptor ROR-beta-like
LOC128222366	XIAP	1.24E-160	12220	-27.30946245	-6.075393034	**baculoviral IAP repeat-containing protein 7-B-like**
LOC128224275	RNF44	2.07E-146	1838	-26.07877692	-5.797941204	uncharacterized LOC128224275
LOC128226112	CACT	1.74E-141	10349	-25.63561381	-5.159978832	ankyrin-2-like
LOC128218330	MEST	6.41E-137	1225	-25.21817022	-4.655959882	mesoderm-specific transcript protein-like
LOC128213945	NA	1.15E-130	7232	-24.63708196	-4.743030431	uncharacterized LOC128213945
LOC128205093	AAC4	3.02E-129	13680	-24.50053341	-5.565685489	AAC-rich mRNA clone AAC4 protein-like
LOC128221568	NA	6.92E-128	2455	-24.36906241	-6.689581521	uncharacterized LOC128221568
LOC128211938	TNIP2	1.06E-127	25950	-24.34809667	-5.495253227	TNFAIP3-interacting protein 1-like
LOC128231777	ELF2	1.34E-120	5209	-23.66512684	-5.740349216	uncharacterized LOC128231777
LOC128231727	NFYA	2.20E-116	19133	-23.24927545	-5.807451225	nuclear transcription factor Y subunit alpha-like
LOC128234544	NFKI1	2.11E-106	70492	-22.23784122	-4.484777399	uncharacterized LOC128234544
LOC128227623	SQSTM	4.62E-102	35732	-21.78190305	-5.482926888	sequestosome-1-like
LOC128218206	LENG9	6.63E-102	26698	-21.76271734	-4.775789378	leukocyte receptor cluster member 9-like
LOC128202975	ATF4	4.56E-101	19415	-21.67161755	-3.975771292	uncharacterized LOC128202975
LOC128222742	NA	5.36E-101	873	-21.66183356	-4.227493469	uncharacterized LOC128222742
LOC128231639	NA	4.69E-100	5505	-21.55944931	-4.073907695	uncharacterized LOC128231639
LOC128246525	ATF4	8.13E-100	13849	-21.53176264	-3.891822468	uncharacterized LOC128246525
LOC128203258	CPTP	4.86E-96	1657	-21.12289367	-4.819214994	ceramide-1-phosphate transfer protein-like
LOC128245632	NA	4.55E-95	2318	-21.01486847	-5.395917193	uncharacterized LOC128245632
LOC128236206	AOX	1.57E-87	16098	-20.17200387	-3.802322678	uncharacterized LOC128236206
LOC128211976	TNIP2	2.71E-87	3272	-20.14308856	-4.954745135	uncharacterized LOC128211976
LOC128238596	RUNX1	3.05E-87	2106	-20.13541511	-4.019533202	runt-related transcription factor 3-like
LOC128213508	NA	1.07E-86	2288	-20.07145223	-3.699102532	uncharacterized LOC128213508
LOC128220115	IRAK4	2.98E-86	14922	-20.01856639	-4.976614765	interleukin-1 receptor-associated kinase 4-like
LOC128212313	CHSTF	4.01E-86	2324	-20.00215019	-4.227886577	carbohydrate sulfotransferase 15-like
LOC128242275	BGB	7.85E-84	2047	-19.73555212	-4.941227547	core-binding factor subunit beta-like
LOC128218161	CP2J2	1.59E-83	819	-19.69740811	-8.031603043	cytochrome P450 2B4-like
LOC128206168	ANKK1	1.59E-83	7839	-19.69653245	-3.731284023	ankyrin repeat and protein kinase domain-containing protein 1-like
LOC128242795	SL9A2	6.07E-83	12237	-19.62718518	-4.104839159	sodium/hydrogen exchanger 3-like
LOC128229194	HS12B	1.28E-79	390	-19.23268357	-5.485154965	heat shock 70 kDa protein 12B-like
LOC128217567	ANR10	3.14E-78	6698	-19.06474496	-3.253584741	26S proteasome non-ATPase regulatory subunit 10-like
LOC128232100	MEGF6	1.05E-76	1578	-18.87871379	-3.307954483	protein draper-like
LOC128225168	NA	2.43E-75	1084	-18.71071405	-5.044410357	phenolphthiocerol/phthiocerol polyketide synthase subunit B-like
LOC128209666	NA	7.32E-75	11501	-18.65044739	-3.477447376	uncharacterized LOC128209666
LOC128207330	NA	3.68E-73	28725	-18.43849865	-2.48933525	uncharacterized LOC128207330
LOC128209840	SLIT1	1.53E-71	1219	-18.23432632	-2.906704681	uncharacterized LOC128209840
LOC128226763	NA	1.04E-70	1752	-18.12794493	-2.977938286	uncharacterized LOC128226763
LOC128241998	PTHD3	5.30E-70	6578	-18.03695492	-3.441888232	uncharacterized LOC128241998
LOC128230810	NA	5.32E-70	546	-18.03541266	-4.226752059	uncharacterized LOC128230810
LOC128222121	LHPL2	1.63E-69	6222	-17.97201312	-3.970681522	LHFPL tetraspan subfamily member 2a protein-like
LOC128213233	GA45A	2.71E-68	6261	-17.81427486	-3.175887784	growth arrest and DNA damage-inducible protein GADD45 alpha-like
LOC128240352	NA	1.41E-66	994	-17.59051039	-6.013606079	uncharacterized LOC128240352
LOC128210003	NA	1.68E-66	23004	-17.57936927	-3.866863979	uncharacterized LOC128210003
LOC128226830	NA	2.17E-66	2215	-17.56385161	-2.972863716	uncharacterized LOC128226830
LOC128217174	SERC	5.67E-65	57381	-17.37643641	-3.609933007	phosphoserine aminotransferase-like

Supplementary Table 2.9: Significantly upregulated gene sets, ASW-treated vs untreated MarBTN

gs_name	gs_exact_source	setSize	enrichmentScore	NES	p.adjust
GOBP_CELLULAR_RESPONSE_TO_INTERLEUKIN_1	GO:0071347	52	0.76	2.23	5.35E-05
GOBP_RESPONSE_TO_INTERLEUKIN_1	GO:0070555	57	0.72	2.15	0.000319503
GOBP_CYTOKINE_PRODUCTION	GO:0001816	383	0.45	1.66	0.002185671
GOMF_DNA_BINDING_TRANSCRIPTION_ACTIVATOR_ACTIVITY	GO:0001216	179	0.52	1.80	0.005012352
GOBP_REGULATION_OF_ALCOHOL_BIOSYNTHETIC_PROCESS	GO:1902930	24	0.82	2.12	0.006018954
GOBP_STEROID_METABOLIC_PROCESS	GO:0008202	145	0.53	1.81	0.006186566
GOBP_RESPONSE_TO_CYTOKINE	GO:0034097	436	0.42	1.57	0.006186566
GOBP_LIPID_LOCALIZATION	GO:0010876	255	0.47	1.68	0.006223428
GOBP_NEGATIVE_REGULATION_OF_CELL_DIFFERENTIATION	GO:0045596	397	0.42	1.56	0.008847163
GOBP_POSITIVE_REGULATION_OF_TRANSCRIPTION_FROM_RNA_POLYMERASE_II_PROMOTER_IN_RESPONSE_TO_STRESS	GO:0036003	17	0.85	2.00	0.009307882
GOBP_REGULATION_OF_INOSITOL_PHOSPHATE_BIOSYNTHETIC_PROCESS	GO:0010919	11	0.91	1.98	0.009484974
GOBP_CYTOKINE_MEDIATED_SIGNALING_PATHWAY	GO:0019221	168	0.51	1.76	0.009484974
GOBP_CELLULAR_RESPONSE_TO_ARSENIC_CONTAINING_SUBSTANCE	GO:0071243	16	0.87	2.04	0.010486159
GOMF_STEROL_BINDING	GO:0032934	57	0.65	1.95	0.010486159
GOMF_DNA_BINDING_TRANSCRIPTION_FACTOR_ACTIVITY	GO:0003700	383	0.42	1.55	0.01233816
GOBP_POSITIVE_REGULATION_OF_TRANSCRIPTION_FROM_RNA_POLYMERASE_II_PROMOTER_IN_RESPONSE_TO_ENDOP	GO:1990440	12	0.90	2.01	0.0137245
GOBP_STEROL_HOMEOSTASIS	GO:0055092	67	0.62	1.90	0.014768071
GOBP_RESPONSE_TO_ARSENIC_CONTAINING_SUBSTANCE	GO:0046685	20	0.82	2.00	0.014880588
GOBP_TOLL_LIKE_RECEPTOR_9_SIGNALING_PATHWAY	GO:0034162	35	0.72	1.97	0.014880588
GOBP_POSITIVE_REGULATION_OF_HEMOPOIESIS	GO:1903708	48	0.67	1.93	0.014880588
GOMF_HEPARIN_BINDING	GO:0008201	111	0.54	1.78	0.014880588
GOMF_EXTRACELLULAR_MATRIX_STRUCTURAL_CONSTITUENT	GO:0005201	196	0.47	1.64	0.015170715
GOBP_T_CELL_ACTIVATION_INVOLVED_IN_IMMUNE_RESPONSE	GO:0002286	31	0.74	1.95	0.015244712
GOBP_T_CELL_DIFFERENTIATION_INVOLVED_IN_IMMUNE_RESPONSE	GO:0002292	31	0.74	1.95	0.015244712
GOBP_VASCULAR_ENDOTHELIAL_GROWTH_FACTOR_PRODUCTION	GO:0010573	22	0.79	1.97	0.018180271
GOBP_AMIDE_BIOSYNTHETIC_PROCESS	GO:0043604	270	0.45	1.64	0.019172575
GOMF_SULFUR_COMPOUND_BINDING	GO:1901681	174	0.49	1.68	0.019536085
GOBP_POSITIVE_REGULATION_OF_ALCOHOL_BIOSYNTHETIC_PROCESS	GO:1902932	12	0.88	1.97	0.022374532
GOBP_REGULATION_OF_DNA_TEMPLATED_TRANSCRIPTION_IN_RESPONSE_TO_STRESS	GO:0043620	25	0.77	1.97	0.024888386
GOBP_T_HELPER_17_CELL_DIFFERENTIATION	GO:0072539	12	0.88	1.96	0.027778048
GOBP_T_HELPER_17_TYPE_IMMUNE_RESPONSE	GO:0072538	12	0.88	1.96	0.027778048
GOBP_CELLULAR_AMIDE_METABOLIC_PROCESS	GO:0043603	471	0.39	1.46	0.027778048
GOBP_RESPONSE_TO_TUMOR_NECROSIS_FACTOR	GO:0034612	120	0.52	1.71	0.028982452
GOCC_COLLAGEN_CONTAINING_EXTRACELLULAR_MATRIX	GO:0062023	294	0.42	1.55	0.032160761
GOBP_LENS_DEVELOPMENT_IN_CAMERA_TYPE_EYE	GO:0002088	44	0.65	1.82	0.039454968
GOBP_CD4_POSITIVE_ALPHA_BETA_T_CELL_DIFFERENTIATION	GO:0043367	25	0.75	1.93	0.039987441
GOBP_CHONDROCYTE_DEVELOPMENT_INVOLVED_IN_ENDOCHONDRAL_BONE_MORPHOGENESIS	GO:0003433	15	0.82	1.90	0.041152314
GOBP_CARBOHYDRATE_BIOSYNTHETIC_PROCESS	GO:0016051	106	0.52	1.71	0.043546804
GOCC_LEADING_EDGE_MEMBRANE	GO:0031256	111	0.51	1.69	0.043546804
GOBP_POSITIVE_REGULATION_OF_CYTOKINE_PRODUCTION	GO:0001819	245	0.43	1.55	0.043546804
GOBP_POSITIVE_REGULATION_OF_LYMPHOCYTE_DIFFERENTIATION	GO:0045621	35	0.69	1.88	0.044165604
GOMF_STEROID_BINDING	GO:0005496	68	0.58	1.77	0.044165604
GOBP_SMALL_MOLECULE_BIOSYNTHETIC_PROCESS	GO:0044283	275	0.42	1.54	0.044165604
GOBP_REGULATION_OF_LIPID_STORAGE	GO:0010883	33	0.71	1.89	0.044207822
GOMF_CIS_REGULATORY_REGION_SEQUENCE_SPECIFIC_DNA_BINDING	GO:0000987	369	0.39	1.46	0.049206775

