Host-Pathogen Coevolution Between Tasmanian Devils (Sarcophilus harrisii) and Devil Facial Tumor Disease

Dylan Garret Gallinson

University of South Florida

Follow this and additional works at: https://digitalcommons.usf.edu/etd

Part of the Bioinformatics Commons, Biology Commons, and the Genetics Commons

Scholar Commons Citation
Gallinson, Dylan Garret, "Host-Pathogen Coevolution Between Tasmanian Devils (Sarcophilus harrisii) and Devil Facial Tumor Disease" (2022). USF Tampa Graduate Theses and Dissertations. https://digitalcommons.usf.edu/etd/9356

This Thesis is brought to you for free and open access by the USF Graduate Theses and Dissertations at Digital Commons @ University of South Florida. It has been accepted for inclusion in USF Tampa Graduate Theses and Dissertations by an authorized administrator of Digital Commons @ University of South Florida. For more information, please contact scholarcommons@usf.edu.
Host-Pathogen Coevolution Between Tasmanian Devils (*Sarcophilus harrisii*) and Devil Facial Tumor Disease

by

Dylan G. Gallinson

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Public Health with a concentration in Genomics
Department of Science
College of Public Health
University of South Florida

Co-Major Professor: Ryan McMinds, Ph.D.
Co-Major Professor: Mark Margres, Ph.D.
Lynn Martin, Ph.D.

Date of Approval:
April 01, 2022

Keywords: GWAS, cancer, genetic population structure, genotype-phenotype, joint modeling

Copyright © 2022, Dylan G. Gallinson
DEDICATION

I dedicate my thesis to my parents, Richard and Candace Gallinson, and my brother, Connor Gallinson. Your tireless support and encouragement have been invaluable to me throughout my entire education, and I am especially appreciative of the extra support and understanding offered during my graduate education. Thanks so much for always taking an interest in my studies, listening to me rant, and inquiring about my research (despite knowing this leads invariably to a long-winded, technical rant). I would also like to thank my friends for their constant support throughout my graduate program. I spent many an outing talking about my research and little else, and I quite appreciate everyone’s patience and genuine interest during these social-gatherings-turned-lectures.
ACKNOWLEDGEMENTS

I thank my committee for all of their help throughout my thesis. The direction and feedback everyone provided at all stages of this research were instrumental in its completion, and I could not have made the progress that I did without such tremendous help. A special thanks to Drs. Ryan McMinds and Mark Margres, our weekly (and sometimes almost daily) meetings made this work possible, and I have learned more about conducting genuine scientific research in this past year than ever before. I consider myself incredibly fortunate for the mentorship both of you have provided me and I am thrilled to continue this opportunity during my PhD program. I would also like to thank the members of the Margres Lab for their helpful feedback in preparing many of my presentations and ensuring such presentations did not devolve into an hour-long digression of technical minutia.

Thanks to the many collaborators who made this work possible: the Storfer Lab (Washington State University) for preparing and sequencing tissues, the Jones Lab (University of Tasmania) and Hamede Lab (University of Tasmania) for trapping devils and collecting metadata, and the McCallum Lab (Griffith University) for aiding in the statistical analyses. I would also like to thank Dr. Douglas Kerlin for digging up missing metadata and helping me in understanding these data and the growth model which was integral to this study.
# Table of Contents

List of Tables .................................................................................................................................. ii
List of Figures ................................................................................................................................ iii
Abstract .......................................................................................................................................... iv
Introduction ......................................................................................................................................1

Methods............................................................................................................................................9
  Sample collection .........................................................................................................................9
  Phenotype data ............................................................................................................................10
  Alignments and variant calling ....................................................................................................13
  DAPC lineage analysis ................................................................................................................15
  BSLMM genome architecture modeling ....................................................................................16
  ATOMM host-pathogen joint modeling .......................................................................................17

Results............................................................................................................................................20
  Genetic population structure of devils and DFTD .................................................................20
  Genome architecture underlying disease-relevant traits .......................................................23
  Disease-relevant candidate genes ............................................................................................24
  Force of infection joint modeling and genotype-by-genotype interactions .........................26

Discussion......................................................................................................................................28
  Genetic population structure differs between devils and DFTD ........................................28
  Disease-relevant traits indicate first-step evolution ..............................................................33
  Devil-DFTD genomic interaction contributes to force of infection .....................................34

Conclusion .....................................................................................................................................37

References.......................................................................................................................................40

Appendix 1: Supplementary Material ............................................................................................50
List of Tables

Table 1: Number of SNPs and samples used in associative tests .............................................15
Table 2: BSLMM genome architecture statistics .......................................................................24
Table 3: Top 5 BSLMM SNPs for host force of infection and virulence .................................25
Table 4: Top 5 ATOMM force of infection genotype-by-genotype interactions ......................27
Table A1: DFTD arrival dates by site ...........................................................................................50
Table A2: Devil lineage Fst estimates ..........................................................................................50
Table A3: DFTD lineage fixation Fst estimates ...........................................................................50
Table A4: GEMMA convergence statistics ..................................................................................51
List of Figures

Figure 1: Devil trapping locations within Tasmania .................................................................10
Figure 2: Devil lineages identified using DAPC ........................................................................21
Figure 3: DFTD lineages identified using DAPC .......................................................................23
Figure 4: Genome architecture identified through BSLMM model fitting .................................24
Figure 5: Force of infection host-pathogen heritability estimates ..............................................26
Figure A1: Tumor area versus tumor volume regression ............................................................51
Figure A2: Host k-means clustering Bayesian information criterion (BIC) .................................52
Figure A3: Host hierarchical k-means clustering Bayesian information criterion (BIC) ..........52
Figure A4: Tumor k-means clustering Bayesian information criterion (BIC) ............................53
Figure A5: BSLMM fitting on ranknorm host force of infection ...............................................53
ABSTRACT

Coevolution is a driving force of rapid evolution, yet the complexity of coevolutionary interactions has made it difficult to characterize the genomic basis of traits mediating such relationships. Coevolutionary dynamics are especially important in host-pathogen systems where the host and pathogen must constantly adapt to one another. The Tasmanian devil and its species-specific transmissible cancer, devil facial tumor disease (DFTD), provide the rare opportunity to study host-pathogen coevolution in a complex natural system. Extensive spatiotemporal devil sampling, high linkage disequilibrium in devils, and a large selective pressure imposed by DFTD facilitate a system tractable for study. Here, we characterized devil and DFTD coevolution by looking at genetic population structures, genome architecture underlying force of infection and virulence, and the contribution of devil-DFTD genome interaction to explaining force of infection. A probe-capture sequence approach was used to sequence 456 devils and 504 tumors at ~197k loci. The genetic structures of devils and DFTD were then identified via clustering and compared. Associative modeling was used to determine genome architecture, and a joint host-pathogen model was used to assess the contribution of genome interactions. Devil and DFTD genetic clustering revealed a decoupled genetic structure, suggesting little evidence of coevolution. Variance in force of infection was attributable primarily to devil genomes (61.1%) and had both large-effect variants (~3 SNPs explained 22.8% total variance) and a polygenic component. Tumor genomes explained a large proportion of virulence (69.8%) and a few large-effect loci contributed to most of this explanatory power (~6 SNPs explained 51.2% total variance). Significant devil-DFTD genomic interactions for force of infection were detectable.
through joint modeling (40.3%), and genotype-by-genotype interaction tests revealed devil and
tumor genes implicated in cancer. Despite the decoupled genetic structure between devils and
DFTD, the significant genome interaction indicates potential coevolution. The identified devil-
DFTD genome interaction represents the first finding providing evidence of coevolution between
devils and DFTD, and the framework used here may be applied to various host-pathogen
systems.
INTRODUCTION

Since the formulation of evolution through natural selection by Darwin and Wallace (1858), advances in science and technology have done much to further characterize the mechanisms of evolution. One major advancement is the discovery that a large proportion of evolution is driven by closely interacting species incurring reciprocal selective pressures on one another, a process called coevolution (Thompson, 1989). In antagonistic relationships, coevolution is often described as an arms race between the participating species (Dawkins & Krebs, 1979; Endara et al., 2017; Langmore et al., 2003; Thompson, 2005), and can lead to highly specialized adaptations such as brood parasitism seen in many avian species (Rothstein, 1990). Coevolution can also act upon mutualistic species, as in the relationship between pollinating insects and flowering plants for which coevolution was likely a driving factor (Hu et al., 2008). The intimate interaction necessary for coevolution to occur leads to spatial variation, as the level of interaction between species may differ by environment, and this can create a mosaic of coevolutionary hot spots and cold spots (Gomulkiewicz et al., 2000). Furthermore, coevolution necessitates reciprocal selective pressures between the participating species, something which is often difficult to observe directly due to the long period of time required for these dynamics to shift (Gaba & Ebert, 2009).