Supplementary Table 2.10: Fifty most significant downregulated genes, ASW-treated vs untreated MarBTN

gene	closest uniprot hit	padj	baseMean	stat	log2FoldChange	NCBI pipeline annotation
LOC128243119	ZNFX1	9.69E-60	1023	16.66	3.27	**NFX1-type zinc finger-containing protein 1-like**
LOC128220373	NA	1.18E-36	1548	13.06	2.16	uncharacterized LOC128220373
LOC128210260	NPFF2	1.29E-36	1306	13.05	2.42	cholecystokinin receptor type A-like
LOC128226839	CCNG1	1.68E-31	395	12.11	2.71	**cyclin-G1-like**
LOC128241732	CX038	1.92E-31	707	12.09	2.19	uncharacterized LOC128241732
LOC128240343	CTDSL	2.60E-31	915	12.07	2.11	**CTD small phosphatase-like protein**
LOC128219566	NA	1.04E-30	214	11.95	3.33	uncharacterized LOC128219566
LOC128208299	BTBD3	7.23E-30	2138	11.79	2.01	BTB/POZ domain-containing protein 6-B-like
LOC128238571	HDC	1.98E-29	3881	11.70	2.73	headcase protein homolog
LOC128232113	MERL	7.76E-29	432	11.58	2.29	merlin-like
LOC128238451	ENPP4	1.45E-28	425	11.53	2.43	bis(5'-adenosyl)-triphosphatase ENPP4-like
LOC128240988	SP3	8.32E-28	549	11.37	2.44	transcription factor Sp4-like
LOC128245448	CASZ1	1.28E-26	423	11.12	2.41	zinc finger protein castor homolog 1-like
LOC128214321	NA	2.08E-25	1528	10.87	1.64	uncharacterized LOC128214321
LOC128240628	NA	4.70E-25	82731	10.79	1.58	uncharacterized LOC128240628
LOC128239332	HDC	9.18E-25	3049	10.73	2.34	headcase protein homolog
LOC128230057	FOXM1	2.01E-24	275	10.65	2.68	uncharacterized LOC128230057
LOC128224801	NA	2.69E-24	557	10.63	2.04	uncharacterized LOC128224801
LOC128230323	OAR	6.00E-24	200	10.55	2.85	uncharacterized LOC128230323
LOC128219750	H6ST2	8.37E-24	408	10.51	2.35	heparan-sulfate 6-O-sulfotransferase 2-like
LOC128229638	NA	1.04E-23	890	10.49	1.98	uncharacterized LOC128229638
LOC128214390	GPR84	1.16E-23	541	10.48	2.80	cholecystokinin receptor type A-like
LOC128213877	NA	3.16E-23	924	10.38	1.80	uncharacterized LOC128213877
LOC128235372	GLBL2	3.25E-23	440	10.38	2.02	beta-galactosidase-1-like protein 2
LOC128209992	NA	8.69E-23	698	10.28	1.84	actin-3-like
LOC128233459	DEDD	9.91E-23	559	10.27	1.90	death effector domain-containing protein-like
LOC128227624	TRIM2	1.03E-22	636	10.27	1.81	tripartite motif-containing protein 2-like
LOC128208515	TRI33	6.60E-22	234	10.08	2.80	E3 ubiquitin-protein ligase TRIM33-like
LOC128228661	TFIP8	9.01E-22	5651	10.05	1.39	tumor necrosis factor alpha-induced protein 8-like
LOC128221235	TRI45	9.40E-22	770	10.05	1.69	tripartite motif-containing protein 45-like
LOC128220204	MED4	1.10E-20	583	9.79	1.82	mediator of RNA polymerase II transcription subunit 4-like
LOC128218787	TCAB1	1.52E-20	726	9.76	1.86	telomerase Cajal body protein 1-like
LOC128230249	FUTC	3.03E-20	442	9.69	1.93	alpha-(1,3)-fucosyltransferase 7-like
LOC128241784	NA	4.57E-20	229	9.65	2.55	uncharacterized LOC128241784
LOC128235494	NA	1.41E-19	2349	9.53	1.56	uncharacterized LOC128235494
LOC128242639	NA	3.28E-19	3148	9.44	1.41	zinc finger protein 62 homolog
LOC128239013	NA	3.30E-19	867	9.44	1.98	uncharacterized LOC128239013
LOC128228804	MB212	3.70E-19	242	9.42	2.26	uncharacterized LOC128228804
LOC128242530	NA	1.18E-18	3500	9.30	1.41	uncharacterized LOC128242530
LOC128227485	NA	1.76E-18	702	9.25	1.59	PWWP domain-containing DNA repair factor 3A-like
LOC128208968	KLH30	2.03E-18	130	9.24	3.06	uncharacterized LOC128208968
LOC128244180	TIF1B	2.28E-18	91	9.22	3.89	protein wech-like
LOC128229010	TFAP4	9.03E-18	228	9.07	2.49	transcription factor AP-4-like
LOC128217374	NA	1.26E-17	1711	9.03	1.38	uncharacterized LOC128217374
LOC128220218	UBIE	1.38E-17	403	9.02	1.81	demethylmenaquinone methyltransferase-like
LOC128227172	LONF3	1.90E-17	473	8.99	1.77	LON peptidase N-terminal domain and RING finger protein 3-like
LOC128242277	NA	2.40E-17	138	8.96	2.72	RNA-binding protein 25-like
LOC128209819	KLH20	4.72E-17	109	8.88	4.53	uncharacterized LOC128209819
LOC128203128	B3GT6	5.33E-17	284	8.87	1.97	beta-1,3-galactosyltransferase 6-like
LOC128231166	TAF8	8.43E-17	508	8.81	1.81	transcription initiation factor TFIID subunit 8-like

Supplementary Table 2.11: Fifty most significant downregulated gene sets, ASW-treated vs untreated MarBTN

gs_name	gs_exact_source	setSize	enrichmentScore	NES	p.adjust
GOBP_MITOTIC_CELL_CYCLE_PROCESS	GO:1903047	418	-0.439470402	-1.83658179	4.78E-07
GOBP_MITOTIC_CELL_CYCLE	GO:0000278	499	-0.4117496	-1.734948239	7.21E-07
GOCC_TRANSFERASE_COMPLEX	GO:1990234	401	-0.430192701	-1.790960277	1.10E-06
GOCC_CULLIN_RING_UBIQUITIN_LIGASE_COMPLEX	GO:0031461	98	-0.614571357	-2.192686117	6.27E-06
GOBP_CELL_CYCLE_PHASE_TRANSITION	GO:0044770	303	-0.452445796	-1.83286045	6.27E-06
GOCC_INTRACELLULAR_PROTEIN_CONTAINING_COMPLEX	GO:0140535	388	-0.419377649	-1.728611729	1.51E-05
GOBP_CELL_DIVISION	GO:0051301	400	-0.411664131	-1.71314055	2.85E-05
GOBP_CHROMATIN_ORGANIZATION	GO:0006325	287	-0.435077645	-1.758305449	5.05E-05
GOBP_EMBRYO_DEVELOPMENT_ENDING_IN_BIRTH_OR_EGG_HATCHING	GO:0009792	318	-0.423265764	-1.727268017	7.40E-05
GOBP_REGULATION_OF_CELL_CYCLE_PHASE_TRANSITION	GO:1901987	253	-0.443624159	-1.773661564	0.000262536
HP_ABNORMALITY_OF_CENTRAL_NERVOUS_SYSTEM_ELECTROPHYSIOLOGY	HP:0030178	318	-0.4075682	-1.663209207	0.000550681
GOBP_MITOTIC_CELL_CYCLE_PHASE_TRANSITION	GO:0044772	228	-0.441511824	-1.735500916	0.000643734
GOBP_REGULATION_OF_CELL_CYCLE_PROCESS	GO:0010564	392	-0.382264559	-1.581099715	0.001123891
GOCC_CUL4_RING_E3_UBIQUITIN_LIGASE_COMPLEX	GO:0080008	21	-0.808474254	-2.11373308	0.001927402
GOBP_ORGANELLE_FISSION	GO:0048285	272	-0.410329676	-1.642976643	0.002411383
GOMF_HISTONE_BINDING	GO:0042393	146	-0.482977255	-1.79750884	0.002456214
GOCC_MICROTUBULE_ORGANIZING_CENTER	GO:0005815	444	-0.357741045	-1.505389273	0.003964917
GOCC_UBIQUITIN_LIGASE_COMPLEX	GO:0000151	170	-0.457353353	-1.739428719	0.00410877
HP_ORAL_CLEFT	HP:0000202	295	-0.394253478	-1.599757679	0.005012352
GOCC_SPINDLE_POLE	GO:0000922	85	-0.540658365	-1.871784083	0.006186566
HP_DEEPLY_SET_EYE	HP:0000490	127	-0.490496023	-1.801017644	0.006186566
GOBP_IN_UTERO_EMBRYONIC_DEVELOPMENT	GO:0001701	184	-0.443202449	-1.704524658	0.006698494
GOBP_SPINDLE_ORGANIZATION	GO:0007051	118	-0.488118278	-1.78004299	0.007157428
HP_INTELLECTUAL_DISABILITY_SEVERE	HP:0010864	263	-0.39997135	-1.601271152	0.008102344
GOBP_NEGATIVE_REGULATION_OF_CELL_CYCLE	GO:0045786	240	-0.404446488	-1.599469992	0.008102344
GOBP_PROTEASOMAL_PROTEIN_CATABOLIC_PROCESS	GO:0010498	282	-0.394180593	-1.585705439	0.008102344
GOBP_MICROTUBULE_CYTOSKELETON_ORGANIZATION	GO:0000226	340	-0.374389049	-1.53449556	0.008102344
GOCC_CENTRIOLE	GO:0005814	102	-0.51246503	-1.826571726	0.009307882
GOMF_METHYLATED_HISTONE_BINDING	GO:0035064	64	-0.578883768	-1.931036114	0.009484974
GOBP_CELL_CYCLE_G2_M_PHASE_TRANSITION	GO:0044839	86	-0.530932646	-1.854364435	0.009484974
GOBP_PROTEIN_COMPLEX_OLIGOMERIZATION	GO:0051259	126	-0.469094732	-1.717595219	0.009909266
GOBP_HISTONE_MODIFICATION	GO:0016570	309	-0.376214811	-1.536662089	0.009909266
GOBP_SISTER_CHROMATID_SEGREGATION	GO:0000819	129	-0.472789659	-1.742164339	0.010486159
HP_ABNORMAL_NERVOUS_SYSTEM_ELECTROPHYSIOLOGY	HP:0001311	373	-0.359362554	-1.489267548	0.010486159
GOMF_CATALYTIC_ACTIVITY_ACTING_ON_A_NUCLEIC_ACID	GO:0140640	390	-0.355957386	-1.472311779	0.010486159
HP_HIGH_PALATE	HP:0000218	418	-0.353736332	-1.478292286	0.010495827
GOBP_DNA_REPAIR	GO:0006281	375	-0.357965064	-1.485515932	0.011925839
GOMF_DAMAGED_DNA_BINDING	GO:0003684	74	-0.549307041	-1.848694178	0.01233816
GOCC_NUCLEAR_CHROMOSOME	GO:0000228	105	-0.496610991	-1.779624627	0.01233816
GOBP_REGULATION_OF_MITOTIC_CELL_CYCLE_PHASE_TRANSITION	GO:1901990	179	-0.42842418	-1.644292498	0.01233816
HP_CLINODACTYLY	HP:0030084	329	-0.36401256	-1.496312671	0.01233816
HP_ABNORMAL_ARACHNOID_MATER_MORPHOLOGY	HP:0100700	67	-0.549826956	-1.829989561	0.012804423
HP_ATROPHY_DEGENERATION_AFFECTING_THE_CEREBRUM	HP:0007369	469	-0.336274404	-1.416832317	0.012865589
HP_FINGER_CLINODACTYLY	HP:0040019	194	-0.415467145	-1.611108267	0.0137245
GOBP_MICROTUBULE_BASED_PROCESS	GO:0007017	458	-0.336372305	-1.41481604	0.0137245
GOBP_MICROTUBULE_ORGANIZING_CENTER_ORGANIZATION	GO:0031023	91	-0.517811939	-1.822318251	0.014143519
HP_INTERICTAL_EEG_ABNORMALITY	HP:0025373	162	-0.436681444	-1.645317465	0.01435951
GOBP_NEGATIVE_REGULATION_OF_CELL_CYCLE_PROCESS	GO:0010948	172	-0.431704762	-1.63884061	0.01435951
GOBP_RNA_SPLICING	GO:0008380	193	-0.416549925	-1.613569537	0.014660567
HP_SYNDACTYLY	HP:0001159	244	-0.385883095	-1.528958872	0.014880588

Chapter 3. Using genetic markers to track bivalve transmissible neoplasia lineages

Contributions to projects as published in:

- Michnowska A., **Hart S.**, Smolarz K., Hallmann A., Metzger M. (2022). *Horizontal transmission of disseminated neoplasia in the widespread clam Macoma balthica from Southern Baltic Sea*. Molecular Ecology, 31(11): 3128–3136.
- Giersch R., **Hart S.**, Reddy S., Yonemitsu M., Rosales M., Korn M., Geleta B., Countway P., Robledo J., Metzger M. (2022). *Survival and detection of bivalve transmissible neoplasia from the soft-shell clam Mya arenaria (MarBTN) in seawater*. Pathogens, 11(3), 283.
- Giersch, R., Sevigny J., Garrett F., Tindbaek K., Yonemitsu M., **Hart S.**, Metzger M. (in prep). *Long-term progression and regression dynamics of Mya arenaria bivalve transmissible neoplasia (MarBTN) in vivo*.

3.1 Abstract

Eight instances of transmissible cancer, clonal lineages of cancer cells that spread from host to host through a population, have been documented in bivalve species. This indicates bivalves are particularly susceptible to contagious cancers, and the development of genetic tools for identifying and tracking these cancers is critical for understanding when and how they spread. We developed a qPCR assay specific to a transmissible cancer lineage in soft-shell clams (*Mya arenaria*) that is effective for quantifying the ratio of cancer vs host cells in infected clams. This assay can be used to confirm a cancer is this same lineage, detect infection at visually undetectable levels, track within-host disease progression, and detect cancer DNA in eDNA samples. We also identified a conserved locus in the Baltic clam (*Macoma balthica*) and used targeted sequencing to confirm a new transmissible cancer lineage in this species. This study provides reproducible methods for tracking a known transmissible cancer and for identifying a novel transmissible cancer that can be applied to other transmissible cancer lineages.

3.2 Introduction

Cancer is typically an evolutionary dead end for cancer cells themselves, ending in either regression or host death. However, several cancer lineages have managed to evade this fate by

transmitting to new hosts and spreading as a contagion through a host population[113]. Transmissible cancer lineages have been documented in Tasmanian devils, arising independently twice in the last ~50 years[4,16,112], in dogs, arising once and spreading worldwide over the past 4000+ years[2,12], and in marine bivalves, arising eight times and spreading through nine different host species[6–10]. While research over the past two decades has illuminated much about the biology of the mammalian transmissible cancers and how they affect their host populations[11–16], much is still unknown about the bivalve transmissible neoplasia (BTN) lineages.

A leukemia-like bivalve cancer was first reported in oysters in 1969[144], characterized by amplification of cells in the circulatory fluid, and has since been identified in at least 15 bivalve species[27,41,145]. This cancer was termed hemic or disseminated neoplasia, and outbreaks affecting up to 90% of certain local populations have been reported[22,146]. In 2015, it was discovered that disseminated neoplasia in the soft-shell clam (*Mya arenaria*) was in fact a single clonal lineage of cells that had spread through the North Atlantic soft-shell clam population[6]. Seven other clonal lineages of bivalve transmissible neoplasia (BTN) have been identified since, and it is likely that many other documented instances of disseminated neoplasia are additional BTN lineages[6–10]. Little is known about how BTN infections progress within individual bivalves or their direct effect on host survival, besides what we can infer from past records of disseminated neoplasia incidences linked with population mortality events[22,146]. Given each BTN lineage is genetically clonal, having arisen from a single founder cancer cell, polymorphisms specific to each BTN lineage are prime targets for genetic assays to track BTN infections, allowing us to study the unique host-pathogen relationships of contagious cancers. Here we report the development of a qPCR assay specific to *Mya arenaria* BTN (MarBTN) that we use to identify and track cancer infections in individual clams, and the identification of a novel BTN lineage in the Baltic clam using targeted sequencing. Each study provides a blueprint that could easily be adapted to identify and track other BTNs in other species.

3.3 Results

3.3.1 Tracking MarBTN with a qPCR assay

A qPCR assay specific to MarBTN DNA is a quick and relatively inexpensive method to detect and track MarBTN infections. Previous studies discovered an expansion of LTR-retrotransposon *Steamer* in the MarBTN lineage[46] and identified loci shared by all cancers but not found in healthy clams[6]. We previously designed primers specific to one of these MarBTN-specific loci (termed "HL03" after the MarBTN isolate the integration site was initially characterized in) using a forward primer in *Steamer*'s 3' long terminal repeat (LTR) and a reverse primer in the genomic DNA just downstream the *Steamer* insertion (**Supplementary Table 3.1**). As control primers we targeted conserved gene *EF1α* with primers conserved in all soft-shell clam DNA (including MarBTN). The MarBTN-specific primers effectively detected MarBTN DNA (**Fig. 3.1A**), but we quickly realized that the ratio of HL03 to *EF1α* did not yield an accurate estimate of the fraction of MarBTN in the hemolymph. This is likely due to copy number differences between the HL03 and *EF1α* loci in MarBTN and variation in copy number across MarBTN samples, unsurprising given the extensive copy number variation previously documented in MarBTN[24,147].

We utilized our recent genome assembly and analysis of the MarBTN genome[147] to identify a *Steamer* insertion site better suited for generating an informative estimate of the fraction of MarBTN cells in hemolymph samples. We identified a *Steamer* insertion 254 bp upstream the gene encoding histone-binding protein N1/N2, designing MarBTN-specific and control *N1N2* primers as done for HL03/*EF1α* (**Fig. 3.1B, Supplementary Table 3.1**) that accurately differentiated MarBTN-infected clams from uninfected clams in a test panel (**Fig. 3.1D**). This locus is tetraploid in the observed MarBTN cells, with *Steamer* insertions in two of four alleles, while healthy clam cells are diploid. Thus, we calculate the fraction of MarBTN in a hemolymph sample given a measured ratio of MarBTN-specific alleles and control *N1N2* alleles (**Fig. 3.1C,** derivation in Methods).

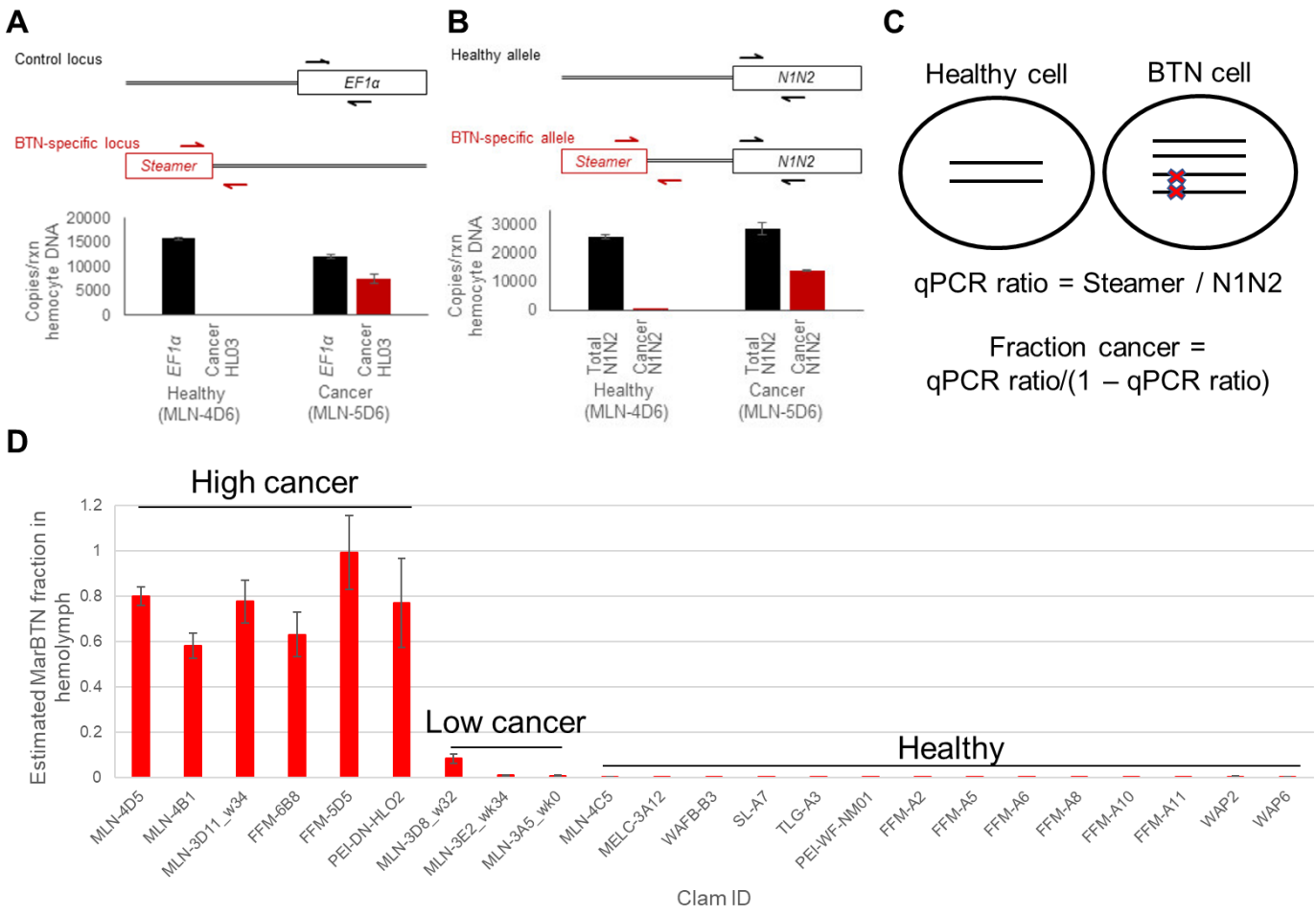


Figure 3.1. MarBTN-specific qPCR marker

Primer binding schematic and example qPCR results for **(A)** the former MarBTN-specific locus "HL03" and **(B)** newly identified locus "*N1N2*" (*figures from Giersch et al.*)[17]. The schematic shows the target control alleles and the MarBTN-associated alleles, with arrows indicating the locations of the control primers (black), used to quantify total clam DNA, and primers specific for the MarBTN lineage (red), used to quantify MarBTN DNA. Amplification results from example healthy and MarBTN-infected hemolymph samples for each primer pair are shown below. **(C)** Representation of *N1N2* locus for healthy host hemocytes and MarBTN cells. Hemolymph samples of MarBTN-infected clams will be a mix of each, with two control alleles for every host cell (diploid) and four control and two BTN-specific alleles (red X) for every BTN cell (tetraploid at *N1N2* locus). The fraction of cancer cells in the sample can be calculated using the accompanying equation. **(D)** Representative samples validating qPCR assay, with high fractions in samples with high cancer infections, low fractions in samples with low cancer infections, and undetectable in healthy samples. For all assays, each sample was run in three reactions, and the values presented are averages of the triplicate results with standard deviation shown as error bars.