The advent of DNA sequencing has proven an invaluable tool to the study of evolution and coevolution. For adaptation through natural selection to occur, it is necessary that variable phenotypes are heritable. Although selection will act upon traits, only those traits which can be transmitted to offspring can increase in frequency and confer adaptive change. Thus, the ability
to query genes, the molecules which facilitate heritable phenotypic variation, has obvious advantages. In particular, genome-wide scans made feasible by high-throughput sequencing now facilitate fine-grained evolutionary studies on a scale previously unimaginable. Yet, despite these advances, the confounding nature of complex regulatory networks and polygenic traits has left fundamental questions such as the genotype-phenotype relationship largely unanswered. Furthermore, these difficulties are compounded in coevolutionary systems, which necessitate the disentangling of two or more genomes and their interactions. Instances of genetically straightforward coevolutionary systems, with strong selective pressures on phenotypes controlled by only a handful of loci, do exist (e.g., rough-skinned newts and their predator, common garter snakes; Feldman et al., 2009) and have provided insights into the genetic underpinnings of coevolution. However, most coevolutionary interactions involve many different traits, each of which likely possess complex genomic architectures. As such, these simpler systems likely represent atypical scenarios and make generalizations to other systems challenging. Thus, it is necessary to select a coevolving system which is both complex enough to be generalizable but not so complex that it is beyond feasibility to study.

Clonally transmissible cancers are exceedingly rare, and only three naturally contagious cancers are currently known: canine transmissible venereal tumor (CTVT) in dogs (Cohen, 1985), bivalve transmissible neoplasia (BTN) which can transmit between multiple bivalve species (Skazina et al., 2021), and devil facial tumor disease (DFTD) in the Tasmanian devil (Pearse & Swift, 2006a). Transmissible cancers avoid immune recognition and are thus able to colonize a new host when cancerous cells are physically transmitted from an infected individual to an uninfected individual (Ujvari et al., 2016). For example, DFTD transmits between devils when an uninfected devil bites the tumor mass of an infected devil and the cancerous cells are
transferred as a clonal allograft to the uninfected individual (Pearse & Swift, 2006b). Although contagious cancers are rare occurrences in nature, the unique host-pathogen relationship between devils and DFTD may provide insight into coevolutionary dynamics as well as the evolution of cancers.

The Tasmanian devil (Sarcophilus harrisii) is the largest living marsupial carnivore and is Tasmania’s top predator, serving a vital role in the Tasmanian ecosystem (Hamede et al., 2015). The extent to which devils and DFTD have been sampled, both geographically and longitudinally, make this system ideal for the study of host-pathogen coevolution. Long-term mark-recapture data collection efforts have been conducted nearly since the discovery of DFTD in 1996 and, after two decades of such efforts, data now exist for sites throughout all of Tasmania (Lazenby et al., 2018). Since its discovery, DFTD has swept from east to west across the island, affecting nearly all devil populations within Tasmania (Woods et al., 2018). The east-west progression of DFTD facilitates a natural experiment such that some eastern sites represent long-diseased devil populations with a reduced number of individuals, and some western sites represent relatively disease-free populations with minimal DFTD-related decline (Woods et al., 2018). Hence, the progression of devil and DFTD evolution can be compared between long-diseased devil populations and those where DFTD emerged relatively recently. Differences in the impact of DFTD on various devil populations also has implications regarding the ecological effects of rapid declines in a top predator through trophic cascades (Hollings et al., 2014).

DFTD was derived from a Schwann cell in a single female devil and is able to avoid the new host’s allogeneic response through the downregulation of MHC (Hamede et al., 2015), the normal expression of which is typically used in allore cognition to initiate the immune response (Afzali et al., 2008). Devils also experienced historic population bottlenecks which severely
reduced their genetic diversity and consequently their immune diversity (Brüniche-Olsen et al., 2014). Ultimately, the downregulation of MHC proteins in DFTD facilitates the evasion of the devil immune system by the tumor, and the low genetic diversity in devils contributed to nearly universal susceptibility (Cheng et al., 2019). The high mortality rate of DFTD, coupled with universal susceptibility in devils, has resulted in the decimation of devil populations throughout Tasmania, with local population declines in excess of 90% and a total population decline of 80% (McCallum et al., 2007). Furthermore, a second diploid strain, DFT2, was discovered in 2014 (Pye et al., 2016). DFT2 presents as symptomatically equivalent to DFTD, but differs histologically and genetically (Pye et al., 2016), and the presence of a Y chromosome indicates that this cancer originated independently within a male devil (Stammnitz et al., 2018). The independent origin of DFT2 in devils appears to indicate a predisposition for transmissible cancers, an otherwise rare phenomenon in natural systems.

Early epidemiological models, which indicated that DFTD transmission was frequency-dependent, predicted devil extinction (McCallum et al., 2009). Despite this, populations long-infected with DFTD, and thus predicted to be extinct, have persisted, and instances of tumor regression have been observed in rare instances (Margres et al., 2018b). The persistence of long-diseased devil populations indicates the possibility of an adaptive response by devils to DFTD, and Epstein et al. (2016) found evidence of DFTD-imposed selection operating on localized regions of chromosome 2 and 3 in the devil genome. Furthermore, these regions showed little evidence of selection pre-DFTD and, of the seven genes found within the two regions, five genes had functions related to cancer risk or the immune system in other mammals (Epstein et al., 2016). Margres et al. (2018a) also found evidence of an evolutionary response to DFTD by devils in the form of a small number of loci explaining the majority of phenotypic variance for
female devil survival after DFTD infection. The high mortality and near ubiquity of DFTD across the devil’s geographic range have made it the driving selective force to devil adaptation, and many genetic signals suggesting a response to abiotic factors (e.g., vegetation cover) seen before DFTD arrival are no longer detectable since its emergence (Fraik et al., 2020).

Although devils are clearly adapting to the massive selective pressure imposed by DFTD, the tumor has also shown signs of evolution. Kwon et al. (2020) identified multiple DFTD clades, with instances of possible lineage competition and replacement in some sites. Further examples of lineage competition were seen in tetraploid DFTD strains, which occurred early within disease progression and resulted in lowered force of infection and tumor virulence (Pearse et al., 2012). A tetraploid strain initially infected devils at West Pencil Pine, but this strain was quickly replaced when a more virulent and infectious diploid strain arrived at the site (Hamede et al., 2015). Tumor regression, although rare, may also partially be the result of differences in tumor genomes rather than in devil genomic differences. RASL11A, a small GTPase which is downregulated in human prostate cancer (Louro et al., 2004) and colon cancer (Weber et al., 2005), is found to be silenced in most tumors but was active in the few instances of tumor regression which were observed, implicating this gene in tumor regression (Margres et al., 2020). Although evolution appears to be occurring both within devils and DFTD, few studies have investigated the extent of host-pathogen coevolution between devils and DFTD.

The effects a pathogen has on its host as it progresses through the host population depend both on short-term dynamics such as transmission rates and the long-term evolution of the host and pathogen (Blanquart, 2019). Despite the initial predictions of devil extinction by early epidemiological models (McCallum et al., 2009), a more recent model predicts devil extinction with only a 21% probability within the next 100 years, whereas DFTD disappearance was 57%
probable and devil-DFTD coexistence was 22% probable (Wells et al., 2019). The newer epidemiological model is an individual-based model with parameters derived from a devil population where DFTD had been present for 10 years, with inputs estimated using approximate Bayesian computation (Wells et al., 2019). Modeling has also revealed that devils with higher fitness tend to have a higher probability of DFTD infection (Wells et al., 2017), which is likely a corollary of socially dominant devils biting more frequently and thus being more likely to bite into a tumor mass (Hamede et al., 2013). Although these models have revealed much regarding DFTD progression, the inclusion of evolutionary information is minimal (Wells et al., 2017, 2019), and long-term predictions of devil and DFTD outcomes are likely influenced by the evolutionary dynamics within this host-pathogen system. Hence, further model refinement necessitates characterizing the genetics underlying devil and DFTD evolution.

A genome-wide association study (GWAS) scans many regions of the target genome for statistical associations with a trait of interest; typically, the sequenced variants (most often SNPs) are not causative but are often in linkage disequilibrium (LD) with a causative variant (Schaid et al., 2018). However, the polygenic nature of most traits can make it difficult to detect trait-associated variants, as most variants will likely be of small effect size. In humans, a GWAS compensates for the difficulty in detecting small-effect variants by utilizing very large sample sizes, typically thousands or tens of thousands of individuals (Crouch & Bodmer, 2020). SNP-capture panels in human studies are also often designed to capture variants from millions of genomic regions, increasing the likelihood that a captured variant is in LD with a causative variant (Bush & Moore, 2012). Although sampling thousands of devils is prohibitively difficult due to declining population sizes and trapping logistics, the devil-DFTD system is amenable to a GWAS. Because DFTD is a novel pathogen and imposes an enormous selective pressure on
devils, variants contributing to disease-relevant phenotypes in both devils and DFTD are likely to be of large effect (Rokyta et al., 2005) and thus easier to detect. Furthermore, the low genetic diversity in devils has resulted in extensive linkage disequilibrium (~200kb; Epstein et al. 2016), facilitating an increased likelihood that a sequenced SNP is in LD with a causative variant without the need for a capture panel targeting millions of loci. Previous work by Margres et al. (2018a) uncovered genotype-phenotype relationships in devils for disease-relevant traits using a GWA approach, thus demonstrating the viability of a GWAS design to further characterize devil and DFTD genetics.

Although much work has been done to uncover the genetics, evolution, and predicted effects of DFTD on devil populations, there remains the need to further elucidate how devils and DFTD are evolving in response to one another. Based on the rapid evolution of devils in response to DFTD, and the concomitant evolution of DFTD, it is probable that many devil and DFTD genes are coevolving. Hence, much could be learned by analyzing these genomes under a coevolutionary framework. Furthermore, understanding host-pathogen evolution between DFTD and devils will better inform empirically-driven epidemiological model building, as previous models have utilized ecological data only (McCallum et al., 2009; Wells et al., 2017, 2019). To this end, determining the genomic architecture (i.e., contribution of genomic variance explaining phenotypic variance and whether a trait is polygenic or controlled by few genes of large effect) of disease-relevant traits in both devils and tumors can help both to better parameterize predictive models and ascertain resistant devils suitable for breeding programs.