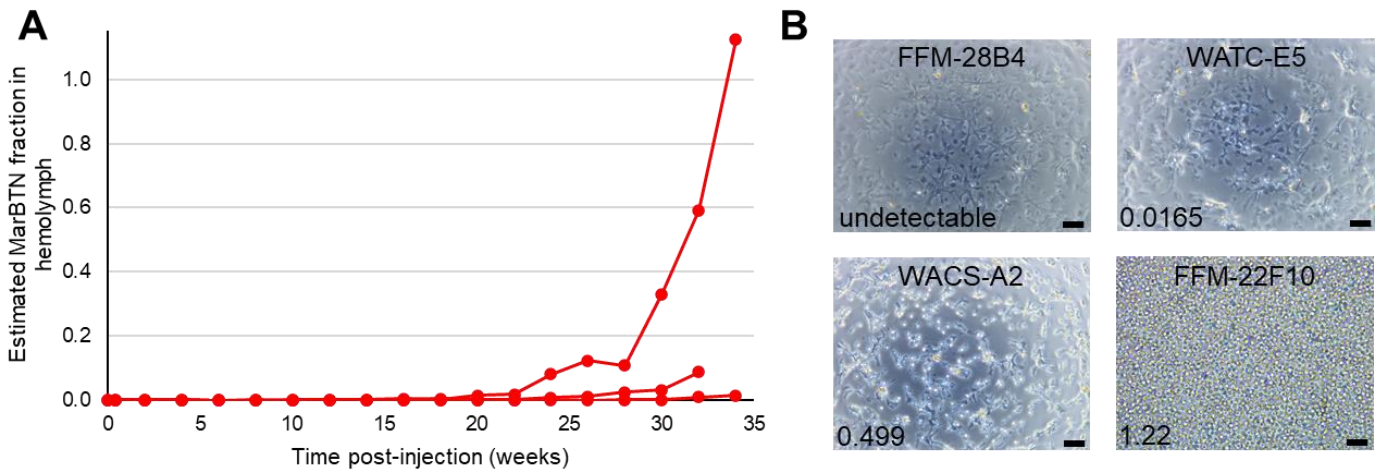


Figure 3.2. MarBTN-specific qPCR tracks percent cancer in hemolymph

(A) Fraction of cancer cells in hemolymph of three clams injected with MarBTN cells at time 0 and tracked until host death. **(B)** Representative hemolymph images with corresponding MarBTN cell fraction calculations from qPCR for clams without infection (upper left), not visibly infected but qPCR-positive (upper right), moderately infected with a mixture of host and cancer cells (lower left), and severely infected without visible host hemocytes. Scale bars are 50 µm.

The *N1N2* locus qPCR assay has multiple valuable applications for the study of MarBTN. The ability of this assay to estimate cancer cell fractions makes it particularly well suited for tracking MarBTN infections longitudinally by periodically sampling hemolymph (**Fig. 3.2A**). The assay is also an effective method for screening clams for low-level MarBTN infection. Early infections may not be visually apparent in hemolymph viewed under the microscope (e.g. WATC-E5, **Fig. 3.2B**), the method of identifying disseminated neoplasia in soft-shell clams before it was discovered to be a single transmissible lineage[19,20,48,95]. Additionally, the assay quickly confirms that a cancer is MarBTN, as opposed to a conventional cancer or independent lineage with a similar cellular phenotype. Confirming naïve and MarBTN-infected clams is critical prior to controlled experiments to understand disease progression. Finally, this assay is effective at detecting MarBTN presence from eDNA in both experimental and natural settings[17].

3.3.2 New BTN lineage in *Macoma balthica*

Eight lineages of BTN have been confirmed since MarBTN was the first to be characterized as a transmitting lineage in 2015[6]. Given the apparent abundance of BTN lineages, it is likely that many other observations of disseminated neoplasia[27,41] are in fact additional BTN lineages. The established method

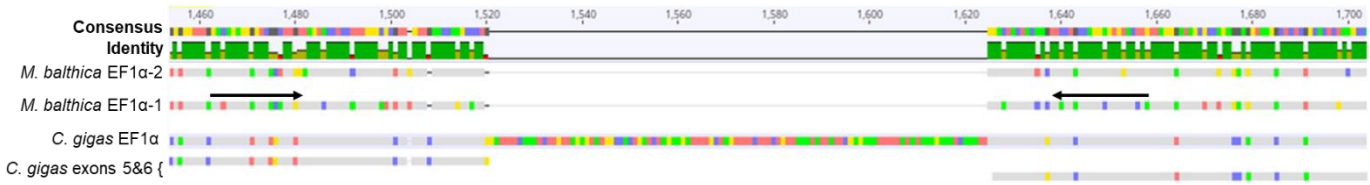


Figure 3.3. Intron-spanning primers for EF1α locus in Macoma balthica

Alignment of the two *M. balthica EF1α* transcripts from Yurchenko et al. to the *C. gigas EF1α* locus. Location of primer binding marked by arrows. Colored nucleotides indicate polymorphic loci (C = blue, T = green, G = yellow, A = red), consensus sequence estimate, and pairwise identity (100% = green, 30-99% = brown, >30% = red) are graphics are from Geneious®.

for confirming cancers belong to a single (or multiple) clonal lineage(s) involves classifying alleles across a cohort of hosts and cancer samples, with cancer alleles that differ from the host but match one another, indicating a clonal lineage. This is especially effective when targeting an intron within a conserved gene, as exons make for conserved primer-binding sites while sequencing across an intron captures greater polymorphism of a non-coding region. We applied this approach to investigate whether a reported disseminated neoplasia in the Baltic clam (*Macoma balthica*, previously *Limecola baltica*) was a BTN lineage.

M. balthica does not have a published reference genome, so we instead queried a *M. balthica* transcriptome data set[148] to identify *M. balthica EF1α*, a conserved gene successfully used in previous BTN identification studies[6,7]. We identified two transcripts with an annotation of EF1α. These transcripts aligned to one another with a high degree of polymorphism, indicating these may represent multiple copies of *EF1α*-like genes in the *M. balthica* genome rather than different isoforms or haplotypes. We assumed intron placement would likely be conserved in the *EF1α* gene among bivalves, so we aligned these transcripts to the Pacific oyster genome (Crassostrea gigas: version GCA_000297895.1[149]) *EF1α* region (**Fig. 3.3**). We then chose primer sites to target coding sequence and amplify across the fifth intron. We chose primers that would only target one of the *EF1α* transcripts (transcript ID: "evgsoapLocus_712298") to control for the possibility of multiple *EF1α* orthologs in the *M. balthica* genome. We used these primers to sequence and deconvolute alleles from four advanced stage disseminated neoplasia samples (hemolymph and host tissue) and two healthy clams as a control.

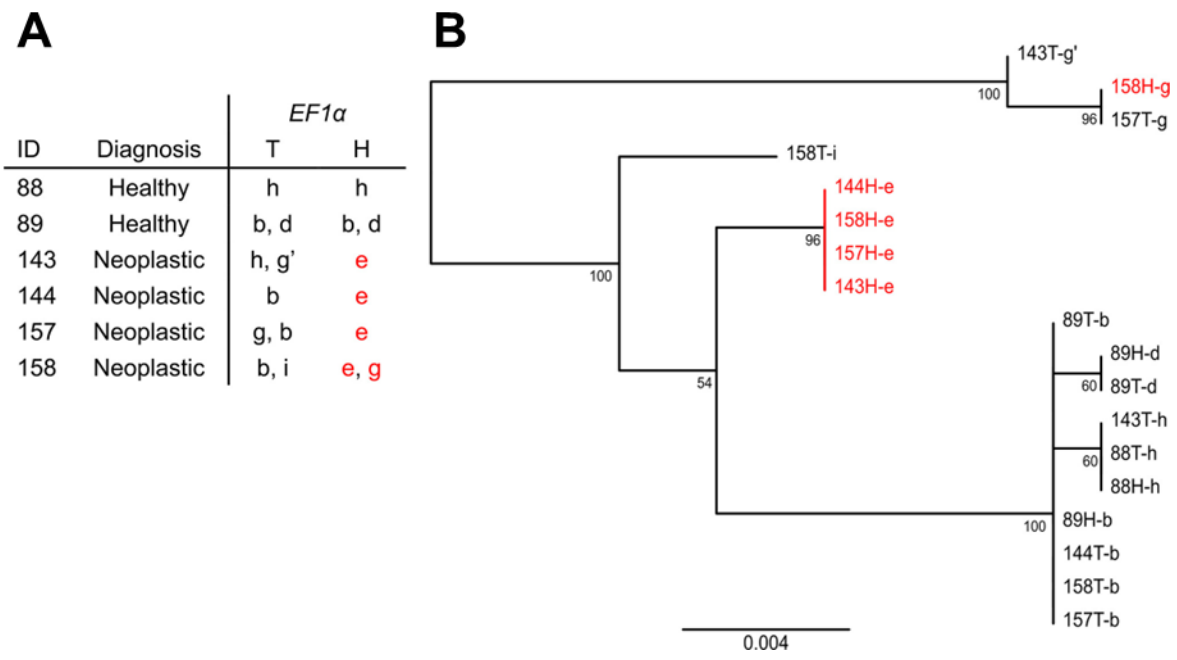


Figure 3.4. EF1α alleles identify a transmissible cancer lineage in *Macoma balthica*

Figure from Michnowska et al.[10] *EF1α* sequences from selected *M. balthica* individuals reveal a common cancer-associated allele. **(A)** *EF1α* alleles associated with hemolymph (H) and solid tissue (T) are listed. Each unique allele is each assigned a letter (b-i). **(B)** Phylogenetic tree of *EF1α* locus across hemolymph and solid tissue across healthy and neoplastic clams. Numbers below the nodes indicate bootstrap values; values below 50 are not shown. Alleles observed in neoplastic cells are highlighted in red. Scale bar represents genetic distance between sequences.

Eight *EF1α* alleles were identified across the sample set, reflecting the pattern of cancer and host alleles that would be expected for a transmissible cancer **(Fig. 3.4A)**. The two healthy clams had allele matches between hemolymph and tissue, as expected for normal animals. Clams diagnosed with disseminated neoplasia all had *EF1α* allele "e" present in hemolymph DNA and absent (or found at a much lower level) in host tissue samples. An additional *EF1α* "g" allele was detected in the cancerous hemolymph sample from individual "158" that was not present in the solid tissue of this specimen, although it was found in tissue DNA of another neoplastic clam. It is possible this represents an allele from the founder clam that has since been lost from the other three samples, a common phenomenon observed in the MarBTN genome[147]. Overall, these results confirm a BTN lineage is present in the Baltic clam, a finding supported by similar analysis performed on a mitochondrial locus for this same sample set[10].

3.4 Discussion

Genetic assays provide a powerful tool to study the spread of clonal lineages of contagious cancer cells. MarBTN has been confirmed at select sites through the soft-shell clam's native range along the east coast of North America[6], but we do not know the full extent of the transmissible cancer's spread. The MarBTN-specific qPCR assay, applied directly to clam population surveys or indirectly via eDNA testing[17], could illuminate the extent of the lineage's spread and incidences within clam populations. The qPCR assay can also be applied to investigate how MarBTN infections progress and their effect on mortality within individual clams, another area that has not been investigated since the discovery that disseminated neoplasia was a transmissible cancer. This assay is a cheap and useful tool for uncovering how MarBTN affects soft-shell clam populations and could easily be adapted to target other BTNs.