Here, we characterized multiple facets of devil and DFTD coevolution. First, the ability to predict DFTD genetic population structure from devil genetic population structure throughout sites in Tasmania was assessed. Tight matching between devil and DFTD genetic structure may
be caused by coevolutionary interactions, although other factors such as dispersal of devils and DFTD can also influence genetic structure. Kozakiewicz et al. (2020) identified differences in dispersal barriers between devils and DFTD, and it was thus expected that dispersal differences may facilitate a decoupled genetic structure between devils and DFTD. Next, the genomic architecture of host force of infection, tumor force of infection, and tumor virulence was identified. Previous work found that few large-effect loci explained a large proportion of variance in case-control and survival in devils (Margres et al., 2018). Hence, based on past work and the recent emergence of DFTD, we hypothesized that selection has favored large-effect variants in both devils and tumors. Finally, the contribution of devil-DFTD genome interaction in explaining force of infection was determined using a recently developed joint modeling approach (Wang et al., 2018). Due to the large selective pressure incurred on devils by DFTD, and hence the possibility for a reciprocal selective pressure on DFTD, we anticipated genomic interactions to be important in explaining force of infection.
METHODS

Sample collection

Devils were trapped from 2006 – 2020 in five regions of Tasmania along an east-west axis representing varying degrees of devil-DFTD coexistence (Fig. 1). A capture-mark-recapture framework was used, whereby devils were tagged with microchips upon first capture and could be identified on subsequent trappings (see Hamede et al., 2015 for details). Ear biopsies were obtained for sequencing of hosts, and DFTD samples were collected from tumor margins and confirmed through histopathological assays. For individuals whose age was unknown, age estimates accurate to the year were made using head width, molar measurements (eruption and wear), and canine over-eruption (Lachish et al., 2009). Because devils typically breed in early March (Bell et al., 2020), April 1st was assumed to be the month and day of devil birth, permitting devil age in days to be estimated. Tumor measurements of length, width, and depth (depth being the least accurate measurement) were also taken on capture to the nearest millimeter, thus allowing for calculations of tumor volume and tumor load (the sum of all tumor volumes; Hamede et al., 2017).
Figure 1. Devil trapping locations within Tasmania. These sites represent a natural gradient of long-diseased populations to populations where DFTD was not present at the time of sampling. Shown are the number of devil generations since DFTD arrival and the percentage of the original population remaining at the site.

**Phenotype data**

Associative modeling and coevolutionary analyses were performed for tumor virulence and force of infection. For both of these phenotypes, it was necessary to estimate the date of first infection for devils and, in some cases, for individual sequenced tumors. A model of tumor growth, based on a logistic growth curve, has been fit to data from West Pencil Pine (Wells et al., 2017), and this model was used to project backwards and estimate the date of first infection (assuming an arbitrarily small tumor load of 0.0001 cm$^3$ as the date of first infection) based on the tumor load of the first trapping where DFTD was observed. The fit model had a maximum tumor load ($M_{\text{max}}$) of 202 cm$^3$ (CI = 198–223 cm$^3$) and a lag phase of \(~60\) days, whereby it is assumed a tumor is not observable upon trapping during the lag phase. Although the logistic growth model was fit to tumor load and samples specific to West Pencil Pine, the model’s parameter estimates are likely generally applicable both to other localities and individual tumor measurements. However, unless specified otherwise, tumor load was used as input to the growth model back calculation.
To minimize potential error from estimating devil age at first infection using the growth model, samples containing uncertainty were removed based on a set of filtering parameters. First, to maximize power, tumor depth was imputed for samples missing a depth measurement. Although the majority of samples possessed depth measurements (~98%), this represents the most difficult tumor dimension to measure and was thus missing more frequently than tumor width or tumor length. Imputation was done by fitting a linear model to individual tumor measurements with no missing measurements to predict tumor volume using tumor area (Fig. A1); hence, imputation was only done for tumors possessing at least width and length measurements. Devils which lacked measurements for any confirmed tumors were removed.

Next, the date of first infection was estimated for devils with multiple tumors on their first trap which contained a mixture of measured and unmeasured tumors. Because tumor load sums the volume of all tumors, a devil missing measurements for some tumors may overestimate the age at first infection if the missing tumors were large. If the estimated infection date was ≤ 60 days from a previous trapping, it was assumed that this previous trap date was close to the date of first infection and that the back calculated estimate was sufficient; samples failing this criterion were removed. Because the $M_{\text{max}}$ for the growth model averaged 202 cm$^3$, accurate estimates could not be made for tumor loads significantly larger than this maximum and back-projections greater than a year were unreliable. As such, devils possessing a tumor load greater than 223 cm$^3$ (the 97.5% tumor load $M_{\text{max}}$ CI estimated by Wells et al., 2017) on their first trap with an estimated infection date that was not ≤ 60 days from a previous trap were removed. A final filtration was done to remove samples with tumors lacking histopathological confirmation. Infection age estimates were made for these samples first with just the confirmed tumors and
then again including the unconfirmed tumors. If the two estimates differed by > 68 days (one standard deviation for the back calculated estimate), the sample was removed.

Tumor virulence was calculated as the difference between the date of first DFTD infection (estimated using the growth model) and the last observed trap date for that devil, necessitating that the devil was trapped two or more times with ≥ 40 days between successive traps. The interval from first infection to last observed trap represents the estimated length of time a devil survives upon being infected with DFTD and is the best proxy for tumor virulence that can be achieved given the data. Because this estimate does not observe the date of devil death, it does not represent the true duration of survival after infection; however, the probability of recapturing a devil does not change significantly based on DFTD status or season (Kery & Schaub, 2011). Thus, attempting to estimate survival beyond the final trapping date would be constant for all individuals and would not alter the relative survival for any devil. Variation in tumor genomes was used to assess variation in devil survival for the virulence phenotype, and although alternative names for this phenotype could be contrived, the term “virulence” is used here for brevity.

A proxy was used for force of infection, which was calculated as the length of time (in days) it took for a devil to become infected with DFTD when infection was possible based on age and date of disease arrival (Table A1). To estimate this proxy, it was first necessary to determine the age at which a devil was first infected with DFTD, and this differed for host versus pathogen force of infection. For devils, the date of first infection was found using the growth model with tumor load at the earliest trap date as input (Wells et al., 2017). For tumors, only the volume of the sequenced tumor at the earliest trap date was used to calculate the date of first infection. Tumor force of infection was calculated in this manner to ensure that downstream
models were fitting the tumor genetic data to the relevant tumor phenotype, rather than to the aggregate of all tumors on a devil. If the sequenced tumor was not the first tumor to infect a devil (considered to be the only tumor on the first trapping or the largest tumor on the first trapping if multiple tumors were present), the sample was removed, a step which was necessary due to the increased susceptibility of devils already infected with DFTD to subsequent infection (Cheng et al., 2019). Because a devil cannot be infected with DFTD before the cancer arrives at the devil’s respective site, the age of the devil when DFTD arrived at its site was subtracted from the devil’s infection age, yielding the number of days it took for the devil to become infected after DFTD arrived. For devils which were less than a year old upon DFTD arrival or were born after arrival, 365 days was subtracted from their infection age because devils younger than one year old are rarely observed with DFTD (Cheng et al., 2019).

**Alignments and variant calling**

To sequence the genomes of devils and DFTD, a hybridization capture panel of 197K loci was developed based on prior studies to (1) target regions of known relevance to disease phenotypes (e.g., Margres et al., 2018a) and (2) encompass the entirety of the devil genome to ensure complete genomic coverage. A total of 456 devils and 504 tumors were sequenced, 338 of which were paired. These libraries were sequenced in five sets; for each set, 192 individual libraries were indexed, pooled, and sequenced 150 PE on an Illumina NovaSeq S4 lane at the North Carolina State University Genomic Sciences Laboratory (https://research.ncsu.edu/gsl/). Reads were demultiplexed, quality-assessed with FastQC (Andrews, 2010), and trimmed with TrimGalore! (Krueger et al., 2021) at default settings to remove adapter contamination. Trimmed reads were then aligned to the reference genome mSarHar1.11 (*Sarcophilus harrisii - Ensembl Genome Browser 105*, n.d.) using BWA MEM (Li, 2013) with the -M flag and all else at default
settings. Picard (Picard Tools - By Broad Institute, n.d.) was used to mark PCR and optical duplicate reads via MarkDuplicates, and Picard CollectHsMetrics was used to obtain probe-capture metrics.

The GATK pipeline (Poplin et al., 2018) was used to call SNPs. HaplotypeCaller, which generates sample-level SNP and indel calls, was run with the -ERC GVCF and --do-not-run-physical-phasing flags for each individual devil and tumor sample. To combine the sample-level files output from HaplotypeCaller, GenomicsDBImport was run separately for devils and tumors with intervals set to each chromosome of the devil reference assembly (including unplaced scaffolds). The tumor and devil databases were updated as new samples were sequenced. Once all samples were sequenced, GenotypeGVCFs was used to extract SNPs and indels from the tumor and devil databases, generating tissue-specific Variant Call Format (VCF) files across all samples. GenotypeGVCFs was run in parallel such that extraction was done per-chromosome and combined in the order of each chromosome once all runs had completed. Next, SelectVariants was used to generate SNP and indel files separately for each tissue and VariantFiltration was run with SNP filters QD < 2.0, FS > 60.0, MQ < 40.0, MQRanksSum < -12.5, and ReadPosRankSum < 8.0, and indel filters QD < 2.0, FS > 200.0, and ReadPosRankSum < -20.0 as recommended by GATK developers. To remove potential host contamination in the cancer data set, tumor SNPs were further filtered using bcftools isec (Danecek et al., 2021) to remove any SNPs common to both the devil and tumor VCF files. These final VCF files contained 7,636,616 devil SNPs and 6,183,694 tumor SNPs and were further filtered based on parameters specific to each analysis as described below (see Table 1 for analysis-specific sample and SNP counts).
Table 1. Number of SNPs and samples used in associative tests. Both SNPs and samples differed between each test based on the number of samples available for the given phenotype.

<table>
<thead>
<tr>
<th>Analysis</th>
<th>SNPs</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAPC</td>
<td>221,965 (hosts)</td>
<td>456 (hosts)</td>
</tr>
<tr>
<td></td>
<td>63,317 (tumors)</td>
<td>504 (tumors)</td>
</tr>
<tr>
<td>GEMMA: force of infection</td>
<td>200,087 (hosts)</td>
<td>315 (hosts)</td>
</tr>
<tr>
<td></td>
<td>19,376 (tumors)</td>
<td>329 (tumors)</td>
</tr>
<tr>
<td>GEMMA: virulence</td>
<td>19,024 (tumors)</td>
<td>158 (tumors)</td>
</tr>
<tr>
<td>ATOMM: force of infection</td>
<td>17,935 (hosts)</td>
<td>314 (hosts)</td>
</tr>
<tr>
<td></td>
<td>13,085 (tumors)</td>
<td>314 (tumors)</td>
</tr>
</tbody>
</table>

DAPC lineage analysis

To determine the population structure of devils and DFTD, Discriminant Analysis of Principal Components (DAPC; Jombart et al., 2010) was implemented in the package adegenet in R version 4.1.2 (R Core Team, 2021). Both devil and tumor SNPs were filtered such that sites with any missing samples were removed before being used as input to DAPC, yielding 221,965 devil SNPs and 63,317 tumor SNPs. The function find.clusters within DAPC identifies groups by running k-means clustering with increasing values of k to identify the optimal number of groups using Bayesian Information Criterion (BIC). The k-means clustering is done after transforming the data via PCA to decrease computation time for large datasets. These clusters are then used as input to DAPC which again transforms the data using PCA before running discriminant analysis to maximize between-group distance while minimizing within-group distance (Jombart et al., 2010). The find.clusters function was thus used to initially identify clusters, retaining all PCs for k-means and selecting the k corresponding to the “elbow” of a BIC plot. The number of principal components selected for the discriminant analysis has important consequences in model fitting, whereby too few PCs results in underfitting and too many PCs yields an overfit model (Jombart et al., 2010). To select the optimal number of PCs, xvalDapc was used to perform cross-validation, randomly splitting the samples into a training set (90%) and a validation set (10%) and running DAPC with a variable number of retained PCs to obtain the accuracy of predicted
group membership (groups determined as above using `find.clusters`). At each PC retention level, training/validation sampling and DAPC were repeated 30 times. The number of PCs retained which yielded the highest predictive accuracy in cross-validation was used in the final DAPC analysis. DAPC was then run on the k-means clusters, retaining the optimal number of PCs and all eigenvalues within the discriminant analysis.

**BSLMM genome architecture modeling**

Genome-wide Mixed Model Association (GEMMA; Zhou & Stephens, 2012) was used to implement a Bayesian Sparse Linear Mixed Model (BSLMM; Zhou et al., 2013) to explore the genomic architecture underlying force of infection (devil and tumor phenotype) and tumor virulence. Because BSLMMs are a combination of linear mixed models (which assume all variants have a small effect) and sparse regression models (which assume a large effect by few variants), they are capable of fitting both polygenic and simpler genetic architectures. The model also accounts for relatedness among samples before model fitting and outputs the proportion of phenotypic variance explained (PVE; the total variance in the phenotype explained by both small- and large-effect SNPs), proportion of genotypic variance explained (PGE; the proportion of PVE explained by only large-effect SNPs), and a posterior inclusion probability (the probability of being a large-effect SNP; PIP) for individual SNPs.

For both traits, the VCF file was first filtered to remove samples with a missing phenotype, and was subsequently filtered on 5% missingness, the removal of sites which were not biallelic, a minor allele frequency (MAF) of 0.05 for devils, and a MAF of 0.01 for tumors. Because DFTD reproduces asexually, this lower tumor MAF was used to capture the increased number of rare variants anticipated in tumors relative to devils. A linear link function was used for all phenotypes (-bslmm 1) with a centered relatedness matrix and default priors. Five
independent chains and 60,000,000 iterations per chain (with a 6,000,000 burn-in) were run for each phenotype. Within and between chain convergence was assessed using Rhat (Table A4), where values near 1 indicate a high degree of within and between-chain convergence (Vehtari et al., 2021), bulk effective sample size (ESS), a measure of sampling efficiency in the bulk of the posterior distribution, and tail ESS, a measure of sampling efficiency in the tails of the posterior distribution, implemented in Rstan (Stan Development Team, 2020) within R version 4.1.0. Convergence was also checked visually by manually inspecting hyperparameter distributions and trace plots. Although individual SNP p-values indicating association with the trait are not output from BSLMM model fitting, SNPs possessing the top five largest PIP (calculated as the mean across the five chains) were further explored. For these SNPs, the closest gene 100kb upstream or downstream from the SNP (representing devil LD of ~200 kb; Epstein et al., 2016) was found using Variant Effect Predictor (VEP; McLaren et al., 2016). Putative gene functions were identified using GeneCards (Stelzer et al., 2016).

**ATOMM host-pathogen joint modeling**

Typical associative models for GWAS consider only a single genome at a time, ignoring potential interactions between the genomes of organisms which are coevolving. Analysis with a Two-Organism Mixed Model (ATOMM; Wang et al., 2018) was developed to jointly model the genomes within a host-pathogen system, estimating the genomic heritability of the marginal host, marginal pathogen, and interaction of the host and pathogen genomes. ATOMM is also capable of looking at genotype-by-genotype interactions between individual SNPs and may thus be used to identify sites with potential signatures of coevolution. Due to the larger number of paired samples available for force of infection (N = 314) relative to tumor virulence (N = 119), only force of infection was analyzed using ATOMM. The force of infection data used as input to
ATOMM were derived in a tumor-centric manner, whereby infection age was estimated based on the volume of the sequenced tumor (rather than tumor load). Tumors which were not the first to infect the devil were retained, but first-infection status was used as a binary fixed effect in the model. ATOMM assumes the response phenotype to be multivariate Gaussian or binomial (Wang et al., 2018), and the force of infection values were thus standardized using the RankNorm function in RNOmni (Zachary McCaw, 2020) within R version 4.1.0. Because ATOMM is incapable of handling missingness and requires haploid genotypes as input, the host and tumor VCF files were first filtered to remove devils and tumors missing phenotype data; sites with missing genotypes and/or that were not biallelic were also removed. Genotypes were then converted into a matrix of zeros and ones, where zero indicated two reference alleles and one indicated the presence of at least one alternate allele (heterozygotes and homozygotes for alternate alleles were considered identical). The haploid genotype matrix was then filtered using a custom script based on a MAF of 0.05 for devils and MAF 0.01 for tumors as described above.

To obtain uncertainty in the genome heritability estimates made by ATOMM, the model was fit using 50 different initialization values for marginal host, marginal pathogen, interaction, and noise genomic estimates. Initializations were chosen using the BSLMM posterior 95% credible intervals as priors, and the initializations yielding the lowest maximum likelihood estimate were used for the genotype-by-genotype interaction tests. ATOMM’s convergence delta was also changed from the default value (0.01) to 0.001 to prevent premature convergence before reaching the optimum. Although model fitting was a computationally trivial task (averaging 2.5 minutes per run), the genotype-by-genotype interaction tests scaled based on the number of host and pathogen SNPs used for model fitting (e.g., 10,000 host and pathogen SNPs requires 100,000,000 individual interaction tests). Thus, to decrease this computational burden, a subset
of SNPs was used for the interaction tests and was selected based on marginal host and pathogen p-values (alpha < 0.05) obtained from the initial ATOMM model fitting. The top five lowest p-value SNPs identified by these interaction tests were further explored using VEP as in the GEMMA analyses.
RESULTS

Genetic population structure of devils and DFTD

Clustering devils based on genetic data (456 individuals with 221,965 SNPs) with k-means implemented in DAPC yielded highest support for two distinct clusters (BIC = 4110; Fig. A2). Clustering analysis revealed the presence of two genetic devil lineages and plotting the geographic distribution of these lineages across Tasmania shows a clear distinction from populations in Freycinet to those in the northwest (West Pencil Pine, Takone, Black River, and Arthur River; Fig. 2A). Genetic divergence between the two devil lineages was low (weighted F_{st} = 0.0551; Table A2), concordant with past studies (Miller et al., 2011). Subsequent hierarchical clustering of just the northwestern sites identified further genetic population structure in devils, with two lineages present (BIC = 3410; Fig. A3) and co-occurring within each northwestern site (Fig. 2C). The separation of genetic clusters via discriminant analysis was greater between the east versus northwest clusters (Fig. 2B) relative to separation of genetic clusters exclusively in the northwest (Fig. 2D), and this was supported by F_{st} estimates (east-northwest weighted F_{st} = 0.0551, northwest-northwest weighted F_{st} = 0.0096).
Figure 2. Devil lineages identified using DAPC. A. The distribution of lineages is shown across Tasmania, with individual devils plotted as dots or pie charts (for densely sampled sites). The site name and number of samples is also shown next to densely sampled sites. Colors correspond to devil lineages shown in 2B. WPP: West Pencil Pine. B. Devil genetic clusters identified using DAPC. C. Hierarchical clustering of the northwestern sites obtained by removing cluster 2 from 2A (i.e., Freycinet). D. Devil genetic clusters identified via hierarchical clustering.