BTN lineages have now been identified in nine host species[6–10], including the lineage we identify here in the Baltic clam. This indicates that bivalves are particularly susceptible to contagious cancers, perhaps due in part to filter feeding and weak non-self rejection mechanisms[113]. Other observations of disseminated neoplasia[27,41,145] likely represent additional BTN cases, a hypothesis that could be tested using the method of host/cancer allele deconvolution and phylogenetic reconstruction established in previous work and used here to identify a *M. balthica* BTN. More information about where we do and do not observe BTNs may illuminate why bivalves are particularly susceptible, and more interestingly, in what cases bivalves are resistant. One example was identified in which the BTN derived from one species (*Venerupis corrugata*), spread through the population of another (*Polititapes aureus*), but is not able to infect the species it came from (*V. corrugata*)[7]. This indicates the original host species may have developed resistance to BTN infection due to the strong negative fitness effect of a transmissible cancer. Identification of additional BTN lineages is the first step in better understanding transmissible cancer susceptibility/resistance, which may illuminate conserved cancer resistance mechanisms we can leverage to fight cancer in humans.

3.5 Acknowledgements

This work was supported by NIH training grants T32-HG000035 and T32-GM007270 (to Samuel Hart), career transition award K22-CA226047 and R01-CA255712 (to Michael Metzger), and National Centre of Science (Poland) grant UMO-2017/26/M/NZ8/00478 (to Katarzyna Smolarz).

3.6 Contributions

Samuel Hart identified target loci and designed/validated primers. Rachael Giersch performed qPCR and prepared figures from Giersch et al. (**Figure 3.1A/B**). Samuel Hart, Marisa Yonemitsu, and Rachael Giersch performed longitudinal MarBTN tracking experiment (**Figure 3.2A**). Karyn Tindbaek provided images and performed qPCR for **Figure 3.2B**. Alicja Michnowska, Katarzyna Smolarz and Anna Hallmann collected Baltic clams. Alicja Michnowska and Samuel Hart performed *EF1α* PCR. Alicja Michnowska and Michael Metzger cloned, sequenced, performed *EF1α* phylogenetic analysis and prepared figure from Michnowska et al. (**Figure 3.4**). Samuel Hart wrote this manuscript, with some text altered from corresponding manuscripts. Michael Metzger supervised this work and contributed to review and editing of the manuscript.

3.7 Methods

3.7.1 M. arenaria collection and processing

Adult soft-shell clams (*M. arenaria*) >50 mm in length were collected by commercial sources from multiple locations in Maine, USA, and animals were housed in 1× artificial sea water (36 g/L Instant Ocean, Blacksburg, VA, USA), in aerated aquaria, supplemented 2–3 times weekly with PhytoFeast or LPB Frozen Shellfish Diet (Reed Mariculture, Campbell, CA, USA). Approximately 0.5–1 mL of hemolymph was collected from the pericardial sinus of each animal using a 0.5 in 26-gauge needle fitted on a 3 mL syringe. 4–5 drops from the syringe (~50 µL, just enough to cover the bottom of the well) were placed in a well of a 96-well plate and incubated at 4–10 °C for 1 hour to allow the cells to settle. Wells were screened for clams with high levels of MarBTN based on morphological differences between

healthy hemocytes and MarBTN cells on an inverted phase-contrast microscope. BTN cells are rounded and refractile and do not adhere to the bottom of the well, while healthy hemocytes adhere tightly to the well and extend multiple pseudopodia.

For the infection tracking experiment (provided here as an example of an application of the qPCR assay, **Figure 3.2A)**, we first identified 12 clams that were negative for MarBTN using the HL03 qPCR assay. We injected six with 250 µL artificial sea water as a negative control, and six with 250 µL MarBTN cells. MarBTN cells were collected from a highly infected clam, with 1 mL hemolymph spun at 1000 g for 10 min and resuspended in 1 mL artificial sea water. We determined cell concentration on an aliquot of 1:1 Erythocin B-stained cells counted on a hemocytometer, diluting the MarBTN cell stock to 100,000 cells per 250 µL for injection. Injected clams were then tracked until death by sampling hemolymph every two weeks. No control clams developed MarBTN infections, while 3/6 MarBTN-injected clams developed qPCR-detectable infections.

3.7.2 M. balthica collection and processing

Clams (approximately 100) were collected in February 2019 from a sampling site H45 located in the Gulf of Gdańsk (southern Baltic Sea) at 45 m depth, an area reported to have the highest prevalence of DN in *M. balthica* in previous studies [150]. Clams exceeding 10 mm in size (large enough for hemolymph withdrawal) were selected from the sediment samples collected with a Van Veen grabber. Transport and laboratory set-up were adjusted to imitate *in situ* conditions. Bivalves were kept in 15 L tanks (approximately 50 clams per one tank) filled with seawater collected at sea bottom from the sampling site for time not exceeding five days. No sediment substrate was added to tanks with animals for purification purposes.

Hemolymph was withdrawn directly from the adductor muscle using a Hamilton microsyringe. Syringes were thoroughly cleaned with 10% hydrochloric acid and washed in 70% ethanol and deionized water between individuals to avoid contamination. Hemolymph volume varied between specimens with a range of approximately 20–100 µl. An equivalent volume of absolute ethanol was added to each

hemolymph sample. DN cells spread through the vessels and sinuses of the circulatory system, thus they were found in a higher concentration in the hemolymph than in the solid tissues of bivalves. Therefore, we selected hemolymph and solid tissue (foot) samples for molecular analysis of the cancer and host, respectively. Foot tissue was dissected from each individual and placed into 200 µl of absolute ethanol. Both hemolymph and solid tissue samples (foot) were stored at −20°C until the time of transport to Seattle, WA, USA, where they were further processed for molecular analysis. Ethanol-fixed samples were transported in the time not exceeding two days via plane in room temperature and after arrival were stored at −20°C.

3.7.3 DNA extraction

DNA was extracted using DNeasy Blood and Tissue Kit (Qiagen). For hemocyte samples, pellets was obtained by spinning the cells down with 1000 g for 10 min before proceeding with the extraction according to the manufacturer protocol. Tissues were lysed according to manufacturer instructions, then P3 buffer (Qiagen) was added to precipitate out polysaccharides that inhibit PCR reactions before proceeding with extraction according to the manufacturer protocol.

3.7.4 MarBTN-specific qPCR assays

To quantify the presence of MarBTN DNA in a sample, allele-specific qPCR was performed using four sets of primers (**Supplementary Table 3.1**). Both cancer-specific primer pairs amplify specific integration sites of the LTR-retrotransposon *Steamer*, found only in MarBTN cells. The primary locus was a MarBTN-specific insertion of *Steamer* upstream the *N1N2* gene. A MarBTN-specific primer pair targeting this insertion junction amplifies half the total amount of *N1N1* alleles in a cancer cell (as the insertion is in two of four copies of the gene in a tetraploid region) and a primer pair in a conserved region of the *N1N2* ORF nearby quantifies the total copies of the *N1N2* locus present. The ratio of the two can be used to determine the fraction of clam hemolymph made up of MarBTN cells. A single plasmid (pCR-SteamerLTR-N1N2) was used for the standard curve. It was made by cloning the *Steamer-N1N2* amplicon, amplified from genomic DNA of MarBTN cells (Zero Blunt TOPO PCR cloning kit,

Invitrogen, Waltham, MA). The secondary marker was a different MarBTN-specific *Steamer* integration site, termed HL03. A plasmid was cloned which includes both the HL03 locus and a separate conserved region of the *EF1α* gene as a control (pIMHL03c2-EF1α). Primers used for cloning control plasmids are listed in **Supplementary Table 3.1**, and sequences have been archived in GenBank (accession numbers OM105837-9). The plasmid concentration was measured (Qubit, Thermo Fisher Scientific) and copy number per µL was calculated based on the plasmid sizes. Plasmids were linearized with 0.25 µL of NotI-HF (NEB, Ipswitch, MA, USA) for 30 min at 37 °C in a 20 µL reaction at 1×10^{10} copies/µL, heat-inactivated 20 min at 65 °C, then diluted to 1×10^{9} with 180 µL Buffer AE (Qiagen). Standard curves were prepared from 1×10^{7} copies/rxn to 1×10^{1} copies/rxn. 2 µL of extracted eDNA was run in 10 µL reactions on a StepOnePlus real-time PCR cycler (Applied Biosystems, Waltham, MA, USA). Reactions were run as follows: 95 °C for 2 min, 40 cycles of 95 °C for 15 s and 60 °C for 30 s, followed by a melt curve using 95 °C for 15 s, 60 °C for 1 min, and ramping 0.3 °C from 60 °C to 95 °C, followed by a 15 s hold at 95 °C. All samples were run in triplicate and values presented are an average of triplicates, treating wells with undetectable amplification as zero copies.

3.7.5 MarBTN fraction equation derivation

- Given:
    - R = Ratio MarBTN-specific allele to control N1N2 alleles
    - H = Fraction Host (healthy) cells
    - C = Fraction Cancer (MarBTN) cells
    - $H = 1 - C$
    - $R = 2C / (4C + 2H)$
- $R = 2C / (4C + 2H) = C / (2C + H) = C / (1+C)$
- $1 / R = (C + 1) / C = 1 + 1/C$
- $1 / R - 1 = 1/C$
- $C = 1 / (1/R - 1) = 1 / (1/R – R/R) = 1 / ((1 -R) / R) = R / (1 – R)$

- **$C = R / (1 - R)$**

There are a couple of limitations to this assay to keep in mind. This equation assumes that healthy cells are diploid and MarBTN cancer cells are tetraploid. While diploidy is likely to be true for healthy cells, we have observed significant copy number variation in MarBTN that indicates this assumption could be violated in MarBTN cells. Indeed, samples collected from Prince Edward Island, Canada appear to be pentaploid in the *N1N2* region, with 2/5 alleles having the MarBTN-specific *Steamer* insertion, so the above equation would result in an inaccurate calculation of MarBTN cell fraction unless it was altered to reflect the real ploidy of tested cancer cells. Since it would severely complicate the assay to determine copy number of the *N1N2* locus for every sample, we assume tetraploidy for samples collected from the USA coast, knowing that the real cancer fraction may slightly deviate from the calculated fraction if this assumption is wrong. We also find that the calculated value can be noisy at high cancer fraction approaching 100%, since small deviations in the MarBTN-specific and control qPCRs have an outsized effect on the fraction, sometimes resulting in a value above 100%. Finally, we set a lower limit of detection for calling low-level MarBTN infections of 1 copy per reaction (when using qubit to quantify the standard curve plasmid).

3.7.6 M balthica EF1α locus identification, sequencing, and phylogenetic analysis

M. balthica EF1α transcripts were identified by querying "ef1" in the gene description annotations from the publicly-available transcriptome[148]. Three transcripts were annotated as *EF1α*, but one was short (346 bp) and had no homology to the other two transcripts or the oyster *EF1α*, so was ignored. *M. balthica* transcripts and *C. gigas EF1α* exons were aligned to the *C. gigas* genome using Geneious® alignment with default settings to determine intron placement. Primers that bound both transcripts sometimes yielded >2 bands in healthy clams, indicating that they derive from different loci. Primers were designed to target one *M. balthica EF1α* transcript for sequencing and phylogenetic analysis (Fwd: GTCTGTGGTGACTCAAAGGT, Rev: CTTGACCTCACCAGGATGGT).