A k-means clustering on DFTD genetic data (504 tumors with 63,317 SNPs) revealed greatest likelihood (BIC = 3220; Fig. A4) for four tumor lineages. After retaining all eigenvalues for the discriminant analysis, the identified clusters once more showed low within-group dispersion but high between-group separation (Fig. 3B). Plotting these tumor lineages based on
their location in Tasmania revealed a population structure differing from that of devils, as each site contained a co-occurrence of three or four lineages (Fig. 3A). Furthermore, these lineages tended to be found within multiple devil trapping sites throughout Tasmania, and lineages 1 and 3 were found at nearly all devil trapping sites. Overall, tumors showed low genetic divergence between lineages (weighted $F_{st}$ range = 0.030 – 0.045; Table A3). Lineage 3 showed lowest genetic divergence between all lineages (weighted $F_{st}$ = ~0.031 for each pairwise test), whereas all other lineages showed greater divergence (weighted $F_{st}$ range = 0.041 – 0.045). The dynamics of these lineages within each major site also shifted over time, although sampling one of the major sites at any given year yielded the presence of at least two tumor lineages (Fig. 3C).

Comparing the genetic structure of devils relative to DFTD over space reveals a decoupled pattern. Devils show clear spatial structuring over an east-west axis (Fig. 2A) with further structure present in northwestern sites (Fig. 2C). However, the co-occurrence of tumor lineages and ubiquity of most lineages throughout Tasmania (Fig. 3A) indicate little genetic structuring within the tumor; hence, devil genetic structure does not appear to be predictive of DFTD genetic structure. The divergence of lineages within devils and tumors was similar, whereby the divergence between DFTD lineage 3 and all other lineages (weighted $F_{st}$ = ~0.031) was slightly closer to the east-northwest devil divergence (weighted $F_{st}$ = 0.0551) than to the northwest-northwest devil divergence (weighted $F_{st}$ = 0.0096), and the pairwise divergence between all other DFTD lineages (weighted $F_{st}$ = ~0.043) was closer to the east-northwest devil divergence.
Figure 3. DFTD lineages identified using DAPC. A. Geographic distribution of tumor lineages across Tasmania, with individual tumors represented as colored dots or pie charts (for densely sampled sites, sample sizes shown in 3C). WPP: West Pencil Pine. B. Tumor lineage clustering along the first two discriminant axes. C. Tumor lineage abundance from the four most densely sampled sites over time.

Genome architecture underlying disease-relevant traits

Model fitting with a BSLMM revealed that devil genomes explained a large proportion of the variance in force of infection (0.611; 95% CI = 0.390 – 0.828; Fig. 4; Table 2), whereas tumor genomes explained less variance in force of infection (0.268; 95% CI = 0.055 – 0.731) but accounted for much of the variance in tumor virulence (0.679; 95% CI = 0.392 – 0.997). Host force of infection was associated with few variants of large-effect (2; 95% CI = 1 – 5) which explained 36.7% of the PVE (or 22.4% of the total force of infection PVE). A few large-effect variants (5; 95% CI = 3 – 17) accounted for 74.6% of tumor virulence PVE (or 50.7% of the total PVE), indicating that these large-effect SNPs were able to explain the bulk of variation in tumor virulence.
Table 2. BSLMM genome architecture statistics. Mean and median variance explained by all genotypes represent the proportion of phenotypic variance explained (PVE) of the BSLMM posterior distribution. Mean variance explained by large-effect SNPs represents the PVE using only large-effect SNPs (e.g., large-effect SNPs explain 22.8% of the total variance in force of infection). Numbers in parenthesis represent Bayesian 95% credible intervals from the posterior distributions.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Phenotype</th>
<th>Mean variance explained by all genotypes (%)</th>
<th>Median variance explained by all genotypes (%)</th>
<th>Number large-effect SNPs</th>
<th>Mean variance explained by large-effect SNPs (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Host</td>
<td>Force of infection</td>
<td>61.1 (39.0–82.8)</td>
<td>61.1</td>
<td>2.5 (1–5)</td>
<td>22.8 (13.0–34.8)</td>
</tr>
<tr>
<td>Tumor</td>
<td>Force of infection</td>
<td>30.0 (5.5–73.1)</td>
<td>26.8</td>
<td>45.2 (0–235)</td>
<td>13.4 (0–29.0)</td>
</tr>
<tr>
<td>Tumor</td>
<td>Virulence</td>
<td>69.8 (39.2–99.7)</td>
<td>67.9</td>
<td>6.4 (3–17)</td>
<td>51.2 (29.0–68.9)</td>
</tr>
</tbody>
</table>

Figure 4. Genome architecture identified through BSLMM model fitting. The y-axis represents the posterior distribution of the PVE (left plots) and PGE (right plots), and each phenotype is shown on the x-axis. Dots in each violin plot represent median values and lines represent Bayesian 95% credible intervals. FOIH: force of infection host; FOIT: force of infection tumor; PGE: proportion genotypic variance explained; PVE: proportion phenotypic variance explained.

Disease-relevant candidate genes

For both host force of infection and tumor virulence, variants with the largest effect sizes were often near a gene (within 100kb upstream or downstream from a gene, representing devil LD of ~200kb; Table 3). All genes nearby force of infection variants lacked a functional annotation, and the variant with the largest PIP (0.978) was found on an unplaced scaffold. The unplaced scaffold variant was not near an annotated gene, although this is likely an artifact of the variant being on a region of the genome which has not been localized to a chromosome. The largest effect variant for tumor virulence (0.856) was a variable chain immunoglobulin (IGV), an
An important component of antibodies and thus of the adaptive immune system (Watson & Breden, 2012). The only other annotated tumor virulence variant was a SNP found within the intron of \textit{SPOCK3} (effect size = 0.190). \textit{SPOCK3} encodes a calcium-binding proteoglycan which can inhibit membrane-type matrix metalloproteinases (MMPs), the expression of which is important for metastasis in T-cell leukemia (Kamioka et al., 2009). Thus, \textit{SPOCK3} appears to play an important regulatory role in tumorigenesis of some human cancers.

\textbf{Table 3.} Top 5 BSLMM SNPs for host force of infection and virulence. The closest gene within a 100kb window is shown with putative functions identified using GeneCards (\url{www.genecards.org}). SNPs are shown from largest to smallest PIP.

<table>
<thead>
<tr>
<th>CHR</th>
<th>Distance</th>
<th>Gene</th>
<th>CHR</th>
<th>Distance</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Unplaced scaffold</td>
<td>2</td>
<td>-4kb</td>
<td>IGV</td>
</tr>
<tr>
<td>6</td>
<td>+13kb</td>
<td>lncRNA</td>
<td>3</td>
<td>Intergenic</td>
<td>N/A</td>
</tr>
<tr>
<td>6</td>
<td>+9kb</td>
<td>Uncharacterized protein</td>
<td>4</td>
<td>-9kb</td>
<td>lncRNA</td>
</tr>
<tr>
<td>6</td>
<td>-12kb</td>
<td>Uncharacterized protein</td>
<td>X</td>
<td>Intergenic</td>
<td>N/A</td>
</tr>
<tr>
<td>2</td>
<td>Intron</td>
<td>Uncharacterized protein</td>
<td>6</td>
<td>Intron</td>
<td>\textit{SPOCK3}</td>
</tr>
</tbody>
</table>

\textbf{Force of infection joint modeling and genotype-by-genotype interactions}

Fitting ATOMM to the host and pathogen genomic data (17,935 host SNPs and 13,085 pathogen SNPs with 314 paired devils and tumors) yielded PVE estimates for the marginal host (0.298), marginal pathogen, (0.076), host-pathogen interaction (0.403), and noise (0.236).

Running ATOMM under default initialization PVEs (host = 0.25, pathogen = 0.25, interaction = 0.25, noise = 0.25) and convergence delta (0.01) resulted in a suboptimal model fit, as indicated by a high MLE which failed to decrease from the initial MLE. However, running ATOMM under 50 different initializations with a lowered convergence delta (0.001) resulted in the same MLE 49 times (MLE = -24.4844), with only a single run differing in its final MLE (MLE = -23.8675) which was higher than the other runs and thus a suboptimal solution. This indicates that, almost irrespective of the initial state of the model, the final fit was identical, and all such runs yielded identical heritability estimates (see above). Heritability estimates from the
concordant model fits are reported here (Fig. 5) both because of their concordance and the greater fit according to MLE.

![Bar chart showing proportions](image)

**Figure 5.** Force of infection host-pathogen heritability estimates. Heritability estimates were identified through model fitting using ATOMM. The y-axis shows the phenotypic variance explained using the marginal host genome, marginal pathogen genome, interaction between the host-pathogen genomes, and random noise.