PCR amplification for *M. balthica EF1α* was done using Q5 Hot Start High-Fidelity DNA polymerase (NEB) with a 30 s extension time and annealing performed at 50°C. PCR reaction mix consisted of buffer (5×, Qiagen, 5 µl), dNTPs (0.5 µl of 10 mM each), forward and reverse primers (1.25 µl of 10 µM for both), 0.25 µl polymerase and ddH2O (to 25 µl). In all cases, 25–50 ng of genomic DNA was amplified for 35 cycles with initial denaturation performed at 98°C for 30 s. PCR products were gel extracted using QIAquick Gel Extraction Kit (Qiagen) and either directly Sanger sequenced, or, when multiple alleles at a locus could not be resolved by direct sequencing, were cloned using the Zero Blunt TOPO PCR Cloning Kit (Invitrogen). Plasmids were transformed into TOP10 or DH5α competent Escherichia coli (Invitrogen) and at least 6 clones were picked for Sanger sequencing using M13F and M13R primers (Azenta, formerly Genewiz). Due to the hairpin found in some of the *EF1α* alleles, DNA samples were processed with Azenta's proprietary GC-rich sequencing protocol and sequence data was collected on an ABI3730xl DNA Analyser. In cases where a single clone was sequenced that was 1 SNP different from another clone found in multiple clones from the same animal, or a single clone was found to be consistent with recombination between two other clones found in multiple clones in the same animal, the single clones were assumed to be PCR artifacts. In cases where differences were found between the sequence results of hemolymph and tissue from the same individual, the alleles found more often in the hemolymph sample were called as the hemolymph alleles and those found more often in tissue were considered the tissue alleles. The primer binding regions were excluded from sequence analysis and all unique alleles were identified.

Sequences were aligned manually in BioEdit (7.1 version) software. Maximum likelihood phylogenetic trees were generated using PhyML (3.0 version), performing 100 bootstrap replicates with automatic model selection through Akaike Information Criterion. Trees were visualized using FigTree (1.4.4 version) software.

3.8 Supplemental tables

Supplementary Table 3.1: Primers for MarBTN qPCR assays

Assay	Specific to	Target	Plasmid	Primer Name	Sequence (5'->3')
N1N2	MarBTN	Steamer-N1N2	pCR-SteamerLTR-N1N2	ClamLTR-F3	TTCAATCATTCAACGCATAACC
N1N2	MarBTN	Steamer-N1N2	pCR-SteamerLTR-N1N2	N1N2can-R3	TCGCTGAGAATTTTTCGGTGT
N1N2	Control	Total-N1N2	pCR-SteamerLTR-N1N2	N1N2-F3	CCCAGGGCAAGAGGAATATGGT
N1N2	Control	Total-N1N2	pCR-SteamerLTR-N1N2	N1N2-R1	GGATACTGCAAGCTTCTTGGAA
HL03	MarBTN	Steamer-HL03	pIMHL03c2-EF1α	ClamLTRF2	ACATGCACATTAAAAGTTATCG
HL03	MarBTN	Steamer-HL03	pIMHL03c2-EF1α	IMHLO3c2-R2	TCTGGGTCATGAATAACGTCA
HL03	Control	EF1α	pIMHL03c2-EF1α	ClamEF1-F3	GGGAAAAGAGGGCAAGGTGAC
HL03	Control	EF1α	pIMHL03c2-EF1α	ClamEF1-R2	TTTCTTCTTCCCACCGACTGC
Cloning primers					
N1N2			pCR-SteamerLTR-N1N2	ClamLTR-F21	ACATGCACATTAAAAGTTATCG
N1N2			pCR-SteamerLTR-N1N2	98171_conR1	GGATACTGCAAGCTTCTTGGAA
HL03			pIMHL03c2-EF1α	ClamLTR-F2[1]	ACATGCACATTAAAAGTTATCG
HL03			pIMHL03c2-EF1α	ClamLTR-R1[1]	TTAGTATAGCCAATACTGTTAC
HL03			pIMHL03c2-EF1α	ClamA-EF1aFor[2]	tagggcccGAAGGATGAGGGAAAAGAGGG
HL03			pIMHL03c2-EF1α	ClamNS-EF1aRev[2]	atGCGGCCGCatcctgcaggCACCTTTTCCTGCTATGGTGC

1 Cloning of the Steamer fragment to generate pIMHL03c2 was done via inverse PCR, as described in Arriagada et al. 2014.

2 Primers were used to amplify EF1α from genomic DNA, the product was cut with ApaI and NotI (NEB), and ligated into pIMHL03c2.

Chapter 4. What can we learn from BTNs about the basic biology of cancer?

4.1 BTN as a cancer model

Soft-shell clam transmissible cancer is a single clonal lineage[6], meaning that much of what we learn about its biology may be specific to that lineage. Even if we were to expand our analyses to include all other confirmed BTNs we would still only have a sample size of eight[6–10], a number that is dwarfed by the scale of human cancer data sets like The Cancer Genome Atlas, which has over 20,000 individual cancer samples[151]. Additionally, bivalves are invertebrates which lack adaptive immune systems[130], meaning that some aspects of the BTN-host relationship may not translate to vertebrate cancers. Despite these limitations, a few key features make BTNs appealing models to investigate cancer biology. For one, transmissible cancers evolve on an extended time scale and are repeatedly challenged by similar evolutionary pressures (e.g. metastasis, immune evasion), allowing greater opportunity for selection to act than in conventional cancers and thus strengthening our ability to detect signals of selection in individual cancer lineages. Additionally, each transmissible cancer lineage represents a natural experiment with each individual infection a clonal replicate, further strengthening our power to detect selection by looking across many individual replicates that started from the same founder. Multiple BTN lineages across the Bivalvia taxonomic class represent additional experiments in different genetic backgrounds, allowing us to identify convergent evolution of the cancers towards conserved adaptive mechanisms. The lack of an adaptive immune system in bivalves allows for a unique opportunity to focus on the role innate immune systems play in anti-tumor response. Finally, since most cancer research focuses on mammalian cancer models, we may be missing key underlying principles of cancer biology that transcend taxonomic classes within the animal kingdom. Here we discuss a few areas where BTNs have potential to inform our understanding about the basic biology of cancer.

4.2 Limits of somatic evolution

Except for germline cells, which go on to form a new organism via sexual reproduction, somatic cells in our bodies are not designed to outlive us or replicate indefinitely[152,153]. Instead, somatic cells have many intrinsic mechanisms to trigger apoptosis when they detect out-of-control growth, which cancers must evolve to evade[140]. However, transmissible cancers must not only evade these mechanisms, but also must adapt to live indefinitely after the death of their original hosts, in addition to surviving transfer to new hosts and evading their immune systems[113]. In our work we identified a high level of genome instability in MarBTN, which we suspect contributed to the ability of this lineage to adapt as a transmissible cancer. One question that arises is whether there is tradeoff to this genome instability that might doom this lineage in the long run. Though an unstable genome may provide the short-term benefit of adaptability, in the long-term it might result in Muller's ratchet – the accumulation of deleterious mutations that asexually reproducing cancer cells cannot purge via sexual reproduction[154]. Given we see so many BTN lineages, future research should investigate the age and genome stability of additional lineages to investigate whether there is a time limit to somatic evolution, or whether any of these lineages are so old and mutated that they must have escaped Muller's ratchet. Possible escape mechanisms could be evolving the ability to uptake host DNA (as seen in some transmissible cancers for mtDNA[8,43,58]), recombination during co-infection by multiple cancer clones (transmissible cancer co-infection has been documented[16,58]), and evolving a more stable genome. Although transmissible cancer evolution occurs at a much shorter scale than the millions of years over which multicellular organism speciate and evolve, centuries to millennia is still a long time for a cancer, or other single-celled organism, to evolve. Though many researchers investigate the evolutionary origins of multicellularity and sex, transmissible cancers offer a unique opportunity to study the reversion to unicellularity and asexual reproduction[155,156].

4.3 Host cancer resistance

MarBTN does not exist in a stable evolutionary environment, instead evolving within a host population that can evolve itself in response. Though we focus on MarBTN evolution in this study, future

work should investigate *Mya arenaria* evolution in response to the challenge of a contagious cancer epidemic. MarBTN would be expected to exert a heavy selective pressure on host populations to develop resistance to cancer infection. Comparing resistant and susceptible individuals may reveal underlying mechanisms of cancer resistance, particularly if these analyses were conducted across multiple host species. Just like how host BTN lineages may convergently evolve to succeed as transmissible cancers, host populations may convergently evolve to resist BTNs. The comparison of hosts from naïve versus chronically exposed populations for differing responses to cancer infection may also reveal evolved cancer resistance mechanisms. Given the abundance of current-day BTN lineages, it is likely that additional BTN lineages that existed in the past have gone extinct. If so, this may have left detectable signatures of selective sweeps in bivalve species' genomes. One noteworthy example of putative resistance evolution is in the pullet shell clam (*Venerupis corrugata*). A BTN lineage derived from the pullet shell clam infects golden carpet shell clams (*Polititapes aureus*) but does not infect the pullet shell clam it was derived from, indicating the pullet shell clam population may have evolved resistance and the cancer only survived by jumping to a new susceptible species[7]. These potential avenues to investigate cancer resistance make BTNs an appealing model, since research in humans has succeeded in identifying many cancer susceptibility markers (e.g. inherited *BRCA* alleles[157]), but little is known about heritable cancer resistance mechanisms[158].

4.4 Innate immune response to cancer

In human cancer, the innate immune system plays a complicated role in cancer initiation and progression, sometimes aiding malignant growth while sometimes working to prevent progression[159]. Elucidating the intricacies of this response is further complicated by the adaptive immune system, which also plays a key role in promoting a tumor-specific immune response[160]. Although mouse models lacking adaptive immune system components have led to discoveries about the role of the innate and adaptive immune systems in cancer[161,162], the lack of an evolutionary history without adaptive immune systems in these "artificial" models limits their informative ability. Bivalves and other invertebrates have instead relied only on the innate immune system to fight infections and cancer throughout their evolutionary history[130],

making them more “natural” models to isolate innate immune responses to cancer. In our transcriptomics investigation, we saw that immune process downregulation plays an important role in MarBTN, indicating that evading the innate immune system was a key aspect of MarBTN evolution. Further investigation of host hemocyte dynamics over the course of MarBTN infection, specifically comparing infections that progress to death versus maintaining a low level of cancer, may illuminate host innate immune response regimens that succeed in preventing cancer progression. These immune responses can be compared across BTNs to identify conserved mechanisms, which can further be compared with vertebrate cancers to identify underlying mechanisms of innate immune cancer defense that transcend all metazoans.

4.5 Drivers of metastasis

Metastasis, the spread of cancer cells from the primary tumor to secondary sites in the host’s body, is a critical step in cancer progression and is the primary cause of death in cancer patients[163]. Despite intense focus of researchers on this cancer progression stage, there is a lack of consensus about how tumors evolve to become metastatic and underlying genetic drivers of metastasis have remained elusive[164]. One factor contributing to this issue is that conventional cancers do not normally have the chance to metastasize many times before the death of the host[165]. Transmissible cancers on the other hand have succeeded in repeatedly metastasizing to new hosts and thus are under continued selection for metastatic ability, which may strengthen our ability to detect genetic driver loci. The BTN process of seawater transfer and new host engraftment mirrors the metastatic process of blood/lymph transfer and new tissue engraftment and thus may share some molecular mechanisms and/or genetic drivers. Identifying conserved metastatic mechanisms across BTNs and comparing these to what we know about human metastasis is another promising avenue for using BTN’s to better understand the basic biology of cancer.

4.6 Conclusions concerning contagious clam cancer

Soft-shell clam transmissible cancer and other BTNs offer a unique opportunity to study cancer outside the usual bounds of cancer research. Though we outline several notable areas where we believe BTN research could inform our mechanistic understanding of cancer, we also note that like most basic

biology research, the potential biological insight gained from a better understanding of BTNs may extend to unpredictable areas across ecology, evolution, or medicine. Though we present our findings, impacts, and future research directions from a cancer biology lens, it is also worth considering the ecological impacts BTNs may have on bivalve populations, the interconnected marine species in their ecosystems, and the shellfish industry. We have been aware of the existence of contagious cancers in bivalves for under a decade, and the work presented in this dissertation is a small early step in characterizing their biology and identifying broader insights to inform our understanding of the natural world.