Sorting the genotype-by-genotype interaction tests by ascending p-value and looking 100kb up- and downstream (i.e., within 200 kb LD for devils; Epstein et al., 2016) from the top 5 lowest p-value host and pathogen SNPs revealed genes showing potential signatures of coevolution (Table 4). For devils, all but one of these SNPs were found within a gene (the *RUSC1* SNP was downstream from the gene), and all genes had an annotated function. Of the SNPs found within genes, most were within non-coding regions (i.e., UTR or intron), but a single nonsynonymous substitution was found within an exon of *SLC4A11*. Only a single SNP in
tumors was within a gene (an intronic substitution in \textit{GLRA3}) and all others were intergenic but nearby a gene. A single gene near a tumor SNP lacked a functional annotation (LOC100917912).

\textbf{Table 4.} Top 5 ATOMM force of infection genotype-by-genotype interactions. Interacting SNPs were identified using only host and pathogen SNPs which were significant based on marginal effects. The closest gene within a 100kb window is shown with putative functions identified using GeneCards (\url{www.genecards.org}). SNPs are shown from smallest to largest p-value.

<table>
<thead>
<tr>
<th>Host Gene</th>
<th>CHR</th>
<th>Distance</th>
<th>Pathogen Gene</th>
<th>CHR</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{RUSC1}: Involved in neurite outgrowth by regulating NGF</td>
<td>4</td>
<td>+7kb</td>
<td>\textit{AGPS}: Upregulated in multiple aggressive human cancers</td>
<td>3</td>
<td>-60kb</td>
</tr>
<tr>
<td>\textit{HNF1A}: Transcription regulator; pancreatic cancer tumor inhibitor</td>
<td>1</td>
<td>Intron</td>
<td>\textit{SKOR2}: Sequence specific double stranded DNA binding activity</td>
<td>1</td>
<td>-57kb</td>
</tr>
<tr>
<td>\textit{PDS5B}: Negative regulator of cell proliferation; tumor suppressor</td>
<td>3</td>
<td>3’ UTR</td>
<td>\textit{LOC100917912}: No annotated function</td>
<td>2</td>
<td>+2kb</td>
</tr>
<tr>
<td>\textit{SLC4A11}: important for cell growth and proliferation</td>
<td>6</td>
<td>Arg → Gln</td>
<td>\textit{AGPS}: Upregulated in multiple aggressive human cancers</td>
<td>3</td>
<td>-60kb</td>
</tr>
<tr>
<td>\textit{DPP6}: Single pass type II membrane protein</td>
<td>5</td>
<td>Intron</td>
<td>\textit{GLRA3}: Glycine receptor subunit</td>
<td>6</td>
<td>Intron</td>
</tr>
</tbody>
</table>
DISCUSSION

Genetic population structure differs between devils and DFTD

The two devil lineages found here (Fig. 2A) corroborates Miller et al. (2011), who identified two major lineages of devils distributed along an east-west axis. Their findings also identified five haplogroups but used a limited number of mitochondrial genotypes (17 mitochondrial sites relative to ~197k nuclear used here), as well as a smaller devil sample size (87 compared to 456 used here). Thus, the existence of two devil lineages appears to find the best support, and the weaker evidence found for additional lineages by previous studies may be an artifact of sparse genetic information or sampling design. Genetic population structure of DFTD was consistent with the four tumor lineages previously identified (Fig. 3A; Kwon et al. 2020). Kwon et al. (2020) identified four tumor clades, co-existence of clades within devil populations throughout Tasmania, and thus the potential for lineage competition within many sites. Tumor lineage dynamics within densely sampled trapping sites over time (Fig. 3C) revealed the possibility for lineage competition, whereby some sites appear to be experiencing intense competition which could lead to lineage replacement (e.g., Freycinet), while lineages appear to be coexisting at roughly constant proportions through time at other sites (e.g., Black River).

Although devils appear to form distinct genetic populations (Fig. 2), indicating minimal gene flow between Freycinet and the northwestern trapping sites (West Pencil Pine, Takone, Black River, and Arthur River), DFTD genetic populations were largely unstructured with reference to geography or time, and multiple lineages co-occurred at each sampled site (Fig. 3A).
If coevolution is a driving force between devil and DFTD adaptation, this decoupling would be less likely under arms race (i.e., recurrent selective sweep) but feasible within trench warfare (i.e., negative frequency dependent selection) dynamics. Under arms race dynamics, a beneficial allele would sweep to fixation within a devil population, and counteradaptation within DFTD would similarly result in allelic fixation. For tumor populations, arms race dynamics is expected to eliminate the coexistence of multiple lineages within a site, as the tumor lineage possessing the beneficial mutation should quickly outcompete other lineages. Recurrent fixation of alleles would likely induce matching of population structure between devils and tumors, whereby sites with little gene flow form discrete units possessing a single tumor lineage co-occurring with region-specific devil lineages.

Assuming trench warfare dynamics, multiple tumor lineages would be maintained within a site. Under these dynamics, devils may adapt to whichever tumor lineage is most common, reducing its fitness and allowing another lineage to rise in frequency. Adaptation in tumors would similarly be against the most common allele in devils, leading to maintenance of polymorphisms in both devils and tumors. If devils and DFTD are coevolving under trench warfare dynamics, maintaining polymorphisms for disease-relevant genes would likely not facilitate multiple devil lineages within a single site. This is because a high degree of gene flow, as is to be expected between devils at a single trapping site, mixes and recombines alleles through sexual reproduction. However, in the asexually reproducing tumor, a single beneficial mutation may give rise to a new lineage which appears genetically distinct from the original lineage after accumulating further neutral mutations. Here, by maintaining polymorphic alleles in the tumor, negative frequency dependent selection would lead to the persistence of multiple tumor strains within a site. However, another possibility is that the decoupled genetic structure
between devils and DFTD indicates a lack of host-pathogen coevolution. Thus, further studies are needed to establish the extent, if any, of devil-DFTD coevolution occurring within each site, and a greater number of devil and tumor samples will need to be collected over a larger timescale within each trapping site to test for arms race versus trench warfare dynamics.

Although the discordance of tumor and devil genetic population structures appears to lend credence to trench warfare dynamics if coevolution is assumed, this argument ignores a temporal component. Potential shifts in tumor lineage dynamics, whereby lineage replacement may be occurring within some sites (e.g., Freycinet), perhaps indicates a selective sweep (Fig. 3C). Thus, it is possible that coevolutionary dynamics are not consistent throughout the geographic range of devils and DFTD, with some sites undergoing negative frequency dependent selection and others experiencing arms race dynamics. The existence of multiple tumor lineages within each site also facilitates tumor lineage competition which may not be influenced by host selective pressures. Such competition may lead to a virulence-transmission tradeoff, as the most transmissible tumors within a site would outcompete other lineages. However, if within-host lineage competition is extensive, tumors may evolve increasing virulence to facilitate growth, as high virulence would grant a fitness advantage by excluding growth of other tumor lineages (Alizon et al., 2009).

Differences in devil and DFTD genetic population structure also indicates differences in dispersal between host and pathogen. Kozakiewicz et al. (2020) found that isolation by resistance, particularly major roads and highways, acted as barriers to gene flow and thus explained devil genetic structure, but that genetic structure in tumors was largely absent and tumors co-existed at many sites. Of note, Kozakiewicz et al. (2020) used fine-grain sampling of devils over a 12,000 km² area in northwest Tasmania, differing from the broader-scale devil
sampling used here (~55,000 km² area). The difference in population structures observed here could be the result of differences in the timings of devil dispersal and reproduction, as devils typically disperse when juveniles (i.e., when least likely to become infected with DFTD) but reproduce as adults (i.e., when most likely to become infected with DFTD; Hamede et al., 2013). However, under this hypothesis gene flow between devils is expected to cover a greater geographic distance than tumor spread, as juveniles tend to travel greater distances than adults (Lachish et al., 2011). The results found here seem to indicate the opposite, that DFTD is capable of greater dispersal than devils. Because DFTD transmits through biting, there exist more opportunities for its spread relative to gene flow between devils, which requires mating between devils and is thus constrained primarily to devil mating season (typically from February to May; Bell et al., 2020). Although biting interactions facilitating DFTD transmission occur most typically during mating season (Andersen et al., 2020; Hamilton et al., 2019), the asexual nature of the tumor may further facilitate its population structure. Despite the restriction in gene flow by major roadways in devils, it is still feasible for a few individuals to successfully cross these barriers. However, for devils a single crossing every few generations provides insufficient gene flow to homogenize demes (Slatkin, 1987), and it is not guaranteed that the individual will reproduce. For DFTD, the successful crossing of a single infected individual could result in proliferation of that lineage in the new population, which is maintained due to its clonal nature.

Despite the apparent support for more rapid dispersal of DFTD relative to devils, this could be an artifact of the sampling scheme used for this study. Devil populations were sampled either in the northwestern region of Tasmania (Arthur River, Black River, Takone, and West Pencil Pine) or in the eastern region (Freycinet). The distance from Freycinet to West Pencil Pine (the nearest western sampling site) is ~215 km, whereas West Pencil Pine is a distance of ~60
km from Takone. Because DAPC minimizes within-group differences and maximizes between-group distances, sampling from geographically distant populations is likely to produce distinct genetic clusters which may not accurately reflect population discreteness (Jombart et al., 2010). Gene flow between Freycinet and the western sites could occur if intermediate devil populations exist between these sites with gene flow occurring along a continuum between the eastern and western sites. Furthermore, clustering only devils from the northwestern sites revealed two devil lineages, indicating that inclusion of the geographically distant Freycinet devils may have swamped the genetic differences between northwestern devils in the DAPC. As such, the disparity between devil and DFTD dispersal found here may be exaggerated, and it will be necessary to design a thorough sampling scheme along a continuum of Tasmania if widespread devil and tumor dispersal patterns are to be elucidated.