References

1. Ní Leathlobhair, M. & Lenski, R. E. Population genetics of clonally transmissible cancers. *Nat Ecol Evol* **6**, 1077–1089 (2022).
2. 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).
3. Rebbeck, C. A., Thomas, R., Breen, M., Leroi, A. M. & Burt, A. Origins and Evolution of a Transmissible Cancer. *Evolution* **63**, 2340–2349 (2009).
4. Pearse, A.-M. & Swift, K. Transmission of devil facial-tumour disease. *Nature* **439**, 549 (2006).
5. Pye, R. J. *et al.* A second transmissible cancer in Tasmanian devils. *PNAS* **113**, 374–379 (2016).
6. 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).
7. Metzger, M. J. *et al.* Widespread transmission of independent cancer lineages within multiple bivalve species. *Nature* **534**, 705–709 (2016).
8. 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).
9. 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).
10. 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).
11. Murchison, E. P. *et al.* Transmissible Dog Cancer Genome Reveals the Origin and History of an Ancient Cell Lineage. *Science* **343**, 437–440 (2014).
12. Baez-Ortega, A. *et al.* Somatic evolution and global expansion of an ancient transmissible cancer lineage. *Science* **365**, eaau9923 (2019).
13. Decker, B. *et al.* Comparison against 186 canid whole-genome sequences reveals survival strategies of an ancient clonally transmissible canine tumor. *Genome Res.* **25**, 1646–1655 (2015).
14. Murchison, E. P. *et al.* Genome Sequencing and Analysis of the Tasmanian Devil and Its Transmissible Cancer. *Cell* **148**, 780–791 (2012).
15. Stammnitz, M. R. *et al.* The Origins and Vulnerabilities of Two Transmissible Cancers in Tasmanian Devils. *Cancer Cell* **33**, 607-619.e15 (2018).
16. Stammnitz, M. R. *et al.* The evolution of two transmissible cancers in Tasmanian devils. *bioRxiv* Preprint at https://doi.org/10.1101/2022.05.27.493404 (2022).
17. 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).
18. Burioli, E. A. V. *et al.* Traits of a mussel transmissible cancer are reminiscent of a parasitic life style. *Sci Rep* **11**, 24110 (2021).
19. Brown, R. S., Wolke, R. E., Saila, S. B. & Brown, C. W. Prevalence of Neoplasia in 10 New England Populations of the Soft-Shell Clam (Mya arenaria). *Annals of the New York Academy of Sciences* **298**, 522–534 (1977).
20. Yevich, P. P. & Barszcz, C. A. Neoplasia in Soft-Shell Clams (Mya arenaria) Collected from Oil-Impacted Sites. *Annals of the New York Academy of Sciences* **298**, 409–426 (1977).

21. Farley, C. A., Plutschak, D. L. & Scott, R. F. Epizootiology and distribution of transmissible sarcoma in Maryland softshell clams, Mya arenaria, 1984-1988. *Environ. Health Perspect.* **90**, 35–41 (1991).
22. Muttray, A. *et al.* Haemocytic leukemia in Prince Edward Island (PEI) soft shell clam (Mya arenaria): Spatial distribution in agriculturally impacted estuaries. *Science of The Total Environment* **424**, 130–142 (2012).
23. Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* **13**, 1050–1054 (2016).
24. 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).
25. 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).
26. Plachetzki, D. C., Pankey, M. S., MacManes, M. D., Lesser, M. P. & Walker, C. W. The Genome of the Softshell Clam Mya arenaria and the Evolution of Apoptosis. *Genome Biology and Evolution* **12**, 1681–1693 (2020).
27. Carballal, M. J., Barber, B. J., Iglesias, D. & Villalba, A. Neoplastic diseases of marine bivalves. *Journal of Invertebrate Pathology* **131**, 83–106 (2015).
28. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Rep* **3**, 246–259 (2013).
29. Nik-Zainal, S. *et al.* Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell* **149**, 979–993 (2012).
30. Gavery, M. R. & Roberts, S. B. A context dependent role for DNA methylation in bivalves. *Brief Funct Genomics* **13**, 217–222 (2014).
31. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
32. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
33. Pilzecker, B. & Jacobs, H. Mutating for Good: DNA Damage Responses During Somatic Hypermutation. *Frontiers in Immunology* **10**, (2019).
34. Poetsch, A. R. The genomics of oxidative DNA damage, repair, and resulting mutagenesis. *Comput Struct Biotechnol J* **18**, 207–219 (2020).
35. 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).
36. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
37. Cagan, A. *et al. Somatic mutation rates scale with lifespan across mammals*. 2021.08.19.456982 https://www.biorxiv.org/content/10.1101/2021.08.19.456982v2 (2021) doi:10.1101/2021.08.19.456982.
38. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).

39. Weghorn, D. & Sunyaev, S. Bayesian inference of negative and positive selection in human cancers. *Nat Genet* **49**, 1785–1788 (2017).
40. Wang, L. *et al.* Pan-Cancer Analyses Identify the CTC1-STN1-TEN1 Complex as a Protective Factor and Predictive Biomarker for Immune Checkpoint Blockade in Cancer. *Frontiers in Genetics* **13**, (2022).
41. Barber, B. J. Neoplastic diseases of commercially important marine bivalves. *Aquat. Living Resour.* **17**, 449–466 (2004).
42. Menghi, F. *et al.* The Tandem Duplicator Phenotype Is a Prevalent Genome-Wide Cancer Configuration Driven by Distinct Gene Mutations. *Cancer Cell* **34**, 197-210.e5 (2018).
43. Rebbeck, C. A., Leroi, A. M. & Burt, A. Mitochondrial Capture by a Transmissible Cancer. *Science* **331**, 303–303 (2011).
44. Strakova, A. *et al.* Recurrent horizontal transfer identifies mitochondrial positive selection in a transmissible cancer. *Nature Communications* **11**, 3059 (2020).
45. Wilson, J. J., Hefner, M., Walker, C. W. & Page, S. T. Complete Mitochondrial Genome of the Soft-shell clam Mya arenaria. *Mitochondrial DNA A DNA MappSeq Anal* **27**, 3553–3554 (2016).
46. 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).
47. Goodier, J. L. Restricting retrotransposons: a review. *Mob DNA* **7**, (2016).
48. Cooper, K. R., Brown, R. S. & Chang, P. W. The course and mortality of a hematopoietic neoplasm in the soft-shell clam, Mya arenaria. *Journal of Invertebrate Pathology* **39**, 149–157 (1982).
49. Smolowitz, R. M., Miosky, D. & Reinisch, C. L. Ontogeny of leukemic cells of the soft shell clam. *Journal of Invertebrate Pathology* **53**, 41–51 (1989).
50. Takata, K., Shimizu, T., Iwai, S. & Wood, R. D. Human DNA Polymerase N (POLN) Is a Low Fidelity Enzyme Capable of Error-free Bypass of 5S-Thymine Glycol*. *Journal of Biological Chemistry* **281**, 23445–23455 (2006).
51. Moldovan, G.-L. *et al.* DNA Polymerase POLN Participates in Cross-Link Repair and Homologous Recombination. *Mol Cell Biol* **30**, 1088–1096 (2010).
52. Arana, M. E., Takata, K., Garcia-Diaz, M., Wood, R. D. & Kunkel, T. A. A unique error signature for human DNA polymerase v. *DNA Repair (Amst)* **6**, 213–223 (2007).
53. Lee, Y.-S., Gao, Y. & Yang, W. How a homolog of high-fidelity replicases conducts mutagenic DNA synthesis. *Nat Struct Mol Biol* **22**, 298–303 (2015).
54. Walker, C., Böttger, S. & Low, B. Mortalin-Based Cytoplasmic Sequestration of p53 in a Nonmammalian Cancer Model. *Am J Pathol* **168**, 1526–1530 (2006).
55. Kwon, Y. M. *et al.* Evolution and lineage dynamics of a transmissible cancer in Tasmanian devils. *PLOS Biology* **18**, e3000926 (2020).
56. Robinson, P. S. *et al.* Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. *Nat Genet* **53**, 1434–1442 (2021).
57. Epstein, B. *et al.* Rapid evolutionary response to a transmissible cancer in Tasmanian devils. *Nat Commun* **7**, (2016).
58. Bruzos et al., A. L. The evolution of two transmissible leukaemias colonizing the coasts of Europe. *bioRxiv* (2022).

59. Andor, N., Maley, C. C. & Ji, H. P. Genomic Instability in Cancer: Teetering on the Limit of Tolerance. *Cancer Research* **77**, 2179–2185 (2017).
60. Hung, W.-Y. *et al.* Tandem duplication/triplication correlated with poly-cytosine stretch variation in human mitochondrial DNA D-loop region. *Mutagenesis* **23**, 137–142 (2008).
61. Yuan, Y. *et al.* Comprehensive molecular characterization of mitochondrial genomes in human cancers. *Nat Genet* **52**, 342–352 (2020).
62. Doyle, J. J. & Doyle, J. L. Isolation of plant DNA from fresh tissue. *Focus* **12**, 13–15 (1990).
63. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460 (2018).
64. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
65. Kronenberg, Z. N. *et al.* FALCON-Phase: Integrating PacBio and Hi-C data for phased diploid genomes. 327064 Preprint at https://doi.org/10.1101/327064 (2018).
66. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
67. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).
68. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
69. Bickhart, D. M. *et al.* Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat Genet* **49**, 643–650 (2017).
70. Burton, J. N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol* **31**, 1119–1125 (2013).
71. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *cels* **3**, 95–98 (2016).
72. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680 (2014).
73. English, A. C. *et al.* Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLOS ONE* **7**, e47768 (2012).
74. Marks, P. *et al.* Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res* **29**, 635–645 (2019).
75. Grabherr, M. G. *et al.* Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat Biotechnol* **29**, 644–652 (2011).
76. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451–9457 (2020).
77. Silliman, K., Spencer, L. H., White, S. J. & Roberts, S. B. Epigenetic and genetic population structure is coupled in a marine invertebrate. 2022.03.23.485415 Preprint at https://doi.org/10.1101/2022.03.23.485415 (2022).
78. Bushnell, B. BBMap. *SourceForge* https://sourceforge.net/projects/bbmap/.
79. AboElkhair, M. *et al.* Reverse transcriptase activity in tissues of the soft shell clam Mya arenaria affected with haemic neoplasia. *Journal of Invertebrate Pathology* **102**, 133–140 (2009).

80. AboElkhair, M. *et al.* Lack of detection of a putative retrovirus associated with haemic neoplasia in the soft shell clam Mya arenaria. *Journal of Invertebrate Pathology* **109**, 97–104 (2012).
81. Brousseau, D. J. & Baglivo, J. A. Field and laboratory comparisons of mortality in normal and neoplastic Mya arenaria. *Journal of Invertebrate Pathology* **57**, 59–65 (1991).
82. Delaporte, M. *et al.* Assessment of haemic neoplasia in different soft shell clam Mya arenaria populations from eastern Canada by flow cytometry. *Journal of Invertebrate Pathology* **98**, 190–197 (2008).
83. Leavitt, D. F. *et al.* Hematopoietic neoplasia inMya arenaria: Prevalence and indices of physiological condition. *Mar. Biol.* **105**, 313–321 (1990).
84. Le Grand, F. *et al.* Disseminated Neoplasia in the Soft-Shell Clam Mya arenaria: Membrane Lipid Composition and Functional Parameters of Circulating Cells. *Lipids* **49**, 807–818 (2014).
85. Lesser, M. P., Thompson, M. M. & Walker, C. W. Effects of Thermal Stress and Ocean Acidification on the Expression of the Retrotransposon Steamer in the Softshell Mya arenaria. *shre* **38**, 535–541 (2019).
86. Mateo, D. R., MacCallum, G. S. & Davidson, J. Field and laboratory transmission studies of haemic neoplasia in the soft-shell clam, Mya arenaria, from Atlantic Canada. *Journal of Fish Diseases* **39**, 913–927 (2016).
87. McLaughlin, S. M., Farley, C. A. & Hetrick, F. M. Transmission studies of sarcoma in the soft-shell clam, Mya arenaria. *In Vivo* **6**, 367–370 (1992).
88. Oprandy, J. J. & Chang, P. W. 5-Bromodeoxyuridine induction of hematopoietic neoplasia and retrovirus activation in the soft-shell clam, Mya arenaria. *Journal of Invertebrate Pathology* **42**, 196–206 (1983).
89. Reinisch, C. L., Charles, A. M. & Troutner, J. Unique antigens on neoplastic cells of the soft shell clam Myaarenaria. *Developmental & Comparative Immunology* **7**, 33–39 (1983).
90. Siah, A., McKenna, P., Berthe, F. C. J., Afonso, L. O. B. & Danger, J.-M. Transcriptome analysis of neoplastic hemocytes in soft-shell clams Mya arenaria: Focus on cell cycle molecular mechanism. *Results in Immunology* **3**, 95–103 (2013).
91. Siah, A., McKenna, P., Danger, J.-M., Johnson, G. R. & Berthe, F. C. J. Induction of transposase and polyprotein RNA levels in disseminated neoplastic hemocytes of soft-shell clams: Mya arenaria. *Developmental & Comparative Immunology* **35**, 151–154 (2011).
92. Siah, A., Delaporte, M., Pariseau, J., McKenna, P. & Berthe, F. C. J. Patterns of p53, p73 and mortalin gene expression associated with haemocyte polyploidy in the soft-shell clam, Mya arenaria. *Journal of Invertebrate Pathology* **98**, 148–152 (2008).
93. Siah, A., McKenna, P., Danger, J. M., Johnson, G. & Berthe, F. C. J. Expression of RAS-like family members, c-jun and c-myc mRNA levels in neoplastic hemocytes of soft-shell clams Mya arenaria using microsphere-based 8-plex branched DNA assay. *Results in Immunology* **2**, 83–87 (2012).
94. Sunila, I. & Farley, C. Environmental limits for survival of sarcoma cells from the soft-shell clam Mya arenaria. *Dis. Aquat. Org.* **7**, 111–115 (1989).
95. Taraska, N. G. & Anne Böttger, S. Selective initiation and transmission of disseminated neoplasia in the soft shell clam Mya arenaria dependent on natural disease prevalence and animal size. *Journal of Invertebrate Pathology* **112**, 94–101 (2013).

96. Walker, C. *et al.* Mass Culture and Characterization of Tumor Cells From a Naturally Occurring Invertebrate Cancer Model: Applications for Human and Animal Disease and Environmental Health. *The Biological Bulletin* **216**, 23–39 (2009).
97. Weinberg, J. R., Leavitt, D. F., Lancaster, B. A. & Capuzzo, J. M. Experimental Field Studies withMya arenaria(Bivalvia) on the Induction and Effect of Hematopoietic Neoplasia. *Journal of Invertebrate Pathology* **69**, 183–194 (1997).
98. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
99. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]* (2013).
100. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
101. Carlson, J., Li, J. Z. & Zöllner, S. Helmsman: fast and efficient mutation signature analysis for massive sequencing datasets. *BMC Genomics* **19**, (2018).
102. Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational signatures. *bioRxiv* (2020) doi:10.1101/372896.
103. Klambauer, G. *et al.* cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res* **40**, e69 (2012).
104. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
105. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
106. Ding, Z. *et al.* Estimating telomere length from whole genome sequence data. *Nucleic Acids Research* **42**, e75 (2014).
107. Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645–3647 (2017).
108. Chu, C., Nielsen, R. & Wu, Y. REPdenovo: Inferring De Novo Repeat Motifs from Short Sequence Reads. *PLOS ONE* **11**, e0150719 (2016).
109. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
110. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
111. 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).
112. Pye, R. J. *et al.* A second transmissible cancer in Tasmanian devils. *PNAS* **113**, 374–379 (2016).
113. 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).
114. 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).
115. Siddle, H. V. & Kaufman, J. Immunology of naturally transmissible tumours. *Immunology* **144**, 11–20 (2015).
116. 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).

117. 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).
118. Hart, S. F. M. *et al.* Centuries of genome instability and evolution in soft-shell clam transmissible cancer. 2022.08.07.503107 Preprint at https://doi.org/10.1101/2022.08.07.503107 (2022).
119. Chang, H.-Y. *et al.* hPuf-A/KIAA0020 Modulates PARP-1 Cleavage upon Genotoxic Stress. *Cancer Research* **71**, 1126–1134 (2011).
120. Cho, H.-C. *et al.* Puf-A promotes cancer progression by interacting with nucleophosmin in nucleolus. *Oncogene* **41**, 1155–1165 (2022).
121. Jiang, H. & Vogt, P. K. Constitutively active Rheb induces oncogenic transformation. *Oncogene* **27**, 5729–5740 (2008).
122. Lerman, M. I. & Minna, J. D. The 630-kb lung cancer homozygous deletion region on human chromosome 3p21.3: identification and evaluation of the resident candidate tumor suppressor genes. The International Lung Cancer Chromosome 3p21.3 Tumor Suppressor Gene Consortium. *Cancer Res* **60**, 6116–6133 (2000).
123. Okpechi, S. C. *et al.* Role of Nischarin in the pathology of diseases: a special emphasis on breast cancer. *Oncogene* **41**, 1079–1086 (2022).
124. Jang, L. K. *et al.* A novel leucine-rich repeat protein (LRR-1): potential involvement in 4-1BB-mediated signal transduction. *Mol Cells* **12**, 304–312 (2001).
125. Falvella, F. S. *et al.* Identification of RASSF8 as a candidate lung tumor suppressor gene. *Oncogene* **25**, 3934–3938 (2006).
126. Wang, J. *et al.* Fibrinogen-like Protein 1 Is a Major Immune Inhibitory Ligand of LAG-3. *Cell* **176**, 334-347.e12 (2019).
127. Qian, W., Zhao, M., Wang, R. & Li, H. Fibrinogen-like protein 1 (FGL1): the next immune checkpoint target. *J Hematol Oncol* **14**, 147 (2021).
128. Yu, J. *et al.* The role of Fibrinogen-like proteins in Cancer. *Int J Biol Sci* **17**, 1079–1087 (2021).
129. Yi, W. *et al.* The regulation role and diagnostic value of fibrinogen-like protein 1 revealed by pan-cancer analysis. *Mater Today Bio* **17**, 100470 (2022).
130. Allam, B. & Raftos, D. Immune responses to infectious diseases in bivalves. *Journal of Invertebrate Pathology* **131**, 121–136 (2015).
131. de la Ballina, N. R., Maresca, F., Cao, A. & Villalba, A. Bivalve Haemocyte Subpopulations: A Review. *Frontiers in Immunology* **13**, (2022).
132. Burioli, E. a. V. *et al.* Transcriptomics of mussel transmissible cancer MtrBTN2 reveals accumulation of multiple cancerous traits and oncogenic pathways shared among bilaterians. 2023.01.03.522559 Preprint at https://doi.org/10.1101/2023.01.03.522559 (2023).
133. 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).
134. 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).

135. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
136. 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).
137. 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).
138. Zitvogel, L., Tesniere, A. & Kroemer, G. Cancer despite immunosurveillance: immunoselection and immunosubversion. *Nat Rev Immunol* **6**, 715–727 (2006).
139. Ujvari, B., Gatenby, R. A. & Thomas, F. The evolutionary ecology of transmissible cancers. *Infection, Genetics and Evolution* **39**, 293–303 (2016).
140. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).
141. Dolgalev, I. msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format. (2022).
142. 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).
143. 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).
144. Fanley, C. A. Probable neoplastic disease of the hematopoietic system in oysters, Crassostrea virginica and Crassostrea gigas. *Natl Cancer Inst Monogr* **31**, 541–555 (1969).
145. Pauley, G. B. A critical review of neoplasia and tumor-like lesions in mollusks. *Natl Cancer Inst Monogr* **31**, 509–539 (1969).
146. Farley, C. A., Plutschak, D. L. & Scott, R. F. Epizootiology and distribution of transmissible sarcoma in Maryland softshell clams, Mya arenaria, 1984-1988. *Environ Health Perspect* **90**, 35–41 (1991).
147. Hart, S. F. *et al.* Centuries of genome instability and evolution in soft-shell clam transmissible cancer. 2022.08.07.503107 Preprint at https://doi.org/10.1101/2022.08.07.503107 (2022).
148. Yurchenko, A. A., Katolikova, N., Polev, D., Shcherbakova, I. & Strelkov, P. Transcriptome of the bivalve Limecola balthica L. from Western Pacific: A new resource for studies of European populations. *Marine Genomics* **40**, 58–63 (2018).
149. Zhang, G. *et al.* The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* **490**, 49–54 (2012).
150. Smolarz, K., Thiriot-Quievreux, C. & Wolowicz, M. Recent trends in the prevalence of neoplasia in the Baltic clam Macoma balthica (L.) from the Gulf of Gdańsk (Baltic Sea). *Oceanologia* **47**, (2005).
151. The Cancer Genome Atlas Pan-Cancer analysis project | Nature Genetics. https://www.nature.com/articles/ng.2764.
152. Hayflick, L. The limited in vitro lifetime of human diploid cell strains. *Experimental Cell Research* **37**, 614–636 (1965).
153. Little, M. P. Cancer models, genomic instability and somatic cellular Darwinian evolution. *Biol Direct* **5**, 19 (2010).
154. Muller, H. J. The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **1**, 2–9 (1964).

155. Aktipis, C. A. *et al.* Cancer across the tree of life: cooperation and cheating in multicellularity. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**, 20140219 (2015).
156. Panchin, A. Y., Aleoshin, V. V. & Panchin, Y. V. From tumors to species: a SCANDAL hypothesis. *Biology Direct* **14**, 3 (2019).
157. King, M.-C., Marks, J. H. & Mandell, J. B. Breast and Ovarian Cancer Risks Due to Inherited Mutations in BRCA1 and BRCA2. *Science* **302**, 643–646 (2003).
158. Klein, G. Toward a genetics of cancer resistance. *Proceedings of the National Academy of Sciences* **106**, 859–863 (2009).
159. Hagerling, C., Casbon, A.-J. & Werb, Z. Balancing the innate immune system in tumor development. *Trends in Cell Biology* **25**, 214–220 (2015).
160. Ribas, A. Adaptive immune resistance: How cancer protects from immune attack. *Cancer Discov* **5**, 915–919 (2015).
161. Ito, M., Kobayashi, K. & Nakahata, T. NOD/Shi-scid IL2rγnull (NOG) Mice More Appropriate for Humanized Mouse Models. in *Humanized Mice* (eds. Nomura, T., Watanabe, T. & Habu, S.) 53–76 (Springer, 2008). doi:10.1007/978-3-540-75647-7_3.
162. Ito, R., Takahashi, T., Katano, I. & Ito, M. Current advances in humanized mouse models. *Cell Mol Immunol* **9**, 208–214 (2012).
163. Steeg, P. S. Tumor metastasis: mechanistic insights and clinical challenges. *Nat Med* **12**, 895–904 (2006).
164. Gui, P. & Bivona, T. G. Evolution of metastasis: new tools and insights. *Trends in Cancer* **8**, 98–109 (2022).
165. Hölzel, D., Eckel, R., Emeny, R. T. & Engel, J. Distant metastases do not metastasize. *Cancer Metastasis Rev* **29**, 737–750 (2010).