Further studies are necessary to resolve the mechanisms underlying genetic population structure of devils and DFTD, and their discordance. Although coevolutionary dynamics often vary in intensity spatially, such variability is likely less in this system due to the overwhelming selective pressure imposed by DFTD (Fraik et al., 2020; Lachish et al., 2009; Lazenby et al., 2018). The short timescale for which DFTD has existed acts as the largest impediment to studying host-pathogen coevolutionary dynamics. In Freycinet, the longest-diseased site sampled, the potential evidence for lineage replacement indicates a selective sweep, and evidence of selective sweeps has also been found in devils (Epstein et al., 2016). Continued monitoring on a site-specific basis is thus necessary to determine if lineage co-existence is stable (i.e., negative frequency dependent selection) or merely an intermediate state of a selective sweep.
Disease-relevant traits indicate first-step evolution

Disease-relevant traits in both devils and DFTD tended to have non-zero PVEs, with many of these traits being attributable to a handful of large-effect variants (Fig. 4). When a novel pathogen first emerges, it is expected that both host and pathogen present suboptimal phenotypes because they have not yet adapted to one another (Berngruber et al., 2013). Therefore, it is anticipated that selection will initially favor variants conferring large effect gains in fitness. For devils, due to the low probability that a novel beneficial mutation arose in such a short period of time, selection is likely to act upon standing genetic variation in the form of previously neutral alleles. The findings for force of infection explained by the devil genome are partly concordant with this hypothesis, as 22.4% of this phenotype is attributable to ~3 variants. However, the remaining 38.7% of the phenotype appears to be polygenic, perhaps indicating selection acting first upon large-effect variants and subsequently variants of a smaller effect size.

Because DFTD is a clonal cell line, tumor cells lack genetic diversity at emergence; hence, diversity between tumor lineages must be the result of de novo mutations. Novel beneficial mutations are expected to be of large-effect, as such mutations are selected for over beneficial mutations of smaller effect (Rokyta et al., 2005). Although DFTD emerged recently, Murchison et al. (2012) found nonsynonymous to synonymous (NS/S) substitution rates of 2.78 and 2.08 in two tumor lineages (relative to NS/S ratios of approximately 1 in male and female devils), including nonsynonymous substitutions and indels in 324 genes specific to the cancer lineages. Large-effect deleterious mutations are purged through purifying selection, indicating that these mutations are marginally deleterious, neutral or beneficial. Given the finding that variability in tumor virulence is explained by a few loci of large effect (Fig. 4), it is likely that some of these mutations represent variants beneficial to tumor virulence.
The high PVE (61.1%) and PGE (37.3%) found in devil force of infection could be attributable to host immune recognition of DFTD and/or genetic changes underlying behavior. Hubert et al. (2018) identified genes under positive selection in behavior, including many of which are implicated in devil social behavior. Because DFTD transmission is so tightly linked to devil behavior (Hamede et al., 2013), it is possible that some of these genes are contributing to force of infection PVE through the devil genome. Although the devil genome explained much of the variance in force of infection, the tumor genome explained less of this variance (26.8%). The PGE for tumor force of infection also formed a nearly uniform distribution with credible intervals negligibly different from zero and one, indicating a high degree of uncertainty in the architecture underlying this trait. The uncertainty found here in both PVE (95% CI = 5.5–73.1%) and PGE (95% CI = 0–96.6%) for the pathogen could be due to genotype-by-genotype interactions occurring between devils and DFTD at many loci. If the majority of tumor loci contributing to force of infection do so primarily through interactions with loci in devils, BSLMM model fitting, which considers host and pathogen genome separately, may be unable to determine if a set of interacting loci are of small or large effect size.

**Devil-DFTD genomic interaction contributes to force of infection**

Joint model fitting using analysis with a two-organism mixed model (ATOMM) revealed that devil-DFTD genomic interactions explained a significant proportion of force of infection PVE (40.2%; Fig. 5). Although genomic interactions identified through associative modeling are not direct evidence of coevolution, this finding implicates coevolution as a viable driver of evolution for force of infection in devils and DFTD. Furthermore, the PVE identified for devil-DFTD genomic interaction was larger than the host, pathogen, and noise estimates, indicating that variability in force of infection is primarily due to interactions between the devil and tumor.
genomes. The large interaction PVE estimated by ATOMM also makes it unlikely that the
interactions are due to spurious associations between variants in the devil and DFTD genome;
rather, it is likely that at least some of this interaction represents signatures of coevolution.

Of note is the smaller host and pathogen PVE estimates made by ATOMM relative to
those made by GEMMA. However, the sum of the GEMMA estimates and ATOMM estimates
(including the interaction PVE for ATOMM) were similar (GEMMA sum = 87.9%, ATOMM
sum = 76.5%), indicating that much of the GEMMA PVE estimates may have represented host-
pathogen interaction. Using ATOMM to characterize host-pathogen relationships in a plant-
bacteria system, Wang et al. (2018) found that the top interacting SNPs differed from the top
host and pathogen SNPs, and in many cases the interaction SNPs with the lowest p-value were
not significant in host and pathogen genomes alone. Such a difference between variants
significant in hosts and pathogens relative to variants significant within genome interactions
indicates that models considering only a single genome have difficulty determining the effect
size of SNPs in genotype-by-genotype interaction. Hence, the low confidence in the pathogen
PGE and PVE for force of infection estimated by GEMMA may have been caused by tumor loci
interacting with variants in devils, something GEMMA is unable to account for by considering
only a single genome.

To further characterize loci in devils and DFTD contributing to coevolutionary
signatures, individual genotype-by-genotype interaction tests were performed, and the top five
most significant interactions were extracted (Table 2). An intronic variant in the devil gene
\textit{HNF1A} was found to interact with a variant 57kb upstream from \textit{SKOR2}. In human pancreatic
and liver cancers, \textit{HNF1A} is found to be downregulated and rescue of this gene’s expression \textit{in vivo}
inhibited tumor growth (Hoskins et al., 2014). \textit{SKOR2} is part of a family of genes producing
Ski proteins, the expression of which was found to be downregulated in tumor metastasis but upregulated in tumor growth for various human cancers (e.g., leukemia, melanoma, pancreatic cancer; Tecalco-Cruz et al., 2018). A DFTD variant 60kb upstream from AGPS, a gene found to be highly expressed in many aggressive human cancers (Benjamin et al., 2013), was also found to interact with two devil variants: a variant 7kb upstream from RUSC1 and a nonsynonymous substitution in SLC4A11. Although RUSC1 does not have direct implications with cancer or immune function, it is involved in regulating nerve growth factor (NGF) which can promote tumor survival in breast cancer and prostate cancer (Molloy et al., 2011). The other gene with which the AGPS variant interacts with, SLC4A11, is upregulated in grade III and IV ovarian cancers and is thus prognostic of these cancers, perhaps aiding in metastasis (Qin et al., 2017).

The signatures of devil-DFTD coevolution found in this study provide the strongest evidence to date that coevolution has contributed to the evolution of a disease-relevant trait in devils and DFTD. The modeling approach used here, being associative, cannot provide direct causative evidence of coevolution and, assuming coevolution to be causative, is unable to determine if coevolution remains ongoing between host and pathogen. However, because devil-DFTD genomic interactions contribute significantly to variation in force of infection, this phenotype represents a viable target for future studies, such as those looking at reciprocal host-pathogen fitness. Furthermore, the variants identified through genotype-by-genotype interactions present promising candidates underlying the genetics of a coevolutionary interaction and may thus be subject to direct genetic manipulation in devil and DFTD cell lines. Hence, the results of this work can generate hypotheses for future studies, ultimately helping to elucidate the genetic basis of complex trait evolution within a host-pathogen system.
CONCLUSION

Although the genetic structure of devils and DFTD provided little evidence for coevolution, genotype-phenotype associative modeling indicates that coevolution may be driving variation in key disease-related traits. Consistent with prior studies (Epstein et al., 2016; Margres et al., 2018), the evolutionary response in devils indicates first-step adaptive changes, as force of infection had a significant large-effect component. Novel to the devil-DFTD system, the results found here represent the first study characterizing the relationship between variation in DFTD genomes and phenotypes. For tumor force of infection, it appears as though the underlying genome architecture is difficult to elucidate without considering genotype-by-genotype interactions with the devil genome. However, variation in DFTD virulence can largely be explained by variation between tumor genomes, specifically through a few (~6) large-effect variants. Because novel beneficial mutations are anticipated to be of large-effect in a recently emerged pathogen (Rokyta et al., 2005), this may indicate an adaptive response in DFTD.

The genetic structure of devils and DFTD, evolutionary responses in both host and pathogen, and evidence suggesting possible coevolution found here indicate multiple complexities underlying host-pathogen dynamics in the devil-DFTD system. For example, although it appears as though tumor virulence is evolving, it remains uncertain if reduced virulence (i.e., virulence-transmission tradeoff) or increasing virulence is favorable to tumor fitness. The co-occurrence of DFTD lineages at each site facilitates lineage competition, whereby between-host competition may favor a virulence-transmission tradeoff, and within-host competition will likely select for increased virulence to competitively exclude other lineages. For
within-host competition to occur, it is necessary for coinfection of multiple DFTD lineages within a single devil. Although evidence of this does exist (Kwon et al., 2020), it is unclear if lineage coinfection is widespread enough to facilitate selection for within-host tumor traits. The recency with which DFTD emerged further confounds determining host-pathogen dynamics (i.e., arms race versus trench warfare), and it may be the case that different sites undergo different dynamics.

Although clonally transmissible cancers are clearly uncommon, the complexities implicit to the devil-DFTD system make it potentially generalizable, as typical coevolutionary relationships involve multiple traits controlled by complex genetics. Furthermore, the findings of this study indicate that the devil-DFTD system is potentially tractable for the study of coevolution. In particular, genotype-phenotype and genotype-genotype-phenotype relationships were established under a GWAS framework given the sufficient sampling used here (e.g., a large enough number of samples and targeted loci considering linkage disequilibrium). Although the associative nature of a GWAS limits causative interpretations and can only provide signals of coevolution rather than establish it as an ongoing driver of evolution, such a study design has the benefit of rapidly scanning a genome to find potential causative variants. The framework used herein may thus be applicable to studying other host-pathogen systems (e.g., humans and SARS-CoV-2) or the evolutionary progression of various cancers in humans. Variants identified both in devils and DFTD, particularly those of large effect, may also represent prognostic or therapeutic targets in human cancers, providing useful avenues for future medical research.

Among the most evident applications of this study are to devil-DFTD epidemiological modeling and devil conservation efforts. The genotype-phenotype relationships identified here for disease-relevant traits can be used to refine the predictive ability of current epidemiological
models by adding an evolutionary component to these models which currently utilize only ecological data (Wells et al., 2017, 2019). Genotype-phenotype relationships can also inform devil selective breeding programs, utilizing large-effect variants as targets for low-cost genotyping to identify individuals with favorable adaptations to DFTD. The presence of a potential coevolutionary relationship may further inform breeding programs by identifying population-specific adaptive alleles which implicate specific devils as being well adapted to specific tumors circulating within a given population. Despite promising signs of devil adaptation to DFTD, continued monitoring and conservation work is necessary to ensure the persistence of this species, and only through science-based decision making may such conservation efforts find success.
REFERENCES


*FelixKrueger/TrimGalore: V0.6.7 - DOI via Zenodo.* Zenodo.

https://doi.org/10.5281/zenodo.5127899


Zachary McCaw. (2020). *RNOmni: Rank Normal Transformation Omnibus Test* (1.0.0) [R]. https://CRAN.R-project.org/package=RNOmni
APPENDIX 1: SUPPLEMENTARY MATERIAL

Table A1. DFTD arrival dates by site. Arrival dates are assumed to be January 1st of the corresponding year.

<table>
<thead>
<tr>
<th>Site</th>
<th>DFTD arrival year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mt. William</td>
<td>1996</td>
</tr>
<tr>
<td>Freycinet</td>
<td>1999</td>
</tr>
<tr>
<td>Fentonbury</td>
<td>2005</td>
</tr>
<tr>
<td>Narawntapu</td>
<td>2007</td>
</tr>
<tr>
<td>West Pencil Pine</td>
<td>2007</td>
</tr>
<tr>
<td>Wilmot</td>
<td>2008</td>
</tr>
<tr>
<td>Takone</td>
<td>2010</td>
</tr>
<tr>
<td>Dip River</td>
<td>2015</td>
</tr>
<tr>
<td>Black River</td>
<td>2015</td>
</tr>
<tr>
<td>Arthur River</td>
<td>2019</td>
</tr>
</tbody>
</table>

Table A2. Devil lineage $F_{st}$ estimates. Pairwise $F_{st}$ values were calculated between devil lineages identified using DAPC. Hierarchical 1 refers to the divergence calculated between Freycinet and the northwestern devil populations (i.e., cluster 1 and 2 in Fig. 2A and 2B), and hierarchical 2 refers to the divergence calculated between individuals in the northwestern populations (i.e., cluster 1 and 2 in Fig. 2C and 2D). $F_{st}$ values calculated as Weir-Cockerham $F_{st}$ estimates (Weir & Cockerham, 1984) using VCFtools (Danecek et al., 2011).

<table>
<thead>
<tr>
<th></th>
<th>Mean $F_{st}$</th>
<th>Weighted $F_{st}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hierarchical 1</td>
<td>0.0388</td>
<td>0.0551</td>
</tr>
<tr>
<td>Hierarchical 2</td>
<td>0.0070</td>
<td>0.0096</td>
</tr>
</tbody>
</table>

Table A3. DFTD lineage fixation $F_{st}$ estimates. Pairwise $F_{st}$ values were calculated between each tumor lineage determined using DAPC (Fig. 3B). $F_{st}$ values calculated as Weir-Cockerham $F_{st}$ estimates (Weir & Cockerham, 1984) using VCFtools (Danecek et al., 2011).

<table>
<thead>
<tr>
<th></th>
<th>Mean $F_{st}$</th>
<th>Weighted $F_{st}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1 &amp; 2</td>
<td>0.012</td>
<td>0.041</td>
</tr>
<tr>
<td>Cluster 1 &amp; 3</td>
<td>0.009</td>
<td>0.030</td>
</tr>
<tr>
<td>Cluster 1 &amp; 4</td>
<td>0.012</td>
<td>0.044</td>
</tr>
<tr>
<td>Cluster 2 &amp; 3</td>
<td>0.009</td>
<td>0.031</td>
</tr>
<tr>
<td>Cluster 2 &amp; 4</td>
<td>0.013</td>
<td>0.045</td>
</tr>
<tr>
<td>Cluster 3 &amp; 4</td>
<td>0.009</td>
<td>0.031</td>
</tr>
</tbody>
</table>
Table A4. GEMMA convergence statistics. Statistics were generated using Rstan (Stan Development Team, 2020) within R version 4.1.0. Rhat values ≤ 1.05 indicate a high degree of within and between chain convergence. Bulk and tail ESS > 100 indicate high efficiency sampling in the bulk and tail of the posterior distribution.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Hyperparameter</th>
<th>Rhat</th>
<th>Bulk ESS</th>
<th>Tail ESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Host force of infection</td>
<td>h</td>
<td>1.001</td>
<td>4021</td>
<td>10638</td>
</tr>
<tr>
<td></td>
<td>PVE</td>
<td>1.000</td>
<td>21129</td>
<td>165697</td>
</tr>
<tr>
<td></td>
<td>Rho</td>
<td>1.001</td>
<td>4846</td>
<td>13848</td>
</tr>
<tr>
<td></td>
<td>PGE</td>
<td>1.000</td>
<td>12518</td>
<td>91415</td>
</tr>
<tr>
<td></td>
<td>Pi</td>
<td>1.002</td>
<td>1568</td>
<td>2055</td>
</tr>
<tr>
<td></td>
<td>N_gamma</td>
<td>1.001</td>
<td>2117</td>
<td>11215</td>
</tr>
<tr>
<td>Tumor force of infection</td>
<td>Rho</td>
<td>1.000</td>
<td>34642</td>
<td>134137</td>
</tr>
<tr>
<td></td>
<td>PGE</td>
<td>1.001</td>
<td>8503</td>
<td>20284</td>
</tr>
<tr>
<td></td>
<td>Pi</td>
<td>1.004</td>
<td>722</td>
<td>646</td>
</tr>
<tr>
<td></td>
<td>N_gamma</td>
<td>1.005</td>
<td>713</td>
<td>635</td>
</tr>
<tr>
<td>Tumor virulence</td>
<td>h</td>
<td>1.000</td>
<td>37094</td>
<td>83063</td>
</tr>
<tr>
<td></td>
<td>PVE</td>
<td>1.000</td>
<td>43378</td>
<td>120909</td>
</tr>
<tr>
<td></td>
<td>Rho</td>
<td>1.000</td>
<td>74326</td>
<td>125592</td>
</tr>
<tr>
<td></td>
<td>PGE</td>
<td>1.000</td>
<td>166157</td>
<td>209845</td>
</tr>
<tr>
<td></td>
<td>Pi</td>
<td>1.001</td>
<td>5118</td>
<td>6334</td>
</tr>
<tr>
<td></td>
<td>N_gamma</td>
<td>1.001</td>
<td>6545</td>
<td>5565</td>
</tr>
</tbody>
</table>

Figure A1. Tumor area versus tumor volume regression. All available samples possessing measurements for tumor length, width, and depth were used. Volumes and areas were calculated in millimeters and the natural log was taken of each before fitting a linear model. This model was used to impute missing depth values.
Figure A2. Host k-means clustering Bayesian information criterion (BIC). Clustering was performed with 456 devils and 221,965 SNPs in DAPC using the `find.clusters` function.

Figure A3. Host hierarchical k-means clustering Bayesian information criterion (BIC). Clustering was performed as in Fig. A2 but by removing individuals from cluster 1 (i.e., Freycinet). Thus, 378 hosts were used in this analysis.
Figure A4. Tumor k-means clustering Bayesian information criterion (BIC). Clustering was performed with 504 tumors and 63,317 SNPs in DAPC using the find.clusters function.

Figure A5. BSLMM fitting on ranknorm host force of infection. A. Distribution of devil force of infection after normalizing using ranknorm. B. PVE and PGE obtained from the posterior distributions after BSLMM model fitting.